Environmental Science and Pollution Research (2022) 29:77157–77187
https://doi.org/10.1007/s11356-022-21201-1

RESEARCH ARTICLE

Cyanobacteria blue-green algae prediction enhancement using hybrid machine learning–based gamma test variable selection and empirical wavelet transform

Salim Heddam[1] · Zaher Mundher Yaseen[2,3,4] · Mayadah W. Falah[5] · Leonardo Goliatt[6] · Mou Leong Tan[7] · Zulfaqar Sa'adi[8] · Iman Ahmadianfar[9] · Mandeep Saggi[10] · Amandeep Bhatia[11] · Pijush Samui[12]

Received: 17 March 2022 / Accepted: 27 May 2022 / Published online: 8 June 2022

Abstract

This study aims to evaluate the usefulness and effectiveness of four machine learning (ML) models for modelling cyanobacteria blue-green algae (*CBGA*) at two rivers located in the USA. The proposed modelling framework was based on establishing a link between five water quality variables and the concentration of CBGA. For this purpose, artificial neural network (ANN), extreme learning machine (ELM), random forest regression (RFR), and random vector functional link (*RVFL*) are developed. First, the four models were developed using only water quality variables. Second, based on the results of the first, a new modelling strategy was introduced based on preprocessing signal decomposition. Hence, the empirical mode decomposition (EMD), the variational mode decomposition (VMD), and the empirical wavelet transform (EWT) were used for decomposing the water quality variables into several subcomponents, and the obtained intrinsic mode functions (IMFs) and multiresolution analysis (MRA) components were used as new input variables for the ML models. Results of the present investigation show that (i) using single models, good predictive accuracy was obtained using the RFR model exhibiting an R and NSE values of ≈0.914 and ≈0.833 for the first station, and ≈0.944 and ≈0.884 for the second station, while the others models, i.e., ANN, RVFL, and ELM, have failed to provide a good estimation of the CBGA; (ii) the decomposition methods have contributed to a significant improvement of the individual models performances; (iii) among the thee decomposition methods, the EMD was found to be superior to the VMD and EWT; and (iv) the ANN and RFR were found to be more accurate compared to the ELM and RVFL models, exhibiting high numerical performances with R and NSE values of approximately ≈0.983, ≈0.967, and ≈0.989 and ≈0.976, respectively.

Keywords Modelling · CBGA · Water quality · ELM · ANN · RVFL · RFR · EMD · VMD · EWT

Highlights

- Cyanobacteria blue-green algae are predicted using machine learning (ML) models.
- Different preprocessing signal decomposition methods are used for data analysis.
- The gamma test was used for input variables selection.
- Signal decompositions improved the prediction capacity of the applied ML models.

Responsible Editor: Marcus Schulz

✉ Salim Heddam
heddamsalim@yahoo.fr

Extended author information available on the last page of the article

Introduction

Background

During the last few years, the degradation of freshwater ecosystems quality has become a serious concern for water resources planning and management (Beretta-Blanco and Carrasco-Letelier 2021). Specific and commonly problems and issues affecting the overall quality of freshwater are certainly the cyanobacteria harmful algal blooms (*HAB*) which has received great deal of attention for water managers (Clercin et al. 2022). Reported as the first and major responsible of the production and proliferation of the cyanotoxins in inland water bodies and aquatic ecosystems, HAB can cause eutrophication and contributes to a significant decrease in the quantity of available drinking water (Choi et al. 2021). In addition, high level of eutrophication was considered as a serious water pollution problem (Zou et al. 2014). Cyanobacteria blue-green algae (*CBGA*) expressed by a concentration in (cells/mL) is one of the well-known HAB and belongs into the category of "photosynthetic organisms" and is expanding rapidly under certain environmental conditions (Gaget et al. 2022). High level of CBGA is considered as "detrimental" to freshwater ecosystems and, with rapid proliferation, the lake or reservoir becomes with a green colour (Sheng et al. 2012). Regarding its importance for water resources management and pollution control, it is necessary to understand the majors and significant factors controlling the growth of CBGA and their production of toxins (Te and Gin 2011).

Factors controlling the concentration of CBGA

It is well recognized that the major issues causing the growth of CBGA in freshwater ecosystems is mainly related to the anthropogenic activities, especially the excessive use of pesticides for agricultural purpose (Bano et al. 2021; Khaleefa and Kamel 2021), and further complicated by climate change (Mahmudi et al. 2020; Sanseverino et al. 2022). However, several environmental factors are responsible for rapid growth and fluctuation over time and space of the CBGA, and several authors worldwide have based on an experimental study conducted into three gorges reservoir (TGR) in China. (Yang et al. 2022) investigated whether or not hydrodynamic factors influence the variation, growth, and proliferation of HAB. It was found that the ratio of mixing depth to euphotic depth (Z_m/Z_e) was a significant factor and significantly affects HAB growth and concentration. In a study conducted by Mahmudi et al. (2020) in the Ambon Bay, Indonesia, the authors reported that water quality variables, i.e., water temperature (T_w), *pH*, salinity (*Sa*), dissolved oxygen (*DO*), nitrate (NO_3), and phosphate (PO_4) significantly affect the abundance of the HAB in marine water. Descy et al. (2016) argued that the level of cyanobacteria abundance was highly linked to environmental conditions, i.e., phosphorus, dissolved inorganic nitrogen, epilimnion temperature, DO, pH, specific conductance (SC) euphotic depth, wind speed, rainfall, and surface irradiance. In another study, García Nieto et al. (2015) reported that water quality variables, i.e., T_w, pH, alkalinity, SC, DO, water turbidity (TU), air temperature (T_a), and Secchi disk depth (*SD*), are the most significant variables affecting the concentration of CBGA in water reservoir. Similarly, Recknagel et al. (2006) demonstrated that NO_3, PO_4, TU, SD, T_w, and pH were the most significant factor controlling the concentration of blue-green algae and diatom populations in lakes freshwater. Indeed, Song et al. (2012) reported that T_w and light are the most significant factors affecting the growth of CBGA. Consequently, due to the high number of factors controlling the growth of HAB and especially CBGA, a complex physical, chemical, and biological process were involved and need robust nonlinear models for its prediction., and generally speaking, modelling CBGA can be achieved using two distinguished approach process–based models and statistically based models (Maier and Dandy 2000; Tiyasha et al. 2020).

Modelling CBGA using machine learning: state of the arts

Mentoring CBGA in River, lakes and reservoirs are mainly based on traditional sampling, laboratory analysis, and cell counting. However, these traditional approaches are laborious and take considerable time to get right (Guo et al. 2021). Modelling using machine learning (ML) is an interesting area of research, and they have proven to be a powerful and credible alternative in the absence of direct in situ measurements (Elzwayie et al. 2016; Sanikhani et al. 2018; Asadollah et al. 2020). For modelling and forecasting the cyanobacteria harmful algal blooms (HAB) based on water quality variables, different ML-based models have been proposed so far. Indeed, numerous ML algorithms are currently being investigated to develop robust predictive models, using suite of predictors. Maier et al. (1998) used an artificial neural network, i.e., the multilayer perceptron neural network (MLPNN) for predicting weekly CBGA measured in (cells/mL) at the River Murray at Morgan, Australia. They used several input variables namely, water colour (CO), TU, T_w, river flow (*Q*), soluble and total Phosphorus (SP, TP), nitrogen, and total iron. High performances were obtained with root-mean-squared error (RMSE) ranging from 318 (cells/mL) to 355 (cells/mL). Maier et al. (2000) applied the B-spline associative memory network (AMN) model for forecasting the concentration of CBGA up to 4 weeks in advance. They used the same input variables reported in Maier and Dandy (1998), and the performances of the AMNs were compared to those of MLPNN and demonstrating its superiority. Vilán Vilán et al. (2013) conducted a comparative study for predicting CBGA based on several water quality variables, i.e., T_a, pH, T_w, DO, TU, SC, alkalinity, and SD. They compared between MLPNN and three support vector regression (SVR) namely, linear (SVR-LN), radial basis function (SVR-RBF), and Pearson VII universal function (SVR-PUK). According to the obtained results, the SVR-RBF was found to be more accurate with coefficient of determination (R^2) equal to 0.92, followed by the SVR-PUK ($R^2 = 0.91$), the MLPNN ($R^2 = 0.64$), and the SVR-LN ($R^2 = 0.57$). Harris and Graham (2017) compared between linear and several ML models for predicting CBGA in the Cheney Reservoir, Kansas, USA. The tested models were respectively ordinary linear regression (Linear), partial least squares (PLS), elastic net (Enet), neural networks (Nnet), multivariate adaptive regression splines (MARS), support vector regression (SVR), single trees (CART), bagged trees (BagT), boosted trees (BT), conditional inference trees (CI-Tree), random forest regression (RF), and Cubist models. For models' development, they used several water quality variables, i.e., T_w, TU, DO, pH, suspended sediment concentration (SSC), and reservoir surface elevation (RL). It was found that the Cubist model was the most accurate exhibiting R value of approximately ≈ 0.87, followed by the random forest regression (RFR) model ($R \approx 0.82$), the BT ($R \approx 0.80$), and the SVR ($R \approx 0.72$), while the other models were failed to correctly predict the CBGA.

Ostfeld et al. (2015) optimized the decision tree model using genetic algorithm (GA-DT) and modelling strategy for CBGA was proposed. Several water quality variables were linked to the CBGA concentration via the GA-DT model, i.e., T_w, T_a, relative humidity (RH %), and wind speed (U_2). It was found that CBGA can be predicted very well with R value of approximately ≈ 0.91. Saboe et al. (2021) applied the long short-term memory (LSTM) neural network for predicting CBGA concentration based on several input water quality variables. From the obtained results, it was found that the LSTM can help in accurately predict CBGA with high performances exhibiting a correlation coefficient (R) of approximately ≈ 0.930 and normalized root mean square error (NRMSE) of ≈ 6.5%. Derot et al. (2020) used RFR for predicting cyanobacteria concentration in the Lake Geneva located in the north of the French Alps. However, they reported that obtaining high forecasting accuracy needs the inclusion of high number of predictors from the combination of several physical, chemical and biological variables, and the R^2 was approximately ≈ 0.90. Su et al. (2022) demonstrated that water T_w and nitrogen were the most significant factors affecting the concentration of the algal blooms in the three gorges reservoir in China. Indeed, the authors compared between several ML namely, extra trees regression (ETR), the RFR, SVR, gradient boosting regression tree (GBRT), classification and regression tree (CART), MLPNN, and the K-neighbors regression (KNR), for predicting algal blooms, and they reported that the high R^2 (≈ 0.60) value was obtained using the ETR model. Park et al. (2021) compared between SVR and MLPNN for predicting the algal concentration in (cells/mL). The authors used several input variables, i.e., nitrogen, nitrate, total dissolved phosphorus, SC, water level of the reservoir, discharge, precipitation, T_a, and WS. It was found that both models were able to accurately predict CBGA without providing any numerical results. Jafarzadeh et al. (2022) compared between four ML models namely, gene expression programming (GEP), SVR, and hybrids wavelet SVR and GEP (W-SVR, W-GEP) for predicting cyanobacterial in Jajrood River, Iran. The models were calibrated using Q, DO, NO_3, PO_4, and biological oxygen demand (BOD), and it was found that the hybrid W-SVR was more accurate exhibiting Nash-Sutcliffe efficiency (NSE) value of approximately ≈ 0.98 compared to the value of ≈ 0.82 obtained using the W-GEP; it was demonstrated that the wavelet algorithm have helped in improving the models' performances. Finally, Pyo et al. (2021) used convolutional neural network (CNN) for predicting CBGA in the Nakdong River in South Korea and reported a NSE value of approximately ≈ 0.76.

Objective, contributions, innovation, and article structure

Based on the reported literature review, it is clear that modelling CBGA using ML models has attracted wide interest and there is high degree of its success (Giere et al. 2020; Nguyen et al. 2020; Rousso et al. 2022). In addition, a wide range of models was proposed and successfully applied exhibiting moderate to high level of accuracies (Park et al. 2021). Indeed, it was found that models for CBGA were based on the use of measured water quality variables without preprocessing, and except the work conducted by Jafarzadeh et al. (2022), the use of signal decomposition for improving the performances of ML models was rarely reported in the literature, which constitutes the major motivation of our present study. Therefore, in the current research, the investigation of how preprocessing signal decomposition contributed significantly to the prediction improvements of ML models for CBGA in river. The literature review has demonstrated on the implementation of signal decomposition algorithms for diverse engineering applications and approved their capacity (Bokde et al. 2020; Wang et al. 2021; Ahmadianfar et al. 2022; Jamei et al. 2022; Tao et al. 2022). Hence, three signal decomposition algorithms namely, empirical mode decomposition (EMD), variational mode decomposition (VMD), and empirical wavelet transform (EWT) were used for the modelling development. These three algorithms were used for decomposing five water quality variables selected as relevant predictors. The specific objectives of the present research are as follows:

(i) The application of four single ML models for predicting CBGA namely: (i) artificial neural network (ANN), (ii) extreme learning machine (ELM), (iii) random forest regression (RFR), and (iv) and random vector functional link (*RVFL*).

(ii) In the second stage of the investigation, new hybrid models were proposed based on the combination of the EMD, EWT, and VMD with the single models.

(iii) The gamma test input variable selection was used for selecting the best input combination and in total seven input combination were adopted for models comparison.

(iv) All models were compared based on numerical and graphical comparisons.

Study area and data

Data used in the present study were collected at two US Geological Survey (USGS) (https://or.water.usgs.gov). The two stations were (i) USGS 14202650 (latitude 45°26′26.31″, longitude 123°07′30.00″NAD83) Wapato creek at SW Gaston road, at Gaston, Washington County, Oregon, USA, and (ii) USGS 14207200 (latitude 45°21′24″, longitude 122°41′02″ NAD27) Tualatin river at Oswego dam, near west Linn, Clackamas County, Oregon, USA (Fig. 1). Six water quality variables were selected and used for developing the models. The modelled variable was the cyanobacteria blue-green algae (CBGA: cells/mL), and five water quality variables were used as independents variables, i.e., the predictors, namely, water temperature (T_w: °C), water pH (std. unit), water dissolved oxygen concentration (DO: mg/L), water specific conductance (SC: uS/cm), and water turbidity (TU: FNU). For the USGS 14202650 station, data were measured at every 30 min (i.e., every half hour) during the period from 12 April 2010 to 29 May 2012 with a total of 9000 data. For the USGS 14207200 station, data were measured at every 60 min (i.e., every hour) during the period from 25 March 2010 to 18 May 2012 with a total of 9000 data. For each station, we split the data into training (70%) and validation (30%). For each station, we provide in Table 1 the statistical description of the five water quality variables and the cyanobacteria blue-green algae, and we highlighted the correlation coefficients (R) for all variables with the CBGA. According to Table 1, at the two stations, very low R values were found and none of the five water quality variables was highly correlated with CBGA, making them an attractive and promising modelling investigation as the simple linear regression does not allow a direct assessment and an effective estimation of the CBGA. Two scenarios were tested in the present study: (i) standalone modelling strategy for which ML models were developed using water quality variables without preprocessing and (ii) three signal decomposition techniques (see details later); i.e., the empirical mode

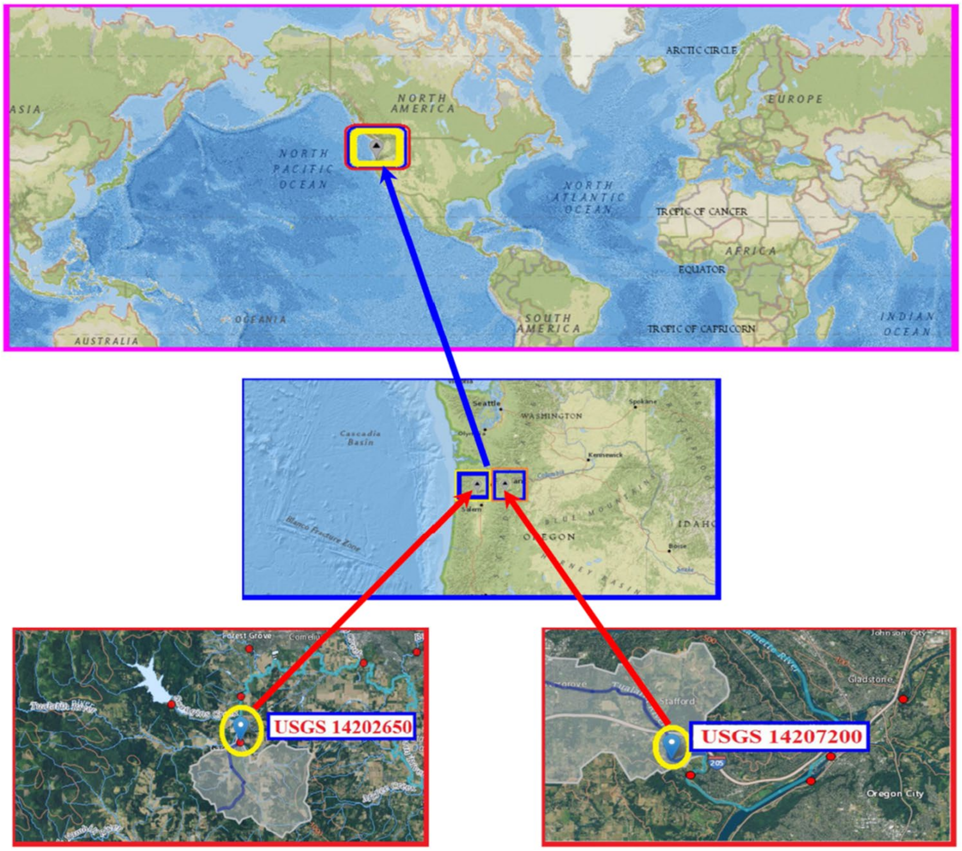


Fig. 1 Maps showings the location of the two USGS stations

Table 1 Summary statistics of cyanobacteria blue-green algae concentration and water quality variables

