A hybrid DE optimized wavelet kernel SVR–based technique for algal atypical proliferation forecast in La Barca reservoir: A case study

P.J. García–Nieto[a,*], E. García–Gonzalo[a], F. Sánchez Lasheras[a], J.R. Alonso Fernández[b], C. Díaz Muñiz[b]

[a]Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain

[b]Cantabrian Basin Authority, Spanish Ministry of Agriculture, Food and Environment, 33071 Oviedo, Spain

Abstract

The atypical proliferation of algae is a consequence of eutrophication, a phenomenon responsible for the deterioration of reservoirs and lakes. Its growth over the last few decades forced different administrations to adopt different solutions, including forecasting and management, with the help of mathematical models. This article presents a model of eutrophication of reservoirs based on a new methodology called multiscale Mexican Hat wavelet as the kernel function for the support vector regression (SVR) method and differential evolution (DE) optimization technique to estimate the abnormal proliferation of algae from physicochemical and biological variables. The present method implies the optimization of the SVR hyperparameters during the training process. In addition, five other SVR models with different nuclei (linear, quadratic, cubic, sigmoid and radial base function) and random forests (RF) were adjusted to experimental data for purposes of comparison. In addition to successfully predicting atypical algae growth (determination coefficients equal to 0.88 and 0.93), the

*Corresponding author. Tel.: +34-985103417; fax: +34-985103354.
E-mail address: lato@orion.ciencias.uniovi.es (P.J. García Nieto).

model shown here can establish the importance of each biological and physicochemical parameter of improved algae growth. Finally, the main conclusions of this research work are presented.

Keywords: Support vector regression (SVR); Wavelet kernel; Differential evolution (DE); Random forests (RF); Eutrophication prediction in reservoirs; Regression analysis

1. Introduction

The atypical growth of algae, a symptom of eutrophication, remains a global environmental problem. It has serious consequences for water quality [1–5]. The fertilization of water has as a consequence a reduction in the percentage of oxygen present in it, the appearance of toxic blooms, eutrophication and finally death of some organisms, which means a reduction in biodiversity [6–9].

The presence of chlorophyll a (Chl-a) is a clear indicator of water eutrophication. Please note that in existing literature the presence of Chl-a is reported as a common way to track the growth of algae [10]. Phosphorus, nitrogen and chlorophyll are commonly used as indicators in reservoirs and other bodies of water [3,11]. Many of these form the basis of the classical approach to classifying trophic status [12], which is considered in the implementation of the Water Framework Directive (WFD) [13,14]. However, biovolumes must be calculated in order to assess this issue more reliably [15]. Please also note that other variables, such as water temperature, pH, dissolved oxygen, Secchi depth, ammonium and nitrogen also have an important role in the growth of algae [16]. Therefore, the development of strategies to prevent algae blooms requires a multivariate

analysis of all the variables detailed above. Any systematic research in this field needs a complete set of data with all these indicators [17].

All around the world, regardless of many other factors, algal atypical growth is a matter of great concern. In small ecosystems such as La Barca reservoir (see Figs. 1(a) and 1(b)) the threat is particularly worrying due to its location and size. In this case, the main threat is due to eutrophication [18–20] that causes the consumption of dissolved oxygen. Indeed, the algal explosion that accompanies the first phase of eutrophication causes a cloudiness that prevents light from penetrating at the bottom of the ecosystem. Consequently, at this bottom, photosynthesis (main producer of free oxygen) is impossible, while at the same time the oxygen-consuming metabolic activity (aerobic respiration) of decomposers is increased, which begin to receive surplus organic matter produced near the surface. In this way, at the bottom of the body of water, oxygen is soon depleted by aerobic activity and the environment soon becomes anoxic. The radical alteration of the environment due to these changes makes unfeasible the existence of most of the species that previously formed the ecosystem [3,19,20].

In this study, a new methodology has been applied using a wavelet kernel SVM–based method combined with the evolutionary optimization method termed Differential Evolution (DE) [21–24], as well as the random forests (RF) technique [25–27] to forecast the growth of phytoplankton atypical in the aforementioned reservoir. All the results obtained are contrasted and compared.

Fig. 1. (a) Large and (b) short scale aerial photographs of the reservoir.

SVR techniques are a new class of methods designed to predict values based on statistical learning from very different fields [27–31], and present high accuracy for almost any multivariate function [27,31,32].

Furthermore, in order to optimize SVR hyperparameters during training the differential evolution (DE) was employed. Differential evolution (SD) is an evolutionary method of global metaheuristics, gleaned from genetic algorithms (GA), and can solve problems to do with continuous variables in multidimensional optimization operations. Similar to other evolutionary computing algorithms, for example particle swarm optimization (PSO) [33–35] or the ant colony optimization [34], DE is an algorithm based on biological processes that makes use of mutation, recombination and selection, among other commonly-used operations [21,22,37,38]. SVR has been used to predict values in many fields, particularly in environmental problems like forest modeling [39], solar radiation prediction [40,41] and air and water quality estimation, to give some examples [42–45].

Random forests (RF) were introduced by Breiman [26] and are used here for purposes of comparison. The RF algorithm presents several advantages [25–27] that are considered of interest to the present research. First of all, it is able to compute large amounts of information, it presents good behaviour in noise situations and it has a relative low number of parameters to set when compared with other algorithms.

In conclusion, several hybrid models based on SVR based on SVR (DE/SVR) were applied with different nuclei [37,38,46] and random forests (RF) in order to model the eutrophication at the reservoir under study. The best model for predicting the

eutrophication of the La Barca reservoir was the hybrid DE model, optimized by the wavelet kernel based on the SVR model.

2. Materials and methods

2.1. Experimental dataset

The data used for DE/SVR and the analyses of random forests RF were collected over 16 years (2001-2016), with 243 samples collected which contained quantitative information on abundance of phytoplankton. Samples were taken at least once every 30 days beginning on 16th January 2001 so that the last sample analysed corresponds to 20th December 2016. Specifically, this reservoir was sampled following the sampling protocols for lakes and reservoirs of the *Spanish Ministry of Agriculture, Food and Environment*, which are consistent with the guidelines established by the European Union and international agencies dealing with these issues [47–49]. The samples were taken with a Niskin hydrographic bottle. The Niskin bottle is a development of the Nansen bottle. Instead of a metal bottle sealed at one end, the bottle is a tube, usually plastic to minimize contamination of the sample, and open to the water at both ends. Each end is equipped with a cap which is either spring-loaded or tensioned by an elastic rope (see Fig. 2(a)) at different depths in the zone corresponding to the depth of the water in the reservoir that is exposed to sufficient sunlight for photosynthesis to occur called the euphotic zone [50]. This zone is determined from the Secchi depth which is the depth at which the pattern on the Secchi disk (see Fig. 2(b)) is no longer visible and it is taken as a measure of the transparency of the water in lakes, reservoirs and oceans. The values of phytoplankton and Chlorophyll were determined from a sample

composed of five homogeneous subsamples obtained with the hydrographic bottle at various equidistant depths in the euphotic zone [3,51,52].

Fig. 2. (a) An example of a Niskin bottle; and (b) Secchi disks.

In this research work, physical–chemical parameters normally used in limnological studies have been measured [3,51–53]. The physical–chemical parameters were analyzed by an ISO 17025 accredited laboratory, following the corresponding methods in the Standard Methods for the Examination of Water and Wastewater [54]. A quality assessment program including internal laboratory control (use of standards, blanks and replicates during analysis) as well as analysis of blanks, replicates and blind samples collected in La Barca reservoir was applied. During the sampling procedure, field blanks were also collected. A total of 10% of samples were replicated to assess variability. Furthermore, analyses of Chlorophyll have been carried out to study the phytoplankton.

The objective of this work was to establish a methodology for estimating abnormal algae growth indicators from easily-measurable variable values. The seven models constructed (six models based on DE/SVR and one based on DE/RF) all use the same input variables. The output variables are chlorophyll in (µg / L) and total phosphorus (mg P / L).

The presence of phytoplankton determines the presence of chlorophyll in the water [54], something which is linked with photosynthesis. The total phosphorus content is obtained as the sum of organically-bound condensed phosphates, phosphates and orthophosphates, taking into account both suspended and dissolved forms.

Predictive models must take into account both biological and physical-chemical variables. In the case of biological variables, the most relevant are *Cyanobacteria*, *Diatoms*, *Euglenophytes*, *Dinophlagellata*, *Chrysophytes* and *Cryptophytes*. In our research, all of them are expressed in mm^3/L. Fig. 3 (a) shows an example of *Cyanobacteria*, while *Diatoms* are presented in Fig. 3 (b). *Euglenophytes* are a kind of autotrophic organism that are capable of producing their own food. An example is shown in Fig. 3 (c). Fig. 4 (d) shows *Dinophlagellata*, while *Chrysophytes* are presented in Fig. 3 (e), and *Cryptophytes* in Fig. 3 (g).

Fig. 3. Biological variables used for this research: (a) *Cyanobacteria*; (b) Diatoms; (c) Euglenophytes; (d) *Dinophlagella*; (e) Chrysophytes; (f) Clorophytes; and (g) Chryptophytes.

The physical-chemical variables employed are [3]: water temperature (ºC); turbidity, expressed in Nephelometric Turbidity Units (NTU), which determines the opacity of water due to suspended solids [55,56]; nitrate concentration (mg NO^{3-}/L), pH (defined

as the decimal logarithm of the reciprocal of the hydrogen ion activity, a_{H^+}, in a solution); conductivity (μS/cm) and concentration of dissolved oxygen (mg O_2/L).

This study builds machine learning models of eutrophication for the La Barca reservoir from the experimental set of data. For comparison purposes, two different techniques have been constructed: support vector regression with optimization of DE parameters (six DE/SVR models) and also random forests with DE parameter optimization (a DE/RF model).

2.2. Computational procedure

2.2.1. Support vector machine (SVM) for regression

Support vector machine (SVM) is a machine learning methodology developed by Vapnik and his colleagues [57,58]. Our training data consists of *N* pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, with $x_i \in \Re^p$ and $y \in \Re$. Firstly, we discuss the linear regression model:

$$f(x) = \langle x, \beta \rangle + \beta_0 \tag{1}$$

where $\langle .,. \rangle$ means the scalar product in $\Re^p$. Later we will deal with non-linear generalizations of this technique. In order to determine β, we carry out the minimization of the functional:

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}\|\beta\|^2 \tag{2}$$

where $\|\beta\|$ means the norm of vector β (i.e., $\|\beta\| = \sqrt{\langle \beta, \beta \rangle}$) and

$$V_\varepsilon(r)=\begin{Bmatrix} 0 & \text{if} & |r|<\varepsilon \\ |r|-\varepsilon & & \text{otherwise} \end{Bmatrix} \tag{3}$$

Eq. (3) describes a function which is an $\varepsilon-$insensitive error measure, that is, one which ignores errors smaller than ε. Therefore, it is in some ways similar to support vector machines for classification, wherein points to the right of the decision boundary and those a long distance from it are not taken into account when optimization occurs. As regards regression, small residuals are symptomatic of these low error points. Eq. (3) produces a support vector error measurement which also presents a linear behavior (beyond ε), but furthermore it equalizes the contributions from any points with small residuals.

If $\hat{\beta}$ and $\hat{\beta}_0$ are the optimal values that minimized the functional *H*, the solution function can be shown to have the form:

$$\hat{\beta}=\sum_{i=1}^{N}\left(\hat{\alpha}_i^*-\hat{\alpha}_i\right)\mathrm{x}_i \tag{4}$$

$$f(\mathrm{x})=\sum_{i=1}^{N}\left(\hat{\alpha}_i^*-\hat{\alpha}_i\right)\langle \mathrm{x},\mathrm{x}_i\rangle+\beta_0 \tag{5}$$

where $\hat{\alpha}_i$, $\hat{\alpha}_i^*$ are positive Lagrange multipliers. Next, the search for a solution of the following expression is required. Please note that it is a quadratic programming problem,

$$\min_{\alpha_i,\alpha_i^*}\ \varepsilon\sum_{i=1}^{N}\left(\alpha_i^*+\alpha_i\right)-\sum_{i=1}^{N}y_i\left(\alpha_i^*-\alpha_i\right)+\frac{1}{2}\sum_{i,j=1}^{N}\left(\alpha_i^*-\alpha_i\right)\left(\alpha_j^*-\alpha_j\right)\langle \mathrm{x}_i,\mathrm{x}_j\rangle \tag{6}$$

taking into account the following conditions:

$$\begin{aligned} & 0 \le \alpha_i, \alpha_i^* \le C\left(=\frac{1}{\lambda}\right) \\ & \sum_{i=1}^{N}\left(\alpha_i^* - \alpha_i\right) = 0 \\ & \alpha_i \alpha_i^* = 0 \end{aligned} \tag{7}$$

Keeping in mind the nature of these constraints, it is possible to observe that usually only a subset of the solution values $\left(\hat{\alpha}_i^* - \hat{\alpha}_i\right)$ are non-zero. In the same way as the classification problem, the solution relies on the input values only through the scalar product $\left\langle \mathrm{x}_i, \mathrm{x}_j \right\rangle$.

