Ecological Modelling 404 (2019) 91–102

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Water eutrophication assessment relied on various machine learning techniques: A case study in the Englishmen Lake (Northern Spain)

P.J. García Nieto[a,*], E. García-Gonzalo[a], J.R. Alonso Fernández[b], C. Díaz Muñiz[b]

[a] Mathematics Department, Facultad de Ciencias, Oviedo University, 33007, Oviedo, Spain
[b] Cantabrian Basin Authority, Ministry of Agriculture, Fishing, Food and Environment, 33071, Oviedo, Spain

ARTICLE INFO

Keywords:
Support vector machine (SVM)
Artificial bee colony (ABC)
Artificial neural networks (ANNs)
M5 model tree
Algal atypical productivity in lakes
Regression analysis

ABSTRACT

Algal atypical productivity, also called eutrophication, is a process where the phosphorus content in the water, together with aquatic flora, increases, causing high Chlorophyll levels and affecting the water quality and its possible applications. Therefore, it is important to be able to anticipate such circumstance to avoid subsequent hazards. In this paper, a model that estimates the conditions where an abnormal growth of algae in reservoirs and lakes takes place is built. This method combines artificial bee colony and support vector machines algorithms to predict the eutrophication taking into account physical-chemical and biological data sampled in the Englishmen Lake and posterior analysis in a laboratory. The support vector machines parameters are tuned by means of the artificial bee colony algorithm, improving the accuracy of the procedure. For comparison sake, two other methods have been used to construct additional models, the M5 model tree and multilayer perceptron network. Two objectives are covered by this study: the forecasting of the algal proliferation by means of the model and, the ranking of the relative importance of the independent variables. Indeed, coefficients of determination of 0.92 for the Chlorophyll and 0.90 for the Total phosphorus concentrations were obtained using this hybrid method that optimizes the regression parameters. Furthermore, the results obtained with M5 model tree and multilayer perceptron network techniques were clearly worse. Finally, conclusions of this work are drawn in the final section.

1. Introduction

Algal atypical proliferation, also known as eutrophication, has become one of the main issues for the water quality standards due to the damage caused by the overabundance of nutrients that modifies the right balance of the natural ecosystem in lakes and rivers (Reynolds, 2006). It has various negative effects: causes the color, taste, transparency, and odor of lake water to change, reduces the lake biodiversity, and also results in supersaturation and deficits of oxygen in the surface water and bottom water layers, respectively. Thus, eutrophication is a type of water pollution due to the increase of nutrients in the water, particularly phosphorus (P) and nitrogen (N), that causes the diminishing of the oxygen dissolved in the water, turbidity, growth of toxic alga and loss of aquatic flora and fauna species (Takaara et al., 2010). Specifically, phosphorus is a necessary nutrient for plants to live, and is the *limiting factor* for plant growth in most freshwater ecosystems (Scholten et al., 2006; Ansari and Gill, 2016). The availability of phosphorus generally promotes excessive plant growth and decay, favoring simple algae and plankton over other more complicated plants, and causes a severe reduction in water quality. Therefore, these blooms are usually due to alterations in the physical–chemical conditions in water bodies (Álvarez Cobelas and Arauzo, 2006; Paerl et al., 2011).

An additional important factor in water algal anomalous productivity is Chlorophyll a (*Chl*-a) because it is an algae and macrophytes pigment and thus, it is relevant in the process of algae growth (Gibson et al., 2000; Chen et al., 2003). The progression of eutrophication symptoms is well described (Bricker et al., 2003; Reynolds, 2006) and most eutrophication assessment methods recognize that the immediate biological response is increased primary production reflected as increased chlorophyll a (*Chl*-a) (Ferreira et al., 2007; Xiao et al., 2007). Apart from these variables, other environmental factors, such as water temperature, pH, dissolved oxygen, secchi depth, ammonium, nitrogen, etc., can also highly affect the degree of algal proliferation (Wang et al., 2008). In light of the complicated origins of algal abnormal growth for real water environment, the development of quantitative relationships between a variety of environmental factors and eutrophic indicators is highly desirable for formulating strategies for the prevention of algae bloom.

* Corresponding author.
E-mail address: lato@orion.ciencias.uniovi.es (P.J. García Nieto).

https://doi.org/10.1016/j.ecolmodel.2019.03.009
Received 17 September 2018; Received in revised form 2 February 2019; Accepted 18 March 2019
Available online 05 April 2019

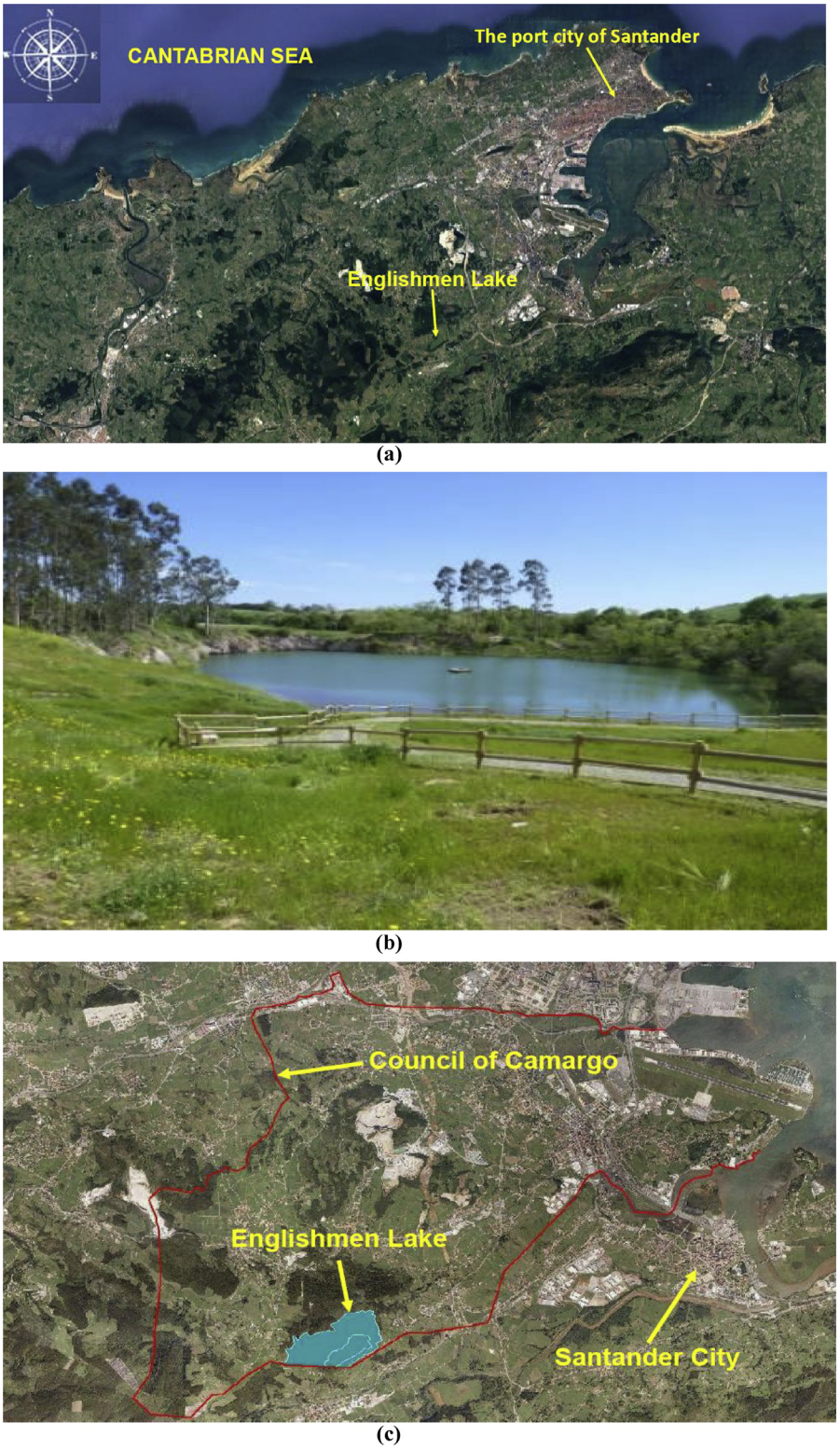

Fig. 1. (a) Aerial photograph of the Englishmen Lake and the port city of Santander (Autonomous Community of Cantabria, Spain); (b) side view of the Englishmen Lake at a larger scale; and (c) study area in the context of Autonomous Community of Cantabria.

The gathering of data for the study of algal abnormal growth can be done within a monitoring program (Kitsiou and Karydis, 2011) or as a particular project. Such data must contain ecological indicators (Karydis, 2009) because they form the basis of a scoring system to characterize the trophic status (Reynolds, 2006), especially for the implementation of the Water Framework Directive (WFD) (Directive 2000/60/EC, 2000Directive, 2000Directive 2000/60/EC, 2000). In this sense, the analysis should include the estimation of the biovolumes (Hillebrand et al., 1999).

Also, algal abnormal productivity is a critical environmental problem in water bodies (e.g. the Englishmen Lake) (see Fig. 1) because eutrophication affects the ecosystem health.

In particular, the Englishmen Lake is prone to algal proliferation that produce hypoxia with all its associates' processes (Díaz and Rosenberg, 2011) and growth of toxic alga. Moreover, the complex mechanism of the algal abnormal growth has not been solved yet due to the nonlinear response of water quality for nutrient loading (Charpa, 1997). However, the fact is that the reduction of nutrient loads is essential to improve the water quality and ecological restoration of water bodies.