Variables	Subset	Unit	X_{mean}	X_{max}	X_{min}	S_x	C_v	R
USGS 14202650 Willamette River at Portland, Oregon, USA								
CBGA	Training	*cells/mL*	3467.696	40130.000	47.000	5688.616	1.640	1.000
	Validation	*cells/mL*	3435.174	39641.000	51.000	5637.976	1.641	1.000
	All data	*cells/mL*	3457.939	40130.000	47.000	5673.178	1.641	1.000
T_w	Training	*°C*	12.556	21.600	4.700	3.963	0.316	0.271
	Validation	*°C*	12.497	21.700	4.800	3.994	0.320	0.270
	All data	*°C*	12.538	21.700	4.700	3.972	0.317	0.271
DO	Training	*mg/L*	9.145	12.800	4.000	1.360	0.149	-0.228
	Validation	*mg/L*	9.121	12.800	4.200	1.357	0.149	-0.248
	All data	*mg/L*	9.138	12.800	4.000	1.359	0.149	-0.234
pH	Training	/	6.815	7.300	5.600	0.248	0.036	0.071
	Validation	/	6.813	7.300	5.500	0.248	0.036	0.072
	All data	/	6.814	7.300	5.500	0.248	0.036	0.071
SC	Training	*uS/cm*	107.261	388.000	68.000	29.880	0.279	-0.060
	Validation	*uS/cm*	107.164	382.000	68.000	28.974	0.270	-0.063
	All data	*uS/cm*	107.232	388.000	68.000	29.609	0.276	-0.061
TU	Training	*FNU*	27.581	149.000	2.600	17.946	0.651	0.115
	Validation	*FNU*	27.590	160.000	2.700	18.298	0.663	0.103
	All data	*FNU*	27.584	160.000	2.600	18.051	0.654	0.111
USGS ID 14207200 Willamette River at Portland, Oregon, USA								
CBGA	Training	*cells/mL*	522.728	2448.000	1.000	454.802	0.870	1.000
	Validation	*cells/mL*	516.548	2448.000	1.000	450.655	0.872	1.000
	All data	*cells/mL*	537.149	2321.000	4.000	464.101	0.864	1.000
T_w	Training	*°C*	16.022	23.100	8.100	3.913	0.244	0.237
	Validation	*°C*	15.986	23.100	8.100	3.922	0.245	0.229
	All data	*°C*	16.106	22.800	8.100	3.894	0.242	0.254
DO	Training	*mg/L*	7.606	10.400	4.600	1.042	0.137	0.171
	Validation	*mg/L*	7.610	10.400	4.600	1.040	0.137	0.177
	All data	*mg/L*	7.597	10.300	4.800	1.048	0.138	0.159
pH	Training	/	7.004	7.300	6.800	0.097	0.014	0.255
	Validation	/	7.003	7.300	6.800	0.097	0.014	0.249
	All data	/	7.006	7.300	6.800	0.098	0.014	0.268
SC	Training	*uS/cm*	231.545	361.000	92.000	72.450	0.313	-0.210
	Validation	*uS/cm*	231.305	361.000	92.000	72.432	0.313	-0.217
	All data	*uS/cm*	232.103	361.000	92.000	72.502	0.312	-0.195
TU	Training	*FNU*	5.829	82.400	0.100	6.532	1.121	0.161
	Validation	*FNU*	5.916	82.400	0.100	6.666	1.127	0.161
	All data	*FNU*	5.625	70.200	0.600	6.204	1.103	0.163

X_{mean} mean, X_{max} maximum, X_{min} minimum, S_x standard deviation, C_v coefficient of variation, R coefficient of correlation with *CBGA*, *Tw* river water temperature, *CBGA* cyanobacteria blue-green algae, *DO* dissolved oxygen, *SC* specific conductance, *TU* turbidity, *FNU* formazin nephelometric unit, *uS/cm* microsiemens per centimetre

decomposition (EMD), the variational mode decomposition (VMD), and the empirical wavelet transform (EWT) were used for decomposing the input variables into several subcomponents. In addition, all data were normalized using the Z-score method as follow:

$$Z = \frac{x - x_{mean}}{x_\sigma} \tag{1}$$

where Z is the normalized score, x_{mean} is the mean value, and x_σ is the standard deviation.

Methodology

Artificial neural network model

One of the most and impressive alternatives to the traditional statistical algorithm which has proved its effectiveness for handling and solving high nonlinear tasks is certainly the artificial neural network (ANN) technique. The ANN is a specific approach inspired from the function of the human brain, which utilizes the concept of learning for providing a suitable response to a specific problem, and successful implementations of the ANN models can be found in the literature (Afan et al. 2014; Karimi et al. 2020; Yaseen et al. 2020). An ANN model based on the backpropagation training algorithm possesses the capability of mapping and establishing a function between the input and output variables using a sequence of measured dataset (Jha et al. 2022; Salman and Kadhum 2022). A simple ANN model with one input, one hidden, and one output layer is depicted in Figure 2. The first layer, i.e., the input layer, contains the predictor variables involved in the modelling of the output variable. The connection between the input and the hidden neurons is established using a matrix of weights and biases (i.e., the W_{ij}). The weighed sum of the input variables multiplied by the connection weights W_{ij} should be moved to the next layer after passing via an activation function, generally the sigmoidal (Oboh et al. 2022). Consequently, the final output of the hidden neurons becomes the input of the single output neuron. According to the Fig. 2, each neuron in the hidden layer calculates the weighted sum of the predictors as follow:

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

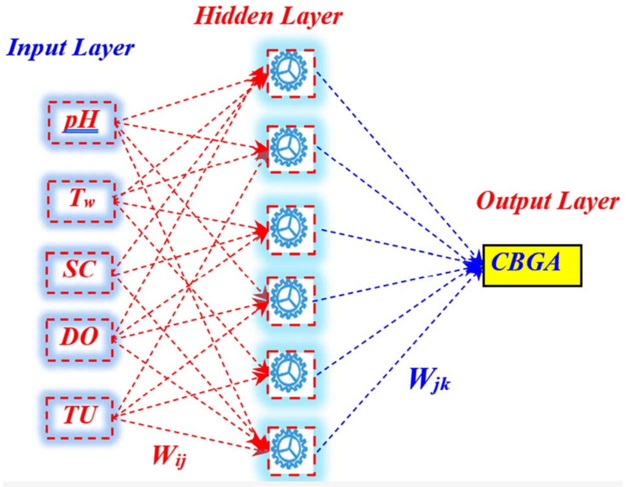


Fig. 2 Flowchart of the ANN model

$$E_j = f\left(A_j\right) \tag{3}$$

$$f\left(A_j\right) = \frac{1}{1 + e^{-A_j}} \tag{4}$$

$$O_K = \sum_{j=1}^{n} \left(W_{jk} \times E_j\right) + b_k \tag{5}$$

The activation function in the hidden layer (i.e., the f) is the sigmoidal, while the output neuron uses a linear activation function; w_{ij} corresponds to the weights between the input and the hidden layer, w_{jk} corresponds to the weights between the hidden and the output layer, b_i is the bias of the ith hidden neuron, and finally, b_k is the bias of the output neuron (Paul et al. 2022).

Extreme learning machine

Extreme learning machine (ELM) was introduced by Huang et al. (2006). Its popularity comes from its fast learning speed and high capacity in handling large dataset (Adnan et al. 2021). This is a result of the randomly generating of the hidden inputs weights and biases (i.e., form the input to the hidden layers) and the analytically calculation of the output weights matrix (Araba et al. 2021; Chen et al. 2022). For any training dataset, N for which the x is the input variable and y corresponds to the output variable (Yan et al. 2022):

$$D = \left\{\left(x_i, y_i\right) | x_i \in R^d, y_i \in R\right\}, i = 1, 2, 3, \cdots N \tag{6}$$

The output of the ELM model with Z hidden neurons can be calculated as follows (Fig. 3):

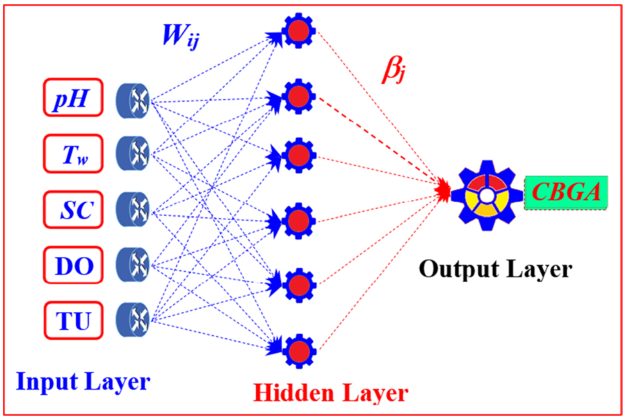


Fig. 3 Flowchart of the extreme learning machine (ELM) architecture

$$Y_j = f(x_j) = \sum_{i=1}^{Z} \beta_i . g(w_i, b_i, x_j), j = 1, 2, 3, \ldots \ldots, Z \tag{7}$$

where $g(.)$ is the sigmoidal activation function, w_{ij} is the weights between the input and the hidden layer, b_i is the bias of the ith hidden node, and β_i is the output weights (i.e., from the hidden to the output layer). The previous equation can be reformulated as follows (Hai et al. 2020):

$$H\beta = T \tag{8}$$

where

$$\boldsymbol{H} = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} g(w_1, b_1, x_1) & \cdots & g(w_Z, b_Z, x_1) \\ \vdots & \iddots & \\ g(w_1, b_1, x_N) & \cdots & g(w_Z, b_Z, x_N) \end{bmatrix}_{N \times Z} \tag{9}$$

H is the output matrix, i.e., the activation of the hidden layer neurons. $\boldsymbol{\beta}$ is the output weights linking the hidden neurons to the output neuron, and $\boldsymbol{T}$ is the target matrix of ELM (Zhao and Chen 2022):

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_Z^T \end{bmatrix}_{Z \times m} \quad \text{and} \quad \beta = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \tag{10}$$

$$\beta = H^{+}T \tag{11}$$

H^{+} corresponds to the Moore-Penrose generalized inverse of matrix $\boldsymbol{H}$.

Random forest regression

Random forest regression (RFR) is a modified version of the original bagging algorithm proposed by Breiman (2001). The RFR uses the concept of trees to solve a classification and regression tasks; hence, the RFR could be viewed as a tree ensemble model (Fig. 4), for which each tree was built depending on the values of sample vector with the respect to the condition of an equal distribution for all calculated trees (Bhagat et al. 2020; Onyelowe et al. 2022). The final output, i.e., response of the RFR model, is calculated as an average of the response provided by the individual trees, leading to a significant improvement in the variance minimization by decreasing the correlation between the trees, which improve its capability to overcome the overfitting problem (Shoar et al. 2022).

According to Fig. 4, for a training dataset having (L) observations composed of one dependent, i.e., the response variable (Y) and an ensemble of independent variables, i.e., the predictors (x), an approximation function using the RFR model should be achieved as follows. First, the RFR generates uniformly (n) sampling, i.e., sample (1) to sample (n) using the bootstrap, i.e., the "bootstrap aggregation" procedure (Elmetwalli et al. 2022). Second, for each sample, grow a tree, and third, proceed by averaging the responses of all constructed trees (Fernández-Habas et al. 2022). It is important to note that the RFR should develop their own internal mechanism to calculate the prediction error designated as the "out-of-bag" error "OOB," equal to the standard deviation (SD) error between calculated and measured values. It is used for ranking the predictors and predictor selection (Rosecrans et al. 2022). More details about random forest can be found in Sharafati et al. (2020).

Random vector functional link

Random vector functional link ($RVFL$) is one of the fewer ML models characterized by their randomization in the training process, and it has a direct link between the input layer and the output layer (Fig. 5; black dashed line) (Almodfer et al. 2022). The basic structure of the $RVFL$ is illustrated in Fig. 5 (PAO et al. 1992; Pao et al. 1994). It is composed of one input layer with a number of neurons equal to the number of predictors (i.e., five), one hidden layer with several neurons also called enhancement nodes, and one output layer (Chauhan and Tiwari 2022). The input to the hidden layer weights are highlighted in red dashed lines, and they are randomly generated and remains fixed and unchangeable during the training process (Hazarika and Gupta 2022), while the weights between the hidden and the output layers (blue dashed line) and the direct weights linking the input to the output layer (black dashed line) need to be trained using the pseudo-inverse or gradient descent algorithms (PAO et al. 1992; Pao et al. 1994; Basilio and Goliatt 2022).

One of the most important features of the RVFL model is its highly efficient training algorithm based on the random initialization of part of their weights, leading to good compromise between the precision, simplicity and illustrating an alleged training cost and very high quality of nonlinear approximation function. For any dataset composed of a pairs if inputs and output variables (Cao et al. 2020):

$$D = \left\{ (x_i, y_i) | x_i \in R^d, y_i \in R \right\}, i = 1, 2, 3, \cdots N \tag{12}$$

Three steps are necessary for achieving the training process of the RVFL model: (i) linear link between input and hidden neurons using the input weights and biases, i.e., W_{ij} and b, (ii) the output of the hidden neurons obtained during the first stage should be nonlinearly transformed using an activation nonlinear sigmoidal function g (.), and (iii) the output weights β was calculated using a lead-square approach (Cao et al. 2020).

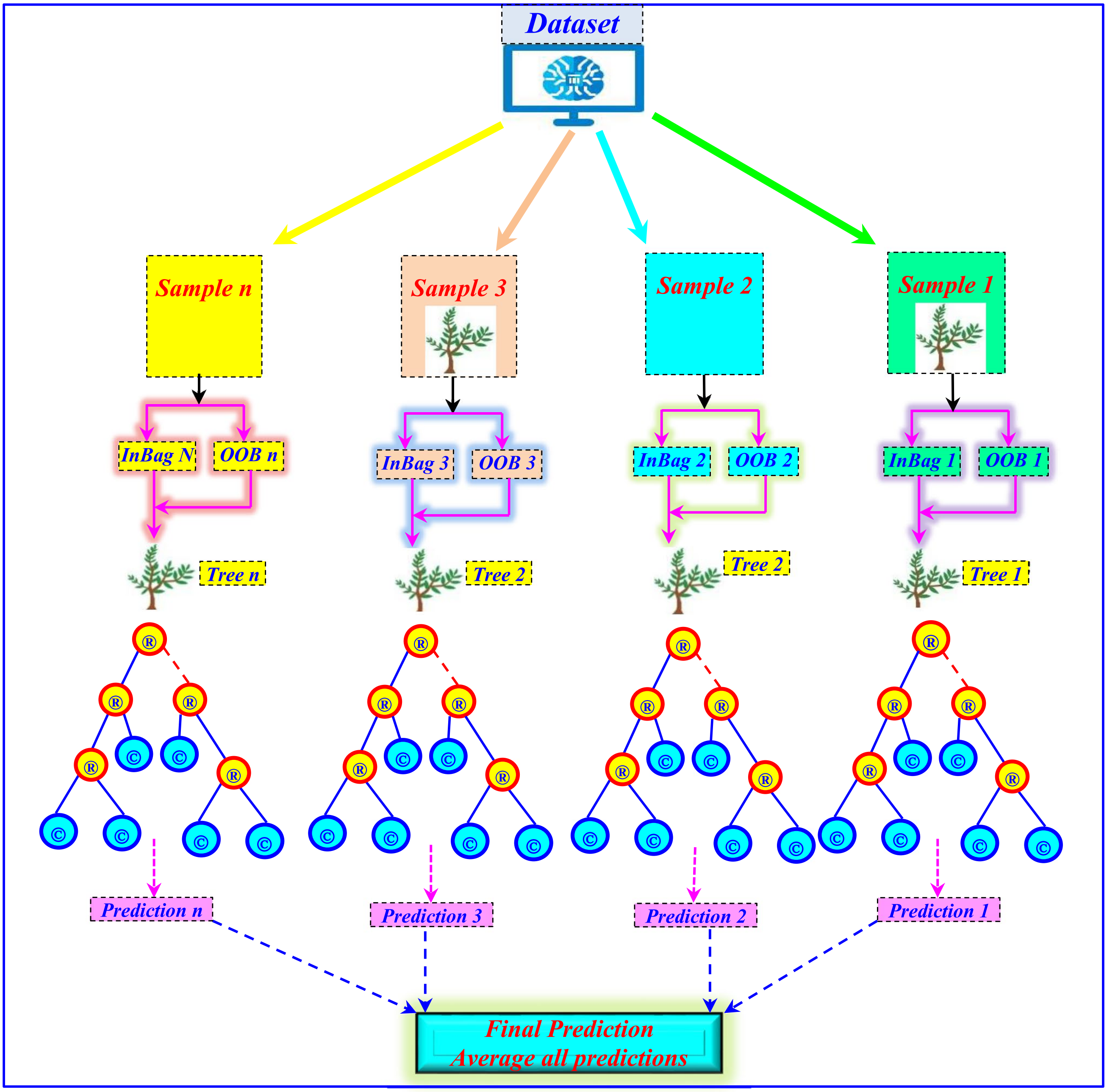
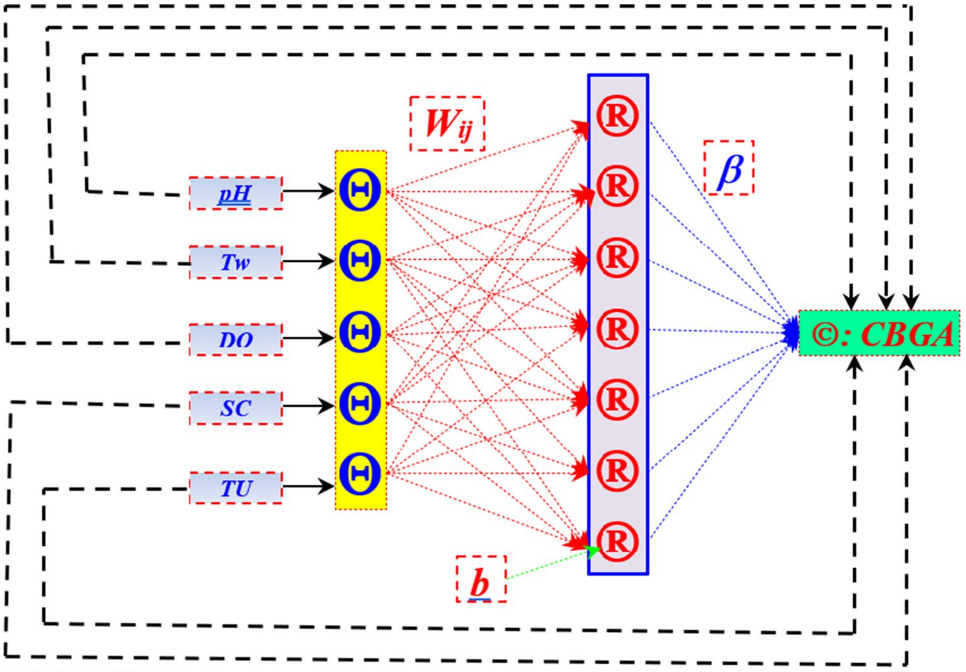


Fig. 4 Random forest regression (RFR) model architecture. The OOB is the out-of-bag

Signal decomposition methods

In the present study, three signal decomposition methods were used namely, (i) the empirical mode decomposition (EMD), (ii) the variational mode decomposition (VMD), and (iii) the empirical wavelet transform (EWT). Hereafter, we provide only a short description of each method without in-depth theoretical description. For more theoretical description, interested readers are referred to those published papers for a full mathematical formulation and for more in-depth information of each algorithm: (Huang et al. (1998) for the EMD algorithm description, Dragomiretskiy and Zosso (2014) for the VMD algorithm, and Gilles (2013) for the EWT algorithm.

