Modeling algal atypical proliferation in La Barca reservoir using L-SHADE optimized gradient boosted regression trees: A case study

Paulino José García Nieto[a,*], Esperanza García–Gonzalo[a], José Ramón Alonso Fernández[b], Cristina Díaz Muñiz[b]

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

[b]Confederación Hidrográfica del Cantábrico, Ministerio Español para la Transición Ecológica, 33071 Oviedo, Spain

Abstract Algal atypical proliferation is a consequence of water fertilization (also called eutrophication) and a worldwide environmental concern since water quality and its uses are seriously compromised. Prevention is the most effective measure given that once the algal proliferation starts, it is too difficult and costly to stop the process. This article presents a nonparametric machine learning algorithm that combines the gradient boosted regression tree (GBRT) model and an improved differential evolution algorithm (L-SHADE) for better understanding and control of the algal abnormal proliferation (usually estimated from Chlorophyll-a and Total Phosphorus concentrations) from physico-chemical and biological variable values obtained in a northern Spain reservoir. This L-SHADE technique involves the optimization of the GBRT hyperparameters during the training process. Apart from successfully estimating algal atypical growth (coefficients of determination equal to 0.91 and 0.93 for Chlorophyll-a and Total Phosphorus concentrations were obtained, respectively), this hybrid model allows here to establish the ranking of each independent biological and physico-chemical variable according to its importance in the algal enhanced growth.

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

Keywords Gradient boosted regression tree (GBRT); Differential evolution (DE); Algal abnormal productivity in reservoirs; Regression analysis

1 Introduction

Algal atypical growth, a primary symptom of *eutrophication*, continues to be a worldwide environmental problem due to the harmful consequences for the water quality and its uses [1–5]. Although, this is a natural process, it can be greatly accelerated by anthropogenically nutrient loadings, especially nitrogen and phosphorus, the most critical factors restricting algal growth [2,4] – phosphorus if the waterbody is a lake or reservoir [3]–.Both natural and artificial water fertilization, that is, excess of nitrogen and phosphorus, usually causes opacity, dissolved oxygen (DO hereinafter) depletion, toxic blooms, eutrophication and loss of biodiversity [6–9].

Since the algal blooms (more precisely, algal and cyanobacterial, i.e. phytoplankton blooms) contain Chlorophyll-a (Chl-a) as the main photosynthetic pigment, is normally used for tracking algae growth [10] and consequently, a good indicator of water eutrophication. Therefore, the eutrophication indicators, Total Phosphorus (Total P), Total Nitrogen (Total N) and Chl-a (directly related to algal, or phytoplankton, quantity), are routinely monitored to characterize the trophic status on water bodies [3,11]. Many of them constitute the basis of the classic approach for classifying the trophic status [12], that is considered in the Water Framework Directive (WFD) implementation [13,14]. However, biovolume calculations are needed to achieve a more reliable evaluation of this complex problem [15]. Other environmental factors, such as water temperature, pH, dissolved oxygen (DO), Secchi depth, ammonium, nitrogen,

etc., can also have a significant influence on the enhanced algal growth [16]. Therefore, quantitative relationships among a variety of environmental factors and eutrophic indicators are highly desirable to develop strategies to prevent algae blooms.

A dataset consisting of ecological indicators and water condition parameters are required to classify eutrophic condition. These data are usually obtained from the data gathered according to a monitoring program [17].

As stated previously, algal atypical growth is a significant environmental issue in water bodies. In reservoirs such as La Barca (see Figs. 1(a) and 1(b)), located in the center-west of Asturias and built for hydroelectric applications but also used for recreational activities, their characteristics make them sensitive to ecosystem threats and thus, the problem can be very serious, Among them, eutrophication is the riskiest one due to the DO consumption and its far-reaching consequences [18], where the harmful or toxic algal blooms occur together with the present scenarios of climate change [7,19,20].

Fig. 1 (a) La Barca reservoir, aerial photograph; and (b) another aerial photograph at a larger scale

Therefore, it is essential to predict the eutrophication and, in turn, to provide advance information for water quality management and facilitate public health risk assessment. Several studies have recently been conducted on water quality prediction models [21,22]. Physical-based water quality modeling approaches are capable of simulating the internal physical processes of the aquatic system, but require extensive information

that is not easily accessible [23]. Moreover, many physical-based water quality models are time-consuming [24].

In this work, we have built a gradient boosted regression tree model (GBRT) [25–29] combined with variants of the evolutionary optimization method known as Differential Evolution (DE) [30–33] to predict phytoplankton atypical growth in La Barca reservoir. This novel hybrid model could be an interesting approach since, at the knowledge of the authors, has not been yet addressed in previous investigations to foretell eutrophication in water bodies (reservoirs and lakes). Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an *ensemble* of weak prediction models, typically decision trees (GBRT models).

GBRTs have been previously used with efficacy to predict biological and physico-chemical output variables in diverse environmental problems such as wind variables forecasting [34]; solar power generation prediction [35], waste estimation in the short term [36] and so on.

Additionally, the differential evolution (DE) technique was used to optimize GBRT hyperparameters in the training stage. Differential evolution (DE) is a metaheuristic evolutionary global method, derived from genetic algorithms (GA), intrinsically capable of solving multidimensional optimization problems involving continuous variables [30–33]. Previous researches show that GBRT is an effective tool in many fields such as bioinformatics, biomedicine, atmospheric environment and civil engineering [37–40].

In conclusion, a hybrid DE optimized GBRT (DE/GBRT) model [41–43] was applied to estimate the eutrophication from the biological and physico-chemical input variables in La Barca reservoir.

2 Materials and methods

2.1 Experimental dataset

The data used were obtained from 243 samples containing quantitative information about the phytoplankton abundance gathered over 16 years (from 2001 to 2016). Specifically, the samples were taken at least once a month beginning on January 16th, 2001 and ending on December 20th, 2016, following the sampling protocols of the *Spanish Ministry of Agriculture, Food and Environment* for lakes and reservoirs. These protocols were devised in accordance with those established by the European Union [44–46]. The reservoir deepest point was determined using a depth gauge [47] and the samples were taken at this spot with a Niskin hydrographic bottle. The Niskin bottle is an evolution of the Nansen bottle where both ends are open, being it more similar to a tube than to a bottle. A cap is provided for each end and it can be closed, once it is filled with the reservoir water, with an elastic rope (see Fig. 2(a)). Secchi depth was obtained by lowering a patterned Secchi disk (see Fig. 2(b)) into the water, until the pattern is no longer visible due to the water turbidity. The Secchi depth limits the euphotic zone, that is, the zone where photosynthesis takes place because it is reached by enough sunlight. Five homogenous subsamples are gathered at different equidistant depths in the euphotic zone [48,49] and the Chl-a and phytoplankton concentrations are obtained composing the values from these samples. The experimental data gathered in this study were the usual physico–chemical variables used in limnological works [3,48,50].

Fig. 2 (a) A Niskin bottle; and (b) examples of Secchi disks

The aim of this work was to establish a way to estimate the algal (phytoplankton) abnormal productivity indicators from values of easy to measure variables. The output variables chosen were Chl-a and phosphorous concentrations –measured in µg/L and mg P/L, respectively– because of their relationship with algal abnormal productivity.

The Chl-a biomolecule is very important because it is directly related to photosynthesis and, consequently, directly linked to phytoplankton abundance [51].

Total P is the sum of every phosphorus forms (condensed phosphates, organically bound phosphates and orthophosphates) in both suspended and dissolved water fractions. The quantity of phosphorus in reservoirs depends on both domestic and industrial discharges, as well as, on agriculture fertilizers run-off and it is a limiting factor in biological productivity in reservoirs.

There are two kinds of predictor variables:

- Biological (all of them referred to phytoplankton composition):

- Cyanobacteria concentration (mm^3/L): they are blue-green algae usually categorized as bacteria nowadays (see Fig. 3(a)).

- Diatoms concentration (mm^3/L): is one of the more prevalent kinds of phytoplankton and is an algae group (see Fig. 3(b)).

- Euglenophytes concentration (mm^3/L): All of them can photosynthesize and thus they are autotrophic organisms, that is, like plants, they produce their own food (see Fig. 3 (c)).

- Dinoflagellates concentration (mm^3/L): they belong to the phytoplankton, are unicellular and can be classified as protist (see Fig. 3(d)).

- Chrysophytes concentration (mm^3/L): they are mostly photosynthetic, heterotrophic types can be found (see Fig. 3(e)).

- Chlorophytes concentration (mm^3/L): they are part of the green algae with both pluricellular and unicellular species (see Fig. 3(f)).

- Cryptophytes concentration (mm^3/L): they are single microscopic cells with a flagellum (see Fig. 3(g)).

Fig. 3 La Barca reservoir biological variables: (a) Cyanobacteria; (b) Diatoms; (c) Euglenophytes; (d) *Dinophlagella*; (e) Chrysophytes; (f) Clorophytes; and (g) Chryptophytes concentrations

- Physico –chemical:

- Water temperature (ºC): is the water body temperature.

- Turbidity (NTU): is a measure the water opaqueness due to suspended solids. The origin of the suspended particles is mainly anthropogenic [52,53] and it is closely related to algal abnormal growth.

- Secchi depth (m). It is a measure of light penetration applied for the estimation of algae concentrations in water bodies. Therefore, it is related to DO concentration. It is also used, together with other parameters as phosphorus, nitrogen, chlorophyll, etc., to assess the trophic condition of a water body.

- Nitrate concentration (mg NO_3^-/L): the nitrogen form in water used by plants to grow. In large quantities can cause, together with high phosphate concentrations, eutrophication processes, strongly linked to cyanobacteria blooms and their toxic metabolites, the cyanotoxins

- Ammonium concentration (mg/L): comes from animal excretion and in oxidizing environments, it can be transformed to nitrate.

- DO concentration (mg O_2/L): oxygen mass dissolved in water. It is essential for the existence of most ecosystems in water. Algal blooms produce much oxygen during the day, but they also can cause its lowering at night or when they die and decompose.

- Conductivity (μS/cm): directly connected with salt concentration in water, that affects phytoplankton concentration and composition [3], it measures the capacity to conduct electricity.

- pH: it is the concentration of hydrogen ions (H^+) given by its logarithm in base ten. It expresses the acidity of a water solution which affects biological and chemical processes and it is an indicator of eutrophication process.

This work constructs a hybrid eutrophication machine learning model for La Barca reservoir from experimental dataset: gradient boosted regression tree with DE parameter optimization (DE/GBRT).

2.2 Computational procedure

2.2.1 Gradient boosting regression tree (GBRT)

Gradient boosting is a machine learning method used for classification and regression that constructs a model from a set of weak models or learners, that are, usually, decision trees. It builds the model by stages, as is typical for boosting methods, and obtains a single strong ensemble model optimizing a differentiable loss function [25–29,54,55].

Let us explain it as a least-squares regression method, where the aim is to teach a model F to predict the values $\hat{y} = F(x)$, minimizing the mean squared error $(\hat{y} - y)^2$, being y the true values from the training set.

At each stage $1 \le m \le M$ of gradient boosting, a weak model F_m is constructed to estimate the training set average value of y. This model is improved building a model $F_{m+1}(x) = F_m(x) + h(x)$. To find h, the gradient boosting method takes into account that, to have a perfect h [25–29,54,55]:

$$F_{m+1}(x) = F_m(x) + h(x) = y \tag{1}$$

that is,

$$h(x) = y - F_m(x) \tag{2}$$

Thus, gradient boosting will perform the fitting of h to the residual $y - F_m(x)$. In each stage, F_{m+1} is constructed as a correction of its predecessor F_m. We can generalize this explanation to other loss functions different from Squared Error, taking into account that residuals $y - F(x)$ are the negative gradients of the loss function $\frac{1}{2}\left(y - F(x)\right)^2$.