One may note immediately that we have two parameters ε and λ connected to the criterion given by Eq. (2). These parameters have quite different functions, namely, that ε is a parameter of the loss function V_ε, while $C\left(=1/\lambda\right)$ is a regularization parameter (sometimes also called a penalty parameter or cost parameter).By means of cross-validation, it is possible to estimate its value.

Where there is a nonlinear behavior of the training dataset, it is possible to convert the SVR approach to this case using a kernel function in order to map the data from the input space to a high-dimensional space (termed feature space) so that we can tackle a problem in linear form [58,59]:

$$\min_{\alpha_i, \alpha_i^*} \varepsilon \sum_{i=1}^{N}\left(\alpha_i^* + \alpha_i\right) - \sum_{i=1}^{N} y_i\left(\alpha_i^* - \alpha_i\right) + \frac{1}{2}\sum_{i,j=1}^{N}\left(\alpha_i^* - \alpha_i\right)\left(\alpha_j^* - \alpha_j\right) K\left(\mathrm{x}_i, \mathrm{x}_j\right) \tag{8}$$

subject to the same constraints indicated by expression (7). In Eq. (8), $K\left(\mathrm{x}_i, \mathrm{x}_j\right)$ is the support vector kernel. The support vector kernel used is a kernel of dot-product type in a feature space and it must satisfy the Mercer's condition [59]. The resulting regression estimates are linear. The fitting function obtained is given via:

$$f\left(\mathrm{x}\right) = \sum_{i=1}^{N}\left(\hat{\alpha}_i^* - \hat{\alpha}_i\right) K\left(\mathrm{x}, \mathrm{x}_i\right) + \beta_0 \tag{9}$$

According to previous research, several different kernel functions have been used whose description can be found in earlier bibliographic studies [27–32,69]. Furthermore, the performance of a SVM model is directly connected to the kernel selected for each problem:

- Radial basis function (RBF kernel):

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = e^{-\sigma\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2} \tag{10}$$

- Polynomial kernel:

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\sigma \mathbf{x}_i \cdot \mathbf{x}_j + a\right)^b \tag{11}$$

- Sigmoid kernel:

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \tanh\left(\sigma \mathbf{x}_i \cdot \mathbf{x}_j + a\right) \tag{12}$$

being *a*, *b* and σ kernel parameters.

Fig. 4. Sketch of the SVM regression model with $\varepsilon-$insensitive tube for a one dimensional problem.

2.2.2. Wavelet procedure

Wavelets are a useful methodology for obtaining information from signals like audio signals or images. They are widely used in geophysics and signal processing. From a mathematical point of view, there will be a correlation of the signal with the wavelet if information of a similar frequency appears in the signal. Indeed, the signal analysed is generated from a set of functions obtained as translations and dilations of the so-called *mother wavelet* or function [69–73]:

$$\psi_{a,c}(x) = |a|^{-\frac{1}{2}} \cdot \psi\left(\frac{x-c}{a}\right) \tag{13}$$

where c is the translation and a is the dilation factor [37]. Thus, the wavelet transform of a function $f(x) \in L^2(\Re)$ is given via [70–73]:

$$W_{a,c}(f) = \langle f(x), \psi_{a,c}(x) \rangle \tag{14}$$

$\langle \cdot, \cdot \rangle$ is the dot product in $L^2(\Re)$. and Eq. (13) is a function $f(x)$ decomposition on the wavelet basis $\psi_{a,c}(x)$. The mother wavelet $\psi(x)$ must meet the condition [72–75]:

$$W_\psi = \int_0^\infty \frac{|H(\omega)|^2}{|\omega|} d\omega < \infty \tag{15}$$

where $H(\omega)$ is the Fourier transform of $\psi(x)$. Then $f(x)$ can be reconstructed as [72–75]:

$$f(x) = \frac{1}{W_\psi} \int_{-\infty}^{\infty} \int_0^\infty W_{a,c}(f) \psi_{a,c}(x) da / a^2 dc \tag{16}$$

Accordingly, if we take the finite terms of Eq. (14), the approximated $\hat{f}(x)$ can be expressed as [70–75]:

$$\hat{f}(x)=\sum_{i=1}^{l} W_i \psi_{a_i,c_i}(x) \tag{17}$$

In the case of a common multidimensional wavelet function $\left\{\mathbf{x}_j, \mathbf{x}\in\Re^N\right\}$, one–dimensional (1–D) wavelet functions produce:

$$\psi(\mathbf{x})=\prod_{j=1}^{N}\psi(\mathbf{x}_j) \tag{18}$$

In this respect, for an in–depth wavelet analysis and theory readers can consult Zhang et al. [72], Daubechies [74] and Zhang and Benveniste [75].

2.2.3. Wavelet kernel and wavelet SVMs

Wavelet kernel SVMs is a particular case involving SVM where the kernel is constructed starting with dot product based on a wavelet [76]. Let $\psi_{a,c}(\mathbf{x})$ be a mother wavelet and $\mathbf{x},\mathbf{x}'\in\Re^N$, then the dot-product wavelet kernel is given as [70–72]:

$$K(\mathbf{x},\mathbf{x}')=\prod_{j=1}^{N}\psi\left(\frac{\mathbf{x}_j-c_j}{a}\right)\cdot\psi\left(\frac{\mathbf{x}'_j-c'_j}{a}\right) \tag{19}$$

The translation invariant wavelet kernel can be expressed as follows [70–75]:

$$K(\mathbf{x},\mathbf{x}')=\prod_{j=1}^{N}\psi\left(\frac{\mathbf{x}_j-\mathbf{x}'_j}{a}\right) \tag{20}$$

A translation invariant wavelet kernel, a so-called Mexican hat wavelet kernel, is given via [70–75]:

$$\psi(\mathbf{x})=\frac{2}{\sqrt[4]{9\pi}}\cdot\left(1-\mathbf{x}^2\right)\cdot\exp\left(-\frac{\mathbf{x}^2}{2}\right) \tag{21}$$

Taking into account the previous expression [74,75]:

$$K\left(\mathbf{x},\mathbf{x}'\right)=\prod_{j=1}^{N}\frac{2}{\sqrt[4]{9\pi}}\left(1-\frac{\left(\mathbf{x}_j-\mathbf{x}'_j\right)^2}{a^2}\right)\cdot\exp\left(-\frac{\left(\mathbf{x}_j-\mathbf{x}'_j\right)^2}{2a^2}\right) \qquad (22)$$

The decision function of wavelet SVMs for regression can be expressed as [70,74,75]:

$$f\left(\mathbf{x}\right)=\sum_{i=1}^{l}\left(\alpha_i-\alpha_i^*\right)\prod_{j=1}^{N}\psi\left(\frac{\mathbf{x}_j-\mathbf{x}_{ij}}{a}\right)+b \qquad (23)$$

Specifically, the multiscale Mexican hat wavelet kernel has been used in this study with success due to its ability to capture abrupt changes of radically-changing functions such as those due to eutrophication. It is obtained subtracting two Gaussian radial basis functions and it is called multiscale Mexican Hat wavelet because it behaves in a similar way to the Mexican Hat wavelet. Indeed, it is defined as [77,78]:

$$K\left(\mathbf{x},\mathbf{x}'\right)=g_2\exp\left(-\sigma_2\left\|\mathbf{x}-\mathbf{x}'\right\|^2\right)-g_1\exp\left(-\sigma_1\left\|\mathbf{x}-\mathbf{x}'\right\|^2\right) \qquad (24)$$

where $g_1=\sigma_1/\left(\sigma_1-\sigma_2\right)$ and $g_2=\sigma_2/\left(\sigma_2-\sigma_1\right)$. This function is shown below in Fig. 5.

Fig. 5. Multiscale Mexican Hat wavelet function.

2.3. Differential evolution (DE) algorithm

The differential evolution (DE) method was initially discovered by Storn and Price [21] and it optimizes a problem iteratively by attemptng to improve a candidate solution concerning a well-known quality measurement. The total population $\mathbf{X}=\left[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_M\right]^T$ involves M individuals in such a way that the *n*-th individual is represented by an objective vector designating an individual's position in

the search space. The objective vector pertaining to the *n*-th individual at the *t*-ith iteration of the optimization is expressed as [21,22]:

$$x_n^p(0) = L^p + \text{rand}(0,1) \times \left(U^p - L^p\right) \tag{25}$$

The differential evolution (DE) technique requires five stages to be able to build the optimization algorithm; these are indicated as follows [21–24]:

- *Initialization*

Firstly, the initial objective vectors of *M* individuals are created (or produced) in the design space *H*. This task is carried out in a random way so that the initial objective vector (0-th iteration: corresponds to the initial location of the particle) of dimension *p* $\left(p \in \{1,...,P\}\right)$ is given via:

$$x_n^p(0) = L^p + \text{rand}(0,1) \times \left(U^p - L^p\right) \tag{26}$$

so that $\text{rand}(0,1)$ represents a random number evenly distributed in the interval [0,1].

- *Mutation*

A benefactor vector is produced using differences of scale among individuals for each individual of the population. The *n*-th benefactor vector is generated by means of the following mutation strategy, written as the expression:

$$\mathbf{v}_n = \mathbf{x}_{r_1}(t) + G\left(\mathbf{x}_{r_2}(t) - \mathbf{x}_{r_3}(t)\right) \tag{27}$$

where r_1, r_2 and r_3 are random integers evenly distributed in the interval $[1, M]$ such that $r_1 \neq r_2 \neq r_3 \neq n$, and *G* is a scaling factor.

- *Crossover*

In order to achieve diversity, the crossover stage must be carried out. Indeed, to generate a trial vector with guarantees, we will do the crossover of the individual elements from the objective vector and benefactor vector. In this study, we have used the binomial crossover to create the *n*-th trial vector in the *p*-th dimension governed by the expression:

$$u_n^p = \begin{cases} v_n^p & \text{if rand}(0,1) < CR \text{ or } r_n = p \\ x_n^p(t) & \text{otherwise} \end{cases} \tag{28}$$

where $CR \in [0,1]$ is a parameter called crossover probability and r_n is a random integer spread evenly in the interval $[1, P]$. Consequently, elements of the trial vector are taken from the benefactor vector with a probability *CR* so that at least one element of the benefactor vector is accepted.

- *Selection*

Next, the trial vector is checked and the *n*-th objective vector is computed at the next iteration via:

$$\mathbf{x}_n(t+1) = \begin{cases} \mathbf{u}_n & \text{if } f(\mathbf{u}_n) \le f(\mathbf{x}_n(t)) \\ \mathbf{x}_n(t) & \text{otherwise} \end{cases} \tag{29}$$

Therefore, the objective vector is replaced by the trial vector if its performance is greater than or equal to the performance of the objective vector.

- *Stopping criterion*

This algorithm is stopped if the permitted maximum number of function evaluations is reached and after all *M* objective vectors have been upgraded. Otherwise, the above steps, from the second one to the fifth, are repeated.

The pseudocode of the DE algorithm can be written as:

Random initialization of the individuals and calculate the objective

while Current_number_of_function_evaluations < Max_function_evaluations **do**

for n = 1:M **do**

Carry out the mutation according to Eq. (27)

Carry out the binomial crossover according to Eq. (28)

Calculate the objective taking into account the constraints of the trial vector

end for

for n = 1:M **do**

Upgrade the *n*-th objective vector according to Eq. (29)

end for

end while

2.4. Random forest regression algorithm

The random forest (RF) algorithm [79–82] presents a number of advantages that are considered of interest for the present research. Firstly, it is capable of computing large quantities of information, it behaves well in noisy situations and has a relatively low number of parameters to set when compared with other algorithms.