In this study, we have used with success the application of support vector machines (SVMs) hybridized with Artificial Bee Colony (ABC) and, for comparative purposes, M5 model tree and Multilayer Perceptron (MLP) to estimate the eutrophication state in Englishmen Lake comparing the results obtained. On the one hand, the SVM approach is and is a method that is based on the statistical learning theory that can be used for the estimation of relevant variables in very different fields (Álvarez Antón et al., 2013; Cortes and Vapnik, 1995; Hansen and Wang, 2005; Hastie et al., 2003; Vapnik, 1998). SVMs is used for classification and regression for multivariate functions with high degree of accuracy. The structural risk minimization and statistical learning theory form the theoretical base of SVM (Kecman, 2005).

Furthermore, a multilayer perceptron is a particular feedforward artificial neural network that is able to perform regressions. The statistical learning theory also provides the foundations of MLPs, which have been used in very different areas (Fine, 1999; Hassoun, 2003; Haykin, 1999). MLP trains the network using backpropagation. MLP is a adaptation of the classic linear perceptron and it is able to process data nonlinearly separable.

The artificial bee colony (ABC) technique was used to optimize the kernel SVM hyperparameters. The artificial bee colony algorithm is a swarm intelligence optimization method founded in the behavior of bee swarms in the search of food. Similar to other evolutionary computation algorithms such as particle swarm optimization (PSO) (Clerc, 2006; Eberhart et al., 2001; Olsson, 2011) or ant colony optimization (Dorigo and Stützle, 2004), ABC exploits the model of social sharing of information (Karaboga and Basturk, 2007; Simon, 2013; Yang et al., 2013). According to previous studies, the SVM method has proven to be effective in the modeling of natural parameters, and has been successfully used in the environmental problems such as the assessment of air quality (García Nieto et al., 2013), the modeling of forest (García Nieto et al., 2012) or the estimation of solar radiation (Zeng and Qiao, 2013).

Additionally, the M5 model tree is a regressor based on decision trees. M5 tree was invented by J.R. Quinlan (Quinlan, 1992). M5 tree falls within the CART tree (Breiman et al., 1984) type of algorithms because it chooses as impurity function the mean squared error. It differs from the CART model in that it builds multivariate linear regression model and it does not attach a constant to the leaf node. The M5 model tree is thus analogous to piecewise linear functions. Another advantage of M5 over CART is that the generated models are much smaller but more accurate (Quinlan, 1992).

In summary, the organization of this paper is as follows: materials, methods and dataset are described in Section 2; the results of the SVM model hybridized with ABC are presented, discussed and compared with the MLP technique and M5 model tree in Section 3; finally, the main conclusions of this research work are exposed in Section 4.

2. Materials and methods

2.1. Study area

The Englishmen Lake is a water body located in the Camargo municipality (Autonomous Community of Cantabria, Spain) 5.5 km away from the port city of Santander (Fig. 1(c)). It was created for an old mining exploitation. It now collects water from the surrounding water table and has naturally evolved into a lake of high ecological value as winter migratory birds (e.g. *Anser anser*) make use of it.

The study area is located on the eastern edge of the *Asturian Massif*. From the stratigraphic point of view, the materials with the highest permeability values in this study area are the Quaternary deposits and the terraces located on the slopes, in addition to the calcareous formations. In the area, two aquifers have been differentiated: the quaternary detritic aquifer and the Jurassic calcareous aquifer. The first is mainly made up of sands, gravels and silts, with an average thickness between 6 and 8 m. The Jurassic calcareous aquifer includes the carbonated formations of the Upper Lias. Between both aquifers there is a direct water connection, so that the discharge of the calcareous aquifer takes place through the Quaternary aquifer. These aquifers belong to the *Puerto del Escudo* underground water body, with minimum storage capacity for the entire water body is about $68.5 \times 10^6\ m^3$. The Englishmen Lake occupies a water area of $13.6 \times 10^4\ m^2$ and has a maximum depth of 6 m.

2.2. Experimental dataset

The data for the analyses were collected from 2006 to 2014 from samples taken from Englishmen Lake, resulting in 244 different processed samples. It gives quantitative information about the phytoplankton species. Specifically, samples were usually taken more than once per month, and the protocols of the Spanish Ministry of Agriculture, Fishing, Food and Environment were followed (World Health Organization, 1998; Willame et al., 2005).

The sampling is taken with a Niskin bottle in the greatest depth point of the lake, being this found using a depth gauge. The Niskin bottle is an evolution of the Nansen bottle. It is a tube open at both ends with a cap at each one that can be applied with a tensioned elastic rope. Samples are taken at different depths within the zone that is exposed to enough sunlight for photosynthesis to happen, which is called the euphotic zone. This zone is defined with the Secchi disk, that determines the Secchi depth as that where the Secchi disk cannot be viewed due to the turbidity of the water. This depth is an indicator of the transparency of the water. The biological data were calculated from five subsamples taken at equal intervals of depth in the euphotic zone (Brönmark and Hansson, 2005).

In this study, the physical–chemical variables measured in limnological studies were used (Gault and Marler, 2009; Negro et al., 2000). They were analyzed in an ISO 17025 accredited laboratory and the standard procedures for the Examination of Water and Wastewater were followed (American Public Health Association, 2005).

The objective of this work is to study the relationship between algal atypical proliferation indicators and the independent variables. There are many physical–chemical and biological variables involved in the eutrophication of a lake. Thus the first step is the selection of the variables for the model. An expert criterion was used in this work. The estimated variables are related to the algae proliferation that are used to evaluate algal anomalous algal growth in water bodies: Chlorophyll concentration (Chl-a) (μg/L) and Total phosphorus (mg P/L). Chlorophyll concentration is biomolecule directly related to photosynthesis, the process where energy from light is obtained. Total phosphorus is a nutrient for aquatic organisms. The independent variables are:

- Biological parameters: euglenophytes (mm^3/L) (see Fig. 2(a)); Cyanobacteria (mm^3/L) (see Fig. 2(b)); dinophlagellata (mm^3/L) (see Fig. 2(c)); chlorophytes (mm^3/L) (see Fig. 2(d)); diatoms (mm^3/L) (see Fig. 2(e)); chrysophytes (mm^3/L) (see Fig. 2(f)); and cryptophytes (mm^3/L) (see Fig. 2(g)).
- Physical–chemical parameters: dissolved oxygen concentration (mg O_2/L); turbidity (NTU); nitrate concentration (mg NO_3^-/L); ammonium ion concentration (mg/L); conductivity (μS/cm); water temperature (ºC) and pH.