Let us call y the dependent variable and x the set of independent variables. The objective is to find an estimate $\hat{F}(x)$ of the function $F^*(x)$ that minimizes the value of some loss function $L\left(y, F(x)\right)$ using a training set $\left\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\right\}$ of already known values of x and their corresponding values of y [25–29,56–59]:

$$\hat{F} = \arg\min_F \mathrm{E}_{x,y}\left[L\left(y, F(x)\right)\right] \tag{3}$$

The gradient boosting method approximates y with a weighted sum of functions $h_i(x)$ from some class H, called weak learners:

$$F(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const} \tag{4}$$

Using the principle of empirical risk minimization, an approximation $\hat{F}(x)$, that makes minimum the mean value for the loss function evaluated for the training

values, is searched. The initial model is a constant function $F_0(x)$, and step by step it expands its value in a greedy way [25,58,59]:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{5}$$

$$F_m(x) = F_{m-1}(x) + \arg\min_{h \in H} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + h(x_i)\right) \tag{6}$$

where $h \in H$ is a weak learner function. But the search of an optimal h for a given loss function L is not computationally feasible, a simplification is applied and a steepest descent method is used to solve this problem. For the continuous problem, H is a collection of differentiable functions on $\Re$, and the model is updated following the equations [25,58,59]:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_{F_{m-1}} L\left(y_i, F_{m-1}(x_i)\right) \tag{7}$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L\left(y_i, F_{m-1}(x_i)\right)}{\partial F_{m-1}(x_i)}\right) \tag{8}$$

where the derivatives are obtained with respect to the functions F_i for $i \in \{1, 2, ..., m\}$. If we are treating a discrete case, where the set *H* is finite, the candidate function *h* that is closest to the gradient of *L* will be chosen and the coefficient γ can then be calculated using line search in equations (7) and (8). This is a heuristic approach and will not give an exact solution to the problem, but only a good approximation.

The generic gradient boosting method can be described in pseudocode [25,27,58,59]:

- Input: differentiable loss function $L\left(y, F\left(x\right)\right)$, training set $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$ and iteration number *M*.

- Algorithm*:*

1. Initialize model using a constant value:
   $F_0\left(x\right) = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, \gamma\right)$
2. For $m = 1$ to M:
   - Compute so-called *pseudo-residuals*:
     $r_{im} = -\left[\frac{\partial L\left(y_i, F\left(x_i\right)\right)}{\partial F\left(x_i\right)}\right]_{F(x)=F_{m-1}(x)}$ for $i = 1, \ldots, n$
   - Fit a weak learner $h_m\left(x\right)$ to the pseudo-residuals using the training set $\left\{\left(x_i, r_{im}\right)\right\}_{i=1}^{n}$.
   - Calculate the multiplier γ_m solving the one-dimensional optimization problem:
     $\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}\left(x_i\right) + \gamma h_m\left(x_i\right)\right)$
   - Update the model:
     $F_m\left(x\right) = F_{m-1}\left(x\right) + \gamma_m h_m\left(x\right)$
3. Output $F_M\left(x\right)$.

Gradient boosting can be used with decision trees, in particular with CART trees, of a given fixed size as weak learners. For this particular situation, Friedman proposes a variation of the gradient boosting method that improves each weak learner quality of fit [25,27,58–60].

In the *m*-th step, a generic gradient boosting fits a decision tree $h_m\left(x\right)$ to the pseudo-residuals. If the leaf number is J_m, the input space is split into J_m separated regions $R_{1m}, \ldots, R_{J_m m}$ and a constant value is obtaind for each region. $h_m\left(x\right)$ for input x is written as the sum [25–29,58–60]:

$$h_m(x) = \sum_{i=1}^{J_m} b_{jm} I\left(x \in R_{jm}\right) \tag{9}$$

where b_{jm} is the constant value calculated for the region R_{jm}. These coefficients b_{jm} are multiplied by some value γ_m, calculated using line search that minimizes the loss function, and then the model is updated:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x); \quad \gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right) \tag{10}$$

Friedman proposed a modification of this algorithm that chooses a different optimal γ_{jm} for each of the regions, instead of only one γ_m for the whole tree. This modified algorithm is called TreeBoost. The model is updated according to [25–29,58,59]:

$$F_m(x) = F_{m-1}(x) + \sum_{i=1}^{J_m} \gamma_{jm} I\left(x \in R_{jm}\right); \quad \gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, F_{m-1}(x_i) + \gamma\right) \tag{11}$$

where the size of trees, *J*, is the number of terminal nodes in trees and it is a parameter that can be adjusted for the training set. The level of variable interaction is controlled by it. If $J = 2$, there is no interaction. $J = 3$ means that interactions between up to two variables can exists and so on.

Typically, a value between 4 and 8 works well and the results are quite insensitive to *J* for these values. $J = 2$ is usually not enough for many applications, and $J > 10$ is often unnecessary.

Overfitting the training set can lead to a bad prediction ability of the model. The *regularization* techniques are intended to reduce this overfitting effect controlling the training process.

There are different approaches to attain this aim that can be read in [25–29,58,59]. In particular, the technique used by the function xgboost is to include in the loss function the so called penalty function whose aim is to limit the overfitting:

$$L(x) = E(x) + \Omega(x) \tag{12}$$

where *E* can be, for instance, the mean squared error, and Ω is the penalty function that controls the model complexity, aiding to avoid overfitting by means of increasing the value of the loss function when the complexity of the model grows, thus penalizing it.

The GBRT method is greatly affected by its parameters [39,61], in particular:

- Nrounds: is the maximum number of iterations performed by the algorithm.
- η: controls the learning rate, that is to say, scales the contribution of each tree by a factor η when it is added to the current approximation. It is used to prevent overfitting by making the boosting process more conservative. Lower value for η implies larger value for Nrounds.
- γ: is the minimum loss reduction required to perform another partition on a leaf node of the tree. As it grows, the algorithm is more conservative.
- Minimum child weight: minimum sum of weight (hessian) required by a child.
- Maximum Δ step: is a cap value over each tree weight estimation.

- Subsample ratio: is the ratio between the training and testing instances.

Therefore, it is convenient to use some technique that adjusts these parameters. 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, grid search is a brute force method and, as such, almost any optimization method improves its efficiency. In this study, in order to avoid these problems associated with the grid search method, the differential evolution (DE) metaheuristic technique described below was used [30–33,41,42] with success.

2.2.2 Differential evolution (DE) algorithm

In evolutionary computation, differential evolution (DE) is a metaheuristic method that optimises a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. Furthermore, DE does not require for the optimized function to be differentiable. DE optimises a problem by maintaining a population of candidate solutions and creating new candidate solutions by combining existing ones according to its simple formulae, and then keeping whichever candidate solution has the best fitness on the optimization problem at hand [33].

The algorithm assumes that the variables of the problem to be optimised are encoded as a vector of real numbers. The length n of these vectors is equal to the number of variables of the problem, and the population is composed of NP vectors (number of parents). A vector $\mathbf{x}_p^g$ is defined, where p is the index of the individual in the population

$(p = 1,...,NP)$ and g is the corresponding generation. Each vector is composed in turn by the variables of the problem $x_{p,m}^{g}$, where m is the index of the variable in the individual $(m = 1,...,n)$. It is assumed that the domain of the problem variables is constrained between minimum and maximum values $\mathbf{x}_m^{\min}$ and $\mathbf{x}_m^{\max}$, respectively. Hence, DE technique is basically composed of four steps [30–33]:

- Initialization;
- Mutation;
- Recombination; and
- Selection.

Initialization is performed at the beginning of the search, and the mutation-recombination-selection steps are performed repeatedly, until a termination condition or stopping criterion is satisfied (number of generations, elapsed time, or quality of solution reached, etc.).

Initialization

The population is initialised (first generation) randomly, considering the minimum and maximum values of each variable:

$$\mathbf{x}_{p,m}^{1} = \mathbf{x}_m^{\min} + rand\,(0,1)\cdot\left(\mathbf{x}_m^{\max} - \mathbf{x}_m^{\min}\right) \text{ for } p = 1,...,NP \text{ and } m = 1,...,n \qquad (11)$$

where $rand\,(0,1)$ is a random number in the range $[0,1]$.

Mutation

The mutation is the construction of *NP* noisy random vectors, which are created from three individuals chosen at random, called target vectors $\mathbf{x}_a$, $\mathbf{x}_b$ and $\mathbf{x}_c$. The noisy random vectors $\mathbf{n}_p^t$ are obtained as follows:

$$\mathbf{n}_p^g = \mathbf{x}_c + F \cdot \left(\mathbf{x}_a - \mathbf{x}_b \right) \text{ for } p = 1, ..., NP \tag{12}$$

with *p*, *a*, *b* and *c* different from each other. *F* is a parameter that controls the mutation rate and is in the range $[0, 2]$.

Recombination

After obtaining the *NP* noisy random vectors, the recombination is performed in a random manner, comparing them with the original vectors $\mathbf{x}_p^g$, obtaining the trial vectors $\mathbf{t}_m^g$ as follows:

$$t_{p,m}^g = \begin{cases} n_{p,m}^g & \text{if} \quad rand(0,1) < GR \\ x_{p,m}^g & \text{otherwise} \end{cases} \text{for } p = 1, ..., NP \text{ and } m = 1, ..., n \tag{13}$$

GR is a parameter that controls the recombination rate. Note that the comparison is carried out variable by variable so that the test vector will be a mixture of the noisy random vectors and original vectors.

Selection

Finally, the selection is made simply by comparing the test vectors with the original ones, so that the vector of the next generation will be the one that has the best value of the fitness function fit:

$$\mathbf{x}_p^{g+1} = \begin{cases} \mathbf{t}_p^g & \text{if} \quad fit\left(\mathbf{t}_p^g\right) > fit\left(\mathbf{x}_p^g\right) \\ \mathbf{x}_p^g & \text{otherwise} \end{cases} \tag{14}$$

Variants of the Differential Evolution algorithm

Differential Evolution is not only a highly effective algorithm very adaptable to different optimization problems but it is also comparatively simple to implement. This has led to numerous variants that try to adapt to the problem that it is being tackled because there is not an algorithm that performs best for all problems. The approach in this article has been trying out and comparing different DE variants, choosing the one that performs better. Along with "vanilla" DE, we have tried:

- JADE: Adaptive Differential Evolution with Optional External Archive [62].
- SaDE: Differential Evolution Algorithm with Strategy Adaptation for Global Numerical Optimization [63].
- SHADE: Success-History based Adaptive DE [64].
- L-SHADE: SHADE with linear population size reduction [65].
- iL-SHADE: Improved L-SHADE algorithm [66].
- JSO: Single objective real-parameter optimization. It is an improved version of iL-SHADE [67].
- MPEDE: Differential evolution with multi-population based ensemble of mutation strategies [68].

2.2.3 The goodness–of–fit

The physico-chemical and biological independent variables used in this work are shown in Tables 1 and 2, respectively [3,69]. Thus, the number of predictors in DE/GBRT model was 16. The estimated variables (Chl-a and Total P concentrations) units are μg/L and mg P/L [69,70], respectively.