Classification and regression trees are methods that satisfy both predictive and explanatory objectives. There are two cases in which these modeling techniques should be used: classification trees, which are useful for clarifying and predicting whether

individuals belong to categories based on quantitative and qualitative variables. Also, a regression tree can be employed to create an explanatory and predictive model for a quantitative dependent variable based on quantitative and qualitative explanatory variables.

The regression tree-splitting criterion is based on choosing the input variable with the lowest Gini Index:

$$I_G\left(t_{X(x_i)}\right)=1-\sum_{j=1}^{m} f\left(t_{X(x_i)}, j\right)^2 \tag{30}$$

so that $f\left(t_{X(x_i)}, j\right)$ is the proportion of samples from the leave j as node t with the value x_i [81,82]. In order to calculate the predicted value of an observation, we then carried out an averaging over all the trees. To this end, two parameters must be optimized in the RF approach:

- ntree: is the number of regression trees (its default value is 500 trees); and
- mtry: is the number of input variables per node (its default value is 1/3 of the complete number of variables).

2.5. The goodness–of–fit

All the variables of the study are presented in Tables 1 and 2. Table 1 shows all the physical-chemical variables, while the biological variables are listed in Table 2 [3,83]. The number of predictors employed by the DE/SVM and RF models was 16. The estimated variables (*Chl*-a and Total phosphorus) units are μg/L and mg P/L [83,84], respectively.

Table 1

Biological variables employed in this research with their mean, median, standard deviation (STD) and mean absolute deviation (MAD).

Table 2

Physical-chemical variables employed in this research with their mean, median, standard deviation (STD) and mean absolute deviation (MAD).

The variable importance order in this research has been determined with the help of the goodness-of-fit criterion [85,86]. This may be defined as a parameter by means of which any variation in the variable produced by the model can be quantified, as can the variability in the same variable across the set of data. That is to say, this variability is expressed thus:

- $SS_{err} = \sum_{i=1}^{n} (t_i - y_i)^2$
- $SS_{tot} = \sum_{i=1}^{n} (t_i - \bar{t})^2$

where the average value of the *n* observed samples is defined as:

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i \tag{31}$$

Then, the quantity R^2, the coefficient of determination, is given as:

$$R^2 \equiv 1 - \frac{SS_{err}}{SS_{tot}} \tag{32}$$

To fix ideas, it can be said that this coefficient gives us an idea about how well the regression values approximate the actual values. Please note that the closer its value to one, the better.

Two additional criteria considered in this study were the root mean square error (RMSE) and mean absolute error (MAE) [27,85,86]. These statistics are also used frequently to evaluate the forecasting capability of a mathematical model. Indeed, the root mean square error (RMSE) and mean absolute error (MAE) are given by the expressions [85,86]:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(t_i - y_i\right)^2}{n}} \tag{33}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\left|t_i - y_i\right|}{n} \tag{34}$$

If the root mean square error (RMSE) has a value of zero, it means that there is no difference between the predicted and observed data. Mean Absolute Error (MAE) is the average vertical distance between each point and the identity line. MAE is also the average horizontal distance between each point and the identity line. MAE has a clear interpretation as the average absolute difference between t_i and y_i.

Different models were constructed at this stage, (specifically in this study, six hybrid models DE/SVM and one RF model) with variables which predicted the variables Chl-a and Total phosphorus as well as the other sixteen biological and physical-chemical

parameters (input variables), by employing the determination coefficient as a criterion to assess whether each model was successful.

In addition, as mentioned above, the success of SVM models depends to a large extent on their parameters. An adequate adjustment of these parameters is therefore essential. Therefore, the fitting process involves calculating the suitability of different models and is often the most demanding task from a computational point of view [25,27,31,32]. Completely different methods can be found in existing literature for the optimization of these parameters as random search, grid search, genetic algorithms, particle swarm optimization (PSO) and so on [29,31,59,70]. Usually, the traditional way of performing hyperparameter optimization in most computational codes has been *grid search*, or a parameter sweep, which is simply an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm. Indeed, the grid search is a *brute force* method and, as such, almost any optimization method improves its efficiency. Specifically, in the present research, DE optimization technique was chosen for the adjustment of SVM hyperparameters of the different kernels with success. At this point, the flow diagram of the best DE/SVM–based model used in this work can be seen in Fig. 6.

Fig. 6. Flowchart of the new hybrid DE/SVM–based model with multiscale Mexican Hat wavelet kernel.

It is to be noted that the coefficient of determination $\left(R^2\right)$ was computed by means of cross–validation [86,87]; to be precise, with a ten–fold cross–validation algorithm.

Specifically, the DE technique employed the coefficient of determination as the objective function for the optimization of the SVM hyperparameters.

An important aspect of the current formulation is that the population is made up of $\mathbf{x}_i$ vectors that are composed of the parameters of the kernel (e.g., in case of multiscale Mexican hat wavelet kernel function, $\mathbf{x}_i = (C_i, \varepsilon_i, \sigma_{1i}, \sigma_{2i})$). Initially, we assume 20 different random sets of four parameters (members of the population) within the search space. In other words, we look for the values of C in $\left[10^{-6}, 10^{4}\right]$, e values in $\left[10^{-10}, 10^{4}\right]$, and σ_1 y σ_2 in $\left[10^{-6}, 10^{4}\right]$. That is, the search space is the four-dimensional space $[-6,4]\times[-10,4]\times[-6,4]\times[-6,4]$. Accordingly, we look for the space of exponents as the SVR algorithm changes more significantly as the order of magnitude in turn changes. We begin by constructing a model with each of these sets of parameters and then its corresponding ten-fold cross-validation R^2 is computed. With the help of mutation and recombination operators, a new population is created and the process is repeated. When they are compared with those of the previous generation, the best individuals in the population survive. The procedure is repeated until either the iteration 200 is reached or until the fitness of one generation improves by less than 10^{-8} when compared with the previous one. Finally, the best performing model is chosen. From a computational point of view, the SVM$-\varepsilon$ regression has been carried out using the LIBSVM library [88] in combination with the DE function implemented in MATLAB code [89,90].

3. Results and discussion

The optimal parameters for models based on DE/SVM found with the differential evolution technique (DE) to total phosphorus and chlorophyll a, respectively, appear in Tables 3 and 4.

Table 3

Total phosphorus optimal hyperparameters obtained with the DE/SVM models.

Table 4

Chorophyll concentration optimal hyperparameters obtained with the DE/SVM models.

In order to have a benchmark model, another model based on DE/RF has also been trained. This model uses total phosphorus and *Chl*-a [84] respectively as output variables and the physical-chemical parameters as the input ones. Table 5 shows the optimal parameters for the RF–based model found with the differential evolution (DE) technique for chlorophyll and Table 6 gives the same information for the total phosphorus variable.

Table 5

Optimal hyperparameters obtained with the DE/RF–based model for total phosphorus.

Table 6

Optimal hyperparameters obtained with the DE/RF–based model for chorophyll concentration.

In addition, the correlation and determination coefficients for DE/SVM and DE/RF based models for total phosphorus and chlorophyll a, respectively, are shown in Tables 7 and 8.

Table 7

Cross–validation coefficients of determination (R^2) and correlation coefficient (r), and root mean square error (RMSE) and mean absolute error (MAE) for the DE/SVM–based and DE/RF–based models for total phosphorus.

Table 8

Cross–validation coefficients of determination (R^2) and correlation coefficient (r), and root mean square error (RMSE) and mean absolute error (MAE) for the DE/SVM–based and DE/RF–based models for the Chorophyll concentration.

Taking into account the previous statistical calculations, the SVM–based technique with a wavelet kernel in combination with DE optimization is the best model for estimating total phosphorus and chlorophyll (specifically, using the multiscale Mexican hat wavelet kernel). Models relied on DE/SVM have determination coefficients equal to 0.93 and 0.88, and correlation coefficients equal to 0.96 and 0.94, respectively. A

computer with a CPU Intel Core i7-4770 @ 3.40 GHz with eight cores and 15.5 GB RAM memory was used, taking 1,272 seconds (approximately 21 min) to obtain the Chlorophyll model and 1,555 seconds (approximately 26 min) for the Total phosphorus model.

The classification of significance of the biological and physical-chemical parameters (input variables), taking as dependent variables total phosphorus and chlorophyll (Chl-a) (output variables), are shown in Tables 9 and 10, and Figs. 7 and 8, respectively.

As a consequence, Secchi depth is the most significant variable in total phosphorus prediction for the SVM–based model of the optimized DE wavelet kernel., followed by turbidity, water temperature, dissolved oxygen concentration, cyanobacteria concentration, chlorophyll concentration, dinophlagellata concentration, ammonium concentration, chlorophytes concentration, chrysophytes concentration, nitrate concentration, pH, euglenophytes concentration, conductivity, chryptophytes concentration and finally, diatoms.

Table 9

Weights for the DE/SVM–based model for the Total phosphorus.

Table 10

Weights for the DE/SVM–based model for the Chlorophyll concentration.

Fig. 7. Comparative significance of the predictor variables in the total phosphorus DE/SVM–based model.

Fig. 8. Comparative significance of the predictor variables in the Chlorophyll DE/SVM–based model.

Secchi depth is the first most significant variable in the total phosphorus prediction. This parameter gives an idea of the turbidity of the water. Turbidity is the second most significant variable in the total phosphorus prediction. Indeed, turbidity increases with phytoplankton growth [3,56] affecting the eutrophication process.

By virtue of their relevance, water temperature and dissolved oxygen are the following variables in the model. In the case of temperature, from our point of view this is due to its influence on the growth of phytoplankton, while in the case of oxygen it by dint of its importance over those organisms that live in the reservoir water.

Furthermore, *Cyanobacteria* are one of the most common consequences of the abnormal algal blooms [52,91–93], causing a particularly serious problem for the water quality [7,50,52,94].

When the concentration of chlorophyll in water is high (eutrophic environment), cyanobacteria proliferate. Their concentration is the sixth most significant variable in the prediction of total phosphorus. The concentration of chlorophyll is related to the concentration of phytoplankton [54].

The concentration of Dinophlagellata is the seventh most important variable due to the photosynthetic nature of these organisms.

The eighth most significant variable in total phosphorus prediction is the concentration of ammonium [3]., followed by the concentration of chlorophytes. This may be due to the fact that La Barca reservoir is notable for chlorophytes being present in large numbers, as La Barca is a eutrophic ecosystem [6,95].

Concentrations of chrysophytes and nitrates come last among variables in the ranking of importance for total phosphorus prediction [96, 97].

The relatively low importance of nitrate concentration may be explained by the fact that although nitrates are sources of *nitrogen*, there are other sources, such as for example the atmosphere [98,99], which makes nitrogen a non-growth-limiting nutrient.

In twelfth place, we find the pH of the water. This result could be due to the relationship of pH with the excessive growth of plants and algae.

The concentration of Euglenophytes is the thirteenth most significant variable in the prediction of total phosphorus (output variable) because dammed waters are usually rich in Euglenophytes.

Conductivity is the fourteenth most significant variable in importance in the prediction of total phosphorus, since ionic phosphate is the main component of total phosphorus in eutrophic environments.

Cyanobacterial concentration is the most important input variable in predicting chlorophyll concentration. In fact, cyanobacteria are a group of photosynthetic bacteria,

some of which fix nitrogen, living in a wide variety of moist soils and water freely or in a symbiotic relationship with lichen-forming plants or fungi.

In addition, there is a relationship between dissolved oxygen and chlorophyll-containing organisms present in water, as has been noted [100]. In fact, dissolved oxygen is the second most significant variable in Chl-a prediction.

The fourth most significant variable for predicting chlorophyll is the concentration of Dinophlagellata. As is well known, this is related to the photosynthetic nature of these organisms. Many dinoflagellates are known to be photosynthetic, but a significant number of them are myxotrophic, combining photosynthesis with the ingestion of prey.

The fifth most significant variable for predicting chlorophyll is conductivity [101], while the sixth is the depth of Secchi, which is used as an indicator of water turbidity.
The eighth most significant variable for predicting chlorophyll is nitrate concentration. It should be noted that nitrate concentration is more important for predicting chlorophyll than for predicting total phosphorus. The concentration of chlorophyll a (Chl-a) was used here as an indicator of algal density. Excessive nitrate concentrations in reservoirs and lakes can cause accelerated eutrophication and loss of dissolved oxygen. In addition, high nitrate concentrations can cause severe algal blooms, creating a risk to humans and animals.