Huang et al. (1998) proposed the empirical mode decomposition (EMD). The EMD decomposes nonlinear and nonstationary signal into a sum of subcomponents called intrinsic mode functions (IMFs) and a residue R_N using Hilbert transformation. The calculated IMFs were used as new input variables for the ML models. The variational mode decomposition (VMD) is a signal decomposition method introduced by Dragomiretskiy and Zosso (2014). The VMD uses an adaptive decomposition process for extracting a series of



Red dashed line: *Weights between Input and Hidden Neurons:* ***Randomly Generated***
Blue dashed line: *link with Weights between Input and output Neurons:* ***Adaptable Parameter.***
Black dashed line: *direct link with Weights between Input and output Neurons:* ***Adaptable Parameter.***
Hidden Neurons also called: ***Enhancement Nodes (ENs)***

Fig. 5 Random vector functional link (*RVFL*) neural network architecture

IMFs. The VMD is considered as an adaptive, quasi-orthogonal, and non-recursive decomposition method (Cannizzaro et al. 2021). The provided IMFs were sent back to the ML as new predictors variables. Finally, Gilles (2013) introduced the empirical wavelet transform (EWT). The EWT uses a robust algorithm to extract the subcomponent, i.e., the multiresolution analysis (MRA) components, by performing spectrum segmentation of the Fourier spectrum of the $x(t)$ into a set of segments (Wang and Hu 2015).

Gamma test input variable selection

In general, models based on ML paradigm have proved their efficacies and robustness. However, the implementation of the ML faces a number of challenges and clashes against several major difficulties, notably. Among a serious of problems, stemming the correct and effective use of ML is certainly the input variables selection (IVS), which becomes a challenging task. The idea behind the use of IVS is to select a suitable number of input variables among a large number of candidates. While several approaches have been proposed and available in the literature, in the present study, we use the famous gamma test method, simply called GT proposed and supported by Končar 1997; Stefánsson et al. (1997). Details and description of the GT algorithm is presented in the Text S1, and the obtained results are reported in Tables S1 and S2.

Performance assessment of the models

All models used in the present study calibrated during the training phase and their accuracies were evaluated using root-mean-square error (RMSE), mean absolute error (MAE), correlation coefficient (*R*), and Nash-Sutcliffe efficiency (NSE) (Yaseen 2021). Expressions are given as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(CBGA_{obs,i}) - (CBGA_{est,i})_i]^2}, (0 \le RMSE < +\infty) \tag{13}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|CBGA_{obs,i} - CBGA_{est,i}|, (0 \le MAE < +\infty) \tag{14}$$

$$R = \left[\frac{\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\left(CBGA_{obs,i} - \overline{CBGA_{obs}}\right)\left(CBGA_{est,i} - \overline{CBGA_{est}}\right)}{\sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{n}\left(CBGA_{obs,i} - \overline{CBGA_{obs}}\right)^2}\sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{n}\left(CBGA_{est,i} - \overline{CBGA_{est}}\right)^2}}\right], (-1 < R \le +1) \tag{15}$$

Table 2 The input combinations of different models

ELM	RVFL	ANN	RFR	Input combination	Output
ELM1	RVFL1	ANN1	RFR1	T_w, DO, pH, SC, TU	CBGA
ELM2	RVFL2	ANN2	RFR2	T_w, pH, SC, TU	CBGA
ELM3	RVFL3	ANN3	RFR3	T_w, DO, SC, TU	CBGA
ELM4	RVFL4	ANN4	RFR4	T_w, SC, TU	CBGA
ELM5	RVFL5	ANN5	RFR5	T_w, pH, SC	CBGA
ELM6	RVFL6	ANN6	RFR6	T_w, SC	CBGA
ELM7	RVFL7	ANN7	RFR7	T_w, TU	CBGA

$$NSE = 1 - \left[\frac{\sum_{i=1}^{N} \left[CBGA_{obs} - CBGA_{est} \right]^2}{\sum_{i=1}^{N} \left[CBGA_{obs,i} - \overline{CBGA_{obs}} \right]^2} \right], (-\infty < NSE \leq 1) \tag{16}$$

$\overline{CBGA}_{obs}$ and $\overline{CBGA}_{est}$ are the mean measured, and mean forecasted $\boldsymbol{CBGA}$, respectively, $\boldsymbol{CBGA}_{obs}$ and $\boldsymbol{CBGA}_{est}$ specify the observed and forecasted cyanobacteria blue-green algae for ith observations, and N shows the number of data points.

Modeling development

Determination of the best input combination using GT algorithm

The purpose of this work is to develop predictive models for an accurate estimation of CBGA concentration based on several water quality variables. We refer to these inputs variables as water T_w, DO, pH, SC, and TU. Because large number of predictors leads to high number of possible input combination (i.e., 2^5–1 = 31), meaningful input combination, it is necessary to use an input variables selection strategy to select the most significant input that contains the most predictive information. In this study, the GT algorithm was applied for each station separately and the obtained results are reported in Tables

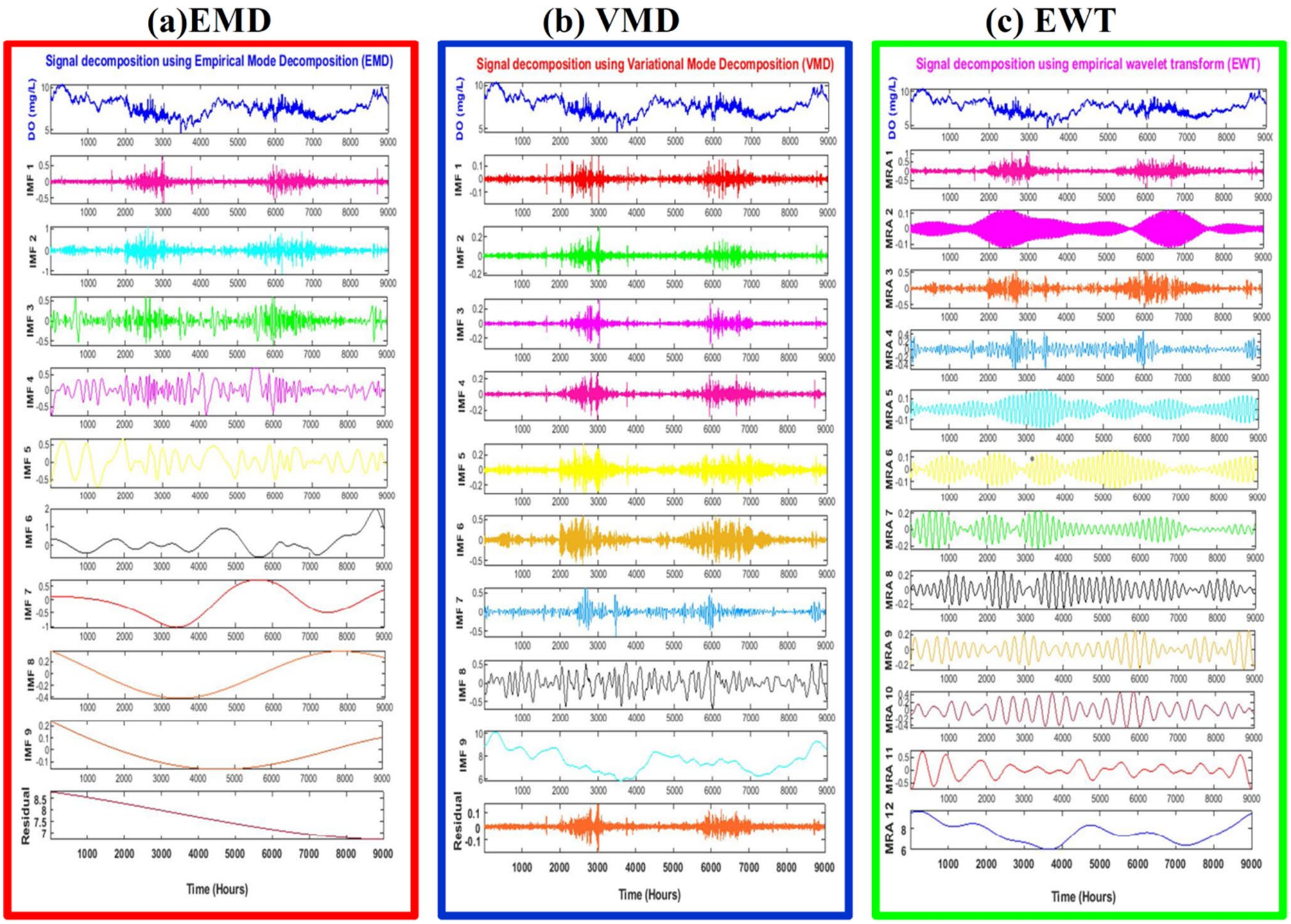


Fig. 6 Resulting intrinsic mode functions (IMFs) and multiresolution analysis (MRA) components for one quality variable using **a** the empirical model decomposition (EMD), **b** the variational mode decomposition (VMF), and **c** empirical wavelet transform (EWT)

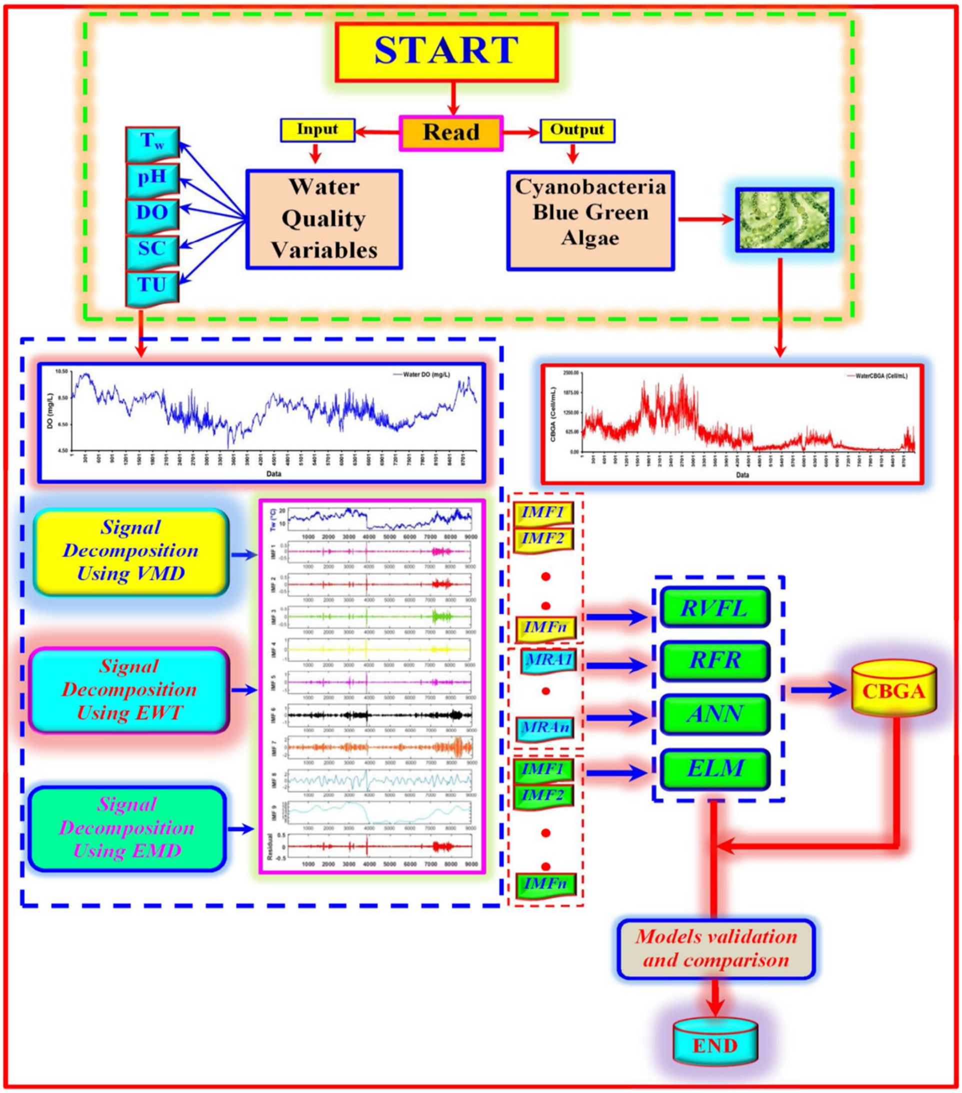


Fig. 7 Flowchart of the proposed modelling framework for cyanobacteria blue-green algae prediction

S1–S2 using a full embedding, i.e., examining all possible input combinations (i.e., 31). The input variable combinations were reported as a "Mask," for which the value (*1*) means that the variable was included and the value (0) means that the variable was excluded. According to Table S1–S2, the influence of the five water quality variables on CBGA was evaluated, and it is clear that the first input combination corresponds to the Mask (11111) and it is selected as the best input variable combination and determined by observing the gamma (Γ) value. Hence, our strategy was to compare between several models having several input combinations, i.e., five, four, three, and two input variables. It is important to note that increasing the number of omitted variable leads to an increase of the gamma (Γ) value.

Using only four input variables, it is clear from Table S1 and S2 that the Mask (11011) is suitable for the two stations, for which the water pH was excluded and the T_w, DO, SC, and TU were selected as the relevant input variables. A second Mask was selected and corresponds to the (10111) for which DO was omitted and T_w, pH, SC, and TU were selected as the relevant input variables. Subsequently, the input combination based on only three input variables were also analysed and it is shown that the Mask (10011) is suitable for the two stations (Tables S1 and S2), and the choice should be made among another Mask, and we have selected the Mask (10110) was adopted. Finally, using only two input variables, we have selected two Masks (10010) and (10001), with respect to the statistical values reported in Tables S1

Table 3 Performances of different standalone models at the USGS 14202650 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	*R*	NSE	RMSE	MAE
ELM1	0.756	0.571	3724.046	2219.280	0.741	0.549	3784.016	2254.217
ELM2	0.776	0.602	3588.702	2177.305	0.751	0.563	3725.382	2231.259
ELM3	0.747	0.558	3782.586	2159.998	0.729	0.531	3860.679	2181.597
ELM4	0.729	0.531	3894.050	2179.729	0.713	0.509	3951.264	2219.505
ELM5	0.619	0.383	4469.077	2483.625	0.591	0.349	4547.912	2500.810
ELM6	0.496	0.246	4938.190	2791.742	0.437	0.178	5111.601	2814.819
ELM7	0.595	0.354	4573.296	2434.374	0.540	0.289	4752.398	2530.584
RVFL1	0.777	0.604	3579.656	2085.049	0.759	0.575	3673.355	2131.864
RVFL2	0.792	0.627	3473.904	2048.362	0.765	0.585	3630.485	2124.916
RVFL3	0.771	0.595	3620.124	2030.167	0.749	0.561	3734.190	2074.307
RVFL4	0.752	0.566	3746.968	2029.327	0.736	0.542	3815.839	2092.055
RVFL5	0.633	0.401	4403.269	2429.416	0.595	0.353	4533.783	2477.578
RVFL6	0.553	0.305	4740.788	2544.218	0.468	0.190	5074.883	2606.888
RVFL7	0.611	0.374	4501.689	2292.672	0.497	0.208	5015.713	2431.017
ANN1	0.741	0.548	3824.826	2164.916	0.728	0.528	3873.537	2175.145
ANN2	0.744	0.554	3799.886	2044.296	0.723	0.522	3896.126	2108.762
ANN3	0.738	0.540	3856.448	2196.682	0.716	0.508	3955.319	2222.072
ANN4	0.752	0.565	3750.753	2079.567	0.731	0.535	3845.899	2155.935
ANN5	0.626	0.392	4434.440	2321.133	0.618	0.382	4429.797	2296.433
ANN6	0.540	0.292	4785.949	2634.546	0.517	0.267	4825.720	2584.096
ANN7	0.626	0.392	4436.373	2332.856	0.581	0.338	4587.742	2426.060
RFR1	0.972	0.940	1393.107	549.011	0.944	0.884	1923.787	762.349
RFR2	0.964	0.925	1554.692	587.973	0.930	0.861	2099.472	800.156
RFR3	0.970	0.935	1445.419	571.553	0.932	0.861	2097.951	821.071
RFR4	0.933	0.858	2143.568	910.691	0.878	0.756	2784.587	1198.149
RFR5	0.884	0.740	2900.974	1321.575	0.828	0.656	3306.579	1463.450
RFR6	0.789	0.610	3553.996	1689.790	0.704	0.493	4013.104	1881.958
RFR7	0.820	0.659	3322.150	1555.781	0.643	0.409	4333.611	2012.144

and S2. Finally, the retained models' structure is reported in Table 2 and it is important to note that among the five water quality variable, water T_w was included into all input combination (i.e., 7 combinations), followed by water SC which was included into six input combination, water TU was included into five input combination, while DO and pH were the input variables having the less significant contribution.