Table 1 Biological independent variables in this work with their mean and standard deviation

Table 2 Mean and standard deviation of the physico-chemical input values

It is necessary to use some criterion to choose a model over others when predicting the output variables from the remaining variables. In this paper, the goodness–of–fit criterion is the coefficient of determination R^2 [71,72]. It is a ratio that measures the relationship between the variation in the predicted variable explained by the model and the variability in the same variable through the dataset. If we term the observed values t_i, and the respective model predicted value y_i, these variabilities can be measured with the following sums of squares [71,72]:

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

where $\bar{t}$ is the average value of the *n* observed samples:

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i \tag{16}$$

Then, the coefficient of determination is:

$$R^2 \equiv 1 - \frac{SS_{err}}{SS_{tot}} \tag{17}$$

The coefficient of determination indicates how well the regression values approximate the actual values. The closer its value to one, the better.

2.2.4 The hybrid DE/GBRT–based model

Additionally, as previously mentioned, the GBRT technique success relies heavily on its hyperparameters such as the maximum number of iterations (Nrounds), learning rate, minimum loss reduction, minimum child weight, maximum step and subsample ratio. Different methods have been used to optimize the parameters [41,42,73,74]: random search, grid search, genetic algorithms, particle swarm optimization and so on. In this study, we have tried different variants of DE algorithm and selected the best one for tuning the GBRT parameters using a hybrid DE/GBRT–based model to estimate the Chl-a and Total P concentrations from the other sixteen biological and physico-chemical parameters (input variables). At this point, the flowchart of DE/GBRT method, where DE stands for a variant of Differential Evolution algorithm, used in this research work can be seen in Fig. 4.

Fig. 4 Flowchart of the new hybrid DE/GBRT–based model (DE stands for a variant of Differential Evolution algorithm)

Furthermore, the technique used here for computing the coefficient of determination $\left(R^2\right)$ was cross–validation [71,75]. Indeed, in order to assess the predictive capacity of the DE/GBRT–based model, a thorough 10-fold cross-validation algorithm was

implemented in this study [75]. To this end, the regression modeling has been performed with Extreme Gradient Boosting algorithm, using the *Xgboost* package [60,61] along with different DE techniques with the *PyADE* package [76], both in python. Table 3 shows the space search used in the DE method.

Table 3 Search space for each of the GBRT parameters in the DE tuning process

The GBRT parameters were optimized with the DE module. DE searches for the best GBRT parameters (maximum number of iterations, learning rate, minimum loss reduction, minimum child weight, maximum step and subsample ratio) for these data using the mean value of the coefficient of determination for the ten-fold cross-validation process. That is, this R^2 is the objective function in this problem. The search space is six-dimensional, with one dimension per each parameter.

We performed 20 runs for each pair optimization algorithm-objective function and obtained the corresponding 20 different R^2. We studied the normality of the distribution of these R^2 sets using the Shapiro-Wilk test and established what pairs of objective function-algorithm were best using parameters as the mean, the maximum or minimum value of each set of 20 R^2. As the data is non-paired, we performed the Kruskal-Wallis test, using the corresponding statistical function from the scipy module of python, to ascertain that the results were significantly different: a *p*-value of 1.3×10^{-7} was obtained from the R^2 values obtained for the Total P concentration model while a *p*-value of 1.6×10^{-7}. As both *p*-values are smaller than the threshold *p*-value 0.05, it can be concluded that there is a significant difference between the performances of the

algorithms. Additionally, to provide more detail, the algorithms are compared pair-wise using the Mann-Whitney test to check that the results of two particular algorithms were significantly different from the statistical point of view.

3 Results and discussion

The convergence curves for the different DE variants for optimizing the parameters of the GBRT model for the prediction of the Chl-a and Total P concentrations are shown in Figs. 5(a) and 5(b), respectively. Our objective function is the ten-fold cross-validation coefficient of determination (R^2) and we want to maximize it (the closer to one, the better) and thus, we have minimized $-R^2$. We can see that in both cases the L-SHADE algorithm is the one that performs best.

Fig. 5 Convergence curves for the different DE variants for optimizing the parameters of the GBRT model for the prediction of: (a) Chl-a concentration; and (b) Total P concentration.

The parameters that characterize each set from the 20 runs of an algorithm are shown in Tables 4 and 5. In both Tables, the algorithms are ranked, mainly (but not only) by the mean value for the 20 runs. Mann-Whitney test results are also taken into account to rank the algorithms. Even though the ranking order is not the same, the best algorithm in both cases is L-SHADE. In the last column of Tables 4 and 5 appears the p-value from the Shapiro-Wilk test for normality.

Table 4 Mean, median, minimum, maximum and p-value of the Shapiro-Wilk test the 20 coefficients of determination (R^2) obtained for each of the different hybrid models for the Chl-a using different parameter tuning optimization algorithms. The algorithms are ranked from worst to best. The best model is indicated in bold font

Table 5 Mean, median, minimum, maximum and p-value of the Shapiro-Wilk test the 20 coefficients of determination (R^2) obtained for each of the different hybrid models for the Phosphorus concentration using different parameter tuning optimization algorithms. The algorithms are ranked from worst to best. The best model is indicated in bold font

For all the sets of 20 R^2, the *p*-value is less than 0.05 which means that the distribution is not normal and thus an adequate test to ascertain if the results for each pair of algorithms are different enough is the Mann-Whitney rank test. The *p*-values obtained comparing two sets of R^2 appear in Tables 6 and 7.

Table 6 *p*-values of the Mann-Whitney rank test comparing the 20 coefficients of determination (R^2) obtained for each of the different hybrid models for the Chl-a using different parameter tuning optimization algorithms. The algorithms are ranked from worst to best

Table 7 *p*-values of the Mann-Whitney rank test comparing the 20 coefficients of determination (R^2) obtained for each of the different hybrid models for the Phosphorus

concentration using different parameter tuning optimization algorithms. The algorithms are ranked from worst to best

The algorithms are ranked from worst to best in Tables 4 to 7. The order is different when the objective function is different. That is, we have an objective function for the Chl-a model and another one for the Total P concentration. We can see that, in many cases, two algorithms overlap as the *p*-value from the Mann-Whitney test that compares the two algorithms is greater than 0.05. For example, in the case of Chl-a, the algorithms that perform worse are iL-SHADE and JSO, which are not very different as the *p*-value from the Mann-Whitney test obtained from the samples of the results of these two algorithms, 0.2930, is greater than 0.05. But JSO is slightly better because its mean, median and maximum values are higher. Also, JSO overlaps with DE, which is a little better. But in any case, DE is significantly better than iL-SHADE, as the *p*-value from the Mann-Whitney test that involves the two algorithms is equal to 0.03, less than 0.05. The next in the ranking is SaDE, which is better than DE but does not differ significantly from it, as the *p*-value is 0.2. But it is significantly better than iL-SHADE and JSO. And so on. The algorithms that top the ranking are JADE and L-SHADE. Nevertheless, L-SHADE is slightly better in this case. Thus, we have chosen L-SHADE for the final model. On the other hand, in the case of the model for Total P concentrations, L-SHADE is the absolute winner, being the best and also differing significantly with all the other algorithms.

Thus, we have used the L-SHADE algorithm that was the best in both cases and have picked the results for the best run obtaining an R^2 of 0.91 for the Chl-a model and 0.93 for the Total P concentration model. (see Table 8).

Table 8 Coefficients of determination (R^2) and correlation coefficients (r) for the hybrid L-SHADE/GBRT–based models fitted in this study for the Chl-a and Total P concentrations

The optimal parameters for the best-fitted GBRT models found with the L-SHADE technique for the Chl-a and Total P concentrations, respectively, appear in Tables 9 and 10.

Table 9 Optimal hyperparameters of the best-fitted GBRT model for Chl-a concentration found with the L-SHADE technique

Table 10 Optimal hyperparameters of the best-fitted GBRT model for Phosphorous concentration found with the L-SHADE technique

The importance measures are relative and the addition of all the values for each criterion amounts to one. They are:

- Gain: it is computed taking into account each variable contribution to each tree that appears in the model.
- Cover: it is the relative number of observations of the variable in the model.
- Frequency: it is the relative number of times an independent variable appears in the trees of the obtained model.

The most significant one is the Gain and thus it has been used to construct the graph of the relative importance of the variables.

As an additional result of these calculations, the significance ranking of the biological and physico-chemical parameters (input variables), taking as dependent variables Chl-a and Total P concentrations (output variables) in this complex study, are shown in Tables 11 and 12, and Figs. 6 and 7, respectively.

Table 11 Relative importance of the variables in the optimal L-SHADE /GBRT model for the Chl-a concentration estimation using the criteria Gain, Cover and Frequency

Table 12 Relative importance of the variables in the optimal L-SHADE /GBRT model for the Phosphorous concentration estimation using the criteria Gain, Cover and Frequency

Fig. 6 Relative importance of the input operation variables to predict the Chl-a concentration.

Fig. 7 Relative importance of the input operation variables to predict the Total P concentration.

In La Barca reservoir, the Cyanobacteria concentration is, by far, the most significant variable to predict Chl-a concentration according to the GBRT model. Its contribution to Chl-a concentration in the model is about 60%. The reason for this high influence can be because as algal grow (phytoplankton biomass increases), Chl-a concentration raises and also cyanobacteria density (the most common side effect of the algal abnormal proliferation in freshwater ecosystems) and cyanobacteria only contain this form of

chlorophyll. Therefore, the phytoplankton enrichment in cyanobacteria contributes to a higher increase in Chl-a concentration. Cyanobacteria proliferation leads to water deterioration since this kind of phytoplankton can produce a range of potent toxins, the cyanotoxins [49,77–79], in freshwater ecosystems that causes serious environmental problems in lakes and reservoirs [9,47,49,80]. Therefore, the measure of cyanobacteria concentration seems to be a good estimation of both Chl-a concentration in La Barca reservoir and water toxicity caused by cyanotoxins.

Markedly less important predictors seem to be Water Temperature and Turbidity (less than 9% contribution). Nevertheless, they are the second and the third variables, respectively, in the rank for Chl-a prediction. This is probably because Water Temperature is an important factor in the reproduction, frequency and distribution of phytoplankton and Turbidity is related to phytoplankton growth [53]. Water temperature was found to be the one of the main limiting factors for phytoplankton growth [81].

Secchi depth is the fourth most significant variable in the Chl-a prediction. It is a measure of transparency and its value goes down as Chl-a concentration increases with algal growth, so Secchi depth is a reasonable indicator of algal abundance, except in high colored lentic waters bodies where turbidity is due to clay, carbonate but not algae [82]. However, it must be kept in mind that there is a nonlinear relationship between Secchi depth and Chlorophyll [83].

The other input variables contributions to Chl-a prediction are much smaller. Thus, the fifth and the sixth variables in the Chl-a prediction importance are the Conductivity and Phosphorus, respectively. The first one is a measure of the salt content in water [84] and

salinity affects phytoplankton concentration and composition [85] and, hence, the Chl-a content in water. Total phosphorus is an essential nutrient and the main responsible for algal abnormal productivity –it is the limiting nutrient in many freshwater bodies [86]–. Total phosphorus is another one of the main limiting factors for phytoplankton growth [81]. Similar contribution has Dinoflagellates concentration, the seventh most significant variable. Dinoflagellates are related to eutrophic environments [87], as La Barca reservoir, and to Harmful Algal Blooms [88]. However, not all dinoflagellates are photosynthetic, some of them are heterotrophs, have not chlorophyll, and other members are able to use both trophic strategies depending on the environmental conditions [89]. The dinoflagellate photosynthetic species seem to be able to tolerate lower inorganic nutrients levels than diatoms and some of them tolerate brackish conditions [90]. The non-photosynthetic dinoflagellates may use organic sources of nitrogen [91].