The eleventh most significant variable for predicting chlorophyll is pH. There is a direct relationship between pH and excessive algae growth, due to high rates of photosynthesis. The next important variable is the concentration of chlorophytes. The

Barca is a eutrophic ecosystem, so the concentration of chlorophytes contributes significantly to this state. Similarly, the concentration of chrysophytes is the thirteenth most important variable when predicting the concentration of chlorophyll in this reservoir. Chrysophytes are characteristically golden brown, and have two types of chlorophyll largely masked by the fucoxanthin pigment. They are also termed golden-brown alga.

Ammonium concentration is the fourteenth most significant variable in the prediction of Chl-a . This significance may be owing to those processes which occur in water when photosynthetic organisms increase abundantly, given the fact that green algae and plants have photosynthetic activity. When these organisms grow too much in water, the concentration of dissolved oxygen nearest the surface consequent to this photosynthetic activity also increases significantly. At this stage, these plants sink as they begin to die, and the decomposition of microbes causes dissolved oxygen to deplete and thus, dead zones are formed [102].

The fifteenth most significant variable when predicting Chlorophyll concentration is turbidity. Turbidity is mainly due to waste of human, agricultural and industrial origin [55] and has a great influence on eutrophication [56]. The last important input variable in chlorophyll concentration prediction is the concentration of hryptophytes. These are mostly photosynthetic and able to live in low light conditions due to a combination of photosynthetic pigments. Therefore, they are found relatively deeply in the water column and can also survive under the ice during the winter, thereby taking advantage of the low light that filters through.

Finally, this research work makes it possible to predict the first dependent variable, total phosphorus. The results agree with the actual experimental values. Indeed, Fig. 9 compares total phosphorus observed with predicted values, using DE/RF–based (see Fig. 9(a)) and DE/SVM–based (see Fig. 9(b)) models. The wavelet kernel DE/SVM–based model shows a better agreement in the results.

Fig. 9. Predicted vs. observed total phosphorus values with: (a) DE/RF–based model ($R^2 = 0.92$) and (b) Wavelet kernel DE/SVM–based model ($R^2 = 0.93$).

Similarly, Fig. 10 compares total phosphorus concentration observed and predicted values using the DE/RF and wavelet kernel DE/SVM models. Again, the wavelet kernel DE/SVM–based model obtains the best results.

Fig. 10. Predicted vs. observed Chorophyll concentration values with: (a) DE/RF–based model ($R^2 = 0.84$) and (b) Wavelet kernel DE/SVM–based model ($R^2 = 0.88$).

4. Conclusions

It is possible to model the eutrophication in the reservoir under study by means of the new model proposed in the present work. A high coefficient of determination $\left(R^2 = 0.93\right)$ was achieved as this hybrid wavelet kernel DE/SVM–based model was trained and then checked with the experimental dataset corresponding to the total phosphorus. The estimated values for this model agree with the dataset values of chlorophyll observed (see Fig. 9). Similarly, this mixed model for the experimental

dataset of the concentration of chlorophyll-a also achieved a high coefficient of determination ($R^2 = 0.88$). The predicted results for the algal untypical production coincide with the dataset of observed values of chlorophyll-a concentration (see Fig. 10).

This innovative method also makes it possible to classify the input variables involved in the forecasting of eutrophication. Furthermore, the Secchi depth is the most influential factor in the total phosphorus model, whilst the concentration of cyanobacteria is the one most closely connected to the concentration of chlorophyll.

The wavelet kernel DE/SVM–based regression method improved the generalization ability of the SVR. From our point of view, it is important to mention that this wavelet DE/SVM–based model is data-driven. In other words, extrapolation for other conditions could lead to innovation. Therefore, an effective wavelet kernel DE/SVM–based model could be an attractive instrument for water management.

Acknowledgements

This research work was made possible due to the cooperation of the Cantabrian Basin Authority (Spanish Ministry of Agriculture, Food and Environment) who gave us access to the experimental data required for this study. Moreover, this research was funded by the Foundation for the Promotion of Applied Scientific Research and Technology in Asturias (FICYT) through the GRUPIN project Reference IDI/2018/000221, co-financed with EU FEDER funds.

References

[1] R.D. Grundy, Strategies for control of man-made eutrophication, Environ. Sci. Tech. 5 (1971) 1184–1190.

[2] V.H. Smith, Low nitrogen to phosphorus ratios favor dominance by blue-green algae in Lake Phytoplankton, Science 221 (4611) (1983) 669–671.

[3] C.S. Reynolds, Ecology of Phytoplankton, Cambridge University Press, New York, 2006.

[4] M.H. Alexandrov, J. Bloesch, Eutrophication of lake Tasaul, Romania—proposals for rehabilitation, Environ. Sci. Pollut. R. 16 (1) (2009) 42–45.

[5] X. Xue, A. Landis, Eutrophication potential of food consumption patterns, Environ. Sci. Technol. 44 (2010) 6450–6456.

[6] M. Álvarez Cobelas, M. Arauzo, Phytoplankton responses to varying time scales in a eutrophic reservoir, Arch. Hydrobiol. Ergebn. Limnol. 40 (2006) 69–80.

[7] M.R. Texeira, M.J. Rosa, Comparing disolved air flotation and conventional sedimentation to remove cyanobacterial cells of *Microcystis aeruginosa*: part I: the key operating conditions, Sep. Purif. Technol. 52 (2006) 84–94.

[8] Y. Liu, H. Guo, Y. Yu, Y. Dai, F. Zhou, Ecological–economic modeling as a tool for watershed management: A case study of Lake Qionghai watershed, China, Limnologica 38 (2) (2008) 89–104.

[9] T. Takaara, D. Sano, Y. Masago, T. Omura, Surface–retained organica matter of *Microcystis aeruginosa* inhibiting coagulation with polyaluminum chloride in drinking water treatment, Water Res. 44 (2010) 3781–3786.

[10] G. Gibson, R. Carlson, J. Simpson, E. Smelzer, Nutrient criteria technical guidance manual: lakes and reservoirs, in: EPA-822-B-00-001, United States Environment Protection Agency (USEPA), Office of Water, Washington DC, USA, 2000.

[11] M. Karydis, Eutrophication assessment of coastal waters based on indicators: a literature review, Global NEST J. 11 (2009) 373–390.

[12] S. Spatharis, G. Tsirtsis, Ecological quality scales based on phytoplankton for the implementation of Water Framework Directive in Eastern Mediterranean. Ecol. Indic. 10 (4) (2010) 840–847.

[13] Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000, Establishing a framework for community action in the field of water policy, L-327, Luxembourg.

[14] A. Borja, D.M. Dauer, Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices, Ecol. Indic. 8 (2008) 331–337.

[15] H. Hillebrand, C.–D. Dürselen, D. Kirschtel, U. Pollinger, T. Zohary, Biovolume calculation for pelagic and benthic microalgae, J. Phycol. 35 (1999) 403–424.

[16] S. Wang, X. Jin, Q. Bu, L. Jiao, F. Wu, Effects of dissolved oxygen supply level on phosphorus release from lake sediments, Colloid Surf. A 316 (2088) 245–252.

[17] D. Kitsiou, M. Karydis, Coastal marine eutrophication assessment: a review on data analysis, Environ. Int. 37 (2011) 778–801.

[18] K. Karlson, R. Rosenberg, E. Bonsdorff, Temporal and spatial large-scale effects of eutrophication and oxygen deficiency on benthic fauna in Scandinavian waters–a review, Oceanogr. Mar. Biol. Ann. Rev. 40 (2002) 427–489.

[19] S.C. Charpa, Surface Water-quality Modelling, McGraw-Hill, New York, 1997.

[20] R.J. Díaz, R. Rosenberg, Introduction to environmental and economic consequences of hypoxia, Int. J. Water Resour. Dev. 27 (2011) 71–82.

[21] R.M. Storn, K. Price, Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces, J. Global Optim. 11 (1997) 341–359.

[22] K. Price, R.M. Storn, J.A. Lampinen, Differential Evolution: A Practical Approach to Global Optimization, Springer, Berlin, 2005.

[23] V. Feoktistov, Differential Evolution: In Search of Solutions, Springer, New York, 2006.

[24] P. Rocca, G. Oliveri, A. Massa, Differential evolution as applied to electromagnetics, IEEE Antennas Propag. 53 (1) (2011) 38–49.

[25] T.M. Mitchell, Machine learning, McGraw-Hill Company Inc, New York, 1997.

[26] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[27] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer–Verlag, New York, 2003.

[28] B. Schölkopf, A.J. Smola, R. Williamson, P. Bartlett, New support vector algorithms, Neural Comput. 12 (5) (2000) 1207–1245.

[29] T. Hansen, C.J. Wang, Support vector based battery state of charge estimator, J. Power Sources 141 (2005) 351–358.

[30] X. Li, D. Lord, Y. Zhang, Y. Xie, Predicting motor vehicle crashes using Support Vector Machine models, Accident Anal. Prev. 40 (2008) 1611–1618.

[31] I. Steinwart, A. Christmann, Support Vector Machines, Springer, New York, 2008.

[32] S. Kulkarni, G. Harman, An Elementary Introduction to Statistical Learning Theory, Wiley, New York, 2011.

[33] R.C. Eberhart, Y. Shi, J. Kennedy, Swarm Intelligence, Morgan Kaufmann, San Francisco, 2001.

[34] M. Clerc, Particle Swarm Optimization, Wiley-ISTE, London, United Kingdom, 2006.

[35] A.E. Olsson, Particle Swarm Optimization: Theory, Techniques and Applications, Nova Science Publishers, New York, 2011.

[36] M. Dorigo, T. Stützle, Ant Colony Optimization, Bradford Publisher, The MIT Press, Cambridge, Massachusetts, USA, 2004.

[37] D. Simon, Evolutionary Optimization Algorithms, Wiley, New York, 2013.

[38] X.–S. Yang, Z. Cui, R. Xiao, A.H. Gandomi, M. Karamanoglu, Swarm Intelligence and Bio-inspired Computation: Theory and Applications, Elsevier, London, 2013.

[39] P.J. García Nieto, J. Martínez Torres, M. Araújo Fernández, C. Ordóñez Galán, Support vector machines and neural networks used to evaluate paper manufactured using *Eucalyptus globulus*, Appl. Math. Model. 36(12) (2012) 6137–6145.

[40] V.H. Quej, J. Almorox, J.A. Arnaldo, L. Saito, ANFIS, SVM and ANN soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment, J. Atmos. Sol-Terr. Phy. 155 (2017) 62–70.

[41] Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters, Renew. Energ. 121 (2018) 324–343.

[42] A. Suárez Sánchez, P.J. García Nieto, P. Riesgo Fernández, J.J. del Coz Díaz, F.J. Iglesias–Rodríguez, Application of an SVM-based regression model to the air

quality study at local scale in the Avilés urban area (Spain), Math. Comput. Model. 54 (5-6) (2011) 1453–1466.

[43] P.J. García Nieto, E.F. Combarro, J.J. del Coz Díaz, E. Montañés, A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study, Appl. Math. Comput. 219 (17) (2013) 8923–8937.

[44] P.J. García Nieto, E. García–Gonzalo, J.R. Alonso Fernández, C. Díaz Muñiz, Hybrid PSO–SVM-based method for long-term forecasting of turbidity in the Nalón river basin: A case study in Northern Spain, Ecol. Eng. 73 (2014) 192–200.

[45] X. Xu, Y. Liu, S. Liu, J. Li, G. Guo, K. Smith, Real-time detection of potable-reclaimed water pipe cross-connection events by conventional water quality sensors using machine learning methods, J. Environ. Manage. 238 (2019) 201–209.

[46] I. Fister, D. Stranad, X.–S. Yang, I. Fister Jr., Adaptation and hybridization in nature-inspired algorithms, in: I. Fister, I. Fister Jr. (Eds.), Adaptation and Hybridization in Computational Intelligence, Springer, New York, 2015, vol. 18, pp. 3–50, 2015.

[47] G.E. Fogg, W.D.P. Stewart, P. Fay, A.E. Walsby, The Blue-green Algae, Academic Press, London, 1973.

[48] World Health Organization, Guidelines for drinking-water quality: health criteria and other supporting information, vol. 2, Geneva, World Health 408 Organization, 1998.

[49] M.J. Smith, G.R. Shaw, G.K. Eaglesham, L. Ho, J.D. Brookes, Elucidating the factors influencing the biodegradation of cylindrospermopsin in drinking water sources, Environ. Toxicol. 23 (2008) 413–421.