Models' configuration

CBGA was modelled using four machines learning, i.e., ELM, ANN, RFR, and RVFL models. First, the four models were applied and compared according to the input variable combinations reported in Table 2, and the obtained results were discussed and deeply analysed for each station separately; thus, during this first part of the investigation, the models were designated as single models. Second, to improve the performances of the single models, a new modelling framework was proposed and based on combining single ML with signal decomposition algorithms, i.e., the EMD, VMD, and the EWT, and the new models were designated as hybrid models, i.e., ELM_EMD, ELM_VMD, and ELM_EWT. The overall procedure of the second stage of the investigation was achieved by dividing the original water quality signal, i.e., pH, T_w, DO, SC and TU into a number of individual subseries, i.e., the IMF using the EMD and VMD, and the MRA using the EWT. An example of obtained IMF and MRA for one quality variable, i.e., the DO concentration, is shown in Fig. 6. Hence, the new subseries were used as new inputs variables. In order to demonstrate the effectiveness of the proposed model approaches, in the next section, the predictive performances of the proposed methods were presented,

Table 4 Performances of hybrid models based on EMD at the USGS 14202650 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	*R*	NSE	RMSE	MAE
ELM_EMD1	0.994	0.987	644.488	422.250	0.973	0.946	1313.095	802.604
ELM_EMD2	0.992	0.985	704.269	447.748	0.972	0.944	1339.559	829.133
ELM_EMD3	0.993	0.987	659.622	435.769	0.971	0.941	1368.372	859.778
ELM_EMD4	0.969	0.939	1410.203	952.895	0.944	0.890	1869.931	1205.565
ELM_EMD5	0.981	0.963	1098.431	699.261	0.963	0.927	1521.912	929.303
ELM_EMD6	0.978	0.956	1198.320	761.604	0.945	0.891	1860.696	1086.612
ELM_EMD7	0.993	0.986	665.845	424.681	0.960	0.920	1599.109	903.677
RVFL_EMD1	0.896	0.802	2529.747	1667.964	0.882	0.778	2655.929	1731.709
RVFL_EMD2	0.904	0.817	2431.898	1591.195	0.891	0.793	2563.386	1641.212
RVFL_EMD3	0.881	0.776	2690.292	1777.901	0.873	0.761	2754.160	1809.043
RVFL_EMD4	0.863	0.744	2880.468	1853.009	0.856	0.733	2911.785	1872.922
RVFL_EMD5	0.901	0.812	2465.227	1640.989	0.892	0.795	2553.041	1690.006
RVFL_EMD6	0.866	0.749	2848.195	1802.260	0.855	0.730	2929.980	1843.650
RVFL_EMD7	0.845	0.713	3046.617	2075.063	0.834	0.696	3108.749	2093.451
ANN_EMD1	0.992	0.984	711.822	370.580	0.989	0.977	851.589	496.193
ANN_EMD2	0.991	0.981	778.167	405.638	0.985	0.970	981.589	572.009
ANN_EMD3	0.992	0.984	713.404	380.000	0.988	0.976	864.837	491.387
ANN_EMD4	0.989	0.978	835.863	449.556	0.983	0.966	1043.880	637.653
ANN_EMD5	0.987	0.975	905.708	460.619	0.981	0.962	1098.774	596.446
ANN_EMD6	0.975	0.949	1286.431	666.125	0.967	0.935	1439.661	763.967
ANN_EMD7	0.988	0.976	881.872	490.599	0.983	0.965	1048.478	580.265
RFR_EMD1	0.997	0.994	455.424	195.892	0.971	0.943	1346.478	538.713
RFR_EMD2	0.997	0.994	457.959	198.301	0.970	0.940	1379.663	555.325
RFR_EMD3	0.997	0.994	454.314	197.417	0.970	0.941	1368.145	554.676
RFR_EMD4	0.997	0.993	462.284	201.682	0.965	0.932	1473.354	592.718
RFR_EMD5	0.997	0.993	466.121	205.298	0.966	0.932	1470.332	592.917
RFR_EMD6	0.997	0.993	474.844	209.029	0.960	0.921	1587.074	647.103
RFR_EMD7	0.997	0.993	476.591	207.763	0.978	0.956	1183.160	535.149

analysed, and discussed. The flowchart of the proposed modelling framework is shown in Fig. 7.

Prediction results and analysis

USGS 14202650 station

Four error indexes were employed to validate the performances of the proposed models and to evaluate the predictive accuracies of the CBGA, i.e., the RMSE, MAE, *R*, and NSE indexes. Seven input combination were evaluated and compared and the prediction results using single models are presented in Table 3 for the USGS 14202650 station. Hereafter, only the results during the validation stage are presented and discussed. According to Table 3, the four aforementioned models yielded different performances ranging from very poor predictive accuracy to excellent predictive accuracy. From Table 3, it can be found that: (i) numerical results for the ANN, ELM, and RFVL show that all models may no yielded satisfactory results and none of them was able to accurately and effectively predict CBGA concentration. In addition, we find that increasing the number of input variables from two to five does not help in improving the models performances. Indeed, the mean *R* and NSE values calculated using the ELM models were ≈0.643 and ≈0.424, respectively, showing the limited performances of the ELM models. In addition, high mean RMSE and MAE were obtained using the ELM models with values of ≈4247.61(cells/mL) and ≈2390.40(cells/mL). Among the seven input combinations, i.e., ELM1 to ELM7, the high R (≈0.751), and NSE (≈0.563) values were obtained using the ELM2 for which DO was omitted from the input variables. The performances of ELM2 were slightly higher than those

Table 5 Performances of hybrid models based on VMD at the USGS 14202650 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	*R*	NSE	RMSE	MAE
ELM_VMD1	0.730	0.533	3889.033	2625.000	0.562	0.278	4789.549	3200.234
ELM_VMD2	0.688	0.474	4126.800	2714.739	0.525	0.232	4941.580	3268.259
ELM_VMD3	0.711	0.506	3998.456	2574.238	0.525	0.224	4966.578	3221.449
ELM_VMD4	0.696	0.484	4085.491	2631.810	0.518	0.219	4981.886	3249.790
ELM_VMD5	0.593	0.351	4581.836	2900.268	0.364	0.004	5624.692	3504.735
ELM_VMD6	0.505	0.256	4907.964	3037.953	0.283	0.127	5985.237	3720.452
ELM_VMD7	0.667	0.445	4237.913	2658.036	0.439	0.118	5295.082	3279.287
RVFL_VMD1	0.583	0.340	4622.337	2948.430	0.583	0.340	4621.544	2960.903
RVFL_VMD2	0.792	0.627	3473.904	2048.362	0.533	0.284	4812.460	3060.406
RVFL_VMD3	0.771	0.595	3620.124	2030.167	0.562	0.316	4704.845	2867.417
RVFL_VMD4	0.752	0.566	3746.968	2029.327	0.525	0.276	4841.616	3029.019
RVFL_VMD5	0.633	0.401	4403.269	2429.416	0.450	0.203	5078.705	3054.987
RVFL_VMD6	0.553	0.305	4740.788	2544.218	0.413	0.170	5180.897	3101.944
RVFL_VMD7	0.611	0.374	4501.689	2292.672	0.522	0.272	4852.350	2977.028
ANN_VMD1	0.979	0.959	1152.938	797.367	0.982	0.963	1087.129	748.496
ANN_VMD2	0.967	0.934	1457.740	934.100	0.965	0.931	1490.826	943.076
ANN_VMD3	0.979	0.959	1151.094	789.243	0.974	0.949	1283.820	882.770
ANN_VMD4	0.958	0.918	1628.581	1024.296	0.959	0.920	1607.665	975.427
ANN_VMD5	0.839	0.704	3095.363	1858.944	0.767	0.582	3675.997	2263.905
ANN_VMD6	0.720	0.513	3971.501	2299.887	0.689	0.474	4125.546	2404.928
ANN_VMD7	0.882	0.778	2679.757	1524.059	0.883	0.779	2676.106	1511.628
RFR_VMD1	0.997	0.992	507.592	224.967	0.997	0.992	507.592	224.967
RFR_VMD2	0.964	0.925	1554.692	587.973	0.996	0.992	518.836	231.308
RFR_VMD3	0.970	0.935	1445.419	571.553	0.996	0.990	566.032	248.897
RFR_VMD4	0.933	0.858	2143.568	910.691	0.995	0.987	645.156	290.285
RFR_VMD5	0.884	0.740	2900.974	1321.575	0.992	0.977	860.470	354.871
RFR_VMD6	0.789	0.610	3553.996	1689.790	0.989	0.967	1027.514	461.803
RFR_VMD7	0.820	0.659	3322.150	1555.781	0.991	0.973	942.880	436.983

of the ELM1 and ELM3, while the ELM6 having only water T_w and pH as input variables was the poorest model showing very poor predictive accuracy. Results obtained the RVFL and ANN models were relatively equal to those obtained using the ELM models with negligible differences demonstrating the real limitations of the models to accurately predict the CBGA. The mean *R* and NSE values obtained using the RVFL and ANN were ≈0.653, ≈0.431, ≈0.659, and ≈0.440, respectively. In addition, the means RMSE and MAE were very high exhibiting the caps of ≈4211.61(cells/mL), ≈2276.94(cells/mL), ≈4202.02(cells/mL), and ≈2281.21(cells/mL); indeed, the comparison between the ELM, ANN, and RVFL models is not as obvious.

Comparisons of the overall results obtained using the RFR with respect to the *R*, NSE, RMSE, and MAE revealed interesting finding. From the results reported in Table 3, the mean *R* and NSE values obtained using the RFR models were ≈0.837 and ≈0.703, respectively, showing improvement rates of about ≈30.14% and ≈65.76%, compared to the values obtained using the ELM models, and improvement rates of about ≈28.23% and ≈63.23%, compared to the values obtained using the RVFL models, and ≈26.98% and ≈59.74% compared to the ANN models, respectively. Thus, the comparisons demonstrate that modelling CBGA using the RFR is more effective that the other models and only the RFR was able to provide an acceptable and robust predictive accuracy. For further highlighting the superiority of the RFR models, we provide a term-by-term comparison of the RMSE and MAE errors and we found that the RFR improves the mean RMSE and MAE of the ELM, RVFL, and ANN models by ≈30.85% and ≈46.57%, ≈30.25% and ≈43.91%, and ≈30.10% and ≈44.01%, respectively. From the seven RFR models (Table 3), it is clear that the

Table 6 Performances of hybrid models based on EWT at the USGS 14202650 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	*R*	NSE	RMSE	MAE
ELM_EWT1	0.993	0.986	673.100	454.585	0.980	0.959	1137.207	788.185
ELM_EWT2	0.993	0.986	663.265	439.916	0.975	0.950	1258.937	831.618
ELM_EWT3	0.994	0.987	640.278	431.335	0.977	0.953	1217.009	796.498
ELM_EWT4	0.994	0.989	601.068	379.798	0.973	0.946	1309.420	816.295
ELM_EWT5	0.994	0.988	616.020	388.213	0.968	0.936	1429.369	870.072
ELM_EWT6	0.994	0.989	599.768	375.476	0.938	0.870	2031.542	1106.553
ELM_EWT7	0.995	0.989	595.060	376.998	0.982	0.964	1062.231	720.912
RVFL_EWT1	0.724	0.524	3925.903	2778.922	0.705	0.496	4002.203	2848.155
RVFL_EWT2	0.640	0.409	4373.054	3041.736	0.599	0.358	4516.229	3088.598
RVFL_EWT3	0.662	0.438	4263.271	2879.771	0.629	0.395	4384.247	2926.445
RVFL_EWT4	0.646	0.417	4342.258	3023.918	0.628	0.394	4386.850	3029.546
RVFL_EWT5	0.542	0.293	4782.286	3101.718	0.507	0.257	4859.247	3148.378
RVFL_EWT6	0.585	0.341	4615.839	3008.928	0.562	0.316	4662.909	3056.566
RVFL_EWT7	0.609	0.370	4513.428	3050.766	0.570	0.325	4632.946	3108.969
ANN_EWT1	0.988	0.975	891.379	591.896	0.983	0.967	1027.648	709.931
ANN_EWT2	0.965	0.930	1499.617	1073.255	0.961	0.923	1564.811	1137.849
ANN_EWT3	0.987	0.974	911.002	619.762	0.983	0.965	1057.018	757.996
ANN_EWT4	0.993	0.986	667.734	360.429	0.990	0.979	807.412	502.223
ANN_EWT5	0.993	0.985	686.651	372.868	0.988	0.976	865.358	539.564
ANN_EWT6	0.989	0.977	856.605	529.356	0.984	0.967	1019.300	690.072
ANN_EWT7	0.990	0.981	784.495	489.536	0.986	0.972	943.812	639.421
RFR_EWT1	0.997	0.994	450.718	200.783	0.989	0.976	866.691	389.270
RFR_EWT2	0.997	0.993	459.578	203.705	0.984	0.966	1041.408	426.202
RFR_EWT3	0.997	0.994	453.484	202.117	0.989	0.976	865.080	388.734
RFR_EWT4	0.997	0.993	469.933	206.888	0.980	0.958	1157.757	456.720
RFR_EWT5	0.997	0.993	472.779	212.686	0.972	0.940	1385.062	565.853
RFR_EWT6	0.997	0.993	482.622	215.867	0.971	0.936	1422.422	555.555
RFR_EWT7	0.996	0.991	526.521	231.071	0.987	0.971	954.844	419.747

RFR1 was the strongest model slightly higher than the RFR2 and RFR3, and highly than the RFR4 and RFR5, while the RFR7 was the poorest model showing the lowest predictive accuracy. As expected, we found that using all input variables (i.e., RFR1) yielded the best performances and more accurate than the models having less input variables, and using only two input variables is not suitable for predicting CBGA.

In the second part of the present study, we use the combined models based on EMD, VMD, and the EWT signal decomposition algorithms to decompose the original water quality signal and then employ the obtained subsignal as new input variables. All these models are based on the same input structure. In total, three modelling strategies are tested and compared as shown in Tables 4, 5 and 6. From Table 4, based on the EMD decomposition, it is clear that all models improve their performances and all hybrid models performed best compared to the single models, showing the high contribution of the EMD in improving the models accuracies.

From Table 4, it is shown that the EMD approach achieves high decreases in MAE and RMSE, and high increase in *R* and NSE values in comparing with the single models. Compared to the single ELM models, the ELM_EMD obtains reductions of approximately ≈63.43% and ≈60.45% in terms of means RMSE and MAE, and an increase of approximately ≈49.44% and ≈117.62% in terms of means *R* and NSE, respectively. It is clear that the achievement in terms of NSE was the most notable and the most remarkable exceeding the rate of 100%. Similarly, the RVFL_EMD models improve the performances of the performances of the single RVFL models by decreasing the means RMSE and MAE by ≈33.92% and ≈20.43%, respectively, while the

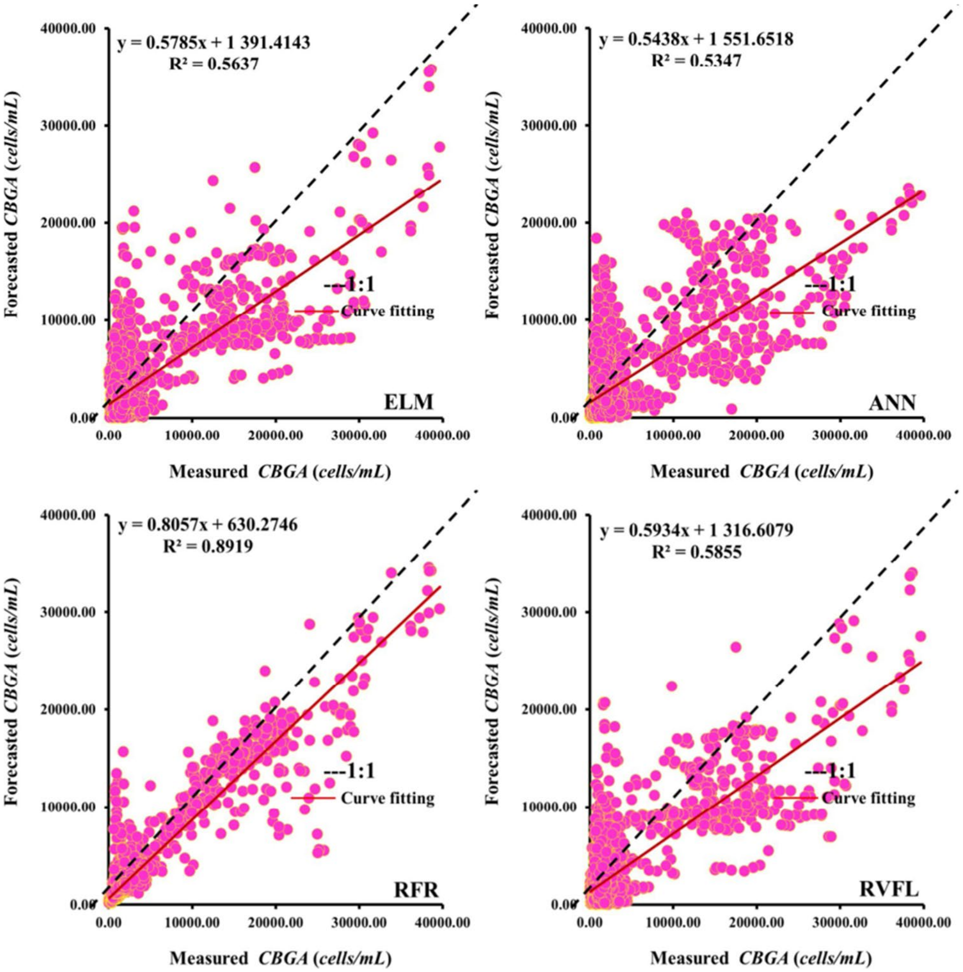
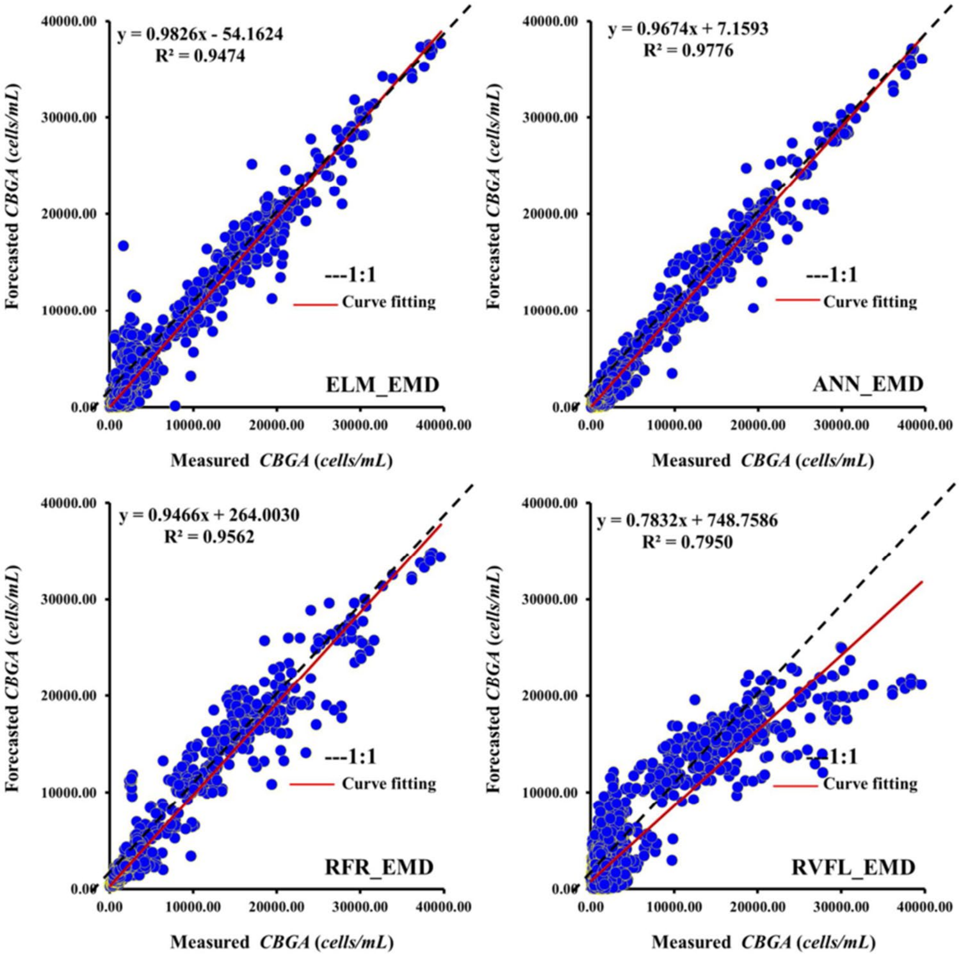