The chlorophytes concentration is the eighth most significant variable in Chl-a prediction. Although chlorophytes concentration contribution to Chl-a prediction is very small, they are also bound to eutrophic ecosystems [92,93], as dinoflagellates, and they dominate over cyanobacteria when effluents are rich in Nitrogen [94]. Therefore, both dinoflagellates and chlorophytes are favor over cyanobacteria in rich Nitrogen environments. Hence, their presence in water should involve less cyanobacteria concentration, so a negative contribution to Chl-a as it was shown.

Then, the DO concentration is the ninth most significant variable in Chl-a prediction. In the DO case, much of the reason of its influence in Chl-a prediction lies in two processes: the photosynthetic activity that releases oxygen into the environment (the

cause why algal blooms produce much oxygen during the day and its lowering at night) and the extended decomposition processes occurring when significant algae amounts die and sink causing DO depletion and creating *dead zones* [95]. Consequently, there is a relationship between DO and chlorophyll-containing organisms present in water as it was pointed out for *Microscystis* blooms [96].

The next significant variables to predict Chl-a concentration in the studied reservoir, according to the hybrid DE/GBRT model, were nitrate concentration [97,98], diatoms concentration, Euglenophytes concentration –high in calm waters–, cryptophytes concentration, Chrysophytes concentration –Chrysophytes have different environmental demands [99], some grow well under unproductive lake conditions, while others are typical of eutrophic environments [100]–, Ammonium concentration – in general, phytoplankton prefer ammonium over other nitrogen sources [101]– and, finally, pH, the last variable in the importance ranking for the Chl-a prediction.

In the case of Total P concentration model the independent variables ranking is shown in Table 12 and Fig. 7. For this second GBRT model the most relevant input variable is the Turbidity –in good agreement with published results [102–104]–, with a contribution of almost 65% to the Total P concentration, more than 5 times the second variable contribution and about 22.4 times the third variable contribution.

Turbidity is mainly due to urban, agricultural and industrial wastes [52] that diminish the water clarity, directly and, in lentic water bodies, indirectly by incorporating nutrients, as phosphorus, that promote phytoplankton growth diminishing water transparency –eutrophication [53]–. Although Phosphorus is the limiting nutrient in the

case of lentic water bodies as reservoirs or lakes [105,106], and was a relevant variable in the Chl-a prediction shown before, phytoplankton growth does not seem to be the main reason of the relationship between turbidity and phosphorus in La Barca reservoir since turbidity was not relevant in the mentioned Chl-a prediction (about 2% contribution).

Secchi depth is the second most significant variable, and it is more relevant here than for the Chl-a prediction shown before, in agreement with observations in some U.S. lakes [107]. Its contribution to the Total P in the model is smaller than Turbidity (about 13% versus 65%) and can be explained because Secchi depth accounts for color, i.e., it could be expressed as a function of turbidity and water color [108] by an indirect relationship through algal growth: phosphorus concentration rise stimulates algal growth that decreases water clarity and consequently, Secchi depth [83].

The next variable in the ranking to predict Total P is Chl-a concentration (about 3% contribution). Chl-a concentration is related to the concentration of phytoplankton [51] whose growth is stimulated by increases in phosphorus concentration. It is the third significant variable in Total P prediction but its contribution is the same that the Conductivity contribution, so that the next variable in the ranking to predict Total P is Conductivity (about 3% contribution). Ion phosphate is the main component of Total P in eutrophic environments as La Barca reservoir since it is the form that autotrophs can assimilate [103] so the more Total P the more ion phosphate concentration and, consequently, the more conductivity.

The next input variables have contributions to Total P prediction less than 3%. In general, these contributions are based on the direct or indirect relationships between the input variables and phytoplankton species growth due to Total P.

Figs. 8 and 9 show the comparison between observed and predicted values by using the DE/GBRT–relied model, for Chl-a and Total P concentrations, respectively.

Fig. 8 Comparison between the observed and predicted Chl-a concentration by using the L-SHADE/GBRT–based model ($R^2 = 0.91$)

Fig. 9 Comparison between the observed and predicted Total P concentration by using the L-SHADE/GBRT–based model ($R^2 = 0.93$)

4 Conclusions

The eutrophication in La Barca reservoir can be accurately modelled using the new hybrid L-SHADE/GBRT–based model developed in this research. This new lower-priced method is a good alternative to costly traditional methods.

A high coefficient of determination ($R^2 = 0.91$) was obtained when the hybrid L-SHADE/GBRT–based model was trained and tested with the experimental dataset corresponding to the Chl-a concentration. The estimated values for this model agree with the Chl-a observed dataset values (see Fig. 8). In the same way, the model for the experimental dataset of the Total P concentration obtained a high coefficient of

determination ($R^2 = 0.93$). The predicted results for the algal atypical proliferation agree with the Total P concentration dataset of observed values (see Fig. 9).

The method also ranked the input variables involved in the eutrophication prediction. Thus, the Cyanobacteria concentration was the most influential parameter in the Chl-a model, whilst the Turbidity was the most related one to the Total P concentration.

The combination of factors that trigger and sustain an algal bloom is not well understood at present, and it is not possible to attribute algal blooms to any specific factor. Our regression approach aims to be a contribution in this regard, identifying the variables or factors that most influence the algae proliferation, without making any assumptions and letting the model extract the information from the data. Therefore, although nature and natural processes are highly complex and difficult to predict [109], our method allowed a better understanding of the interrelationship among the water eutrophication variables that could help for prevention and remediation. Thus, taking into account the importance ranking obtained from the model (see Tables 11 and 12, and Figs. 6 and 7), some contributions can be made on the dynamics of Algal atypical proliferation:

- The Cyanobacteria (blue-green algae) have a competitive advantage over other types of algae in the environment that we have studied.
- The high temperatures are optimal for the growth of Cyanobacteria.
- The phosphorus is the main nutrient that promotes the growth of Cyanobacteria in the studied reservoir.
- In turbid water, blue-green algae have higher growth rates than any other group

of algae. In addition, the main cause of turbidity appears to be runoff from surface waters, which favors phosphorus enrichment of waterways. The phosphorus runoff encourages the growth of Cyanobacteria.

These variables (Cyanobacteria, Temperature, Phosphorus, and Turbidity) appear like those that have the greatest influence on the dynamics of the atypical proliferation of algae in the reservoir waters in which we have applied our regression approach. In turn, Turbidity appears as the factor most related to phosphorus in the reservoir.

Future research works will focus on the application of this model to new results and also to other reservoirs in order to determine the spatial and temporal differences.

Finally, hybrid L-SHADE/GBRT regression method seems to improve significantly the generalization capability obtained with the only GBRT–based regressor, without optimizing its parameters.

Summing up, an effective L-SHADE/GBRT–based model could be an interesting tool to estimate and prevent the eutrophication in water bodies.

Acknowledgements The authors wish to thank the Cantabrian Basin Authority (Spanish Ministry of Agriculture, Food and Environment) for providing the experimental dataset used in this research. Additionally, we would like to thank Anthony Ashworth for his revision of English grammar and spelling of the manuscript.

Compliance with ethical standards

Conflicts of Interest The authors declare no conflict of interest.

References

1. Alexandrov MH, Bloesch J (2009) Eutrophication of lake Tasaul, Romania—proposals for rehabilitation. Environ Sci Pollut R 16(1):42–45
2. Grundy RD (1971) Strategies for control of man-made eutrophication. Environ Sci Tech 5:1184–1190
3. Reynolds CS (2006) Ecology of phytoplankton. Cambridge University Press, New York
4. Xue X, Landis A (2010) Eutrophication potential of food consumption patterns. Environ Sci Technol 44:6450–6456
5. Smith VH (1983) Low nitrogen to phosphorus ratios favor dominance by blue-green algae in Lake Phytoplankton. Science 221(4611):669–671
6. Álvarez Cobelas M, Arauzo M (2006) Phytoplankton responses to varying time scales in a eutrophic reservoir. Arch Hydrobiol Ergebn Limnol 40:69–80
7. Liu Y, Guo H, Yu Y, Dai Y, Zhou F (2008) Ecological–economic modeling as a tool for watershed management: A case study of Lake Qionghai watershed, China. Limnologica 38(2):89–104
8. 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
9. Texeira MR, Rosa MJ (2006) Comparing disolved air flotation and conventional sedimentation to remove cyanobacterial cells of Microcystis aeruginosa: part I: the key operating conditions. Sep Purif Technol 52:84–94
10. Gibson G, Carlson R, Simpson J, Smelzer E (2000) 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
11. Karydis M (2009) Eutrophication assessment of coastal waters based on indicators: a literature review. Global NEST J 11:373–390
12. Spatharis S, Tsirtsis G (2010) Ecological quality scales based on phytoplankton for the implementation of Water Framework Directive in Eastern Mediterranean. Ecol Indic 10(4):840–847
13. Borja A, Dauer DM (2008) Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. Ecol Indic 8:331–337
14. 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
15. Hillebrand H, Dürselen C–D, Kirschtel D, Pollinger U, Zohary T (1999) Biovolume calculation for pelagic and benthic microalgae. J Phycol 35:403–424
16. 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

17. Kitsiou D, Karydis M (2011) Coastal marine eutrophication assessment: a review on data analysis. Environ Int 37:778–801
18. Karlson K, Rosenberg R, Bonsdorff E (2002) Temporal and spatial large-scale effects of eutrophication and oxygen deficiency on benthic fauna in Scandinavian waters–a review. Oceanogr Mar Biol Ann Rev 40:427–489
19. Charpa SC (1997) Surface water-quality modelling. McGraw-Hill, New York
20. Díaz RJ, Rosenberg R (2011) Introduction to environmental and economic consequences of hypoxia. Int J Water Resour Dev 27:71–82
21. Chibole OK (2013) Modeling River Sosiani's water quality to assess human impact on water resources at the catchment scale. Ecohydrol Hydrobiol 13:241–245
22. Wu G, Xu Z (2011) Prediction of algal blooming using EFDC model: case study in the Daoxiang Lake. Ecol Modell 222:1245–1252
23. Dogan E, Sengorur B, Koklu R (2009) Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. J Environ Manag 90:1229–1235
24. Singh KP, Basant A, Malik A, Jain G (2009) Artificial neural network modeling of the river water quality—a case study. Ecol Modell 220:888–895
25. Hastie T, Tibshirani R, Friedman J (2016) The elements of statistical learning. Springer-Verlag, New York
26. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth and Brooks, Monterrey, CA
27. Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data An 38(4):367–378
28. Schapire RE (2003) The boosting approach to machine learning an overview. In: Denison DD, Hansen MH, Holmes CC, Mallick B, Yu B (eds) Nonlinear estimation and classification, Lecture notes in statistics, Springer, Germany, vol 171, pp 149–171
29. Bühlmann P, Hothorn T (2007) Boosting algorithms: regularization, prediction and model fitting. Stat Sci 22(4):477–505
30. Feoktistov V (2006) Differential evolution: in search of solutions. Springer, New York
31. Price K, Storn RM, Lampinen JA (2005) Differential evolution: a practical approach to global optimization. Springer, Berlin
32. Rocca P, Oliveri G, Massa A (2011) Differential evolution as applied to electromagnetics. IEEE Antennas Propag 53(1):38–49
33. Storn R, Price K (1997) Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 11:341–359
34. Landry M, Erlinger TP, Patschke D, Varrichio C (2016) Probabilistic gradient boosting machines for GEFCom2014 wind forecasting. Int J Forecasting 32(3):1061–1066
35. Persson C, Bacher P, Shiga T, Madsen H (2017) Multi-site solar power forecasting using gradient boosted regression trees. Sol Energy 150:423–436
36. Johnson NE, Ianiuk O, Cazap D, Liu L, Starobin D, Dobler G, Ghandehari M (2017) Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City. Waste Manage 62:3–11
37. Xiong D, Gui Q, Hou W, Ding M (2018) Gradient boosting for single image super-resolution. Information Sciences 454–455:328–343