[50] R. Willame, T. Jurckzak, J.F. Iffly, T. Kull, J. Meriluoto, L. Hoffman, Distribution of hepatotoxic cyanobacterial blooms in Belgium and Luxembourg, Hydrobiologia 551 (2005) 99–117.

[51] C. Brönmark, L.–A. Hansson, The Biology of Lakes and Ponds, Oxford University Press, New York, 2005.

[52] A. Quesada, E. Moreno, D. Carrasco, T. Paniagua, L. Wormer, C. de Hoyos, A. Sukenik, Toxicity of *Aphanizomenon ovalisporum* (*Cyanobacteria*) in a Spanish water reservoir, Eur. J. Phycol. 41 (2006) 39–45.

[53] A.I. Negro, C. de Hoyos, J.C. Vega, Phytoplankton structure and dynamics in Lake Sanabria and Valparaíso reservoir (NW Spain), Hydrobiologia 424 (2000) 25–37.

[54] American Public Health Association, American Water Works Association, Water Environment Federation, Standard Methods for the Examination of Water and Wastewater, no. 21. APHA/AWWA/WEF, Washington, 2005.

[55] R.L. France, R.H. Peters, Predictive model of the effects on Lake Metabolism of decreased airborne litter fall through riparian deforestation, Conserv. Biol. 9 (6) (1995) 1578–1586.

[56] K.H. Nicholls, R.J. Steedman, E.C. Carey, Changes in phytoplankton communities following logging in the drainage basins of three boreal forest lakes in north-western Ontario, Can. J. Fish. Aquat. Sci. 60 (2003) 43–54.

[57] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1999.

[58] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.

[59] N. Cristianini, J. Shawe–Taylor, An Introduction to Support Vector Machines and Other Kernel–based Learning Methods, Cambridge University Press, New York, 2000.

[60] N.K. Shrestla, S. Shukla, Support vector machine based modeling of evapotranspiration using hydro–climatic variables in a sub–tropical environment, Agr. Forest Meteorol. 200 (2015) 172–184.

[61] J.–L. Chen, G.–S. Li, S.–J. Wu, Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration, Energ. Convers. Manage. 75 (2013) 311–318.

[62] M.R. Nikoo, N. Mahjouri, Water quality zoning using probabilistic support vector machines and self–organizing maps, Water Resour. Manag. 27 (7) (2013) 2577–2594.

[63] R. Ziani, A. Felkaoui, R. Zegadi, Bearing fault diagnosis using multiclass support vector machines with binary particle swarm optimization and regularized Fisher's criterion, J. Intell. Manuf. 28 (2017) 405–417.

[64] J. Zeng, W. Qiao, Short–term solar power prediction using a support vector machine, Renew. Energ. 52 (2013) 118–127.

[65] E.G. Ortiz–García, S. Salcedo–Sanz, A.M. Pérez–Bellido, J.A. Portilla–Figueras, L. Prieto, Prediction of hourly O3 concentrations using support vector regression algorithms, Atmos. Environ. 44 (35) (2010) 4481–4488.

[66] M. Pal, A. Goel, Estimation of discharge and end depth in trapezoidal channel by support vector machines, Water Resour. Manag. 21 (10) (2007) 1763–1780.

[67] F.J. de Cos Juez, P.J. García Nieto, J. Martínez Torres, J. Taboada Castro, Analysis of lead times of metallic components in the aerospace industry through a supported vector machine model, Math. Comput. Model. 52 (2010) 1177–1184.

[68] R. De Leone, M. Pietrini, A. Giovannelli, Photovoltaic energy production forecast using support vector regression, Neural Comput. Appl. 26 (2015) 1955–1962.

[69] J. Shawe–Taylor, N. Cristianini, Kernel methods for pattern analysis. Cambridge University Press, New York, 2004.

[70] Q. Wu, The forescasting model based on wavelet support vector machine, Expert Syst. Appl. 36 (4) (2009) 7604–7610.

[71] C.–C. Wei, Wavelet kernel support vector machines forecasting techniques: Case study on water-level predictions during typhoons, Expert Syst. Appl. 39 (2012) 5189–5199.

[72] L. Zhang, W. Zhou, L. Jiao, Wavelet support vector machine, IEEE T. Syst. Man Cy. B 34 (1) (2004) 34–39.

[73] P.J. García Nieto, E. García–Gonzalo, J.R. Alonso Fernández, C. Díaz Muñiz, A hybrid wavelet kernel SVM–based method using artificial bee colony algorithm for predicting the cyanotoxin content from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain), J. Comput. Appl. Math. 309 (2017) 587–602.

[74] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE T. Inform. Theory 36 (2009) 961–1005.

[75] Q.H. Zhang, A. Benveniste, Wavelet networks, IEEE T. Neural Networ. 3 (1992) 889–898.

[76] G.Y. Chen, W.F. Xie, Pattern recognition with SVM and dual-tree complex wavelets, Image Vision Comput. 25 (2007) 960–966.

[77] N. Kingsbury, D.B.H. Tay, M. Palaniswami, Multi-Scale Kernel Methods for Classification, in: IEEE Workshop on Machine Learning for Signal Processing, Mystic, Connecticut, USA, 2005, pp. 43–48.

[78] M.S. Dalwani, Machine Learning in neuroimaging based modalities using support vector machines with wavelet kernels, PhD Dissertation, University of Colorado, 2017.

[79] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth and Brooks, Monterrey, CA, 1984.

[80] V. Rodriguez-Galiano, M.P. Mendes, M.J. Garcia-Soldado, M. Chica-Olmo, L. Ribeiro, Predictive modeling of groundwater nitrate pollution using random forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (southern Spain), Sci. Total Environ. 476–477 (2014) 189–206.

[81] L. Wang, X. Zhou, X. Zhu, Z. Dong, W. Guo, Estimation of biomass in wheat using random forest regression algorithm and remote sensing data, Crop J. 4 (2016) 212–219.

[82] R. Genuer, J.–M. Poggi, C. Tuleau–Malot, N. Villa–Vialaneix, Random forests for big data, Big Data Res. 9 (2017) 28–46.

[83] E.S. Allman, J.A. Rhodes, Mathematical Models in Biology: An Introduction, Cambridge University Press, New York, 2003.

[84] D.J. Barnes, D. Chu, Introduction to Modeling for Biosciences, Springer, New York, 2010.

[85] L. Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, New York, 2003.

[86] D. Freedman, R. Pisani, R. Purves, Statistics, W.W. Norton & Company, New York, 2007.

[87] R. Picard, D. Cook, Cross-validation of regression models, J. Am. Stat. Assoc. 79 (387) (1984) 575–583.

[88] C.–C. Chang, C.–J. Lin, LIBSVM: a library for support vector machines, ACM T. Int. Syst. Technol. 2 (2011) 1–27.

[89] K.–L. Du, M.N.S. Swamy, Search and Optimization by Metaheuristics: Techniques and Algorithms Inspired by Nature, Birkhäuser, New York, 2016.

[90] A. Kaveh, T. Bakhshpoori, Metaheuristics: Outlines, MATLAB Codes and Examples, Springer, Berlin, 2019.

[91] I. Chorus, J. Bartram, Toxic Cyanobacteria in Water: a Guide to Their Public Health Consequences, Monitoring and Management, E & FN Spon, London, 1999.

[92] M. Aboal, M.A. Puig, Intracellular and dissolved microcystins in reservoirs of the river Segura basin, Murcia, SE Spain, Toxicon 45 (4) (2005) 509–518.

[93] P.M. Gault, H.J. Marler, Handbook on *Cyanobacteria*: Biochemistry, Biotechnology and Applications, Nova Science Publishers, New York, 2009.

[94] C.S. Dow, U.K. Swoboda, Cyanotoxins, in: B.A. Whitton, M. Potts (Eds.), The Ecology of Cyanobacteria: Their Diversity in Time and Space, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000, pp. 613–632, 2000.

[95] C. Pérez–Martínez, P. Sánchez–Castillo, Temporal occurrence of *Ceratium hirundinella* in Spanish reservoirs, Hydrobiologia 452 (2004) 101–107.

[96] B.A. Zeeb, C.E. Christie, J.P. Smol, D.L. Findlay, H.J. Kling, H.J.B. Birks, Responses of diatom and Chrysophyte assemblages in Lake 227 sediments to experimental eutrophication, Can. J. Fish. Aquat. Sci. 51 (10) (1994) 2300–2311.

[97] E. Ortega–Mayagoitia, C. Rojo, Phytoplankton from the Daimiel National Park. II. Cyanophytes, dinoflagellates, cryptophytes, chrysophytes and xanthophytes, Anal. Jardín Bot. Mad. 57 (2) (199) 251–266.

[98] S. Fields, Global nitrogen: cycling out of control. Environ, Health Persp. 112 (10) (2004) A556–A563.

[99] X. Zhan, Y. Bo, F. Zhou, X. Liu, H.W. Paerl, J. Shen, R. Wang, F. Li, S. Tao, Y. Dong, X. Tang, Evidence for the importance of atmospheric nitrogen deposition to eutrophic lake Dianchi, China, Environ. Sci. Technol. 51 (12) (2017) 6699–6708.

[100] Y. Huang, M. Chen, Variation of dissolved oxygen in the experiments of occurrence & disappearance for Microcystis bloom, Procedia Environ. Sci. 18 (2013) 559– 566.

[101] V.K. Gadi, Y.–R. Tang, A. Das, C. Monga, A. Garg, C. Berretta, L. Sahoo, Spatial and temporal variation of hydraulic conductivity and vegetation growth in green infrastructures using infiltrometer and visual technique, Catena 155 (2017) 20–29.

[102] K.K. Arend, D. Betelsky, J.V. DePinto, S.A. Ludsin, J.J. Roberts, D.K. Rucinski, D. Scavia, D.J. Schwab, T.O. Hook, Seasonal and interannual effects of hypoxia on fish habitat quality in central Lake Erie, Freshwater Biol. 56 (2011) 366–383.

Table 1

Biological variables employed in this research with their mean, median, standard deviation (STD) and mean absolute deviation (MAD).

Biological input variables	Name of the variable	Mean	Median	STD	MAD
Cyanobacteria (mm^3/L)	*Cyanobacteria*	4.092	1.100	5.513	4.361
Diatoms (mm^3/L)	Diatoms	1.099	1.071	0.584	0.352
Euglenophytes (mm^3/L)	Euglenophytes	0.535	0.545	0.227	0.193
Dinophlagellata (mm^3/L)	*Dinophlagellata*	0.139	0.047	0.176	0.144
Chrysophytes (mm^3/L)	Chrysophytes	0.259	0.221	0.177	0.149
Chlorophytes (mm^3/L)	Chlorophytes	0.120	0.112	0.091	0.072
Chryptophytes (mm^3/L)	Chryptophytes	0.985	0.994	0.369	0.316

Table 2

Physical-chemical variables employed in this research with their mean, median, standard deviation (STD) and mean absolute deviation (MAD).

Physical-chemical input variables	Name of the variable	Mean	Median	STD	MAD
Water temperature (ºC)	Water_temp	17.057	17.000	4.103	3.252
Turbidity (NTU)	Turbidity	5.656	4.000	4.825	3.132
Nitrate concentration (mg NO^{3-} /L)	Nitrate	0.832	0.7100	0.407	0.299
Ammonium concentration (mg/L)	Ammonium	0.118	0.110	0.059	0.077
Dissolved oxygen concentration (mg O_2/L)	DOC	9.020	8.800	1.785	1.412
Conductivity (μS/cm)	Conductivity	268.222	275.000	42.944	30.903
pH values	pH_values	7.779	8.000	0.406	0.327
Secchi depth (m)	Secchi_depth	2.018	1.900	0.962	0.874

Table 3

Total phosphorus optimal hyperparameters obtained with the DE/SVM models.