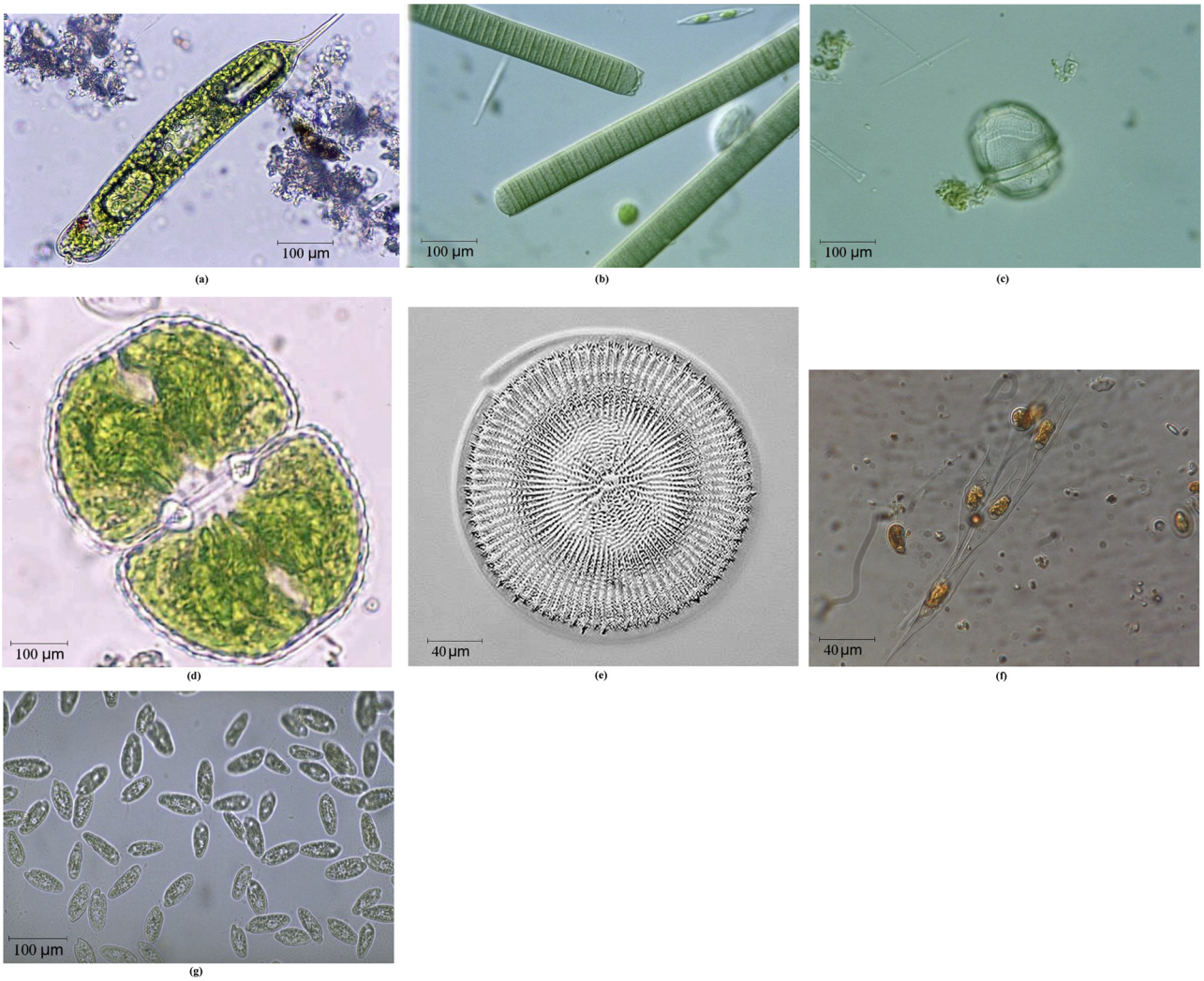

Fig. 2. Biological independent variables in the Englishmen Lake: (a) Euglenophytes; (b) Cyanobacteria; (c) Dinophlagella; (d) Clorophytes; (e) Diatoms; (f) Chrysophytes; and (g) Chryptophytes.

2.3. Computational procedure

2.3.1. Support vector machine (SVM) method

Support vector machines are used for classification and regression and belong to the supervised learning methods (Cristianini and Shawe-Taylor, 2000; Schölkopf et al., 2000). The SVMs were intended for classification, and only later they were generalized for regression problems (Chen et al., 2013; Nikoo and Mahjouri, 2013; Ortiz-García et al., 2010; Pal and Goel, 2007; Shrestla and Shukla, 2015; Zeng and Qiao, 2013), being then called *support vector regression* (SVR). The SVR model does not use the whole training data, because the cost function ignores (within a thresholdε) the samples too close to the model. If the data is non-linear but separable, the *kernel trick* can be used. Many regression problems cannot be linearly treated in the space of the independent values X, but can be translated to a higher dimensionality feature space if a suitable mapping $\mathbf{x} \mapsto \psi(\mathbf{x})$ is given (Cristianini and Shawe-Taylor, 2000; Ortiz-García et al., 2010; Schölkopf et al., 2000; Shrestla and Shukla, 2015) where an adequate regression can be obtained.

In SVR we wish to predict a real-valued output y for the observed value t so that our training data is a set of L points of the form $\{\mathbf{x}_i, t_i\}$, where $i = 1, 2, ...,L$, $y \in \Re$, $\mathbf{x} \in \Re^D$ (Cristianini and Shawe-Taylor, 2000; Nikoo and Mahjouri, 2013; Ortiz-García et al., 2010; Pal and Goel, 2007; Schölkopf et al., 2000; Shrestla and Shukla, 2015; Steinwart and Christmann, 2008) so that:

$$y_i = \mathbf{w}\cdot\mathbf{x}_i + b \tag{2}$$

where ". " denotes the dot product, $\mathbf{x}_i$ is a D–dimensional real input vector, $\mathbf{w}$ is a vector normal to the maximum-margin hyperplane and y_i is the predicted output value. The parameter $\frac{b}{||\mathbf{w}||}$ is used to determine the offset of this plane from the origin and along the normal vector $\mathbf{w}$. The SVR uses a penalty function where the error of corresponding to the predicted value y_i is not taken into account if it is less than a distance ε away from the observed value t_i, i.e., if $|t_i - y_i| < \varepsilon$. We can see in Fig. 3, the region where $y_i \pm \varepsilon \ \ \forall\ i$ that is called $\varepsilon-$ insensitive tube. Also, the dependent variables outside the tube are placed in one of two slack variable penalties, depending on whether they are above (ξ^+) or below (ξ^-) the tube, with $\xi^+ > 0$, $\xi^- > 0 \ \ \forall\ i$:

$$t_i \le y_i + \varepsilon + \xi^+ \tag{3}$$

$$t_i \ge y_i - \varepsilon - \xi^- \tag{4}$$

We are trying to find a function that can accurately estimate new values from a set of new independent variables. To this aim, the SVM model is trained in the so called training set. At this stage the error function is sequentially optimized. Two different types of error functions are defines and, consequently, to kind of SVM models can be defined:

a) Regression SVM Type 1 (also known as $\varepsilon-$ SVM regression): for this

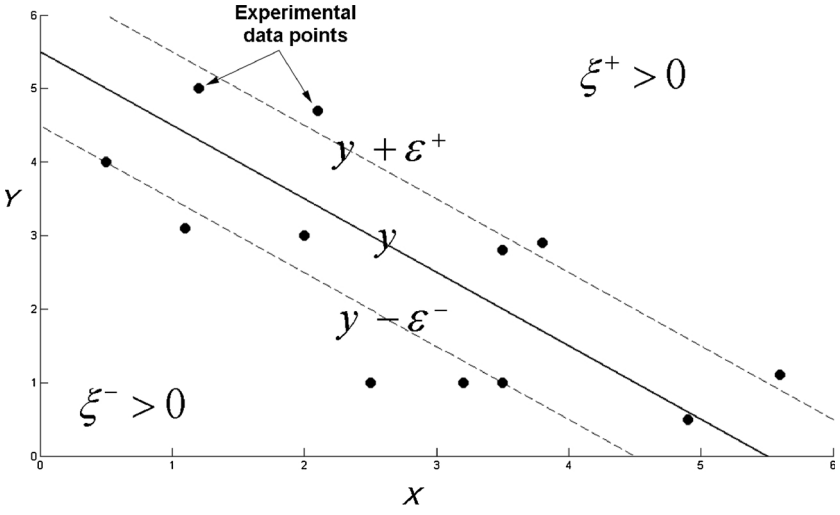

Fig. 3. Regression with ε^- insensitive tube for one–dimensional problem.

type of SVM, we have to solve an optimization problem minimizing the following general risk function (Cristianini and Shawe-Taylor, 2000; Hansen and Wang, 2005; Heddam and Kisi, 2018; Schölkopf et al., 2000; Steinwart and Christmann, 2008):

$$R[\mathbf{w}, b, \xi] = \frac{1}{2} ||\mathbf{w}||^2 + C \sum_{i=1}^{L} (\xi_i^+ + \xi_i^-) \tag{5}$$

where C is the *penalization* or *cost parameter* to control the training errors, that is to say, C measures a trade-off between the empirical error and the model complexity, and subject to:

$$\left\{ \begin{array}{c} \langle \mathbf{w}, \psi(\mathbf{x}_i) \rangle + b - y_i \leq \varepsilon + \xi_i^+ \\ y_i - \langle \mathbf{w}, \psi(\mathbf{x}_i) \rangle - b \leq \varepsilon + \xi_i^- \\ \xi_i^+, \xi_i^- \geq 0 \end{array} \right\} \; i = 1, \; ..., L \tag{6}$$

- Regression SVM Type 2 (also known as $\nu-$ SVM regression): for this SVM model, it is necessary to solve the following optimization problem, minimizing the following general risk function (Cristianini and Shawe-Taylor, 2000; Hansen and Wang, 2005; Heddam and Kisi, 2018; Schölkopf et al., 2000; Steinwart and Christmann, 2008):

$$R[\mathbf{w}, b, \xi] = \frac{1}{2} ||\mathbf{w}||^2 - C \left[\nu\varepsilon + \frac{1}{L} \sum_{i=1}^{L} (\xi_i^+ + \xi_i^-) \right] \tag{7}$$

being C again the *penalty parameter*. Indeed, C determines the trade-off between the margin (model flatness) and the magnitude of the slack variables (training error), and subject to:

$$\left\{ \begin{array}{c} \langle \mathbf{w}, \psi(\mathbf{x}_i) \rangle + b - y_i \leq \varepsilon + \xi_i^+ \\ y_i - \langle \mathbf{w}, \psi(\mathbf{x}_i) \rangle - b \leq \varepsilon + \xi_i^- \\ \xi_i^+, \xi_i^- \geq 0 \end{array} \right\} \; i = 1, \; ..., L \tag{8}$$

where $\psi : X \rightarrow Z$ is a mapping of the independent variables space into a larger dimension space Z, where an inner product is defined using a positive definite function k (kernel trick) (Hansen and Wang, 2005; Schölkopf et al., 2000; Steinwart and Christmann, 2008):

$$\langle \psi(\mathbf{x}), \; \psi(\mathbf{x}') \rangle = \sum_i \psi_i(\mathbf{x}) \, \psi_i(\mathbf{x}') = k(\mathbf{x}, \mathbf{x}') \tag{9}$$

In fact, if a function satisfies Mercer's condition (Chen et al., 2013; Hansen and Wang, 2005; Steinwart and Christmann, 2008), then it is an allowable support vector kernel. Furthermore, as this problem is quadratic with linear constraints, the optimality conditions of Karush-Kuhn-Tucker are necessary and sufficient and the solution, which is a linear combination of a subset of sample points denominated support vectors (s.v.) can be obtained from the dual problem, as follows:

$$\mathbf{w} = \sum_{\text{s.v.}} \beta_i \psi(\mathbf{x}_i) \Rightarrow$$
$$f_{\mathbf{w}, b}(\mathbf{x}) = \sum_{\text{s.v.}} \beta_i \langle \psi(\mathbf{x}_i), \; \psi(\mathbf{x}) \rangle + b = \sum_{\text{s.v.}} \beta_i \, k(\mathbf{x}_i, \; \mathbf{x}) + b \tag{10}$$

It is important to select an appropriate kernel function because the kernel function defines the space in which the dataset is regressed. Examples of kernel functions:

- RBF (Radial basis function):

$$k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\sigma ||\mathbf{x}_i - \mathbf{x}_j||^2} \tag{11}$$

- Polynomial:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\sigma \mathbf{x}_i \cdot \mathbf{x}_j + a)^b \tag{12}$$

- Sigmoid:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\sigma \mathbf{x}_i \cdot \mathbf{x}_j + a) \tag{13}$$

being a, b and σ the kernel parameters.