Fig. 8 Scatterplots of measured against predicted (*CBGA*) using single models for the validation stage: USGS 14202650 station

mean *R* and NSE values were improved by ≈33.13% and ≈75.92%, respectively. In addition, ANN_EMD models yielded improvement rates of approximately ≈75.08% and ≈74.08%, in terms of means RMSE and MAE, and increase the mean *R* and NSE values by ≈49.02% and ≈119.18%, respectively, exhibiting the high improvements rates among all proposed hybrid models. Finally, the RFR_EMD models have also contributed to significant improvements rates of the single RFR models performance metrics, for which the means RMSE and MAE were decreased by ≈52.29% and ≈55.06%, respectively, while the mean R and NSE values were improved by ≈15.71% and ≈33.42%, respectively. From the results reported in Table 4, it is clear that the ELM_EMD and RFR_EMD models are quite alike and none of them was able to significantly surpassed the other exhibiting negligible differences in terms of models performances. Comparing all models reported in Table 4, the ANN_EMD models outperform all other models and yielded the high mean *R* and NSE values and the lowest mean RMSE and MAE values, and among the seven input combination, i.e., the ANN_EMD1 to ANN_EMD7, it is shown that the first five models (i.e., ANN_EMD1 to ANN_EMD5) have a value of *R* and NSE higher than 0.980 and 0.960, respectively. Further comparison between the hybrid models based on EMD signal decomposition revealed that (i) even with the EMD, all models have shown their numerical performances significantly improved, the ANN_EMD were the most models in terms of improvement rates, while the RVFL_EMD models are those on which the improvement was less sensitive, and (ii) taking into account fewer input variables, it is interestingly shown in Table 4 that ANN_EMD7 and RFR_EMD7 were very accurate and exhibiting very high predictive accuracy; indeed, RFR_EMD7 was the best accurate random forest model with *R* and NSE values of approximately ≈0.978 and ≈0.956, respectively, while the ANN_EMD7 was remarkably interesting model exhibiting very

Fig. 9 Scatterplots of measured against predicted (*CBGA*) using hybrid models based on empirical mode decomposition (EMD) for the validation stage: USGS 14202650 station

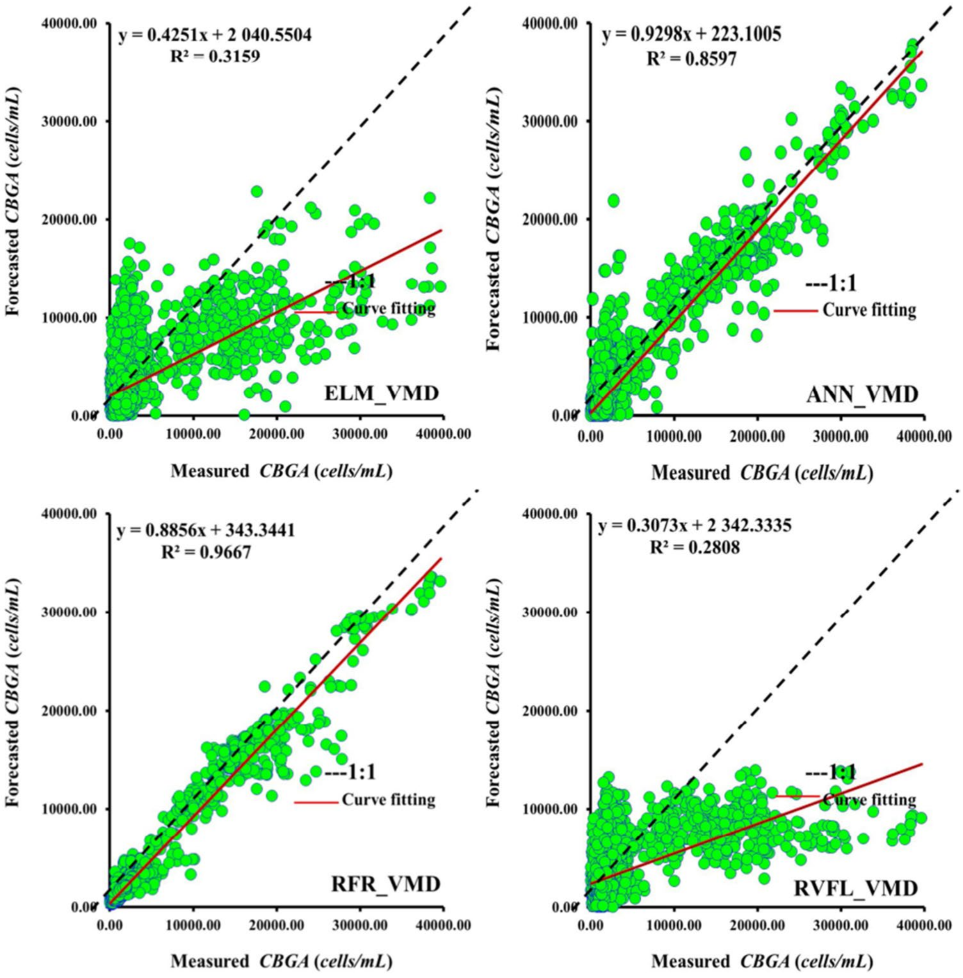


high *R* and NSE values of approximately ≈0.983 and ≈0.965, respectively.

Figure 8 illustrates the scatterplot of measured and predicted CBGA using single models, and it is clear that the RFR model was the only model for which the data were less scattered compared to the ELM, ANN, and RVFL models. Similarly, Fig. 9 illustrates the scatterplot of measured and predicted CBGA using hybrid models based on the EMD decomposition algorithm, and it is clear that the ANN_EMD model was the only model for which the data were less scattered compared to the ELM_EMD, RFR_EMD, and RVFL_EMD models.

The performances of all models based on variational model decomposition (VMD) are shown in Table 5. The analysis of the results in Table 5 revealed some important finding. First, it is clear that the predictive performances of the ELM and RVFL models degrade significantly with the use of the VMD algorithm. It can be observed that the measured CBGA was poorly fitted to the calculated data showing high RMSE and MAE values and very poor *R* and NSE values, and compared to the single model, the mean values of all performance metrics were deteriorated demonstrating the limitations of the VMD algorithm in improving the performances of these two kind of ML models. According to Table 5, the mean RMSE and MAE of the single ELM and RVFL models were increased by ≈18.72% and ≈28.62%, respectively, while the mean *R* and NSE values were decreased by ≈40% and ≈146.92%, respectively. In contrast to the poor results obtained using the ELM and RVFL models, the performances of the single ANN and RFR were significantly improved using the VMD algorithm, and more precisely, the obtained results using the RFR_VMD were very strong. More precisely, it can be observed from Table 5 that the RFR_VMD models were able to maintain a high means *R*, an NSE values, and yielded best performances compared to the all other models. An outstanding means *R* and NSE of approximately ≈0.994 and ≈0.983 were obtained using

Fig. 10 Scatterplots of measured against predicted (*CBGA*) using hybrid models based on variational mode decomposition (VMD) for the validation stage: USGS 14202650 station



the RFR_VMD, and a slightly difference between the seven input combination was very small, and the *R* and NSE values were ranged between ≈0.989 to ≈0.997 and ≈0.967 to ≈0.992, respectively. However, regarding the ANN_VMD models, it is clear that the ANN_VMD5 and ANN_VMD6 were not able to significantly improve their accuracies showing a very limited improvement rates for which the *R* and NSE values does not surpassed the values of ≈0.770 and ≈0.580, respectively. Overall, using the VMD algorithm, the best performances were obtained using the RFR_VMD1 (*R* ≈ 0.997 and NSE ≈ 0.992) and followed by the ANN_VMD1 (*R* ≈ 0.997, NSE ≈ 0.992). Figure 10 illustrates the scatterplot of measured and predicted CBGA using hybrid models based on the VMD decomposition algorithm, and it is clear that the RFR_VMD model was the only model for which the data were less scattered compared to the ANN_VMD, and it is clear that the ELM_VMD and RVFL_VMD models were very poor and showing high scattered data with very low R^2 values.

The model performance among the different models based on the EWT is displayed in Table 6, in which the RMSE, MAE, *R*, and NSE are calculated and displayed. Generally speaking, the proposed ANN_EWT method performs best equally with RFR_EWT, following by the ELM_EWT, while the RVFL_EWT was failed to improve its performances showing very poor numerical indexes. Specifically, the ANN_EWT shows the minimal average for RMSE and MAE, which decreases by ≈75.23% and ≈68.83% compared with the original single ANN with respect to the seven input combinations, respectively. The average reduction of RMSE and MAE are ≈62.58% and ≈64.18% of the RFR_EWT models, which are less than the values of the ANN_EWT. Moreover, the proposed ELM_EWT models also help to decrease the means RMSE and MAE of the

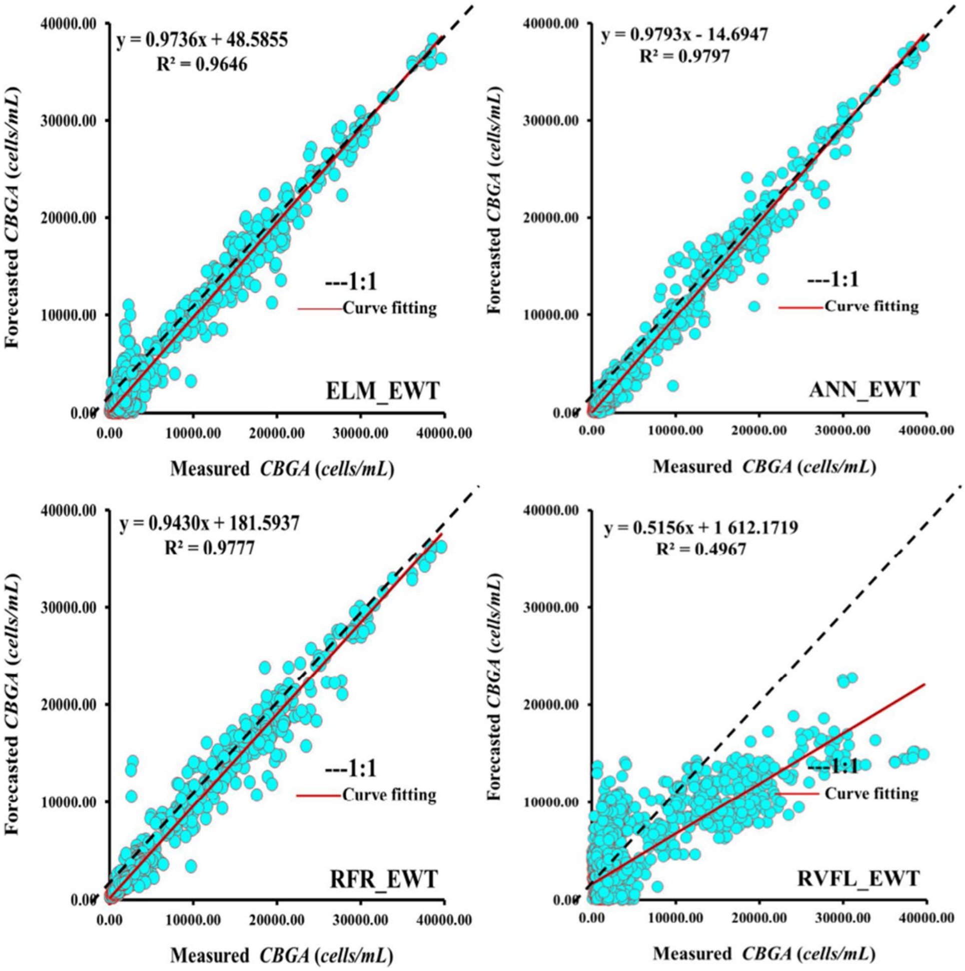


Fig. 11 Scatterplots of measured against predicted (*CBGA*) using hybrid models based on empirical wavelet transform (EWT) for the validation stage: USGS 14202650 station

original single ELM models by ≈68.23% and ≈64.56%, respectively. All of these findings indicate that the proposed EWT decomposition algorithm was found to be an effective and robust decomposition algorithm

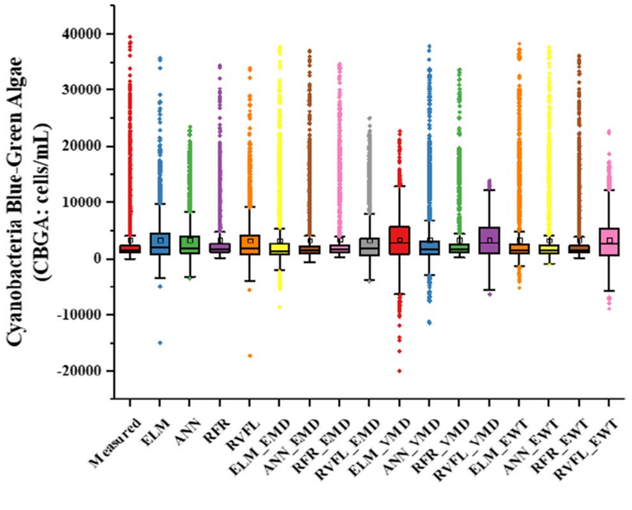


Fig. 12 Boxplots of measured and calculated river cyanobacteria blue-green algae (*CBGA*: Cells/mL) at the USGS 14202650 (validation stage)

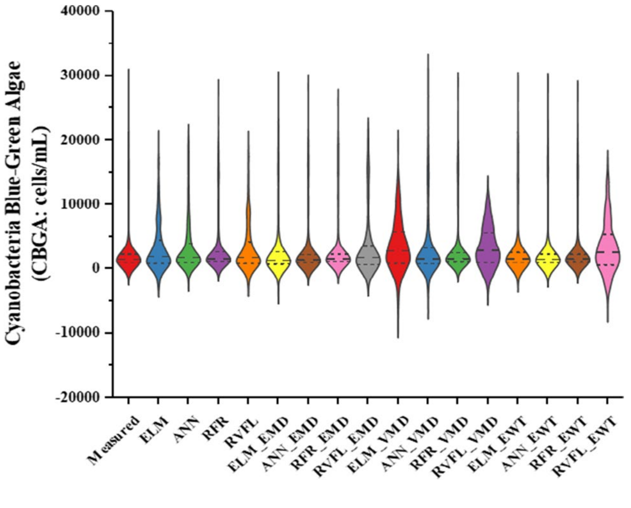


Fig. 13 Violin plot showing distributions of the measured and calculated river cyanobacteria blue-green algae (*CBGA*: Cells/mL) at the USGS 14202650 (validation stage)

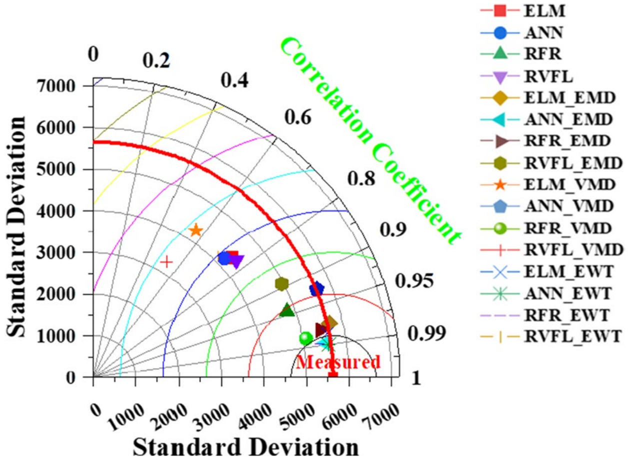


Fig. 14 Taylor diagram of cyanobacteria blue-green algae (*CBGA*: cells/mL) illustrating the statistics of comparison between the proposed models at the USGS 14202650 (validation stage)

leading to significant predictive accuracy of the CBGA concentration.