Kernel	Optimal hyperparameters
Linear	Regularization factor $C = 1.5576\times10^{-1}$, $\varepsilon = 4.1438\times10^{-2}$
Quadratic	Regularization factor $C = 2.0778\times10^{1}$, $\varepsilon = 4.6164\times10^{-2}$, $\sigma = 1.5205\times10^{-1}$, $a = 3.4586\times10^{1}$, $b = 2$
Cubic	Regularization factor $C = 1.0000\times10^{-10}$, $\varepsilon = 5.3076\times10^{-2}$, $\sigma = 8.2509\times10^{2}$, $a = 2.5900\times10^{3}$, $b = 3$
Sigmoid	Regularization factor $C = 1.0000\times10^{4}$, $\varepsilon = 4.3521\times10^{-2}$, $\sigma = 1.6669\times10^{-5}$, $a = 1.9239\times10^{-1}$
RBF	Regularization factor $C = 8.9996\times10^{-1}$, $\varepsilon = 3.5670\times10^{-6}$, $\sigma = 8.6992\times10^{-1}$
Mult. Mexican Hat	Regularization factor $C = 4.5890\times10^{-1}$, $\varepsilon = 4.1286\times10^{-6}$, $\sigma_1 = 3.2224\times10^{-1}$, $\sigma_2 = 9.7516\times10^{-1}$

Table 4

Chorophyll concentration optimal hyperparameters obtained with the DE/SVM models.

Kernel	Optimal hyperparameters
Linear	Regularization factor $C = 2.4054\times10^{0}$, $\varepsilon = 2.6940\times10^{-2}$
Quadratic	Regularization factor $C = 1.0000\times10^{4}$, $\varepsilon = 1.0000\times10^{-10}$, $\sigma = 6.5403\times10^{-3}$, $a = 1.0000\times10^{-6}$, $b = 2$
Cubic	Regularization factor $C = 3.6585\times10^{-9}$, $\varepsilon = 2.8255\times10^{-2}$, $\sigma = 2.5188\times10^{2}$, $a = 2.3917\times10^{-6}$, $b = 3$
Sigmoid	Regularization factor $C = 8.8941\times10^{3}$, $\varepsilon = 2.6798\times10^{-2}$, $\sigma = 3.0068\times10^{-4}$, $a = 2.9717\times10^{-3}$
RBF	Regularization factor $C = 3.2195\times10^{0}$, $\varepsilon = 1.5907\times10^{-2}$, $\sigma = 9.9853\times10^{-1}$
Mult. Mexican Hat	Regularization factor $C = 1.4790\times10^{0}$, $\varepsilon = 1.2531\times10^{-2}$, $\sigma_1 = 4.2358\times10^{-1}$, $\sigma_2 = 1.2483\times10^{0}$

Table 5

Optimal hyperparameters obtained with the DE/RF–based model for total phosphorus.

Parameters	Values
Number of trees	111
Number of variables tried at each split	7

Table 6

Optimal hyperparameters obtained with the DE/RF–based model for chorophyll concentration.

Parameters	Values
Number of trees	37
Number of variables tried at each split	6

Table 7

Cross–validation coefficients of determination (R^2) and correlation coefficient (r), and root mean square error (RMSE) and mean absolute error (MAE) for the DE/SVM–based and DE/RF–based models for total phosphorus.

Model	Coeff.of det.(R^2)	Corr. Coeff. (*r*)	RMSE	MAE
Linear–SVM	0.8215	0.9064	0.010960	0.008467
Quadratic–SVM	0.8976	0.9474	0.005770	0.003513
Cubic–SVM	0.8968	0.9470	0.005861	0.004821
Sigmoid–SVM	0.8214	0.9063	0.010970	0.008468
RBF–SVM	0.9239	0.9612	0.003462	0.001456
Multiscale Mexican Hat Wavelet–SVM	**0.9255**	**0.9620**	**0.003436**	**0.001448**
Random Forest	0.9155	0.9568	0.008040	0.006017

Table 8

Cross–validation coefficients of determination (R^2) and correlation coefficient (r), and root mean square error (RMSE) and mean absolute error (MAE) for the DE/SVM–based and DE/RF–based models for the Chorophyll concentration.

Model	Coeff.of det.(R^2)	Corr. Coeff. (*r*)	RMSE	MAE
Linear–SVM	0.7264	0.8523	5.997	4.596
Quadratic–SVM	0.8356	0.9141	3.512	2.156
Cubic–SVM	0.8523	0.9232	3.106	2.131
Sigmoid–SVM	0.7263	0.8522	5.995	4.595
RBF–SVM	0.8803	0.9382	1.197	0.888
Multiscale Mexican Hat Wavelet–SVM	**0.8839**	**0.9402**	**1.047**	**0.741**
Random Forest	0.8418	0.9175	5.123	4.007

Table 9

Weights for the DE/SVM–based model for the Total phosphorus.

Variables	Weights
SecchiDepth	–1.2697
Turbidity	1.0654
Temperature	0.8112
Oxygen	–0.8068
Cyanobacteria	0.5719
Chlorophyll	0.5457
Dinophlagellata	0.4719
Ammonium	0.3878
Chlorophytes	–0.2297
Chrysophytes	0.1878
Nitrate	0.0754
pH	–0.0607
Euglenophytes	0.0410
Conductivity	0.0364
Chryptophytes	–0.0240
Diatoms	–0.0228

Table 10

Weights for the DE/SVM–based model for the Chlorophyll concentration.

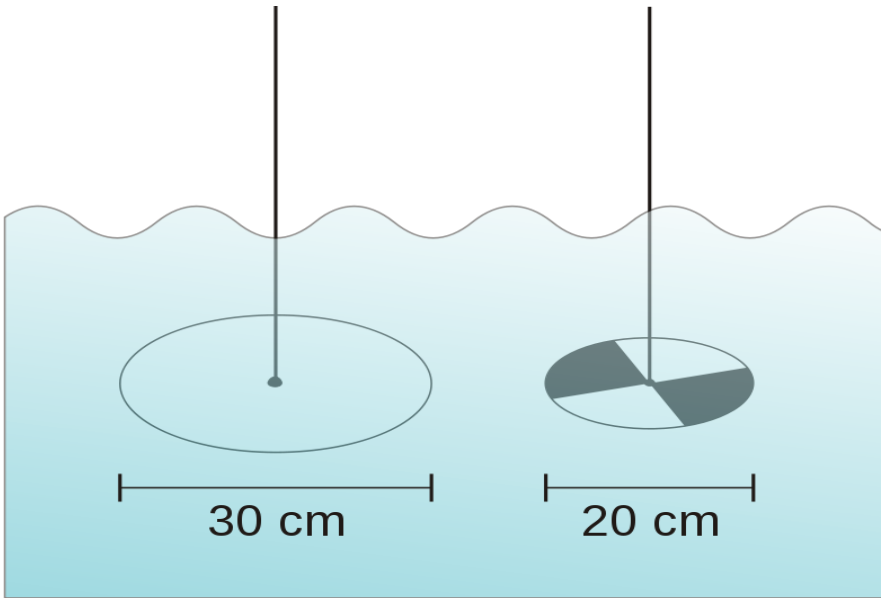
Variables	Weights
Cyanobacteria	2.1035
Oxygen	–1.0707
Phosphorus	1.0084
Dinophlagellata	0.9634
Conductivity	0.5363
SecchiDepth	–0.4296
Temperature	0.4008
Nitrate	0.3773
Diatoms	0.3591
Euglenophytes	0.3483
pH	–0.2760
Chlorophytes	–0.1184
Chrysophytes	–0.1071
Ammonium	–0.0817
Turbidity	0.0718
Chryptophytes	–0.0512

(a)

(b)

Fig. 1. (a) Large and (b) short scale aerial photographs of the reservoir.

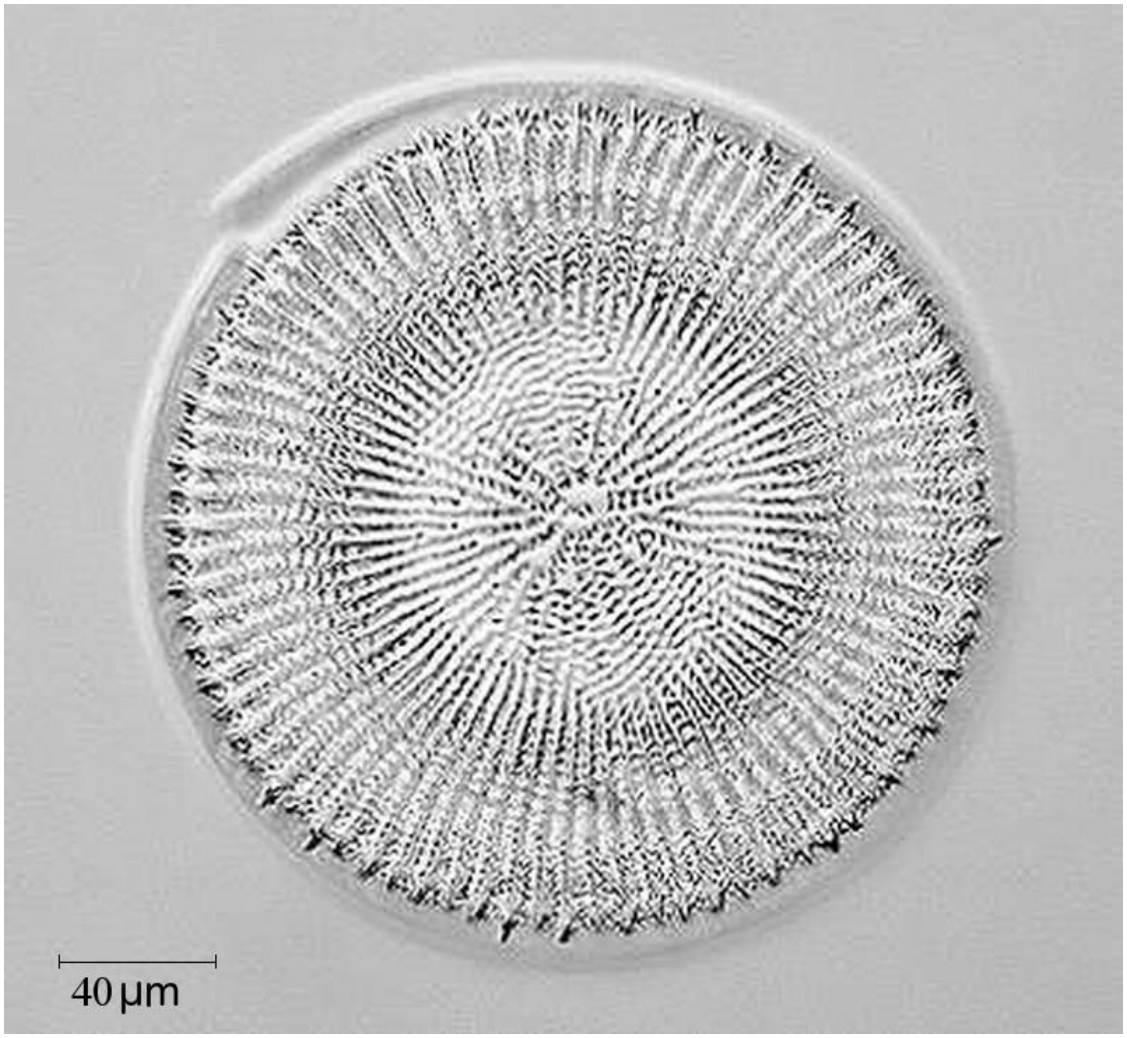
(a)



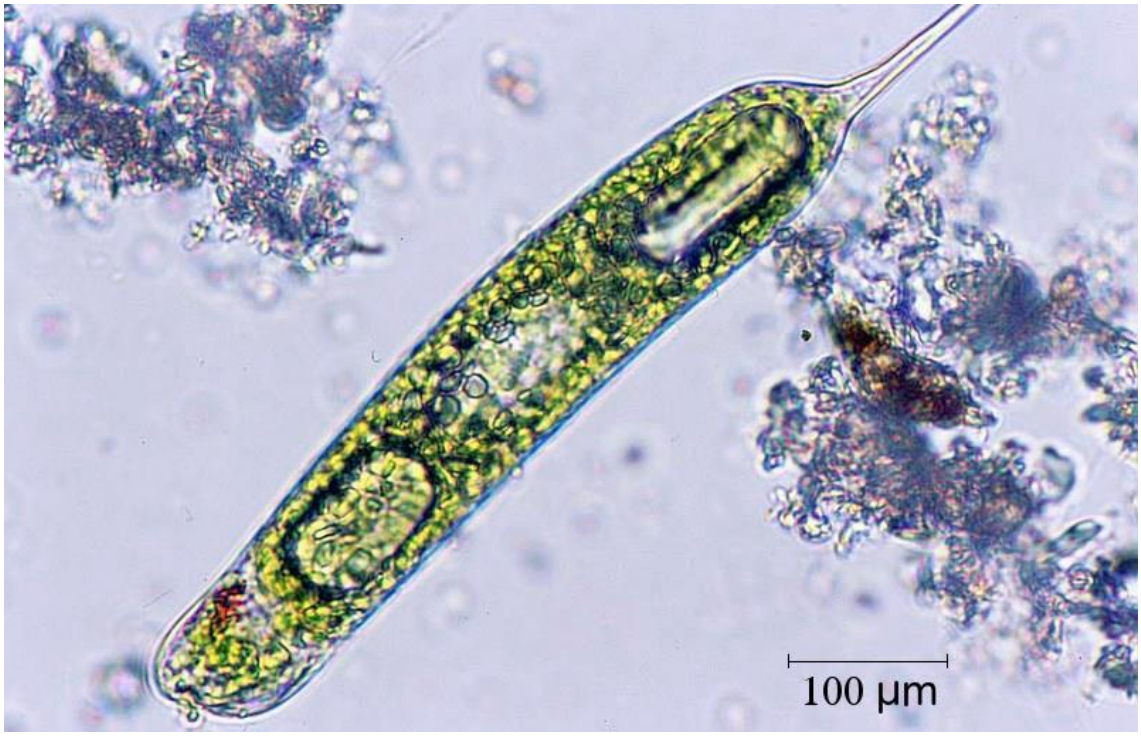
(b)

Fig. 2. (a) An example of a Niskin bottle; and (b) Secchi disks.