2.3.2. The artificial bee colony (ABC) algorithm

The algorithm Artificial Bee Colony (ABC) is Swarm Intelligence optimization algorithm that relies in the behaviour of bee swarms searching for food sources (Karaboga and Basturk, 2007; Karaboga and Akay, 2009; Karaboga and Gorkemli, 2014; Simon, 2013; Yang et al., 2013). In this evolutionary algorithm, a population of possible solutions evolves with the iterations toward the optimum using a strategy that involves some random component (Karaboga and Basturk, 2007; Karaboga and Akay, 2009; Karaboga and Gorkemli, 2014). Artificial Bee Colony algorithm is a swarm–based algorithms which is characterized by the sharing of information between the individuals in the swarm or population. Indeed, the bee colony has three groups of bees: employed, onlookers and scouts. An employed bee is assigned to a food. The so called employed bees fly to their food source and return to inform the hive with a dance. Onlookers choose food sources depending on dances. If the food source is not good enough, it is abandoned and the employed bee becomes a scout and searches a new food source. Therefore, ABC considers three phases (Karaboga and Basturk, 2007; Simon, 2013; Yang et al., 2013):

- Searching (or the employee bee phase): The process to find nectar source, that is to say, each food source is foraged by employee bees.
- Recruiting (or the onlooker bee phase): they choose a food source watching the dance of employed bees within the hive. The foraging is supervised and sometimes corrected by the onlooker.

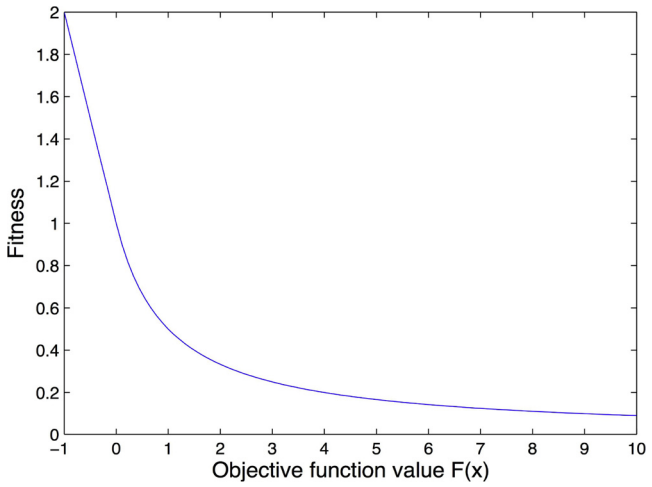

Fig. 4. Relationship between the objective function and the fitness function of a food source.

- Abandon (or the scout bee phase): the depleted sources are discarded and the scout bees search for new food sources.

The N food sources, the vectors $\mathbf{p}_i$, represent the possible set of parameters. It is a position in the space of possible solutions. The dimension of the food source is the number of parameters we are optimizing. The food sources are initialized randomly in a plausible hypercube and their fitness is evaluated. If f is the objective function, the fitness of a food source is (see Fig. 4):

$$Fitness(F(\mathbf{p}_i)) = \begin{cases} \frac{1}{1+F(\mathbf{p}_i)} & if \quad F(\mathbf{p}_i) > 0 \\ 1 + |F(\mathbf{p}_i)| & if \quad F(\mathbf{p}_i) \leq 0 \end{cases} \quad i = 1, \ldots, N \tag{14}$$

As the objective function decreases, the fitness function increases (we are minimizing). If we want to maximize a function g the objective function must be $F = -g$ and then a maximum of g is a minimum of F.

2.3.2.1. The employee bee phase. In this phase the employee bees searches food sources introduces a tentative variation of every i food source following the equation (Karaboga and Basturk, 2007; Simon, 2013; Yang et al., 2013):

$$v_{ij} = p_{ij} + R_{ij}(p_{kj} - p_{ij}) \tag{15}$$

where j is one of the parameters, chosen randomly, that we are trying to modify, k a food source different from i, also randomly chosen, and R_{ij} a real number drawn from $[-1, 1]$. Once v_{ij} has been computed, its fitness is calculated. If it is higher than the *fitness* $(F(p_{ij}))$, its value is swapped with v_{ij} and the trial counter is reinitialized. If it is smaller, the trial counter is increased and there is no change in this food source.

2.3.2.2. The onlooker bee phase. We draw a number r_i in $[0, 1]$, for each food source $\mathbf{p}_i$. We try to change one parameter in the food source, if $r_i < prob_i$, We compute $prob_i$ from the fitness of the food source (Karaboga and Basturk, 2007; Simon, 2013; Yang et al., 2013):

$$prob_i = \frac{0.9\, Fitness(F(\mathbf{p}_i))}{\max\limits_{k=1,\ldots,N}(Fitness(F(\mathbf{p}_k)))} + 0.1 \tag{16}$$

2.3.2.3. The scout bee phase. A food source is abandoned if, given a number of trials, it does not improve its fitness. Once a food source is discarded, a new one is chosen randomly. For each iteration, the provisional optimum is the food source with the highest fitness (Karaboga and Basturk, 2007; Simon, 2013; Yang et al., 2013).

This continues until a stopping condition is satisfied. In the present case, the stopping condition has been a maximum number of iterations and a difference of fitness for the best food source between consecutive iterations under a threshold for a fixed number of iterations.

2.4. Neural network: multilayer perceptron

Neural network models were inspired by the biological neuron model, based on the work of the psychiatrist Warren McCulloch and the mathematician Walter Pitts in the 1940s (Fine, 1999; Haykin, 1999). The multilayer perceptron (MLP) consists of an input layer, an output layer and also, in between, some hidden layers with nonlinearly-activating nodes (Fine, 1999). The nodes in each layer connect with a certain weight w_{ij} (synaptic weight) to all the nodes in the next layer. Perceptron learning occurs through changes in the connection weights, after items of data are processed, that depend on the error in the output compared to the expected result. This is an example of supervised learning, and it is performed through backpropagation, a generalization of the least mean squares algorithm in the linear perceptron.

A multilayer perceptron (MLP) is a feedforward artificial neural network able to perform regressions and is a modified linear perceptron. In studying the functional model of neural networks, we focus on feedforward networks, whose architecture can be represented in an acyclic manner so that each node is not backpropagated (as shown in Fig. 5), with specific activation functions and weights with fixed values.

The network create the function $\mathbf{f} : X \subset \mathrm{R}^n \to Y \subset \mathrm{R}^c$, that can be expressed (Fine, 1999; Haykin, 1999):

$$\begin{aligned} &\mathbf{f}(\mathbf{x}) = \boldsymbol{\phi}(\boldsymbol{\psi}(\mathbf{x})) = (\boldsymbol{\phi} \circ \boldsymbol{\psi})(\mathbf{x}) \\ &\boldsymbol{\phi}: X \subset \mathrm{R}^n \to U \subset \mathrm{R}^m \\ &\boldsymbol{\psi}: U \subset \mathrm{R}^m \to Y \subset \mathrm{R}^c \end{aligned} \tag{17}$$

where U the characteristics space, that is, the space of hidden variables. Given the architecture (Fine, 1999; Haykin, 1999):

- $\psi_j(x) = \psi(w_j^T x + w_{j0})$, ψ is the hidden layer unit activation function, where $\mathbf{w}_j \in \mathrm{R}^n$ is the vector of parameters for the units and $w_{j0} \in \mathrm{R}$ is its threshold value. The ψ function can be hyperbolic tangent, logistical or sigmoid.
- $\phi_j(u) = \phi(c_j^T u + c_{j0})$, ϕ is the activation function for the output layer units and $\mathbf{c}_j \in \mathrm{R}^m$ is the vector that contains the weights for the units, being $c_{j0} \in \mathrm{R}$ is its threshold value. The identity function, any dichotomous or Heaviside function are usually used as activation function ϕ

and thus, the MLP implemented function is (Fine, 1999; Haykin, 1999):

$$\mathrm{f}(\mathbf{x}) = \sum_{j=1}^{m} c_j \psi(\mathbf{w}_j^T \mathbf{x} + w_{j0}) + c_0 \tag{18}$$

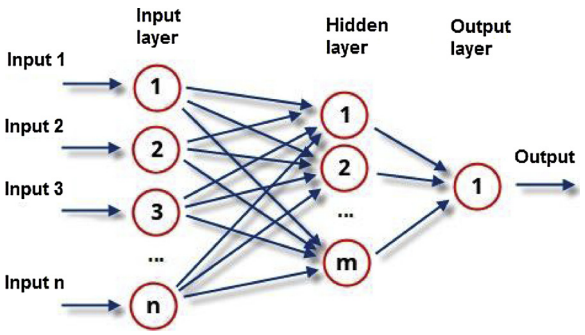

Fig. 5. Graph of an MLP network with n neurons in the input layer, a neuron in the output layer and m neurons in the hidden layer.