38. Lu H, Wang H, Yoon SW (2019) A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. Expert Systems with Applications 116:340–350
39. Chen Z–Y, Zhang T–H, Zhang R, Zhu Z–M, Yang J, Cheng P–Y, Ou C–Q, Guo Y (2019) Extreme gradient boosting model to estimate $PM_{2.5}$ concentrations with missing-filled satellite data in China. Atmospheric Environment 202:180–189
40. Liu B, Yan S, You H, Dong Y, Li Y, Lang J, Gu R (2018) Road surface temperature prediction based on gradient extreme learning machine boosting. Computers in Industry 99:294-302
41. Simon D (2013) Evolutionary optimization algorithms. Wiley, New York
42. Yang X–S, Cui Z, Xiao R, Gandomi AH, Karamanoglu M (2013) Swarm intelligence and bio-inspired computation: theory and applications. Elsevier, London
43. Fister I, Stranad D, Yang X–S, Fister Jr I (2015) Adaptation and hybridization in nature-inspired algorithms. In: Fister I, Fister Jr I (eds) Adaptation and hybridization in computational intelligence, Springer, New York, vol 18, pp 3–50
44. Fogg GE, Stewart WDP, Fay P, Walsby AE (1973) The blue-green algae. Academic Press, London
45. Smith MJ, Shaw GR, Eaglesham GK, Ho L, Brookes JD (2008) Elucidating the factors influencing the biodegradation of cylindrospermopsin in drinking water sources. Environ Toxicol 23:413–421
46. World Health Organization (1998) Guidelines for drinking-water quality: health criteria and other supporting information, vol 2, Geneva, World Health 408 Organization
47. Willame R, Jurckzak T, Iffly JF, Kull T, Meriluoto J, Hoffman L (2005) Distribution of hepatotoxic cyanobacterial blooms in Belgium and Luxembourg. Hydrobiologia 551:99–117
48. Brönmark C, Hansson L–A (2005) The biology of lakes and ponds. Oxford University Press, New York
49. Quesada A, Moreno E, Carrasco D, Paniagua T, Wormer L, de Hoyos C, Sukenik A (2006) Toxicity of Aphanizomenon ovalisporum (Cyanobacteria) in a Spanish water reservoir. Eur J Phycol 41:39–45
50. Negro AI, de Hoyos C, Vega JC (2000) Phytoplankton structure and dynamics in Lake Sanabria and Valparaíso reservoir (NW Spain). Hydrobiologia 424:25–37
51. American Public Health Association, American Water Works Association, Water Environment Federation (2005) Standard Methods for the Examination of Water and Wastewater, no. 21. APHA/AWWA/WEF, Washington
52. France RL, Peters RH (1995) Predictive model of the effects on Lake Metabolism of decreased airborne litter fall through riparian deforestation. Conserv Biol 9(6):1578–1586
53. Nicholls KH, Steedman RJ, Carey EC (2003) Changes in phytoplankton communities following logging in the drainage basins of three boreal forest lakes in north-western Ontario. Can J Fish Aquat Sci 60:43–54
54. Friedman JH, Hastie T, Tibshirani R (2000) Additive logistic regression: A statistical view of boosting. Ann Stat 28(2):337–407
55. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. Ann Stat 29(5):1189–1232

56. Mayr A, Binder H, Gefeller O, Schmid M (2014) The evolution of boosting algorithms: From machine learning to statistical modelling. Method Inform Med 6(1):419–427
57. Mayr A, Binder H, Gefeller O, Schmid M (2014) Extending statistical boosting: An overview of recent methodological developments. Method Inform Med 6(2):428–435
58. Taieb SB, Hyndman RJ (2014) A gradient boosting approach to the kaggle load forecasting competition. Int J Forecasting 30(2):382–394
59. Döpke J, Fritsche U, Pierdzioch C (2017) Predicting recessions with boosted regression trees. Int J Forecasting 33:745–759
60. Ridgeway G (2007) Generalized boosted models: a guide to the GBM package. http://www.saedsayad.com/docs/gbm2.pdf. Accessed 3 August 2007
61. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, California, USA, pp 785–794
62. Brest J, Maucec M S, Boskovic B (2017) Single objective real-parameter optimization: Algorithm jSO. In: Proceedings of the 2017 IEEE Congress on Evolutionary Computation, IEEE Publisher, Donostia-San Sebastián, Spain, pp 1311–1318
63. Brest J, Maucec MS, Boskovic B (2016) iL-SHADE: Improved L-SHADE algorithm for single objective real-parameter optimization. In: Proceedings of the 2016 IEEE Congress on Evolutionary Computation, IEEE Publisher, Vancouver, BC, Canada, pp 1188–1195
64. Tanabe R, Fukunaga AS (2013) Evaluating the performance of SHADE on CEC 2013 benchmark problems. In: Proceedings of the 2013 IEEE Congress on Evolutionary Computation, IEEE Publisher, Cancun, Mexico, pp 1952–1959
65. Qin AK, Huang VL, Suganthan PN (2009) Differential evolution algorithm with strategy adaptation for global numerical optimization. IEEE T Evolut Comput 13(2):398–417
66. Zhang J, Sanderson AC (2009) JADE: Adaptive differential evolution with optional external archive. IEEE T Evolut Comput 13(5): 945–958
67. Tanabe R, Fukunaga AS (2014) Improving the search performance of SHADE using linear population size reduction. In: Proceedings of the 2014 IEEE Congress on Evolutionary Computation (CEC), IEEE Publisher, Beijing, China, pp 1658–1665
68. Wu G, Mallipeddi R, Suganthan PN, Wang R, Chen H (2016) Differential evolution with multi-population based ensemble of mutation strategies. Inform Sciences 329:329–345
69. Allman ES, Rhodes JA (2003) Mathematical models in biology: an introduction. Cambridge University Press, New York
70. Barnes DJ, Chu D (2010) Introduction to modeling for biosciences. Springer, New York
71. Freedman D, Pisani R, Purves R (2007) Statistics. W.W. Norton & Company, New York
72. Wasserman L (2003) All of statistics: a concise course in statistical inference. Springer, New York
73. Bishop CM (2011) Pattern recognition and machine learning. Springer, Berlin
74. Berk RA (2016) Statistical learning from a regression perspective. Springer, Berlin

75. Picard R, Cook D (1984) Cross-validation of regression models. J Am Stat Assoc 79(387):575–583
76. Criado Ramón D (2017) PyADE: Multiple differential evolution algorithms. https://github.com/xKuZz/pyade. Accessed 23 October 2017
77. Chorus I, Bartram J (1999) Toxic cyanobacteria in water: a guide to their public health consequences, monitoring and management. E & FN Spon, London
78. Aboal M, Puig MA (2005) Intracellular and dissolved microcystins in reservoirs of the river Segura basin, Murcia, SE Spain. Toxicon 45(4):509–518
79. Gault PM, Marler HJ (2009) Handbook on Cyanobacteria: biochemistry, biotechnology and applications. Nova Science Publishers, New York
80. Dow CS, Swoboda UK (2000) Cyanotoxins. In: Whitton BA, Potts M (eds) The ecology of Cyanobacteria: their diversity in time and space, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 613–632
81. Lv J, Wu H, Chen M (2011) Effects of nitrogen and phosphorus on phytoplankton composition and biomass in 15 subtropical, urban shallow lakes in Wuhan, China. Limnologica 41(1):48–56
82. Brezonik PL, Menken KD, Bauer ME (2005) Landsat-based remote sensing of lake water quality characteristics, including chlorophyll and colored dissolved organic matter (CDOM). Lake Reserv Manage 21:373–382
83. Rast W, Thornton JA (2005) Phosphorus loading concept and the OECD eutrophication programme: origin, application and capabilities. In: O'Sullivan P, Reynolds CS (eds) The lakes handbook: lake restoration and rehabilitation, Blackwell Science, Oxford, UK, vol 2, pp 354–385
84. Gadi VK, Tang Y–R, Das A, Monga C, Garg A, Berretta C, Sahoo L (2017) Spatial and temporal variation of hydraulic conductivity and vegetation growth in green infrastructures using infiltrometer and visual technique. Catena 155:20–29
85. Redden AM, Rukminasari N (2008) Effects of increases in salinity on phytoplankton in the Broadwater of the Myall lakes, NSW, Australia. Hydrobiologia 608:87–97
86. Schindler DW, Hecky RE, Findlay DL, Stainton MP, Parker BR, Paterson MJ, Beaty KG, Lyng M, Kasian SEM (2008) Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. P Natl Acad Sci USA 105(32):11254–11258
87. Bužančić M, Gladan ŽN, Marasović I, Kušpilić G, Grbec B (2016) Eutrophication influence on phytoplankton community composition in three bays on the eastern Adriatic coast. Oceanologia 58(4):302–316
88. Watson SB, Whitton BA, Higgins SN, Paerl HW, Brooks BW, Wehr JD (2015) Harmful algal blooms. In: Wehr J, Sheath R, Kociolek JP (eds) Aquatic ecology, freshwater algae of North America, Academic Press, Cambridge, Massachusetts, USA, chpt 20, pp 873–920
89. Dagenais-Bellefeuille S, Morse D (2013) Putting the N in dinoflagellates. Front Microbiol 4(369):1–14
90. Taylor FJR (1976) Dinoflagellates from the International Indian Ocean Expedition. Biblioth Bot 132:1–234
91. Rissik D, van Senden D, Doherty M, Ingleton T, Ajani P, Bowling L, Gibbs M, Gladstone M, Kobayashi T, Suthers I, Froneman W (2009) Plankton-related environmental and water-quality issues. In: Suthers IM, Rissik D (eds) Plankton: a guide to their ecology and monitoring for water quality, CSIRO Publishing, Melbourne, Australia, chpt 3, pp 39–72