(a)



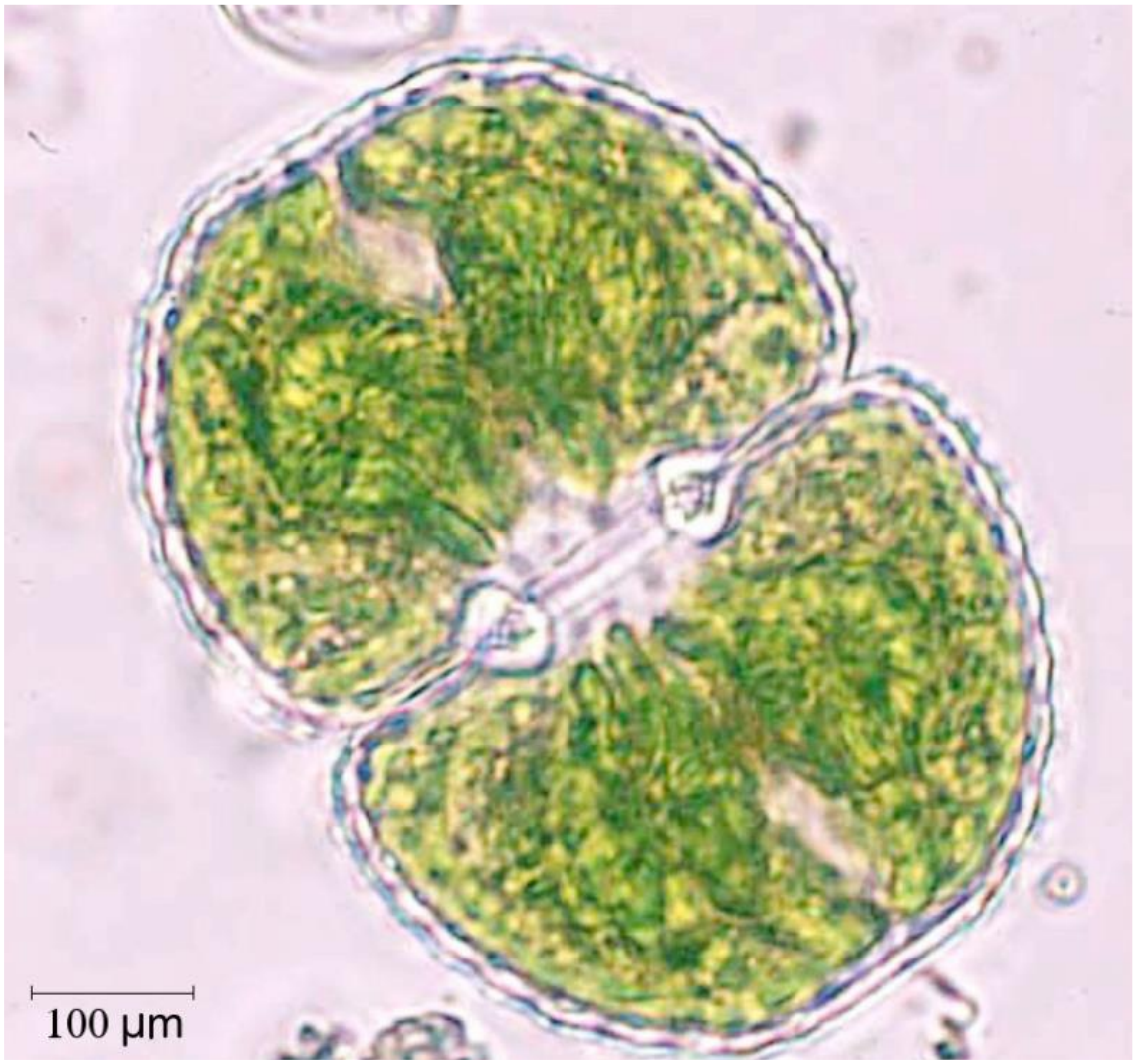
(b)



(c)

(d)

(e)



(f)

(g)

Fig. 3. Biological variables used for this research: (a) *Cyanobacteria*; (b) Diatoms; (c) Euglenophytes; (d) *Dinophlagella*; (e) Chrysophytes; (f) Clorophytes; and (g) Chryptophytes.

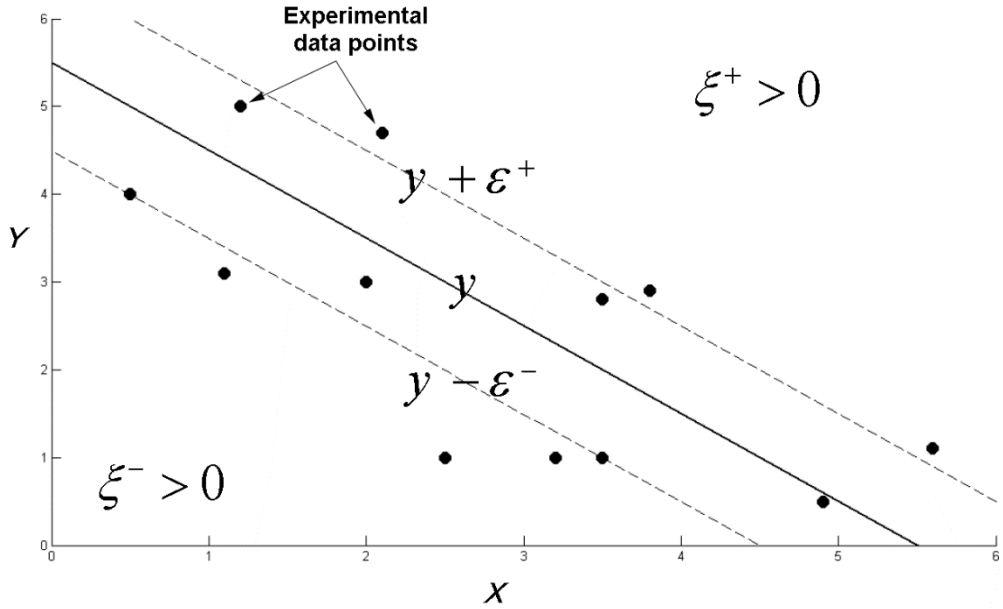


Fig. 4. Sketch of the SVM regression model with $\varepsilon-$insensitive tube for a one dimensional problem.

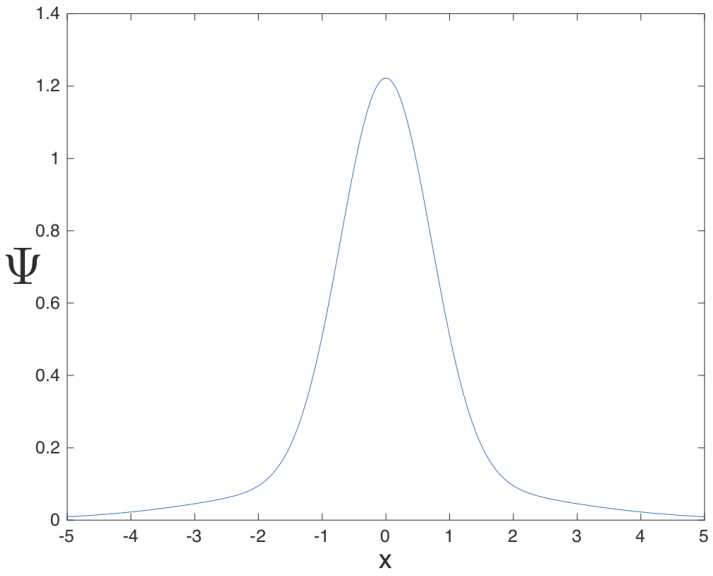


Fig. 5. Multiscale Mexican Hat wavelet function.

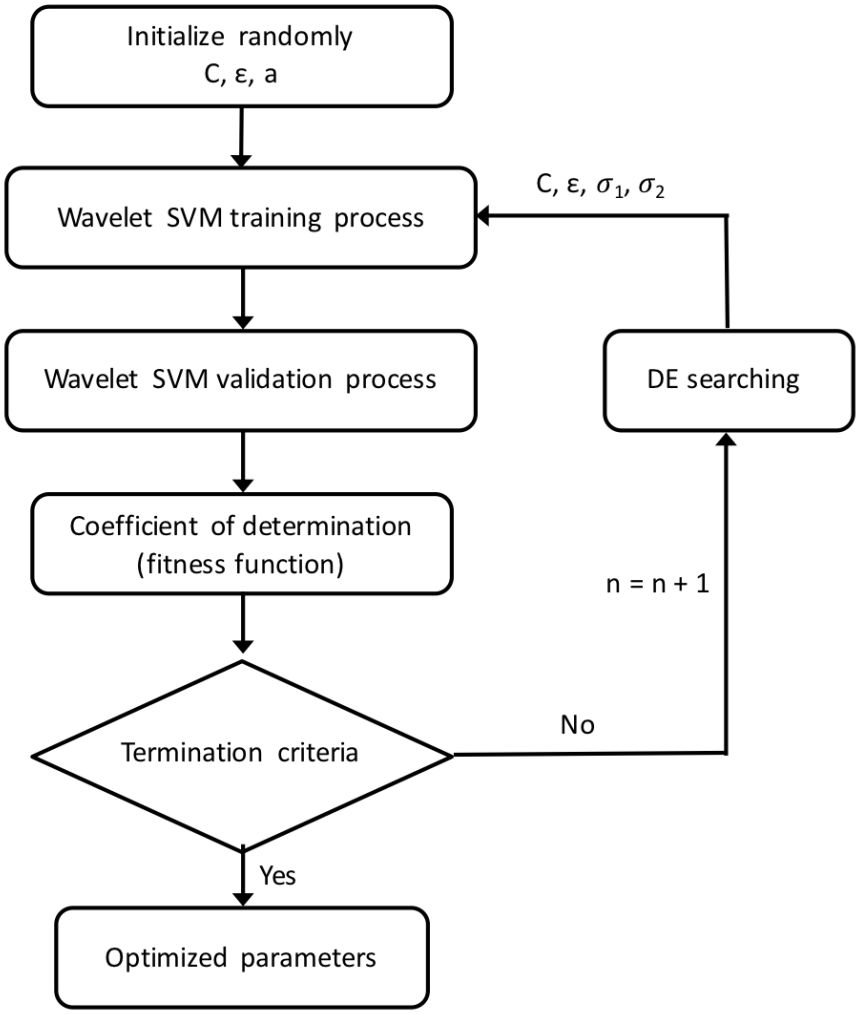


Fig. 6. Flowchart of the new hybrid DE/SVM–based model with multiscale Mexican Hat wavelet kernel.