2.5. M5 model tree

M5 Model Tree was developed from binary decision tree. Linear regression functions are found at the terminal (leaf) nodes, with relationship between predictor and predicted variables (Quinlan, 1992). It differs from decision tree in that it can also be used for numerical data (Mitchell, 1997; Quinlan, 1992). M5 model tree generation has two steps (Rahimikhoob et al., 2013; Solomatine and Xue, 2004). First, data are split into subsets and then, a decision tree is created. The standard deviation reduction (SDR) is used for splitting the data and can be calculated as (Pal and Deswal, 2009; Heddam and Kisi, 2018):

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \tag{19}$$

where sd is the standard deviation T_i is a subset of samples with the ith outcome of the set and T is the set of samples in the node (Rahimikhoob et al., 2013). The splitting criterion depends on calculating the expected reduction in this error due to the testing of each attribute at that node. After the splitting, the data in child nodes have less error in compared to parent nodes. M5Tree keeps the one that optimizes the expected error reduction after trying all the possible splits. This division can generate a huge tree-like structure with poor generalization ability. To overcome this problem, the over-grown tree is pruned and the linear regression functions replaced. This method increases the variables space into areas (subspaces) and constructs a linear regression model for each one (Pal, 2006). Further details of an M5Tree can be obtained from Quinlan (Quinlan, 1992).

2.6. The goodness-of-fit of this approach

The coefficient of determination R^2 was the metric used for the goodness-of-fit. For each observed value t_i in the data there is an associated modelled value y_i, called predicted values. The variability of the data is measured with different formulas (Wasserman, 2003):

- $SS_{tot} = \sum_{i=1}^{n} (t_i - \bar{t})^2$: the total sum of squares directly related with the sample variance.
- $SS_{err} = \sum_{i=1}^{n} (t_i - y_i)^2$: the residual sum of squares.
- $SS_{reg} = \sum_{i=1}^{n} (y_i - \bar{t})^2$: the explained sum of squares.

Where, $\bar{t}$ is the mean of the n samples:

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i \tag{20}$$

The coefficient of determination is defined as:

$$R^2 \equiv 1 - \frac{SS_{err}}{SS_{tot}} \tag{21}$$

If the coefficient of determination equals 1.0, it would mean that the predicted values coincide with the observed ones.

Two additional criteria considered in this study were the root mean square error (RMSE) and mean absolute error (MAE) (Hastie et al., 2003; Wasserman, 2003). 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 (Freedman et al., 2007; Wasserman, 2003):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (t_i - y_i)^2}{n}} \tag{22}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |t_i - y_i|}{n} \tag{23}$$

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.

Table 1
Mean and standard deviation of the biological variables.

Biological input variables	Name of the variable	Mean	Std
Cyanobacteria (mm^3/L)	*Cyanobacteria*	0.6900	0.4426
Diatoms (mm^3/L)	Diatoms	1.3505	1.1873
Euglenophytes (mm^3/L)	Euglenophytes	1.1774	0.5727
Dinophlagellata (mm^3/L)	*Dinophlagellata*	0.1842	0.1790
Chrysophytes (mm^3/L)	Chrysophytes	0.2580	0.1766
Chlorophytes (mm^3/L)	Chlorophytes	0.1196	0.0910
Chryptophytes (mm^3/L)	Chryptophytes	0.9889	0.3575

Table 2
Mean and standard deviation of the biological variables of the physical-chemical variables.

Physical-chemical input variables	Name of the variable	Mean	Std
Water temperature (ºC)	Water_temp	17.0381	4.1057
Turbidity (NTU)	Turbidity	5.5844	4.7626
Nitrate concentration (mg NO^{3-}/L)	Nitrate	0.8310	0.4070
Ammonium concentration (mg/L)	Ammonium	0.1640	0.0876
Dissolved oxygen concentration (mg O_2/L)	DOC	9.0201	1.7814
Conductivity (μS/cm)	Conductivity	277.1844	31.1170
pH values	pH_values	7.7754	0.4012

3. Analysis of results and discussion

The physical-chemical can be seen in Table 1 and the biological variables are shown in Table 2 (Allman and Rhodes, 2003). The number of input variables in the ABC–SVM, M5 model tree and MLP model was 15 (see Tables 1 and 2). The predicted variables, *Chl*-a and Total phosphorus, have been measured in μg/L and mg P/L (Allman and Rhodes, 2003; Barnes and Chu, 2010), respectively.

In this study, three different models have been constructed (specifically in this research, the novel hybrid ABC–SVM, M5 model tree and MLP) with predicted variables the *Chl*-a and Total phosphorus and predictor variables the other sixteen biological and physical-chemical parameters (input variables), using the coefficient of determination R^2 as criterion to assess the success of each model.

Also, the SVM methods depend on the SVM hyperparameters: ε that characterizes the ε –insensitive tube; the regularization factor C (see Eqs. (5) and (7)); a, b and σ that are the parameters of kernel. To improve the defect hyperparameters some of the methods commonly used are (Cristianini and Shawe-Taylor, 2000; Hansen and Wang, 2005; Wu, 2009): random search, grid search, genetic algorithms, particle swarm optimization (PSO) and so on. Usually, the traditional way of performing hyperparameter optimization 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. In this research work, the Artificial Bee Colony (ABC) algorithm was applied (Olsson, 2011).

Thus, we have chosen ABC optimization technique as an efficient and simple technique (Álvarez Antón et al., 2013; Fine, 1999; Haykin, 1999; Kecman, 2005) for adjusting the SVR parameters. Fig. 6 shows the flowchart of this ABC-SVM model.

The food sources $\mathbf{x}_i$ are vectors that consist in sets of hyperparameters to adjust. For instance, $\mathbf{x}_i = (C_i, \varepsilon_i, \sigma_i)$ is the set of parameters associated with SVM with the RBF kerne. In this study, we have 20 food sources. We initialize their values randomly in the first iteration. Following the equations that regulate the ABC algorithm, described

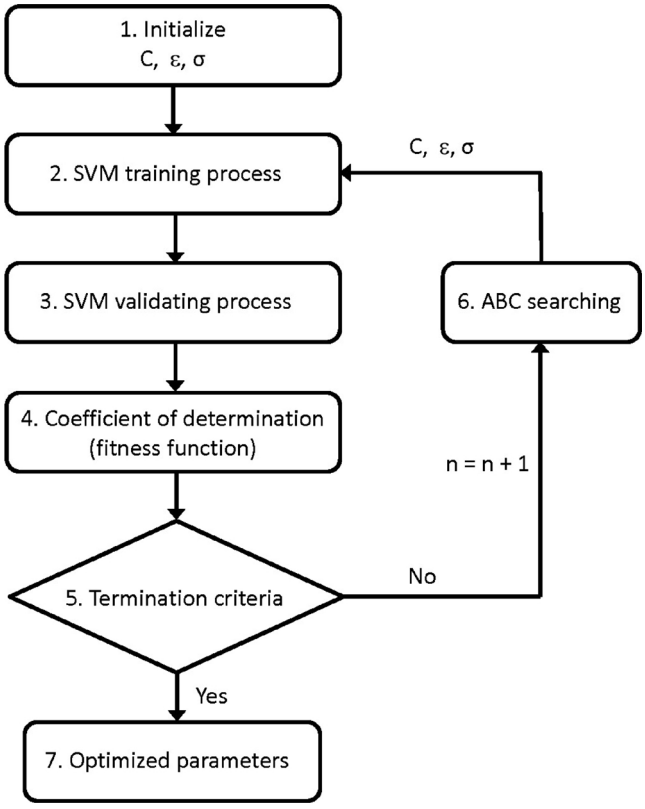

Fig. 6. Flowchart of the ABC–SVM model.

previously, the food sources are updated. In each step, the cost function value for each set of hyperparameters is computed. The cost function value is the minus ten-cross-validation (Efron and Tibshirani, 1997; Picard and Cook, 1984). With this aim, the set of samples is split in 10 sets. The SVM is trained with a set consisting of 9 of this subsets and a model is built. This model is tested with the remaining set and the corresponding coefficient of determination is calculated. This process is done for each of the 10 sets and, when the full cycle is completed, each food source is associated with its corresponding mean coefficient of determination. If the termination criteria are satisfied, the global best $\mathbf{x}_i$, that is, the food source with the best fitness, is the set of optimized parameters (Álvarez Antón et al., 2013; Hastie et al., 2003).

LIBSVM library (Chang and Lin, 2011) has been used to construct the regression model and the parameters have been optimized with the ABC technique using the ABC version for MATLAB (Karaboga and Basturk, 2007).