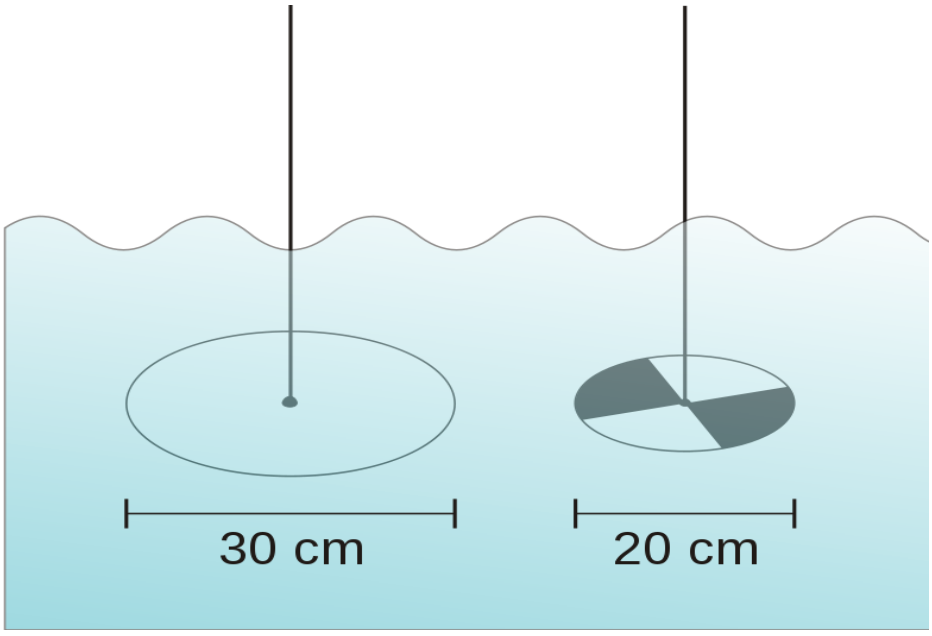
92. Ortega-Mayagoitia E, Rojo C (2000) Fitoplancton del Parque Nacional Las Tablas de Daimiel. III. Diatomeas y clorofitas. Anal Jardin Bot Mad 58(1):17–37
93. Pérez–Martínez C, Sánchez–Castillo P (2004) Temporal occurrence of Ceratium hirundinella in Spanish reservoirs. Hydrobiologia 452:101–107
94. Bogard MJ, Vogt RJ, Haye NM, Leavitt PR (2020) Unabated nitrogen pollution favors growth of toxic Cyanobacteria over Chlorophytes in most hypereutrophic lakes. Environ Sci Technol 54(6):3219–3227
95. Arend KK, Betelsky D, DePinto JV, Ludsin SA, Roberts JJ, Rucinski DK, Scavia D, Schwab DJ, Hook TO (2011) Seasonal and interannual effects of hypoxia on fish habitat quality in central Lake Erie. Freshwater Biol 56:366–383
96. Huang Y, Chen M (2013) Variation of dissolved oxygen in the experiments of occurrence & disappearance for Microcystis bloom. Procedia Environ Sci 18:559–566
97. Fields S (2004) Global nitrogen: cycling out of control. Environ Health Persp 112(10):A556–A563
98. Zhan X, Bo Y, Zhou F, Liu X, Paerl HW, Shen J, Wang R, Li F, Tao S, Dong Y, Tang X (2017) Evidence for the importance of atmospheric nitrogen deposition to eutrophic lake Dianchi, China. Environ Sci Technol 51(12):6699–6708
99. Zeeb BA, Christie CE, Smol JP, Findlay DL, Kling HJ, Birks HJB (1994) Responses of diatom and Chrysophyte assemblages in Lake 227 sediments to experimental eutrophication. Can J Fish Aquat Sci 51(10):2300–2311
100. Ortega–Mayagoitia E, Rojo C (1999) Phytoplankton from the Daimiel National Park. II. Cyanophytes, dinoflagellates, cryptophytes, chrysophytes and xanthophytes. Anal Jardín Bot Mad 57(2):251–266
101. Bronk DA, See JH, Bradley P, Killberg L (2007) DON as a source of bioavailable nitrogen for phytoplankton. Biogeosciences 4:283–296
102. Schilling KE, Kim S-W, Jones CS (2017) Use of water quality surrogates to estimate total phosphorus concentrations in Iowa rivers. J Hydrol Reg Stud 12:111–121
103. Lannergård EE, Ledesma JLJ, Fölster J, Futter MN (2019) An evaluation of high frequency turbidity as a proxy for riverine total phosphorus concentrations. Sci Total Environ 651(1):103–113
104. Villa A, Fölster J, Kyllmar K (2019) Determining suspended solids and total phosphorus from turbidity: comparison of high-frequency sampling with conventional monitoring methods. Environ Monit Assess 191(10):605–620
105. Correll DL (1999) Phosphorus: a rate limiting nutrient in surface waters. Poultry Sci 78(5):674–682
106. Schindler DW, Carpenter SR, Chapra SC, Hecky RE, Orihel DM (2016) Reducing phosphorus to Curb lake eutrophication is a success. Environ Sci Technol 50(17):8923−8929
107. Lambou VW, Hern SC, Taylor WD, Williams LR (1982) Chlorophyll, phosphorus, Secchi disk, and trophic state. J Am Water Resour As 18(5):807−813
108. Brezonik PL (1978) Effect of organic color and turbidity of Secchi disk transparency. J Fish Res Board Can 35(11):1410−1416
109. Stehlík M, Dušek J, Kiseľák J (2016) Missing chaos in global climate change data interpreting? Ecol Complex 25:53–59

(a)

(b)

Fig. 1 (a) La Barca reservoir, aerial photograph; and (b) another aerial photograph at a larger scale

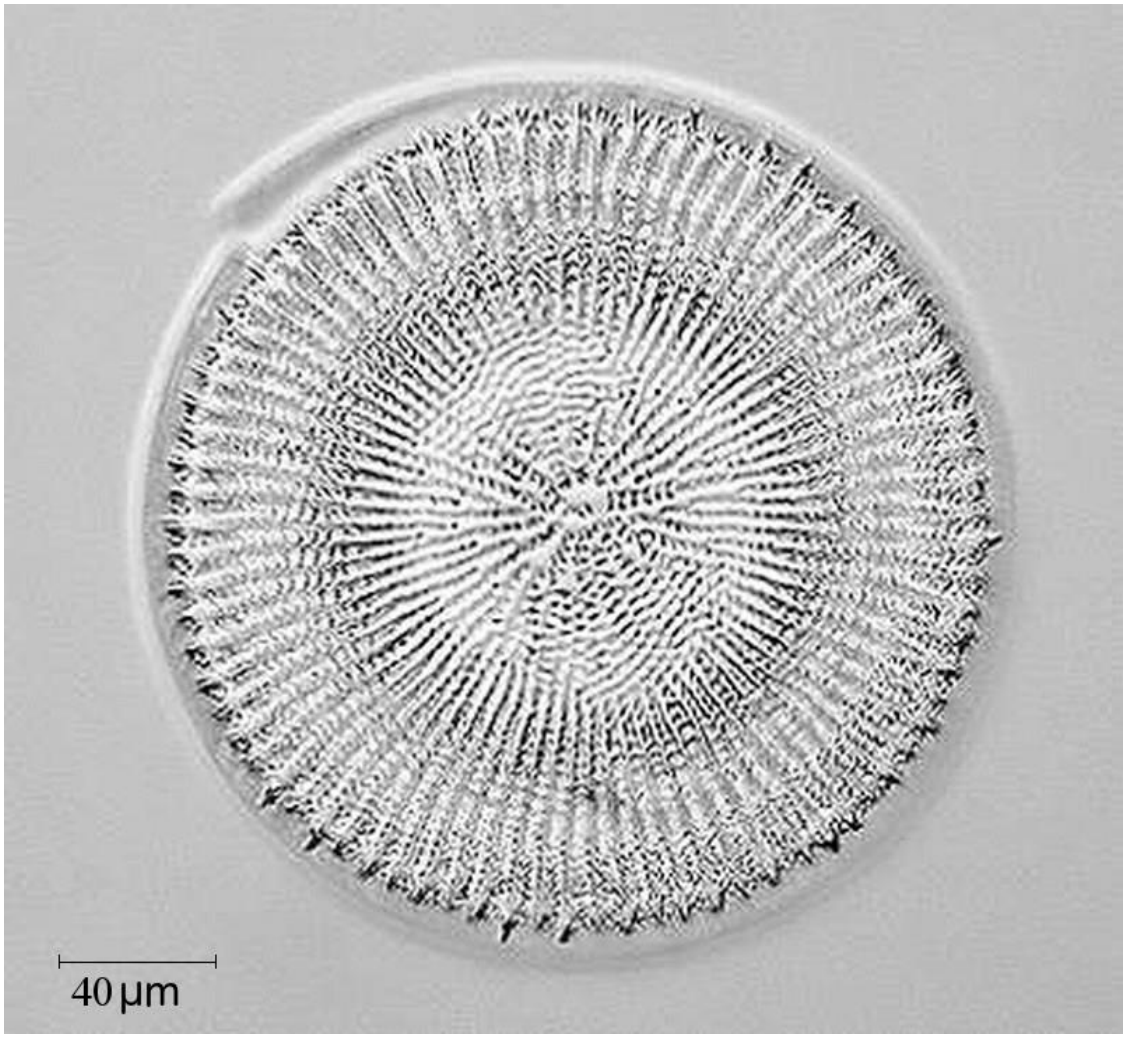
(a)



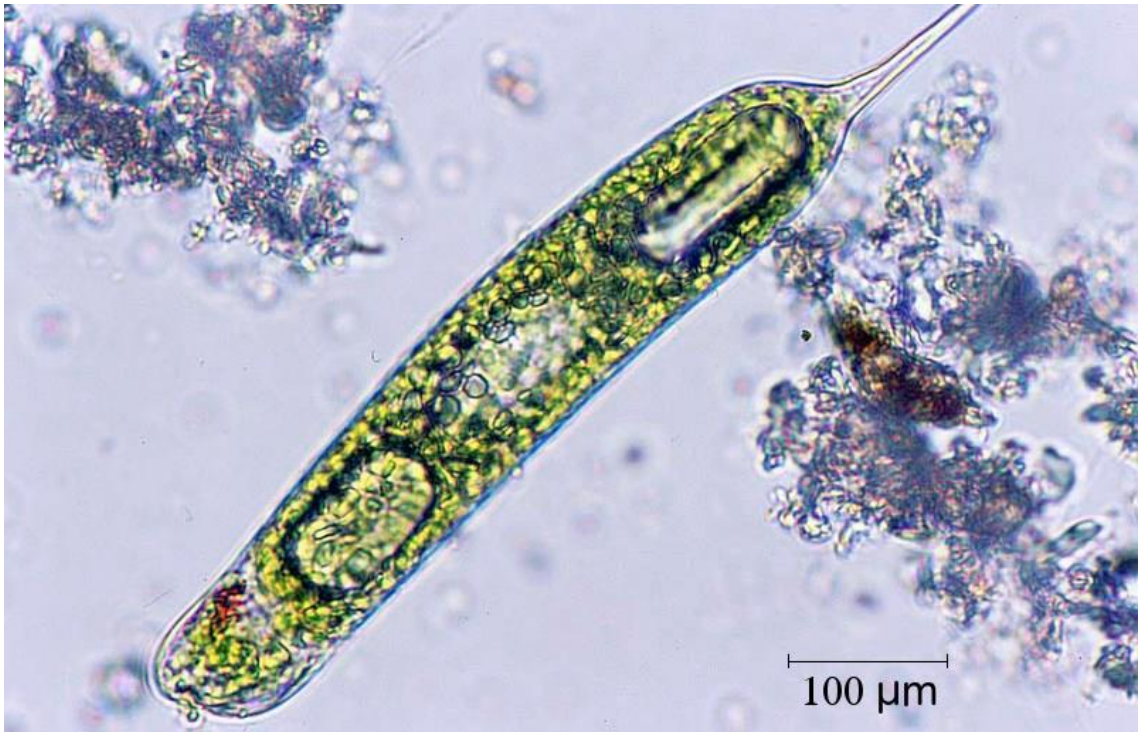
(b)

Fig. 2 (a) A Niskin bottle; and (b) examples of Secchi disks

(a)