Finally, according to the all results reported above (Tables 3, 4, 5 and 6), we can conclude that (i) RFR1 is the most accurate model among all single models, and the difference between its performances and the ANN1, RVFL1, and ELM1 is very larger; thus, neither ANN nor ELM or RVFL are suitable for CBGA prediction, and (ii) among the hybrid models, the ANN_EMD1 is the most accurate model based on EMD signal decomposition, the RFR_VMD1 is the most accurate model based on VDM signal decomposition, and ANN_EWT4 is the most accurate model based on EWT signal decomposition, and in overall and based on the above experiments, the proposed RFR_VMD1 was the most accurate and possesses a more powerful predictive ability than all other models and they produced the lowest RMSE and MAE values of ≈507.59 (cells/mL) and ≈224.96 (cells/mL), respectively, and the

Table 7 Performances of different standalone models at the USGS 14207200 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	*R*	NSE	RMSE	MAE
ELM1	0.836	0.700	246.971	173.017	0.820	0.673	265.556	186.322
ELM2	0.798	0.636	271.691	187.037	0.795	0.632	281.709	196.877
ELM3	0.801	0.642	269.764	195.344	0.776	0.602	292.945	210.425
ELM4	0.766	0.586	289.853	214.812	0.744	0.554	310.323	230.466
ELM5	0.607	0.369	358.042	276.572	0.614	0.379	366.143	280.149
ELM6	0.751	0.564	297.405	218.304	0.727	0.529	318.820	235.594
ELM7	0.532	0.283	381.689	304.322	0.529	0.281	393.904	315.353
RVFL1	0.841	0.707	244.069	169.279	0.825	0.680	262.760	184.621
RVFL2	0.803	0.644	268.710	185.623	0.789	0.623	285.018	201.143
RVFL3	0.808	0.653	265.530	194.519	0.783	0.613	289.020	210.965
RVFL4	0.771	0.594	287.017	210.827	0.748	0.561	307.888	225.622
RVFL5	0.794	0.630	274.087	188.786	0.718	0.482	334.421	256.216
RVFL6	0.754	0.569	295.850	216.128	0.727	0.529	318.857	233.507
RVFL7	0.581	0.335	367.511	294.394	0.585	0.340	377.348	302.543
ANN1	0.802	0.642	269.467	186.366	0.797	0.637	280.031	196.149
ANN2	0.775	0.600	285.121	198.068	0.763	0.582	300.154	210.364
ANN3	0.780	0.608	282.264	206.777	0.754	0.568	305.281	222.929
ANN4	0.735	0.540	305.670	228.682	0.718	0.515	323.341	242.533
ANN5	0.750	0.562	298.202	212.565	0.735	0.541	314.779	226.563
ANN6	0.716	0.513	314.447	235.632	0.701	0.493	330.881	250.762
ANN7	0.611	0.373	356.891	280.040	0.622	0.388	363.281	284.544
RFR1	0.965	0.926	122.558	76.969	0.914	0.833	189.616	117.897
RFR2	0.940	0.878	157.611	96.530	0.870	0.755	229.776	144.469
RFR3	0.957	0.910	135.379	85.974	0.880	0.773	221.275	138.509
RFR4	0.909	0.814	194.183	131.352	0.805	0.646	276.323	189.475
RFR5	0.843	0.681	254.545	184.468	0.803	0.625	284.315	207.159
RFR6	0.873	0.756	222.522	151.461	0.750	0.562	307.487	209.776
RFR7	0.809	0.645	268.502	192.731	0.661	0.438	348.309	250.652

Table 8 Performances of hybrid models based on EMD at the USGS 14207200 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	*R*	NSE	RMSE	MAE
ELM_EMD1	0.964	0.930	119.096	86.349	0.950	0.901	146.443	104.061
ELM_EMD2	0.970	0.941	109.135	78.296	0.947	0.896	149.737	105.582
ELM_EMD3	0.968	0.937	113.462	82.823	0.950	0.901	146.392	104.288
ELM_EMD4	0.967	0.934	115.468	80.894	0.940	0.883	158.740	109.869
ELM_EMD5	0.972	0.944	106.171	75.009	0.954	0.908	140.564	98.636
ELM_EMD6	0.963	0.927	121.653	85.290	0.940	0.883	158.834	109.211
ELM_EMD7	0.961	0.924	124.244	88.476	0.931	0.865	170.434	115.479
RVFL_EMD1	0.943	0.889	150.194	104.863	0.936	0.876	163.569	113.807
RVFL_EMD2	0.939	0.882	155.008	109.343	0.931	0.866	169.828	117.752
RVFL_EMD3	0.941	0.886	152.129	107.055	0.933	0.871	166.980	118.324
RVFL_EMD4	0.931	0.866	164.807	117.326	0.924	0.854	177.514	125.823
RVFL_EMD5	0.931	0.868	163.946	117.913	0.925	0.855	176.980	125.241
RVFL_EMD6	0.923	0.852	173.435	122.918	0.915	0.838	187.095	130.417
RVFL_EMD7	0.926	0.857	170.309	119.071	0.919	0.845	182.880	127.382
ANN_EMD1	0.979	0.959	91.753	62.902	0.964	0.929	124.011	83.732
ANN_EMD2	0.974	0.949	101.297	68.031	0.961	0.922	129.327	86.649
ANN_EMD3	0.978	0.956	94.966	64.920	0.965	0.931	121.755	82.667
ANN_EMD4	0.971	0.943	107.681	71.715	0.958	0.918	133.104	88.366
ANN_EMD5	0.972	0.944	106.739	70.646	0.960	0.920	131.586	84.964
ANN_EMD6	0.966	0.933	116.732	77.241	0.954	0.910	138.996	90.674
ANN_EMD7	0.968	0.937	112.943	74.734	0.955	0.911	138.181	89.825
RFR_EMD1	0.990	0.981	62.793	40.049	0.959	0.920	131.330	84.357
RFR_EMD2	0.990	0.980	64.223	40.994	0.958	0.917	134.159	86.458
RFR_EMD3	0.990	0.980	63.143	40.305	0.959	0.920	131.549	84.918
RFR_EMD4	0.990	0.979	64.933	41.453	0.956	0.914	136.003	87.776
RFR_EMD5	0.989	0.978	67.384	43.033	0.955	0.911	138.746	89.563
RFR_EMD6	0.988	0.977	68.444	43.745	0.954	0.909	140.101	91.092
RFR_EMD7	0.989	0.978	67.186	42.854	0.955	0.911	138.562	89.221

higher *R* and NSE values of ≈0.997 and ≈0.992, respectively. Figure 11 illustrates the scatterplot of measured and predicted CBGA using hybrid models based on the EWT decomposition algorithm, and it is clear that the RFR_EWT model was the only model for which the data were less scattered compared to the ANN_EWT, and it is clear that the ELM_EWT and RVFL_EWT models were very poor and showing high scattered data with very low R^2 values. The boxplot, violin plot, and Taylor diagram for all developed models at the USGS 14202650 were depicted in Figs. 12, 13 and 14, showing the superiority of one model compared to the other models and the improvement gained using the decomposition algorithms is clearly presented.

USGS 14207200 station

The obtained results for the USGS 14207200 station are reported in Tables 7, 8, 9 and 10. According to Table 7, using single models, the predictive accuracy was ranged from poor to moderate and only one model was found to be accurate, i.e., the RFR1. Results indicate that the RFR models were more accurate than the ANN, ELM, and RVFL models, exhibiting the high means *R* (≈0.812) and NSE (≈0.662) values, and the lowest mean RMSE (≈265.30) and MAE (≈179.70) values, respectively. It can be clearly seen from Table 7 that the single RFR1 model can accurately predict the CBGA with very satisfactory performances exhibiting the high *R* (≈0.914) and

Table 9 Performances of hybrid models based on VMD at the USGS 14207200 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	*R*	NSE	RMSE	MAE
ELM_VMD1	0.679	0.461	330.779	258.986	0.540	0.271	396.618	303.383
ELM_VMD2	0.729	0.532	308.366	239.597	0.567	0.293	390.512	296.720
ELM_VMD3	0.711	0.506	316.849	249.294	0.544	0.269	397.123	302.370
ELM_VMD4	0.733	0.537	306.586	240.706	0.547	0.258	400.202	307.308
ELM_VMD5	0.736	0.541	305.183	236.394	0.568	0.288	392.008	299.396
ELM_VMD6	0.725	0.526	310.162	241.655	0.581	0.309	386.194	296.662
ELM_VMD7	0.644	0.414	344.879	270.341	0.530	0.267	397.638	306.388
RVFL_VMD1	0.641	0.410	346.051	278.682	0.633	0.402	359.136	288.938
RVFL_VMD2	0.624	0.390	351.995	279.731	0.601	0.362	371.040	292.209
RVFL_VMD3	0.650	0.422	342.626	269.978	0.638	0.409	357.199	280.505
RVFL_VMD4	0.634	0.401	348.645	274.954	0.626	0.393	362.009	282.945
RVFL_VMD5	0.592	0.350	363.211	284.759	0.573	0.330	380.203	297.444
RVFL_VMD6	0.603	0.363	359.528	280.040	0.589	0.348	374.941	290.652
RVFL_VMD7	0.599	0.358	361.005	285.463	0.586	0.344	376.164	292.147
ANN_VMD1	0.979	0.958	92.037	68.714	0.922	0.846	182.361	126.400
ANN_VMD2	0.964	0.930	119.321	88.316	0.887	0.774	221.042	152.266
ANN_VMD3	0.972	0.945	105.806	80.550	0.910	0.823	195.368	136.284
ANN_VMD4	0.942	0.888	150.787	112.681	0.847	0.702	253.753	172.118
ANN_VMD5	0.952	0.906	138.022	102.135	0.873	0.752	231.329	158.197
ANN_VMD6	0.924	0.854	172.392	127.454	0.843	0.700	254.202	182.575
ANN_VMD7	0.922	0.850	174.236	129.092	0.822	0.667	268.186	193.930
RFR_VMD1	0.992	0.983	59.187	37.045	0.954	0.907	141.822	91.934
RFR_VMD2	0.991	0.982	61.191	38.706	0.949	0.895	150.805	98.743
RFR_VMD3	0.991	0.982	61.027	38.377	0.952	0.901	146.021	94.659
RFR_VMD4	0.990	0.979	65.301	41.984	0.945	0.886	156.709	103.081
RFR_VMD5	0.990	0.979	65.940	42.142	0.936	0.869	168.055	106.699
RFR_VMD6	0.988	0.973	73.738	48.105	0.924	0.847	181.416	116.109
RFR_VMD7	0.987	0.970	78.481	52.193	0.925	0.844	183.493	124.344

NSE (≈0.833) values, and the lowest RMSE (≈189.61) and MAE (≈117.89) values, respectively, but beyond the RFR1 model, there is still a decreasing trend from RFR1 to RFR7 for which the errors metrics between the measured and predicted CBGA concentration were becoming very large. Results indicate that ANN, ELM, and RVFL models were relatively equal showing negligible difference, and all were less accurate compared to the RFR models. RFVL model gave slightly lower mean RMSE (≈310.75) and MAE (≈230.66) values compared to the values obtained using ANN models (RMSE ≈ 316.82, MAE ≈233.40), and the values obtained using ELM models (RMSE ≈ 318.48, MAE ≈ 236.45). Taking into account the number of inputs variables, it can be seen from Table 7 that the best performances for all models were obtained using the first input combination based on the five water quality variables. Figure 15 illustrates the scatterplot of measured and predicted CBGA using single models for USGS 14207200 station, and it is clear that the RFR model was the only model for which the data were less scattered compared to the ANN, ELM, and RVFL models.

Table 8 gives the comparison results between the hybrid models based on the EMD signal decomposition. According to Table 8, it is clear that all single models have shown their performances significantly improved using the EMD algorithm. When comparing the single models with hybrid models, it is remarkable that (i) using the EMD, the means *R*, NSE, RMSE, and MAE of the single ELM models were improved by ≈32.10%, ≈70.87%, ≈51.95%, and ≈54.86%, respectively; (ii) the means *R*, NSE, RMSE, and MAE of the single RVFL models were improved by ≈25.27%, ≈56.87%,

Table 10 Performances of hybrid models based on EWT at the USGS 14207200 station

Models	Training				Validation			
	R	NSE	RMSE	MAE	*R*	NSE	RMSE	MAE
ELM_EWT1	0.981	0.963	87.035	64.182	0.946	0.889	154.628	109.085
ELM_EWT2	0.965	0.932	117.482	87.152	0.943	0.885	157.435	116.521
ELM_EWT3	0.954	0.909	135.671	102.795	0.932	0.865	170.771	129.393
ELM_EWT4	0.967	0.936	114.212	83.823	0.942	0.885	157.452	115.169
ELM_EWT5	0.966	0.934	116.177	85.850	0.940	0.879	161.600	116.254
ELM_EWT6	0.969	0.940	110.760	79.874	0.952	0.904	143.977	105.207
ELM_EWT7	0.972	0.944	106.755	76.752	0.947	0.890	153.735	109.380
RVFL_EWT1	0.711	0.505	316.905	249.226	0.694	0.483	333.966	263.074
RVFL_EWT2	0.719	0.516	313.383	244.207	0.712	0.507	326.128	252.464
RVFL_EWT3	0.716	0.513	314.471	248.214	0.707	0.501	328.264	257.540
RVFL_EWT4	0.770	0.592	287.747	224.240	0.752	0.567	305.668	237.694
RVFL_EWT5	0.702	0.493	320.885	254.348	0.687	0.473	337.127	265.218
RVFL_EWT6	0.713	0.508	316.008	243.664	0.705	0.498	329.203	255.222
RVFL_EWT7	0.729	0.530	308.767	239.515	0.726	0.528	319.064	248.454
ANN_EWT1	0.986	0.972	76.062	53.843	0.959	0.916	134.286	92.522
ANN_EWT2	0.982	0.963	86.320	60.688	0.957	0.910	139.354	95.801
ANN_EWT3	0.984	0.968	80.970	57.889	0.962	0.924	128.251	88.832
ANN_EWT4	0.979	0.958	92.693	64.654	0.960	0.922	129.998	90.003
ANN_EWT5	0.979	0.958	91.989	63.379	0.960	0.921	130.615	90.603
ANN_EWT6	0.971	0.943	107.151	71.659	0.960	0.922	130.108	87.441
ANN_EWT7	0.971	0.943	107.348	73.486	0.958	0.915	135.358	93.061
RFR_EWT1	0.989	0.978	66.110	42.449	0.938	0.875	164.490	103.867
RFR_EWT2	0.989	0.977	67.720	43.465	0.932	0.863	171.823	105.674
RFR_EWT3	0.989	0.978	66.757	42.850	0.945	0.888	155.189	100.594
RFR_EWT4	0.988	0.977	69.009	44.434	0.941	0.882	159.318	100.691
RFR_EWT5	0.988	0.976	69.777	44.891	0.960	0.919	131.989	86.224
RFR_EWT6	0.988	0.975	71.191	46.096	0.962	0.924	128.208	84.366
RFR_EWT7	0.988	0.975	70.717	45.595	0.912	0.830	191.622	112.548

≈43.69%, and ≈46.81%, respectively; (iii) the means *R*, NSE, RMSE, and MAE of the single ANN models were improved by ≈31.96%, ≈72.96%, ≈58.65%, and ≈62.85%, respectively; and (iv) the means *R*, NSE, RMSE, and MAE of the single RFR models were improved by ≈17.82%, ≈38.21%, ≈48.82%, and ≈51.23%, respectively. In addition, it is clear that among the four hybrid models, the most significant improvement was gained by the ANN_EMD slightly higher than the RFR_EMD and much higher than the RVFL_EMD and ELM_EMD models. When comparing the ANN_EMD models with the ELM_EMD models, the predictive accuracy was higher than the later. For instance, promoting percentages of the means *R*, NSE, RMSE, and MAE by the ANN_EMD are ≈1.58%, ≈3.27%, ≈14.39%, and ≈18.77%, respectively. Similarly, the promoting percentages of the means *R*, NSE, RMSE, and MAE by the ANN_EMD compared to the RVFL_EMD models are ≈3.60%, ≈7.26%, ≈25.13%, and ≈29.33%, respectively. Finally, the promoting percentages of the means *R*, NSE, RMSE, and MAE by the ANN_EMD compared to the RFR_EMD models are ≈0.314%, ≈0.61%, ≈3.52%, and ≈1.06%, respectively, demonstrating that in overall the ANN_EMD and RFR_EMD were relatively equal. In conclusion, while all models have benefited from a significant improvement rate using the EMD algorithm, obtained results clearly indicate that adding the EMD to the single models is solid and credible way to improve the predictive accuracy of the CBGA. Figure 16 illustrates the scatterplot of measured and predicted CBGA using hybrid models based on the EMD algorithm, and it is clear that the ANN_EMD model was the only model for which the data were less scattered compared to the RFR_EMD, ELM_EMD, and RVFL_EMD models.

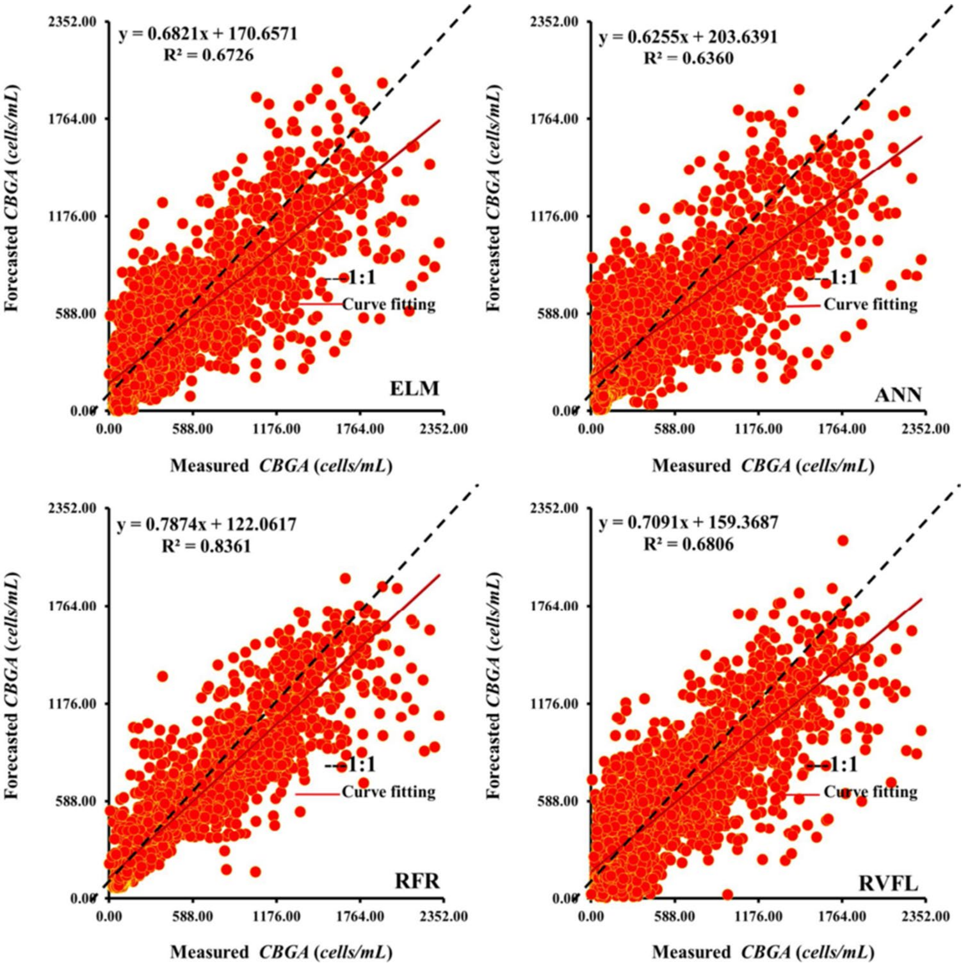


Fig. 15 Scatterplots of measured against predicted (*CBGA*) using single models for the validation stage: USGS 14207200 station

Results obtained using hybrid models based on VMD signal decomposition are reported in Table 9. Overall, two models have shown their performances significantly improved and the two other models were deteriorated and showing a significant decrease in calculated performances. First, combining the single ELM with the VMD algorithm, it is clear that the single ELM models show an average reduction in the means *R* and NSE values of approximately ≈29.09%, and ≈86.70%, respectively, and an increase of the means RMSE and MAE values of approximately ≈19.23%, and ≈21.63%, respectively. For the RVFL_VMD, there is an increase of the means RMSE and MAE values of approximately ≈15.70%, and ≈20.26%, respectively, and a significant decrease in the means *R* and NSE values of approximately ≈21.88%, and ≈47.92%, respectively. Consequently, this statement confirms what is already discussed in previous section regarding the obtained results at the USGS 14202650 station that the VMD algorithm cannot be considered an efficient algorithm for improving the performances of the single ELM and RVFL models.