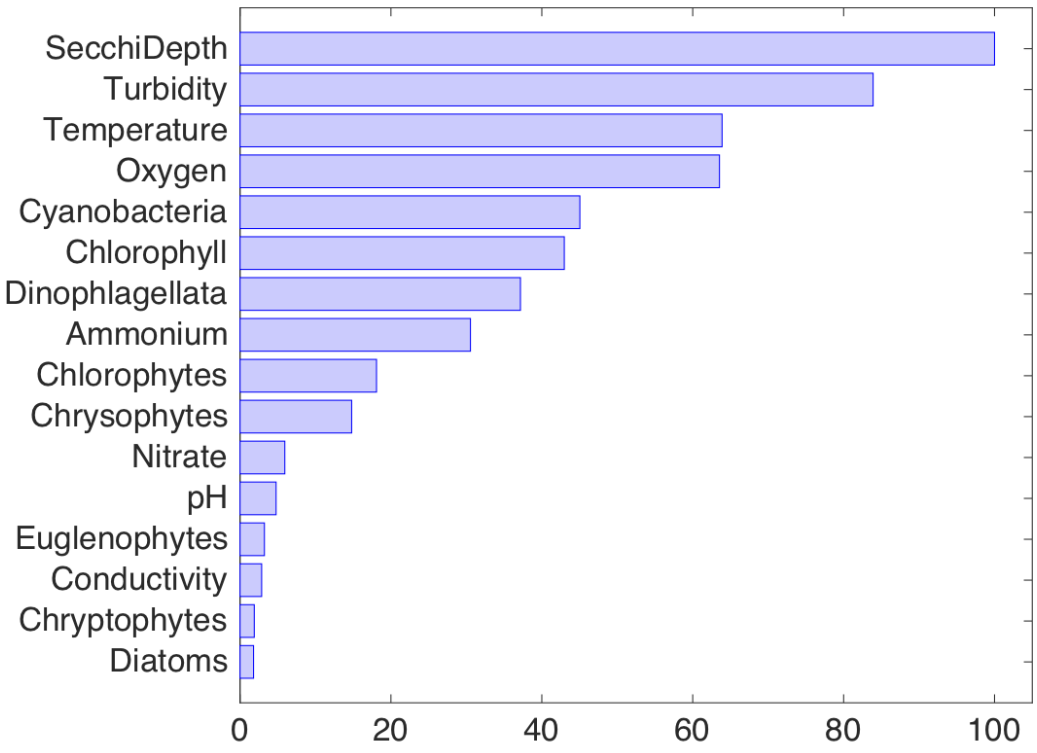


Fig. 7. Comparative significance of the predictor variables in the total phosphorus DE/SVM–based model.

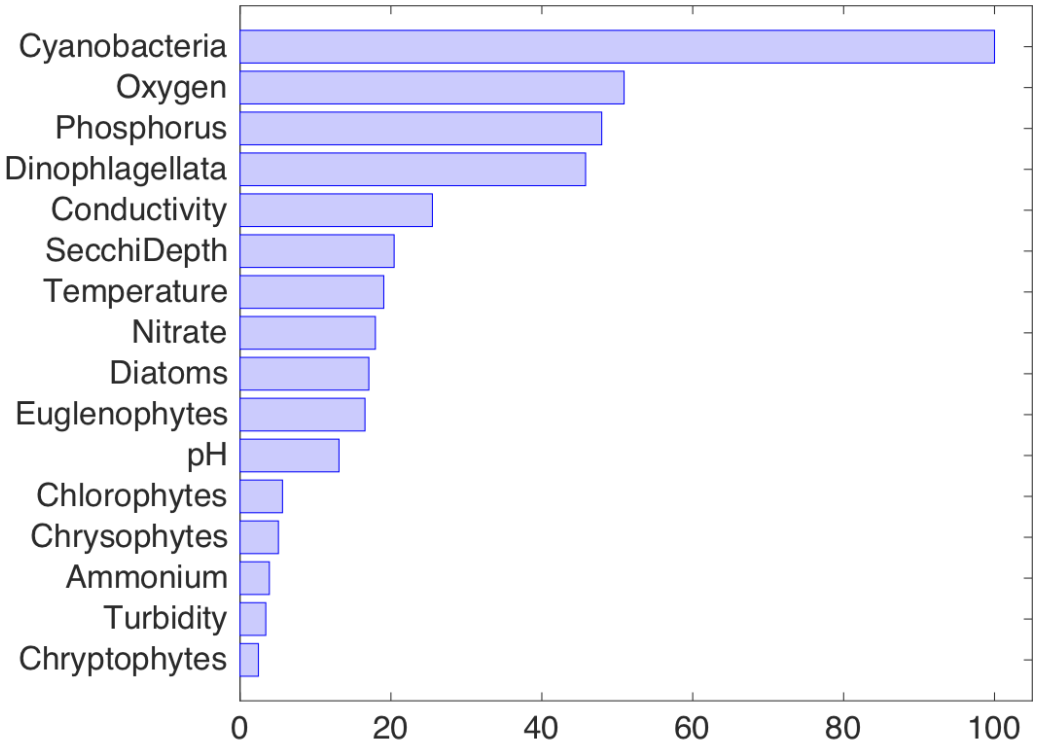
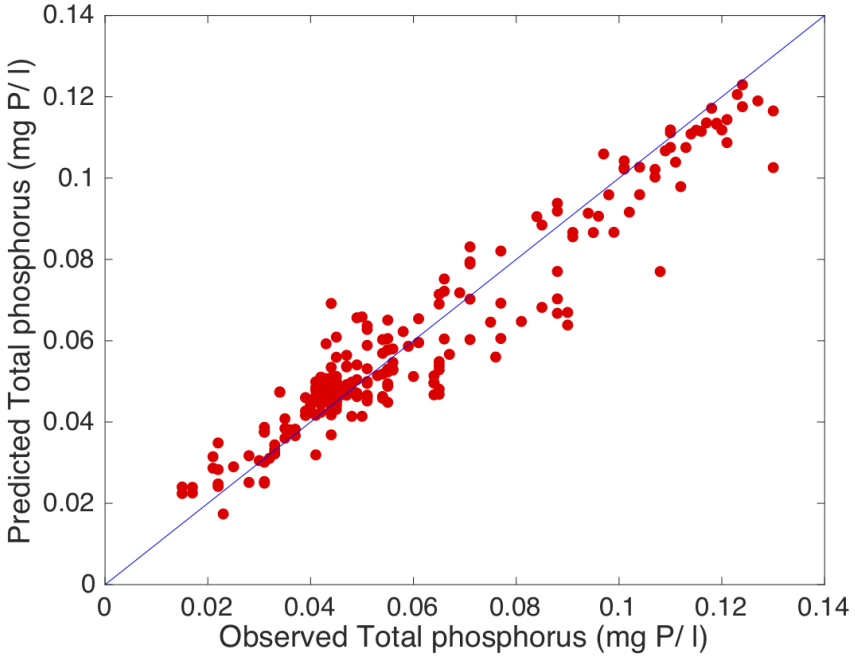
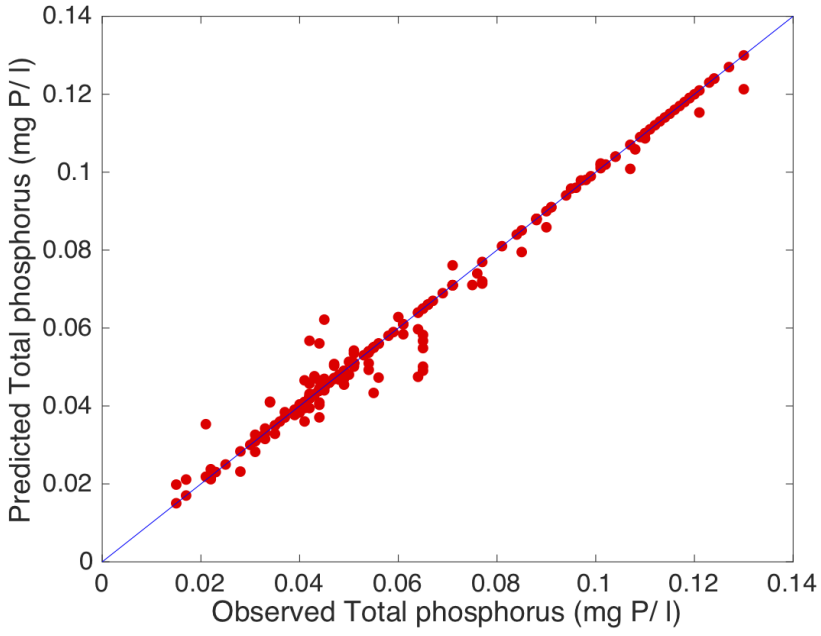


Fig. 8. Comparative significance of the predictor variables in the Chlorophyll DE/SVM–based model.

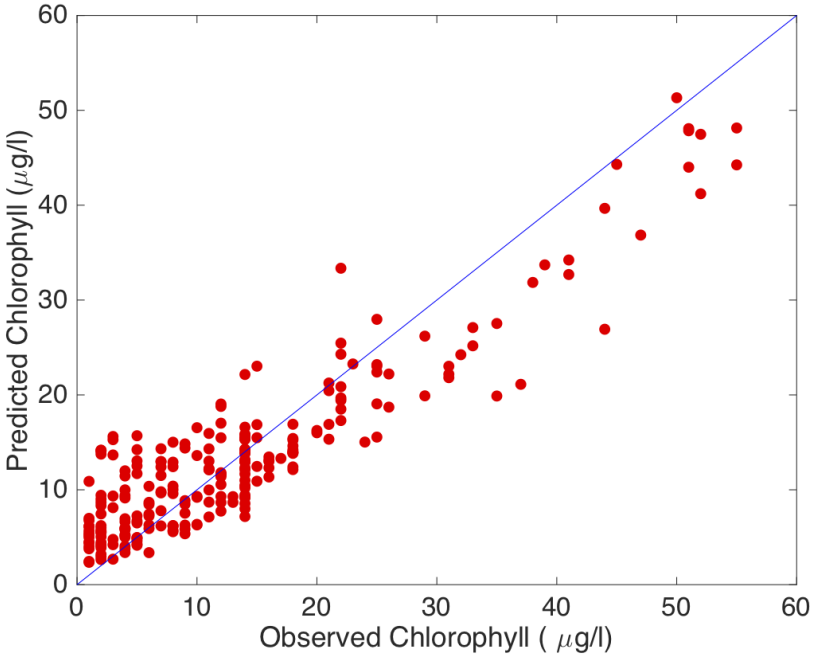


(a)

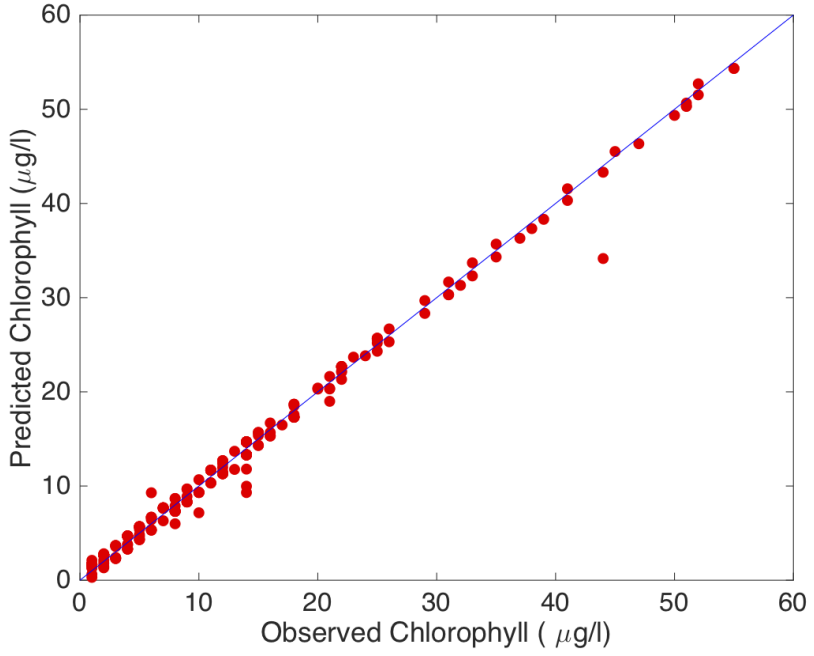


(b)

Fig. 9. Predicted vs. observed total phosphorus values with: (a) DE/RF–based model ($R^2 = 0.92$) and (b) Wavelet kernel DE/SVM–based model ($R^2 = 0.93$).



(a)



(b)

Fig. 10. Predicted vs. observed Chorophyll concentration values with: (a) DE/RF–based model ($R^2 = 0.84$) and (b) Wavelet kernel DE/SVM–based model ($R^2 = 0.88$).