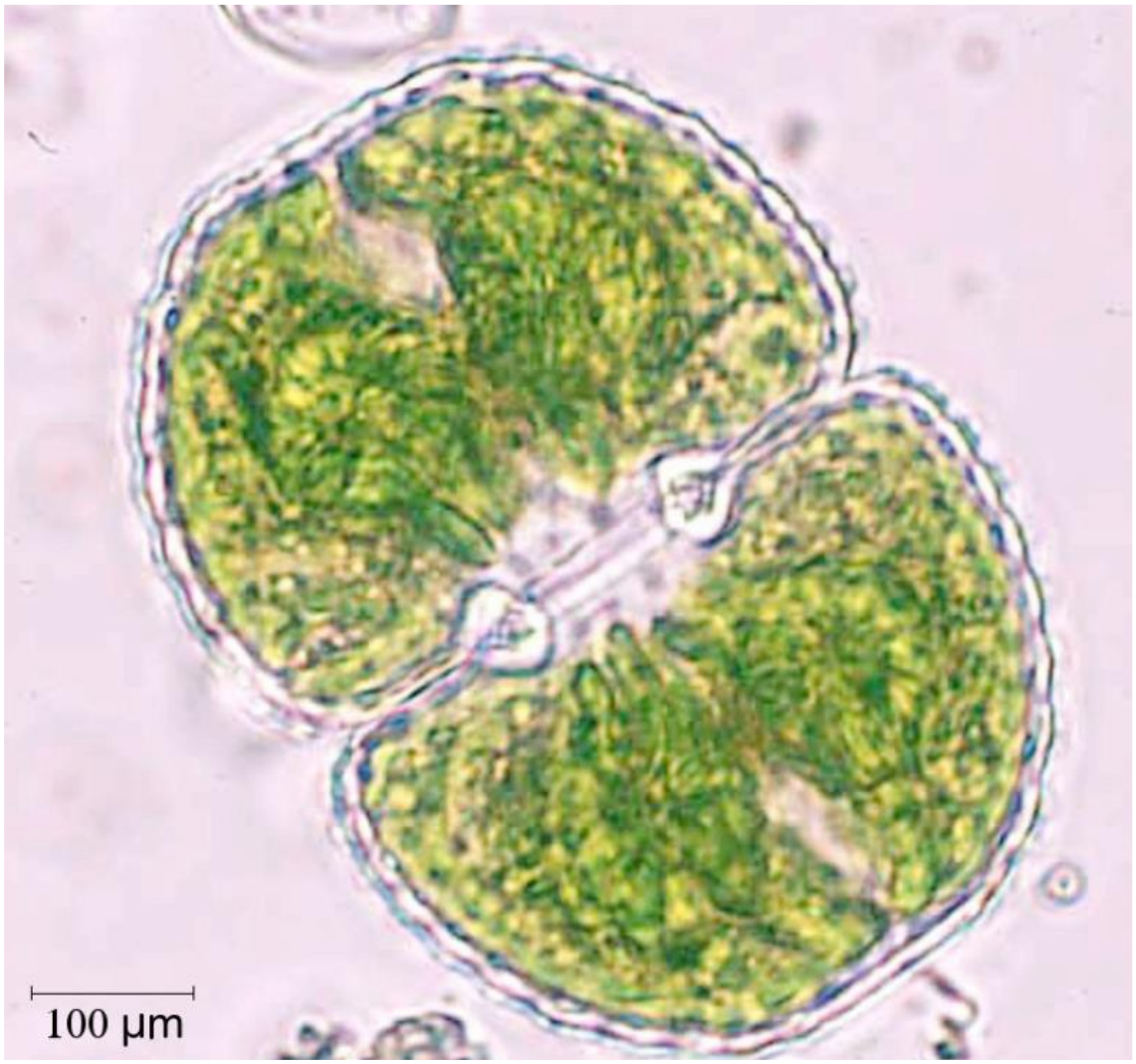
(b)



(c)

(d)

(e)



(f)

(g)

Fig. 3 La Barca reservoir biological variables: (a) Cyanobacteria; (b) Diatoms; (c) Euglenophytes; (d) *Dinophlagella*; (e) Chrysophytes; (f) Clorophytes; and (g) Chryptophytes concentrations

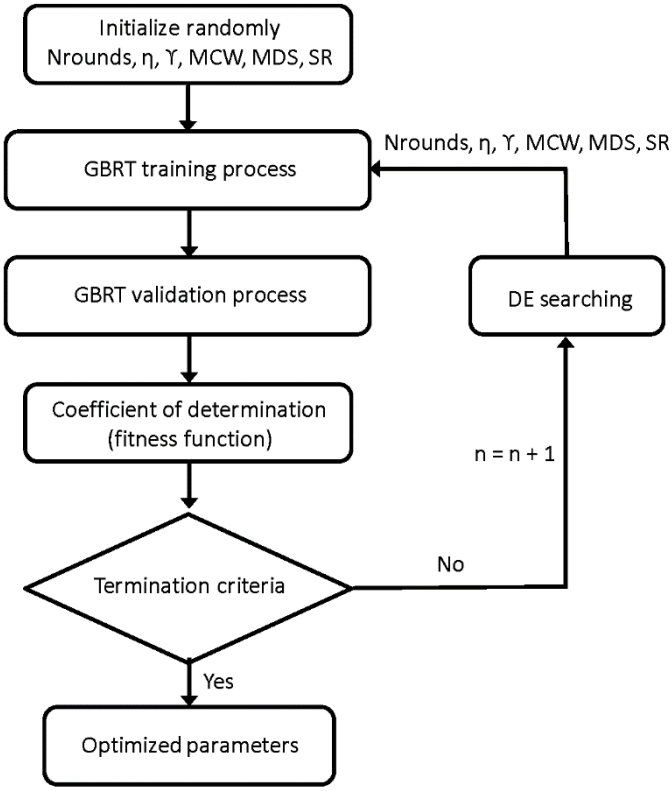
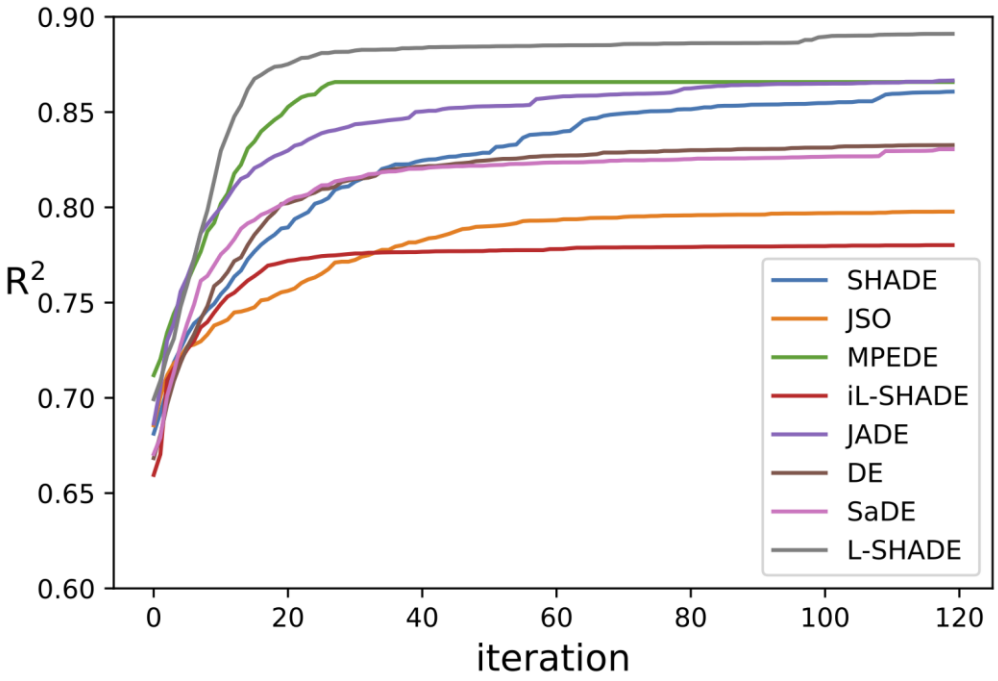
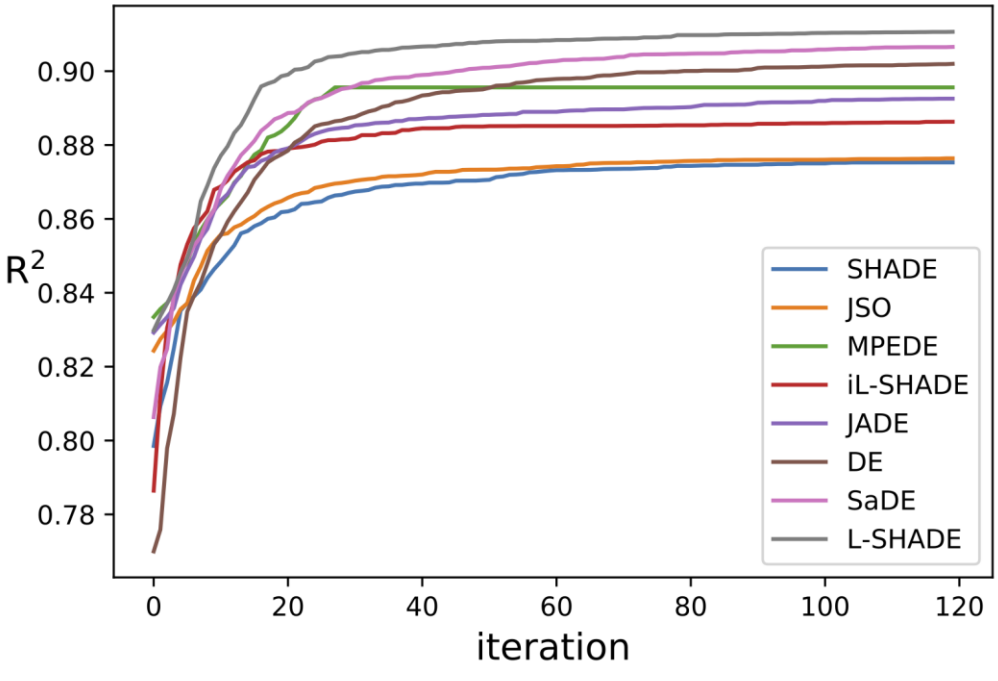


Fig. 4 Flowchart of the new hybrid DE/GBRT–based model. DE stands for a variant of Differential Evolution algorithm.



(a)



(b)

Fig. 5 Convergence curves for the different DE variants for optimizing the parameters of the GBRT model for the prediction of: (a) Chl-a concentration; and (b) Total P concentration

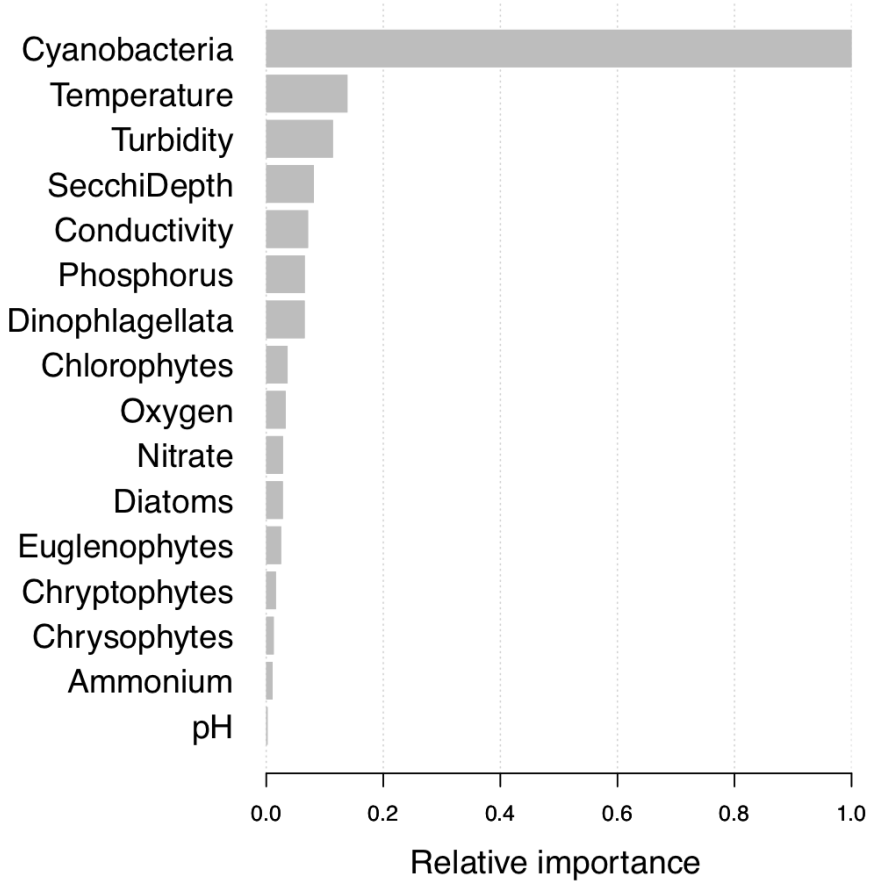


Fig. 6 Relative importance of the input operation variables to predict the Chl-a concentration