Second, compared to the results of the single models reported in Table 7, using the VMD helps in significantly improving the performances of the single ANN and RFR models, for which the ANN_VMD improve the mean RMSE, MAE, *R*, and NSE values of the single ANN by ≈19.92%, ≈41.35%, ≈27.57%, and ≈31.34%, respectively. The RFR_VMD contributed significantly in the improvement of the mean RMSE, MAE, *R*, and NSE values of the single RFR by ≈15.87%, ≈32.75%, ≈39.24%, and ≈41.52%, respectively. Furthermore, the superiority of the RFR_VMD was clearly demonstrated, for which we can see that the RFR_VMD performs better than the ELM_VMD, RVFL_VMD and ANN_VMD exhibiting an improvement of the RMSE and MAE performance metrics of approximately ≈59.12% and ≈65.17%, ≈56.27%, ≈ 63.67%, and ≈29.75% and ≈34.42%, respectively. Figure 17 illustrates the scatterplot of measured and predicted CBGA using hybrid models based on the VMD algorithm, and it is clear that the RFR_VMD model was the only model for which the data were less scattered

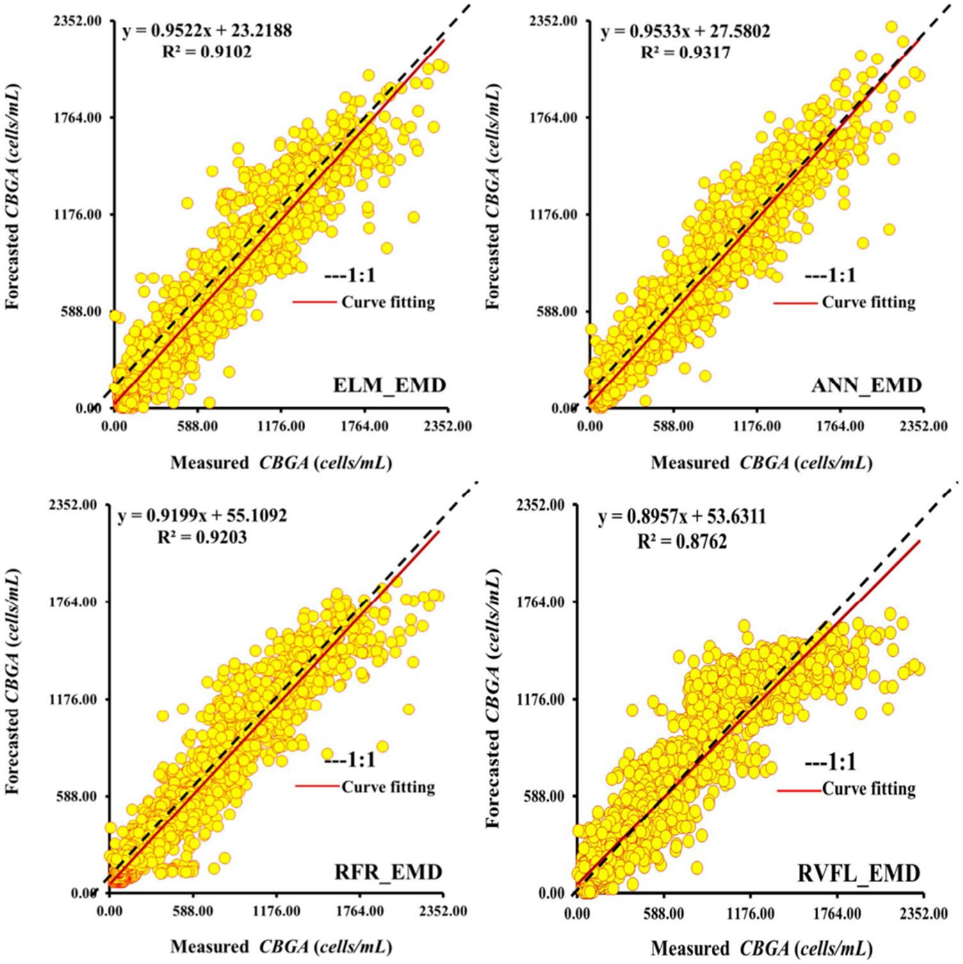
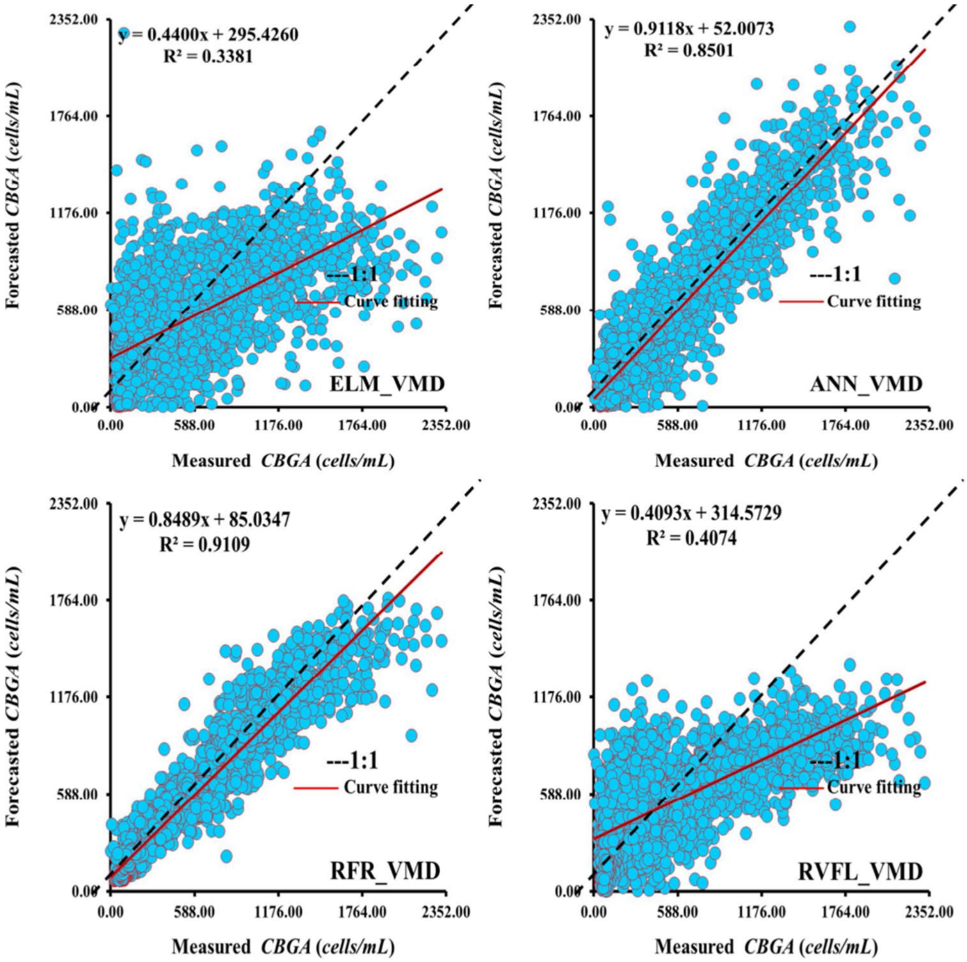


Fig. 16 Scatterplots of measured against predicted (*CBGA*) using hybrid models based on empirical mode decomposition (EMD) for the validation stage: USGS 14207200 station

compared to the ANN_EMD, ELM_EMD, and RVFL_EMD models.

Table 10 provides a comparative study between the hybrid models based on the EWT algorithms. The RMSE and MAE criterions using ANN_EWT model have an average value of only ≈132.65 and ≈91.18, the lowest of all. The RMSE and MAE enhancements between ANN_EWT and the other models are ≈15.61% and ≈20.32% compared to the ELM_EWT, ≈59.29% and ≈61.14% compared to the RVFL_EWT, ≈59.29% and ≈61.14% compared to the RVFL_EWT, and ≈15.84% and ≈8.06% compared to the RFR_EWT, respectively, which is significant. It is clear that all models except the RVFL_EWT have gained significant improvement in terms of numerical performances and the enhancements between ANN_EWT model and the single ANN regarding the RMSE and MAE are ≈58.15% and ≈60.93%, between ELM_EWT model and the single ELM regarding the RMSE and MAE are ≈50.67% and ≈51.60%, and the enhancements between RFR_EWT model and the single RFR regarding the RMSE and MAE are ≈40.62% and ≈44.83%, respectively, always above 40%, again significant. As we noted above for the USGS 14202650, the RVFL model was failed to improve its performances; the situation remains the same as the performances of the single RVFL were decreased using the EWT algorithm. Based on the predictive error results shown in (Tables 7, 8, 9, and 10), it can be observed that (a) for the single models, the RFR1 has obtained the minimum RMSE (≈189.61) and MAE (≈117.89), it achieves the maximal *R* value of 0.914, and it got the maximal NSE value of 0.833, respectively, and (b) the ANN_EMD3 clearly have higher prediction precision than all other hybrid models, i.e., RFR_VMD1 and RFR_EWT6. It denotes that the RVFL_VMD and RVFL_EWT perform worse than the other hybrid models. Figure 18 illustrates the scatterplot of measured and predicted CBGA using hybrid models based on the EWT algorithm, and it is clear that the ANN_EWT model was the only model for which the data were less scattered compared to the RFR_EMD, ELM_EMD, and RVFL_EMD

Fig. 17 Scatterplots of measured against predicted (*CBGA*) using hybrid models based on variational mode decomposition (VMD) for the validation stage: USGS 14207200 station



models. The boxplot, violin plot, and Taylor diagram for all developed models at the USGS 14207200 were depicted in Figs. 19, 20 and 21, showing the superiority of one model compared to the other models and prediction improvement.

Conclusion

This study uses water quality variables to construct robust model for predicting cyanobacteria blue-green algae (CBGA) concentration in river using data collected at two USGS stations. The study focused on establishing a direct link between CBGA and water pH, T_w, DO, TU, and SC using four machines learning, i.e., ANN, ELM, RFR, and RVFL models. In the lights of the results obtained, several conclusions can be drawn. Overall, it appears that ANN, RVFL, and ELM models cannot provide reasonable predictive relationships for CBGA using a variety of input variables combination involving low models' performances with high errors metrics. Conversely, using the RFR, the predictive accuracy has found to significantly increase, showing an excellent improvement in the model performances with R and NSE values reaching the cap of ≈ 0.944 and ≈ 0.884 for USGS 14202650 station and ≈ 0.914 and ≈ 0.833 for USGS 14207200 station. This first concluding remark is important and revealed that RFR which belong to the category of ensemble learning methods is more suitable for CBGA compared to the standalone ML methods, although additional validation data are required to perform and provide a more thorough validation analysis.

As the present study highlighted limits in the applicability of single ML models for CBGA prediction, a new modelling framework was proposed based on preprocessing signal decomposition. Hence, three signal decomposition algorithms were tested, and in overall, the EMD algorithm was found to be more suitable than the VMD and EWT algorithms. Using the EMD algorithm, the ANN model calibrated using the five water quality variables (i.e., water pH, T_w, SC, DO, and TU) was found to be more accurate and yielded high R and NSE values of approximately ≈ 0.989 and ≈ 0.977, followed by the ELM model with the values of ≈ 0.989 and ≈ 0.977, respectively, while the RFR and RVFL were ranked in the third and fourth place,

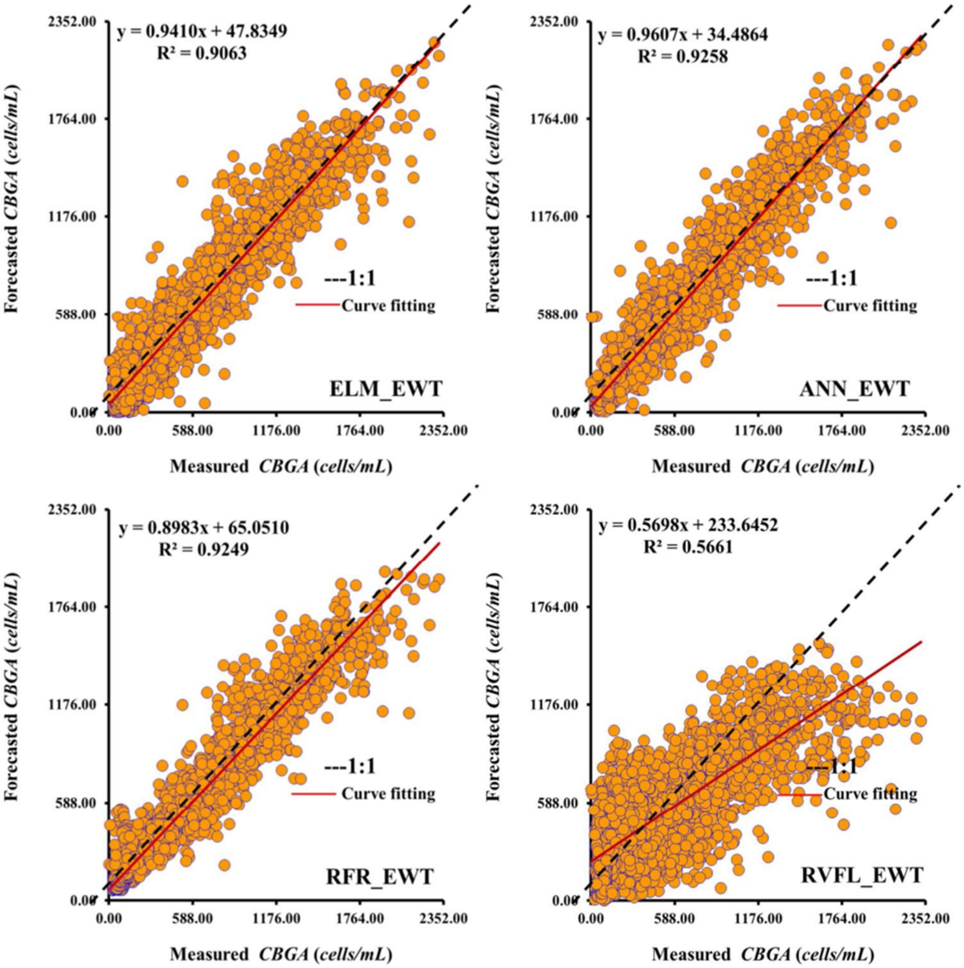


Fig. 18 Scatterplots of measured against predicted (*CBGA*) using hybrid models based on empirical wavelet transform (EWT) for the validation stage: USGS 14207200 station

respectively. Subsequently, the VMD was also used, and in overall, it was found that the improvement gained in models performances was less important compared to the EMD for the ANN, ELM, and RVFL models, with the exception of the combined RFR-VMD who was found to be more accurate compared to the RFR-EMD. Regarding the EWT

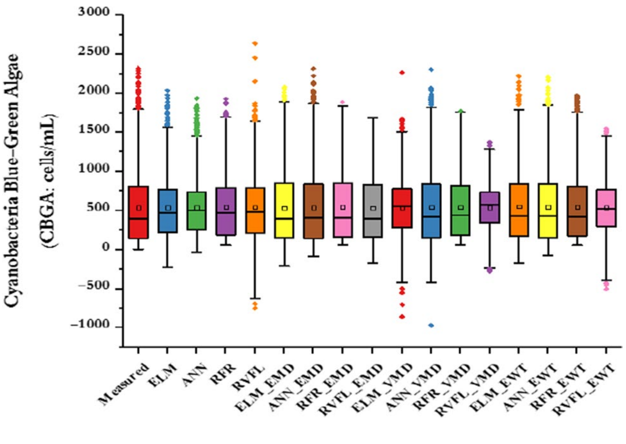


Fig. 19 Box-plots of measured and calculated river cyanobacteria blue-green algae (*CBGA*: *Cells/mL*) at the USGS 14207200 (validation stage)

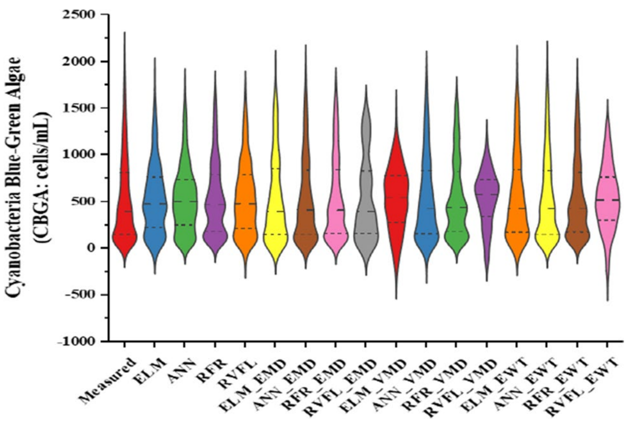


Fig. 20 Violin plot showing distributions of the measured and calculated river cyanobacteria blue-green algae (*CBGA*: *Cells/mL*) at the USGS 14207200 (validation stage)

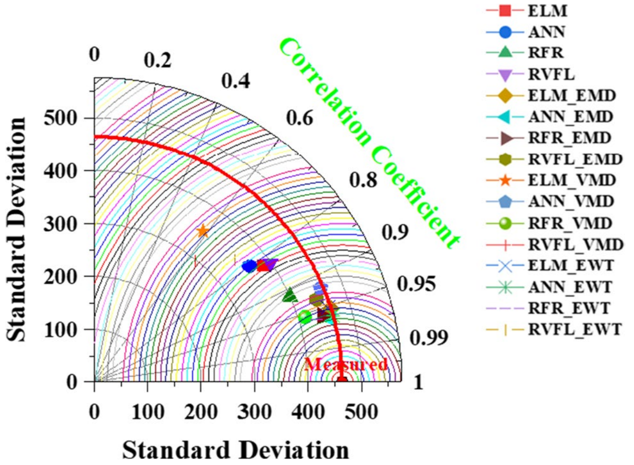


Fig. 21 Taylor diagram of cyanobacteria blue-green algae (*CBGA*: *Cells/mL*) illustrating the statistics of comparison between the proposed models at the USGS 14207200 (validation)

algorithm, it was found to be an excellent algorithm for improving the performances of the ML models and, CBGA estimation using the ELM, ANN, and RFR was significantly increased; on the contrary, the RVFL model does not beneficed to any improvement; then, further losses of predictive accuracy are set to continue, demonstrating the specificity of this kind of ML algorithm. In overall, using the EWT, high performances were obtained and the *R* and NSE values have reached the cap of ≈ 0.989 and ≈ 0.976 using the combined RFR and EWT, the cap of ≈ 0.986 and ≈ 0.972 using the combined ANN and EWT, and the values of ≈ 0.982 and ≈ 0.964 using the combined ELM and EWT, respectively.

In conclusion, the outstanding performances obtained in the present study show the robustness and the credibility of the proposed modelling framework based on the combined ML and signal decomposition. In the future, it is highly recommended to extend the present investigation to other location and using other water quality variables in order to investigate the efficacy of the proposed signal decomposition in improving the estimation of CBGA in river. It is also recommended applying other algorithms, i.e., wavelet transform and the complete ensemble empirical mode decomposition with adaptive noise.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-022-21201-1.

Acknowledgements The authors would like to reveal their appreciation and gratitude to the respected reviewers and handling editors for their constructive comments. The authors would like to thank Al-Mustaqbal University College for providing technical support for this research.