ABC is used to optimize the SVM hyperparameters. The ABC looks for the best hyperparameters (i.e. for RBF kernel C, σ and ε) computing the fitness in every iteration where the fitness factor is constructed using the coefficient of determination (R^2).As significant changes in the regression only take place when there is a change in the magnitude order of the parameters, the parameters have been considered powers of ten and the searching space has been that of the exponents. That is, for instance, in the RBF kernel case, the searching space has been $[-6, 4] \times [-10, 4] \times [-20, 4]$ and then, C values move in the interval $[10^{-6}, 10^{4}]$, σ values vary in $[10^{-10}, 10^{4}]$ and, finally, ε values are drawn from $[10^{-20}, 10^{4}]$. The number of food sources has been 20 and the stopping criterion is met if there is no improvement, within a threshold of 10^{-6}, of in the R^2 after ten iterations, with a limit for the number of iterations equal to 500. Table 3 shows the optimal hyperparameters obtained with the ABC–SVM models for the Total phosphorus. An iMac with a CPU Intel Core i5-4570 at 3.2 GHz with 8 GB of RAM was used, taking 149 s to obtain the Chlorophyll model and 806 s for the Total phosphorus model.

For comparison purposes, M5 tree and a multilayer perceptron (MLP) models have been built for the Total phosphorus as output variable in order to predict the algal anomalous growth in the Englishmen Lake. An artificial neural network (ANN) (see Fig. 5), is constructed using three different parameters (Fine, 1999; Haykin, 1999): the learning rate that regulates the update of the interconnection weights, the number of hidden layers, the momentum factor that controls the oscillation of weight changes of the ANN and the activation function that converts a neuron's weighted input to its output. The obtained ANN optimal parameters for the multilayer perceptron (MLP) appears in Table 4. Table 5 shows the coefficients of determination and correlation for the ABC–RBF–SVM model, multilayer perceptron and M5 tree model fitted for the Total phosphorus in this study.

The SVM with the RBF kernel function gets the highest goodness-of-fit for estimating the Total phosphorus to assess the state of eutrophication, with a coefficient of determination R^2 equal to 0.90 and a correlation coefficient r of 0.95. These results mean a good agreement is obtained between predicted and observed data.

Following the same methodology for the second dependent variable (Chlorophyll concentration), we have obtained the results shown below. Indeed, Table 6 shows the optimal hyperparameters obtained with the ABC–SVM models for the Chlorophyll concentration. Additionally, another multilayer perceptron (MLP) and another M5 tree model were built for the Chlorophyll output variable in order to predict the algal atypical proliferation in the Englishmen. The ANN optimal parameters for this same variable using multilayer perceptron (MLP) are shown in Table 7. Table 8 shows the determination and correlation coefficients for ABC–RBF–SVM, MLP and M5 tree models for the *Chl*-a variable.

Similarly, taking into account our previous statistical calculations, again the ABC– RBF–SVM is the model with higher goodness-of-fit for estimating the Chlorophyll concentration to predict the algal anomalous growth, with a coefficient of determination R^2 equal to 0.92 and a correlation coefficient r equal to 0.96. An also there is a good agreement between the model and the observed data.

The nine input variables for the Total phosphorus (first output variable) have been ranked by their importance in this complex nonlinear problem and this ranking is shown in Table 9 and Fig. 7.

In the same way, the fifteen predictor variables in the Chlorophyll concentration model importance ranking is shown in Table 10 and Fig. 8.

Finally, this work estimates the Total phosphorus using the ABC–RBF–SVM model quite accurately. Fig. 9 shows the comparison between observed and predicted Total with M5 tree, MLP, and ABC–RBF–SVM models. Thus, a SVM model with a ABC optimization method is a good choice to achieve a good approximation to the nonlinearities in this regression problem.

Similarly, Fig. 10 shows the comparison between observed and predicted Chlorophyll concentration values with M5 tree, MLP and ABC–RBF–SVM models. Again, ABC–SVM model is a good choice to achieve a very effective approach to the algal abnormal growth in water bodies such as the Englishmen Lake.

4. Conclusions

From the experimental and numerical results, the following conclusions can be drawn:

- Firstly, algal abnormal growth (or eutrophication) is a very prevailing and severe problem in water bodies such as reservoirs and lakes. The assessment methods that use sampling and the following analysis are expensive. Thus, an alternative technique such as the ABC–RBF–SVM method is a good option to study the algal atypical productivity in the Englishmen Lake.
- Secondly, an ABC–RBF–SVM model successfully predicts the algal anomalous concentration from the other and easier to measure input

Table 3
Optimal hyperparameters obtained with the ABC–SVM models for the Total phosphorus.

Kernel	Values of optimal hyperparameters
Linear	Regularization factor$C = 1.5389 \times 10^{0}$, $\varepsilon = 1.1472 \times 10^{-1}$
Quadratic	Regularization factor$C = 3.3364 \times 10^{-4}$, $\varepsilon = 5.0484 \times 10^{-2}$, $\sigma = 5.4801 \times 10^{1}$, $a = 1.0000 \times 10^{4}$, $b = 2$
Cubic	Regularization factor$C = 1.7402 \times 10^{-1}$, $\varepsilon = 4.6663 \times 10^{-2}$, $\sigma = 6.7252 \times 10^{-1}$, $a = 2.7942 \times 10^{0}$, $b = 3$
Sigmoid	Regularization factor$C = 2.1419 \times 10^{0}$, $\varepsilon = 1.6346 \times 10^{-1}$, $\sigma = 1.0000 \times 10^{-1}$, $a = 9.9563 \times 10^{-4}$
RBF	Regularization factor$C = 1.1094 \times 10^{0}$, $\varepsilon = 1.2709 \times 10^{-2}$, $\sigma = 1.4886 \times 10^{0}$

Table 4
The ANN (MLP) model parameters for the Total phosphorus.

Parameters	Values
Number of hidden neurons	10
Learning rate	0.1
Momentum factor	1.0×10^{-10}
Activation function	Tangent sigmoid transfer function

Table 5
Cross–validation coefficients of determination (R^2) and correlation coefficient (r), and RMSE and MAE for the ABC–SVM model, multilayer perceptron (MLP) and M5 tree models for the Total phosphorus.

Model	Coeff.of det.(R^2)	Corr. Coeff. (r)	RMSE	MAE
Linear	0.69	0.83	0.0141	0.0111
Quadratic	0.86	0.93	0.0073	0.0055
Cubic	0.87	0.93	0.0069	0.0053
Sigmoid	0.66	0.81	0.0153	0.0128
RBF	0.90	0.95	0.0030	0.0020
Multilayer perceptron	0.84	0.92	0.0072	0.0055
M5 Tree	0.84	0.92	0.0085	0.0063

Table 6
Optimal hyperparameters for ABC–SVM model for the Chlorophyll concentration.

Kernel	Values of optimal hyperparameters
Linear	Regularization factor$C = 3.7177 \times 10^{0}$, $\varepsilon = 1.8626 \times 10^{-1}$
Quadratic	Regularization factor$C = 7.7329 \times 10^{0}$, $\varepsilon = 4.3352 \times 10^{-2}$, $\sigma = 4.2702 \times 10^{-1}$, $a = 5.6292 \times 10^{0}$, $b = 2$
Cubic	Regularization factor$C = 3.9046 \times 10^{-3}$, $\varepsilon = 3.4846 \times 10^{-2}$, $\sigma = 3.8883 \times 10^{0}$, $a = 6.8313 \times 10^{0}$, $b = 3$
Sigmoid	Regularization factor$C = 1.3351 \times 10^{0}$, $\varepsilon = 3.6467 \times 10^{-5}$, $\sigma = 1.0000 \times 10^{-1}$, $a = 5.0781 \times 10^{-2}$
RBF	Regularization factor$C = 6.2252 \times 10^{1}$, $\varepsilon = 1.0202 \times 10^{-2}$, $\sigma = 2.0517 \times 10^{0}$

Table 7
The ANN (MLP) parameters for the Chlorophyll concentration.

Parameters	Values
Number of hidden neurons	10
Learning rate	0.1
Momentum factor	0.01
Activation function	Tangent sigmoid transfer function

variables lowering the costs of its assessment.

- Thirdly, a coefficient of determination of 0.90 was obtained for the ABC–RBF–SVM model that estimates the Total phosphorus. Indeed, the predicted values match consistently the values in the dataset of observed Total phosphorus (see Fig. 7). Also, a coefficient of determination of 0.92 was obtained for the ABC–RBF–SVM model that predicts the Chlorophyll concentration. Thus, the estimated results for the algal growth corresponding to this output variable are in

Table 8
Cross–validation coefficients of determination (R^2) and correlation coefficient (r), and RMSE and MAE for the ABC–SVM model, multilayer perceptron (MLP) and M5 tree models for the Chlorophyll concentration.

Model	Coeff.of det.(R^2)	Corr. Coeff. (r)	RMSE	MAE
Linear	0.71	0.85	1.8939	1.4488
Quadratic	0.85	0.92	1.1265	0.6758
Cubic	0.86	0.93	0.9949	0.5750
Sigmoid	0.70	0.83	1.9363	1.3516
RBF	0.92	0.96	0.1174	0.1145
Multilayer perceptron	0.83	0.91	0.7755	0.6189
M5 Tree	0.83	0.91	1.1987	0.8281

Table 9
Weight of the variables in ABC–SVM model for the Phosphorus variable.