Author contribution Salim Heddam: conceptualization, modelling and software, project leader, writing, investigation. Zaher Mundher Yaseen: supervision, writing, investigation, analysis, visualization, revision, and edits. Mayadah W. Falah, Leonardo Goliatt, Mou Leong Tan, Zulfaqar Sa'adi, Iman Ahmadianfar, Mandeep Saggi, Amandeep Bhatia, Pijush Samui: writing, investigation, analysis, visualization, revision, and edits.

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

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

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

Competing interests The authors declare no competing interests.

Institutional review board statement Not applicable.

References

Adnan RM, Mostafa R, Kisi O et al (2021) Improving streamflow prediction using a new hybrid ELM model combined with hybrid particle swarm optimization and grey wolf optimization. Knowl-Based Syst 230:107379

Afan HA, El-Shafie A, Yaseen ZM et al (2014) ANN based sediment prediction model utilizing different input scenarios. Water Resour Manag 29:1231–1245. https://doi.org/10.1007/s11269-014-0870-1

Ahmadianfar I, Shirvani-Hosseini S, He J et al (2022) An improved adaptive neuro fuzzy inference system model using conjoined metaheuristic algorithms for electrical conductivity prediction. Sci Rep 12:1–34

Almodfer R, Zayed ME, Elaziz MA et al (2022) Modeling of a solar-powered thermoelectric air-conditioning system using a random vector functional link network integrated with jellyfish search algorithm. Case Stud Therm Eng 31:101797. https://doi.org/10.1016/j.csite.2022.101797

Araba AM, Memon ZA, Alhawat M et al (2021) Estimation at completion in civil engineering projects: review of regression and soft computing models. Knowl-Based Eng Sci 2:1–12

Asadollah SBHS, Sharafati A, Motta D, Yaseen ZM (2020) River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. J Environ Chem Eng. https://doi.org/10.1016/j.jece.2020.104599

Bano S, Burhan Z-U-N, Nadir M et al (2021) Removal efficiency of marine filamentous Cyanobacteria for Pyrethroids and their effects on the biochemical parameters and growth. Algal Res 60:102546. https://doi.org/10.1016/j.algal.2021.102546

Basilio SA, Goliatt L (2022) Gradient boosting hybridized with exponential natural evolution strategies for estimating the strength of geopolymer self-compacting concrete. Knowl-Based Eng Sci 3:1–16

Beretta-Blanco A, Carrasco-Letelier L (2021) Relevant factors in the eutrophication of the Uruguay River and the Río Negro. Sci Total Environ 761:143299. https://doi.org/10.1016/j.scitotenv.2020.143299

Bhagat SK, Tiyasha T, Tung TM et al (2020) Manganese (Mn) removal prediction using extreme gradient model. Ecotoxicol Environ Saf 204:111059. https://doi.org/10.1016/j.ecoenv.2020.111059

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) Random Forrests. Mach Learn

Cannizzaro D, Aliberti A, Bottaccioli L et al (2021) Solar radiation forecasting based on convolutional neural network and ensemble learning. Expert Syst Appl 181:115167. https://doi.org/10.1016/j.eswa.2021.115167

Cao W, Hu L, Gao J et al (2020) A study on the relationship between the rank of input data and the performance of random weight neural network. Neural Comput Applic 32:12685–12696. https://doi.org/10.1007/s00521-020-04719-8

Chauhan V, Tiwari A (2022) Randomized neural networks for multilabel classification. Appl Soft Comput 115:108184. https://doi.org/10.1016/j.asoc.2021.108184

Chen H, Huang Q, Lin Z, Tan C (2022) Detection of adulterants in medicinal products by infrared spectroscopy and ensemble of window extreme learning machine. Microchem J 173:107009. https://doi.org/10.1016/j.microc.2021.107009

Choi H, Han C, Antoniou MG (2021) Sustainable and green decomposition of cyanotoxins and cyanobacteria through the development of new photocatalytic materials. Curr Opin Green Sustain Chem 28:100444. https://doi.org/10.1016/j.cogsc.2020.100444

Clercin NA, Koltsidou I, Picard CJ, Druschel GK (2022) Prevalence of Actinobacteria in the production of 2-methylisoborneol and geosmin, over Cyanobacteria in a temperate eutrophic reservoir. Chem Eng J Adv 9:100226. https://doi.org/10.1016/j.ceja.2021.100226

Derot J, Yajima H, Jacquet S (2020) Advances in forecasting harmful algal blooms using machine learning models: a case study with Planktothrix rubescens in Lake Geneva. Harmful Algae 99:101906. https://doi.org/10.1016/j.hal.2020.101906

Descy J-P, Leprieur F, Pirlot S et al (2016) Identifying the factors determining blooms of cyanobacteria in a set of shallow lakes. Ecol Inform 34:129–138. https://doi.org/10.1016/j.ecoinf.2016.05.003

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

Elmetwalli AH, Mazrou YSA, Tyler AN et al (2022) Assessing the efficiency of remote sensing and machine learning algorithms to quantify wheat characteristics in the Nile Delta Region of Egypt. Agriculture. https://doi.org/10.3390/agriculture12030332

Elzwayie A, El-shafie A, Yaseen ZM et al (2016) RBFNN-based model for heavy metal prediction for different climatic and pollution conditions. Neural Comput Applic. https://doi.org/10.1007/s00521-015-2174-7

Fernández-Habas J, Carriere Cañada M, García Moreno AM et al (2022) Estimating pasture quality of Mediterranean grasslands using hyperspectral narrow bands from field spectroscopy by Random Forest and PLS regressions. Comput Electron Agric 192:106614. https://doi.org/10.1016/j.compag.2021.106614

Gaget V, Almuhtaram H, Kibuye F et al (2022) Benthic cyanobacteria: a utility-centred field study. Harmful Algae 113:102185. https://doi.org/10.1016/j.hal.2022.102185

García Nieto PJ, Alonso Fernández JR, García-Gonzalo E et al (2015) A new predictive model for the cyanotoxin content from experimental cyanobacteria concentrations in a reservoir based on the ABC optimized support vector machine approach: a case study in Northern Spain. Ecol Inform 30:49–59. https://doi.org/10.1016/j.ecoinf.2015.09.010

Giere J, Riley D, Nowling R et al (2020) An investigation on machine-learning models for the prediction of cyanobacteria growth. Fundam Appl Limnol 194:85–94

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

Guo J, Ma Y, Lee JHW (2021) Real-time automated identification of algal bloom species for fisheries management in subtropical coastal waters. J Hydro-Environ Res 36:1–32. https://doi.org/10.1016/j.jher.2021.03.002

Hai T, Sharafati A, Mohammed A et al (2020) Global solar radiation estimation and climatic variability analysis using extreme learning machine based predictive model. IEEE Access 8:12026–12042. https://doi.org/10.1109/ACCESS.2020.2965303

Harris TD, Graham JL (2017) Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset. Lake Reserv Manag 33:32–48. https://doi.org/10.1080/10402381.2016.1263694

Hazarika BB, Gupta D (2022) Random vector functional link with ε-insensitive Huber loss function for biomedical data classification. Comput Methods Prog Biomed 215:106622. https://doi.org/10.1016/j.cmpb.2022.106622

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 G-B, Zhu Q-Y, Siew C-K (2006) Extreme learning machine: theory and applications. Neurocomputing 70:489–501

Jafarzadeh N, Mirbagheri SA, Rajaee T et al (2022) Using artificial intelligent to model predict the biological resilience with an emphasis on population of cyanobacteria in Jajrood River in The Eastern Tehran, Iran. J Environ Heal Sci Eng. https://doi.org/10.1007/s40201-021-00760-4

Jamei M, Karbasi M, Malik A et al (2022) Long-term multi-step ahead forecasting of root zone soil moisture in different climates: novel ensemble-based complementary data-intelligent paradigms. Agric Water Manag 269:107679

Jha SK, Chishti Z, Ahmad Z, Arshad K-R (2022) Enterobacter sp. SWLC2 for biodegradation of chlorpyrifos in the aqueous medium: modeling of the process using artificial neural network approaches. Comput Electron Agric 193:106680. https://doi.org/10.1016/j.compag.2021.106680

Karimi B, Mohammadi P, Sanikhani H et al (2020) Modeling wetted areas of moisture bulb for drip irrigation systems: an enhanced empirical model and artificial neural network. Comput Electron Agric. https://doi.org/10.1016/j.compag.2020.105767

Khaleefa O, Kamel AH (2021) On the evaluation of water quality index: case study of Euphrates River, Iraq. Knowl-Based Eng Sci 2:35–43

Končar N (1997) Optimisation methodologies for direct inverse neurocontrol. University of London, London

Mahmudi M, Serihollo LG, Herawati EY et al (2020) A count model approach on the occurrences of harmful algal blooms (HABs) in Ambon Bay. Egypt J Aquat Res 46:347–353. https://doi.org/10.1016/j.ejar.2020.08.002

Maier HR, Dandy GC (1998) Understanding the behaviour and optimising the performance of back-propagation neural networks: an empirical study. Environ Model Softw 13:179–191. https://doi.org/10.1016/S1364-8152(98)00019-X

Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ Model Softw 15:101–124

Maier HR, Dandy GC, Burch MD (1998) Use of artificial neural networks for modelling cyanobacteria Anabaena spp. in the River Murray, South Australia. Ecol Modell 105:257–272. https://doi.org/10.1016/s0304-3800(97)00161-0

Maier HR, Sayed T, Lence BJ (2000) Forecasting cyanobacterial concentrations using B-spline networks. J Comput Civ Eng 14:183–189. https://doi.org/10.1061/(asce)0887-3801(2000)14:3(183)

Nguyen HQ, Ha NT, Pham TL (2020) Inland harmful cyanobacterial bloom prediction in the eutrophic Tri An Reservoir using satellite band ratio and machine learning approaches. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-019-07519-3

Oboh IO, Offor UH, Okon ND (2022) Artificial neural network modeling for potential performance enhancement of a planar perovskite solar cell with a novel TiO_2/SnO_2 electron transport bilayer using nonlinear programming. Energy Rep 8:973–988. https://doi.org/10.1016/j.egyr.2021.12.010

Onyelowe KC, Gnananandarao T, Ebid AM (2022) Estimation of the erodibility of treated unsaturated lateritic soil using support vector machine-polynomial and -radial basis function and random forest regression techniques. Clean Mater 3:100039. https://doi.org/10.1016/j.clema.2021.100039

Ostfeld A, Tubaltzev A, Rom M et al (2015) Coupled data-driven evolutionary algorithm for toxic cyanobacteria (blue-green algae) forecasting in Lake Kinneret. J Water Resour Plan Manag 141:4014069. https://doi.org/10.1061/(asce)wr.1943-5452.0000451

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

Park Y, Lee HK, Shin J-K et al (2021) A machine learning approach for early warning of cyanobacterial bloom outbreaks in a freshwater reservoir. J Environ Manag 288:112415. https://doi.org/10.1016/j.jenvman.2021.112415

Paul T, Vainio S, Roning J (2022) Detection of intra-family coronavirus genome sequences through graphical representation and artificial neural network. Expert Syst Appl 194:116559. https://doi.org/10.1016/j.eswa.2022.116559

Pyo J, Cho KH, Kim K et al (2021) Cyanobacteria cell prediction using interpretable deep learning model with observed, numerical, and sensing data assemblage. Water Res 203:117483. https://doi.org/10.1016/j.watres.2021.117483

Recknagel F, Cao H, Kim B et al (2006) Unravelling and forecasting algal population dynamics in two lakes different in morphometry and eutrophication by neural and evolutionary computation. Ecol Inform 1:133–151. https://doi.org/10.1016/j.ecoinf.2006.02.004

Rosecrans CZ, Belitz K, Ransom KM et al (2022) Predicting regional fluoride concentrations at public and domestic supply depths in basin-fill aquifers of the western United States using a random forest model. Sci Total Environ 806:150960. https://doi.org/10.1016/j.scitotenv.2021.150960

Rousso BZ, Bertone E, Stewart RA et al (2022) Automation of species-specific cyanobacteria phycocyanin fluorescence compensation using machine learning classification. Ecol Inform 2022:101669

Saboe D, Ghasemi H, Gao MM et al (2021) Real-time monitoring and prediction of water quality parameters and algae concentrations using microbial potentiometric sensor signals and machine learning tools. Sci Total Environ 764:142876. https://doi.org/10.1016/j.scitotenv.2020.142876

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

Sanikhani H, Deo RC, Samui P et al (2018) Survey of different data-intelligent modeling strategies for forecasting air temperature using geographic information as model predictors. Comput Electron Agric 152:242–260

Sanseverino I, Pretto P, António DC et al (2022) Metagenomics analysis to investigate the microbial communities and their functional profile during cyanobacterial blooms in Lake Varese. Microb Ecol 83:850–868. https://doi.org/10.1007/s00248-021-01914-5

Sharafati A, Haji Seyed Asadollah SB, Motta D, Yaseen ZM (2020) Application of newly developed ensemble machine learning models for daily suspended sediment load prediction and related uncertainty analysis. Hydrol Sci J. https://doi.org/10.1080/02626667.2020.1786571

Sheng H, Liu H, Wang C et al (2012) Analysis of cyanobacteria bloom in the Waihai part of Dianchi Lake, China. Ecol Inform 10:37–48. https://doi.org/10.1016/j.ecoinf.2012.03.007

Shoar S, Chileshe N, Edwards JD (2022) Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: application of random forest regression. J Build Eng 50:104102. https://doi.org/10.1016/j.jobe.2022.104102

Song K, Li L, Li S et al (2012) Hyperspectral retrieval of phycocyanin in potable water sources using genetic algorithm–partial least squares (GA–PLS) modeling. Int J Appl Earth Obs Geoinf 18:368–385. https://doi.org/10.1016/j.jag.2012.03.013

Stefánsson A, Končar N, Jones AJ (1997) A note on the gamma test. Neural Comput Applic 5:131–133

Su Y, Hu M, Wang Y et al (2022) Identifying key drivers of harmful algal blooms in a tributary of the Three Gorges Reservoir between different seasons: causality based on data-driven methods. Environ Pollut 297:118759. https://doi.org/10.1016/j.envpol.2021.118759

Tao H, Hameed MM, Marhoon HA et al (2022) Groundwater level prediction using machine learning models: a comprehensive review. Neurocomputing 489:271–308. https://doi.org/10.1016/j.neucom.2022.03.014

Te SH, Gin KY-H (2011) The dynamics of cyanobacteria and microcystin production in a tropical reservoir of Singapore. Harmful Algae 10:319–329. https://doi.org/10.1016/j.hal.2010.11.006

Tiyasha, Tung TM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. J Hydrol 585:124670. https://doi.org/10.1016/j.jhydrol.2020.124670

Vilán Vilán JA, Alonso Fernández JR, García Nieto PJ et al (2013) Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona Reservoir (Northern Spain). Water Resour Manag 27:3457–3476. https://doi.org/10.1007/s11269-013-0358-4

Wang J, Hu J (2015) A robust combination approach for short-term wind speed forecasting and analysis — combination of the ARIMA (Autoregressive Integrated Moving Average), ELM (extreme learning machine), SVM (support vector machine) and LSSVM (least square SVM) forecasts using a GPR (Gaussian process regression) model. Energy 93:41–56. https://doi.org/10.1016/j.energy.2015.08.045

Wang M, Rezaie-balf M, Naganna SR, Yaseen ZM (2021) Sourcing CHIRPS precipitation data for streamflow forecasting using intrinsic time-scale decomposition based machine learning models. Hydrol Sci J

Yan J, Chen F, Liu T et al (2022) Subspace alignment based on an extreme learning machine for electronic nose drift compensation. Knowl-Based Syst 235:107664. https://doi.org/10.1016/j.knosys.2021.107664

Yang Z, Wei C, Liu D et al (2022) The influence of hydraulic characteristics on algal bloom in three gorges reservoir, China: a combination of cultural experiments and field monitoring. Water Res 211:118030. https://doi.org/10.1016/j.watres.2021.118030

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

Yaseen ZM, Naganna SR, Sa'adi Z et al (2020) Hourly river flow forecasting: application of emotional neural network versus multiple machine learning paradigms. Water Resour Manag 34:1075–1091. https://doi.org/10.1007/s11269-020-02484-w

Zhao Y-P, Chen Y-B (2022) Extreme learning machine based transfer learning for aero engine fault diagnosis. Aerosp Sci Technol 121:107311. https://doi.org/10.1016/j.ast.2021.107311

Zou R, Zhang X, Liu Y et al (2014) Uncertainty-based analysis on water quality response to water diversions for Lake Chenghai: a multiple-pattern inverse modeling approach. J Hydrol 514:1–14. https://doi.org/10.1016/j.jhydrol.2014.03.069

Authors and Affiliations

Salim Heddam[1] · Zaher Mundher Yaseen[2,3,4] · Mayadah W. Falah[5] · Leonardo Goliatt[6] · Mou Leong Tan[7] · Zulfaqar Sa'adi[8] · Iman Ahmadianfar[9] · Mandeep Saggi[10] · Amandeep Bhatia[11] · Pijush Samui[12]

Zaher Mundher Yaseen
yaseen@ukm.edu.my

Mayadah W. Falah
mayadahwaheed@mustaqbal-college.edu.iq

Leonardo Goliatt
goliatt@gmail.com

Mou Leong Tan
mouleong@usm.my

Zulfaqar Sa'adi
zulfaqar@utm.my

Iman Ahmadianfar
Im.ahmadian@gmail.com

Mandeep Saggi
mandeepsaggi90@gmail.com

Amandeep Bhatia
amandeepbhatia.singh@gmail.com

Pijush Samui
Pijush@nitp.ac.in

[1] Laboratory of Research in Biodiversity Interaction Ecosystem and Biotechnology, Hydraulics Division, Agronomy Department, Faculty of Science, University, 20 Août 1955, Route El Hadaik, BP 26, Skikda, Algeria

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

[3] USQ's Advanced Data Analytics Research Group, School of Mathematics Physics and Computing, University of Southern Queensland, QLD, Toowoomba 4350, Australia

[4] New Era and Development in Civil Engineering Research Group, Scientific Research Center, Al-Ayen University, Thi-Qar 64001, Iraq

[5] Building and Construction Engineering Technology Department, AL-Mustaqbal University College, Hillah 51001, Iraq

[6] Computational Modeling Program, Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil

[7] GeoInformatic Unit, Geography Section, School of Humanities, Universiti Sains Malaysia, 11800 Penang, Malaysia

[8] Centre for Environmental Sustainability and Water Security (IPASA), School of Civil Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 UTM, Sekudai, Johor, Malaysia

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

[10] Department of Computer Science, Thapar Institute of Engineering and Technology, Patiala, India

[11] Department of computers science and engineering, Thapar University, Patiala, India

[12] Department of Civil Engineering, National Institute of Technology (NIT), Patna, Bihar 800005, India