Input variable	Weight
Water temperature	1.0066
Turbidity	0.9861
Chlorophyll	0.9665
Dissolved oxygen concentration	–0.7811
Cyanobacteria	0.5615
Dinophlagellata	0.5029
Euglenophytes	0.4535
Nitrate concentration	0.4393
Ammonium concentration	0.3965
Chrysophytes	–0.3734
Chlorophytes	–0.3625
pH	–0.2282
Conductivity	–0.1635
Chryptophytes	0.1211
Diatoms	0.0323

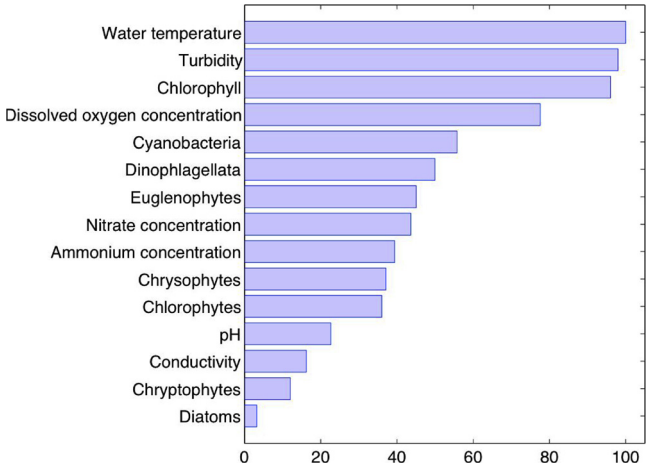

Fig. 7. Relative importance of the predictor variables in the Total phosphorus ABC–SVM model.

Table 10
Weight of the variables in ABC–SVM model for the Chlorophyll variable.

Input variable	Weight
Euglenophytes	2.2856
Dinophlagellata	1.1356
Total phosphorus	1.0227
Turbidity	0.6035
pH	0.5562
Dissolved oxygen concentration	–0.5444
Conductivity	0.4923
Chlorophytes	–0.3899
Cyanobacteria	0.3272
Nitrate concentration	–0.3139
Chrysophytes	0.2747
Diatoms	–0.1038
Ammonium concentration	–0.0071
Water temperature	–0.0029
Chryptophytes	0.0011

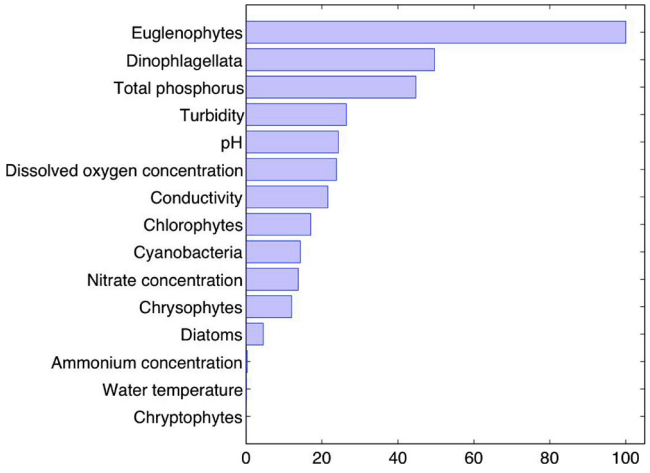

Fig. 8. Ranking for the relative importance of the predictor variables for the Chlorophyll concentration in the ABC–SVM model.

accordance with the dataset.

- Fourthly, predictor variables of the algal atypical growth have been ranked in order of relative importance. This is one of the main findings of this work, being Water temperature variable the most important in the estimation of the Total phosphorus. On the other hand, the Euglenophytes plays a significant role in the Chlorophyll concentration. Also, important variables in the prediction of Total phosphorus and Chlorophyll concentration are the Turbidity and Dinophlagellata concentration, respectively.
- Finally, the hyperparameters affects significantly the final model and its setting is fundamental.

In summary, other eutrophication processes with the same or different pollution sources can benefit from the use of these techniques, but it must always be taken into account the particular circumstances of each location. Regarding the applicability and transferability of the established model, since the Cantabrian Basin Authority manages a specific geographical area with relatively homogeneous environmental conditions (Atlantic climate), for similar lakes, in this geographical area, the predictions are expected to be valid, within the scope of this particular geographical area. Thus, an ABC–SVM model is a good approach to the prediction of algal abnormal productivity in water bodies.

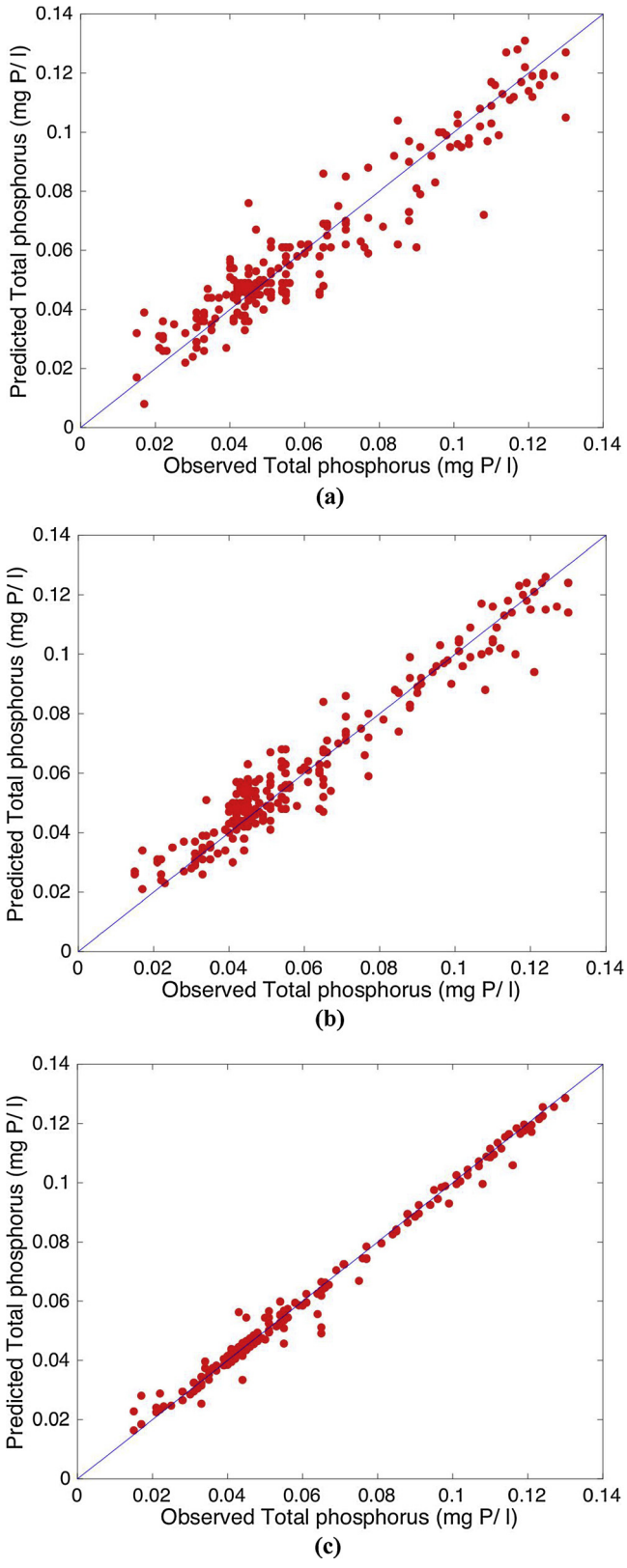

Fig. 9. Predicted vs. observed Total phosphorus values with three different models: (a) M5 tree (R^2 = 0.84); MLP (R^2 = 0.84); and RBF-SVM (R^2 = 0.90).

Acknowledgements

Authors acknowledge the support by the Cantabrian Basin Authority (Spanish Ministry of Agriculture, Fishing, Food and Environment) that provided the pollutant data in the Englishmen Lake located in Autonomous Community of Cantabria (Northern Spain). Additionally, we thank Anthony Ashworth for his revision of the English in the

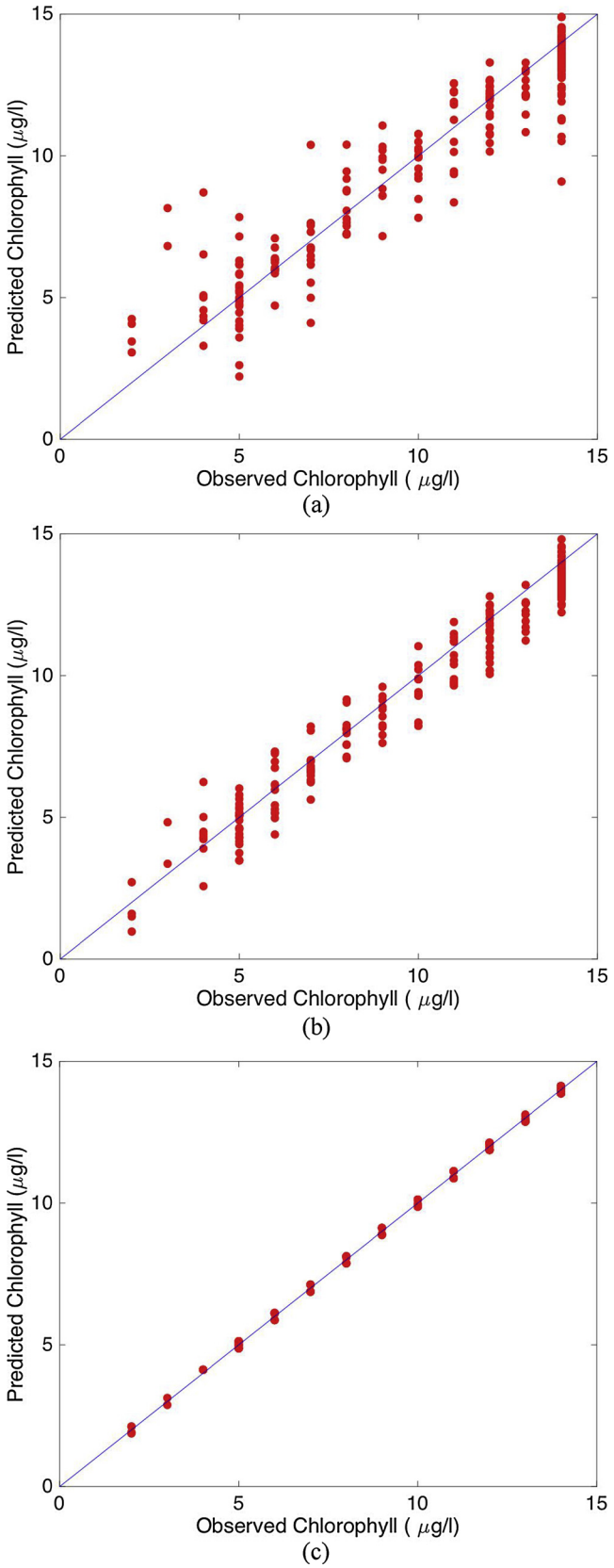

Fig. 10. Predicted vs. observed Chlorophyll concentration values for three different models: (a) M5 tree ($R^2 = 0.83$); MLP ($R^2 = 0.83$); and RBF-SVM ($R^2 = 0.92$).

manuscript.

References

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

Álvarez Antón, J.C., García Nieto, P.J., Blanco Viejo, C., Vilán Vilán, J.A., 2013. Support vector machines used to estimate the battery state of charge. IEEE Trans. Power Electr. 28 (12), 5919–5926.

American Public Health Association, 2005. American Water works association, Water environment Federation. Standard Methods for the Examination of Water and Wastewater, no. 21. APHA/AWWA/WEF, Washington.

Ansari, A.A., Gill, S.S., 2016. Eutrophication: Causes, Consequences and Control. Springer, New York, USA.

Barnes, D.J., Chu, D., 2010. Introduction to Modeling for Biosciences. Springer, New York, USA.

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

Bricker, S.B., Ferreira, J.G., Simas, T., 2003. An integrated methodology for assessment of estuarine trophic status. Ecol. Model. 169, 39–60.

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

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

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

Chen, Y.W., Fan, C.X., Teubner, K., Dokulil, M., 2003. Changes of nutrients and phytoplankton chlorophyll-a in a large shallow lake, Taihu, China: an 8-year investigation. Hydrobiologia 506, 273–279.

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

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

Cortes, C., Vapnik, V., 1995. Support vector networks. Mach. Learn. 20, 273–297.

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

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

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.

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

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

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the .632 + bootstrap method. J. Am. Stat. Assoc. 92 (438), 548–560.

Ferreira, J.G., Bricker, S.B., Simas, T.C., 2007. Application and sensitivity testing of an eutrophication assessment method on coastal systems in the United States and European Union. J. Environ. Manage. 82 (4), 433–445.

Fine, T.L., 1999. Feedforward Neural Networks Methodology. Springer-Verlag, New York.

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

García Nieto, P.J., Combarro, E.F., del Coz Díaz, J.J., Montañés, E., 2013. 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), 8923–8937.

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

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

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

Hastie, T., Tibshirani, R., Friedman, J.H., 2003. The Elements of Statistical Learning. Springer-Verlag, New York.

Haykin, S., 1999. Neural Networks. A Comprehensive Foundation, Prentice Hall, New York.

Heddam, S., Kisi, O., 2018. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J. Hydrol. 559, 499–509.

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

Karaboga, D., Akay, B., 2009. A survey: algorithms simulating bee swarm intelligence. Artif. Intell. Rev. 31 (1), 68–85.

Karaboga, D., Basturk, B., 2007. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. J. Global Optim. 39 (3) 459–171.

Karaboga, D., Gorkemli, B., 2014. A quick artificial bee colony (qABC) algorithm and its performance on optimization problems. Appl. Soft Comput. 23, 227–238.

Karydis, M., 2009. Eutrophication assessment of coastal waters based on indicators: a literature review. Glob. NEST J. 11, 373–390.

Kecman, V., 2005. Support vector machines: an introduction. In: Wang, L. (Ed.), Support Vector Machines: Theory and Applications. Springer-Verlag, Heidelberg, pp. 1–48.

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

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill Company Inc, New York.
Negro, A.I., de Hoyos, C., Vega, J.C., 2000. Phytoplankton structure and dynamics in Lake Sanabria and Valparaíso reservoir (NW Spain). Hydrobiologia 424, 25–37.
Nikoo, M.R., Mahjouri, N., 2013. Water quality zoning using probabilistic support vector machines and self-organizing maps. Water Resour. Manage. 27 (7), 2577–2594.
Olsson, A.E., 2011. Particle Swarm Optimization: Theory, Techniques and Applications. Nova Science Publishers, New York.
Ortiz-García, E.G., Salcedo-Sanz, S., Pérez-Bellido, A.M., Portilla-Figueras, J.A., Prieto, L., 2010. Prediction of hourly O_3 concentrations using support vector regression algorithms. Atmos. Environ. 44 (35), 4481–4488.
Paerl, H.W., Xu, H., McCarthy, M.J., Zhu, G., Qin, B., Li, Y., Gardner, W.S., 2011. Controlling harmful cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): the need for a dual nutrient (N & P) management strategy. Water Res. 45, 1973–1983.
Pal, M., 2006. M5 model tree for land cover classification. Int. J. Remote Sens. 27 (4), 825–831.
Pal, M., Deswal, S., 2009. M5 model tree based modelling of reference evapotranspiration. Hydrol. Process. 23 (10), 1437–1443.
Pal, M., Goel, A., 2007. Estimation of discharge and end depth in trapezoidal channel by support vector machines. Water Resour. Manage. 21 (10), 1763–1780.
Picard, R., Cook, D., 1984. Cross-validation of regression models. J. Am. Stat. Assoc. 79 (387), 575–583.
Quinlan, J.R., 1992. Learning with continuous classes. Proceedings of Australian Joint Conference on Artificial Intelligence 343–348.
Rahimikhoob, A., Asadi, M., Mashal, M., 2013. A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. Water Resour. Manage. 27 (14), 4815–4826.
Reynolds, C.S., 2006. Ecology of Phytoplankton. Cambridge University Press, New York.
Schölkopf, B., Smola, A.J., Williamson, R., Bartlett, P., 2000. New support vector algorithms. Neural Comput. 12 (5), 1207–1245.
Scholten, M.C.T., Foekema, E.M., Dokkum, H.P., Kaag, N.H.B.M., Jak, R.G., 2006. Eutrophication Management and Ecotoxicology. Springer, New York, USA.
Shrestla, N.K., Shukla, S., 2015. Support vector machine based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment. Agric. For. Meteorol. 200, 172–184.
Simon, D., 2013. Evolutionary Optimization Algorithms. Wiley, New York.
Solomatine, D.P., Xue, Y.P., 2004. M5 model trees and neural networks: application to flood forecasting in the upper reach of the Hual River in China. J. Hydrol. Eng. 9 (6), 491–501.
Steinwart, I., Christmann, A., 2008. Support Vector Machines. Springer, New York.
Takaara, T., Sano, D., Masago, Y., Omura, T., 2010. Surface–retained organica matter of *Microcystis aeruginosa* inhibiting coagulation with polyaluminum chloride in drinking water treatment. Water Res. 44, 3781–3786.
Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York.
Wang, S., Jin, X., Bu, Q., Jiao, L., Wu, F., 2008. Effects of dissolved oxygen supply level on phosphorus release from lake sediments. Colloid Surf. A 316, 245–252.
Wasserman, L., 2003. All of Statistics: A Concise Course in Statistical Inference. Springer, New York.
Willame, R., Jurckzak, T., Iffly, J.F., Kull, T., Meriluoto, J., Hoffman, L., 2005. Distribution of hepatotoxic cyanobacterial blooms in Belgium and Luxembourg. Hydrobiologia 551, 99–117.
World Health Organization, 1998. Guidelines for Drinking-Water Quality: Health Criteria and Other Supporting Information, vol. 2 World Health Organization, Geneva.
Wu, Q., 2009. The forecasting model based on wavelet *v*- support vector machine. Expert Syst. Appl. 36 (4), 7604–7610.
Xiao, Y., Ferreira, J.G., Bricker, S.B., Nunes, J.P., Zhu, M., Zhang, X., 2007. Trophic assessment in Chinese coastal systems - Review of methodologies and application to the Changjiang (Yangtze) estuary and Jiaozhou Bay. Estuar. Coast. 30 (6), 1–18.
Yang, X.–S., Cui, Z., Xiao, R., Gandomi, A.H., Karamanoglu, M., 2013. Swarm Intelligence and Bio-inspired Computation: Theory and Applications. Elsevier, London.
Zeng, J., Qiao, W., 2013. Short–term solar power prediction using a support vector machine. Renew. Energy 52, 118–127.