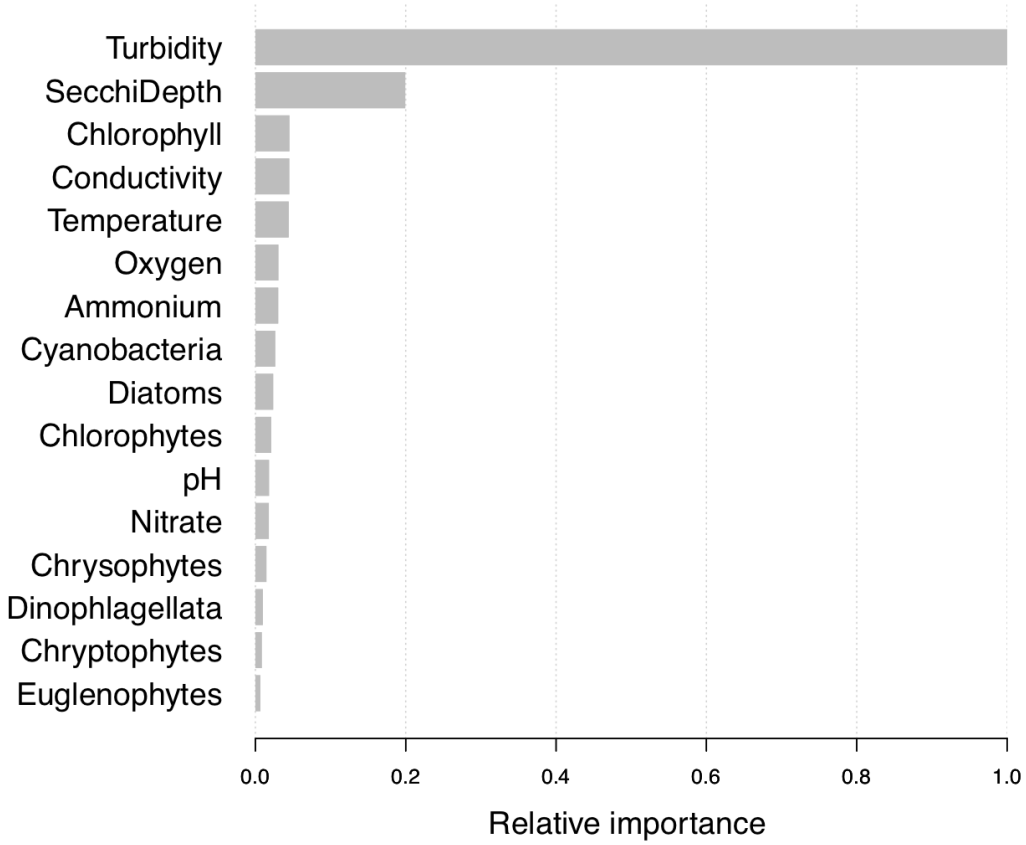


Fig. 7 Relative importance of the input operation variables to predict the Total P concentration

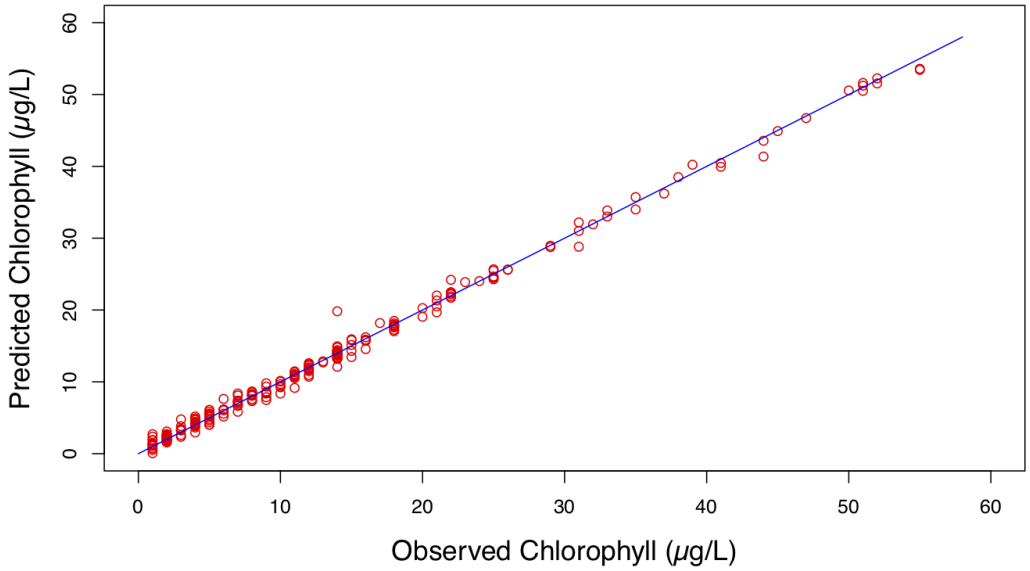


Fig. 8 Comparison between the observed and predicted Chl-a concentration by using the L-SHADE/GBRT–based model ($R^2 = 0.91$)

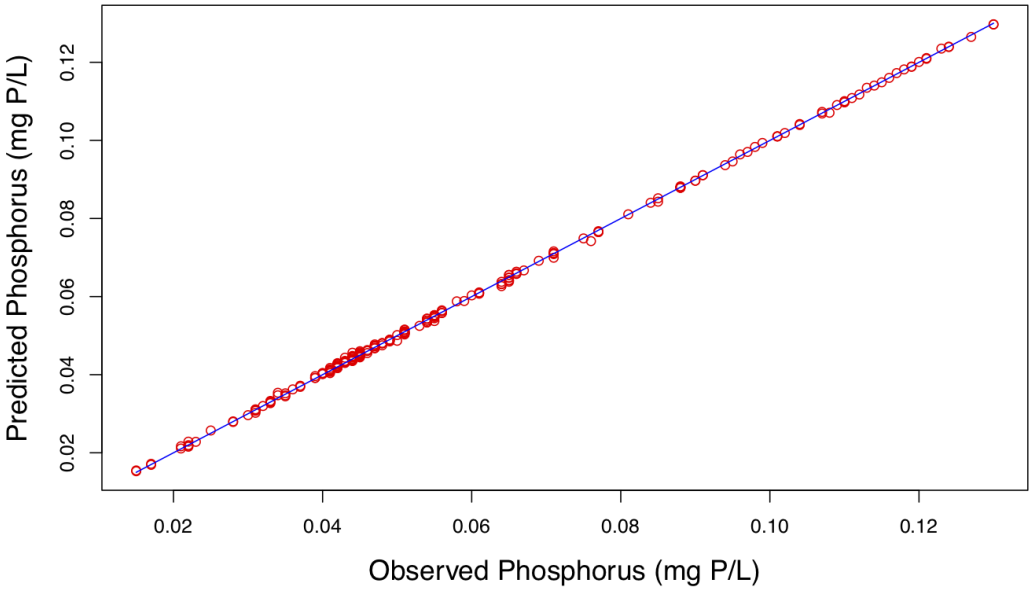


Fig. 9 Comparison between the observed and predicted Total P concentration by using the L-SHADE/GBRT–based model ($R^2 = 0.93$)

Table 1 Biological independent variables in this work with their mean and standard deviation

Biological input variables	Short name of the variable	Mean	Std
Cyanobacteria (mm^3/L)	*Cyano*	4.0917	5.5127
Diatoms (mm^3/L)	Diat	1.0989	0.5840
Euglenophytes (mm^3/L)	Eugl	0.5352	0.2264
Dinophlagellata (mm^3/L)	Dino	0.1388	0.1761
Chrysophytes (mm^3/L)	Chrys	0.2585	0.1768
Chlorophytes (mm^3/L)	Chloro	0.1201	0.0909
Chryptophytes (mm^3/L)	Chryp	0.9851	0.3687

Table 2 Mean and standard deviation of the physico-chemical input values

Physical-chemical input variables	Short name of the variable	Mean	Std
Water temperature (ºC)	WTemp	17.0572	4.1033
Turbidity (NTU)	Tur	5.6555	4.8250
Nitrate concentration (mg NO^{3-} /L)	Total N	0.8324	0.4072
Ammonium concentration (mg/L)	Amm	0.1182	0.0588
Dissolved oxygen concentration (mg O_2/L)	DO	9.0196	1.7851
Conductivity (μS/cm)	Con	268.222	42.9435
pH values	pH	7.7785	0.4058
Secchi depth (m)	SeDe	2.0184	0.9621

Table 3 Search space for each of the GBRT parameters in the tuning process

GBRT hyperparameters	Lower limit	Upper limit
Rounds	1	100
η	0.1	1
γ	0	30
Minimum child weight (MCW)	1	30
Maximum Δ step (MDS)	0	30
Subsample ratio	0.5	1

Table 4 Mean, median, minimum, maximum and p-value of the Shapiro-Wilk test the 20 coefficients of determination (R^2) obtained for each of the different hybrid models for the Chl-a using different parameter tuning optimization algorithms. The algorithms are ranked from worst to best. The best model is indicated in bold font

Algorithm	Mean	Median	Max	Min	p-value
iL-SHADE	0.7855	0.7246	0.8970	0.7184	0.0001
JSO	0.7986	0.7589	0.9030	0.7185	0.0001
DE	0.8327	0.8884	0.8977	0.7185	0.0002
SaDE	0.8305	0.8904	0.9026	0.7206	0.0001
SHADE	0.8607	0.8914	0.9022	0.7246	0.0001
MPEDE	0.8658	0.8772	0.8858	0.7246	<0.0001
JADE	0.8666	0.8973	0.9045	0.7206	<0.0001
L-SHADE	0.8928	0.8972	**0.9051**	0.8149	<0.0001

Table 5 Mean, median, minimum, maximum and p-value of the Shapiro-Wilk test the 20 coefficients of determination (R^2) obtained for each of the different hybrid models for the Phosphorus concentration using different parameter tuning optimization algorithms. The algorithms are ranked from worst to best. The best model is indicated in bold font

Algorithm	Mean	Median	Max	Min	p-value
SHADE	0.8753	0.8568	0.9192	0.8566	<0.0001
JSO	0.8769	0.8569	0.9204	0.8566	<0.0001
MPEDE	0.8956	0.9037	0.9076	0.8551	<0.0001
iL-SHADE	0.8878	0.8868	0.9257	0.8527	<0.0001
JADE	0.8926	0.9166	0.9193	0.8567	0.0001
DE	0.9020	0.9160	0.9181	0.8566	<0.0001
SaDE	0.9066	0.9180	0.9196	0.8568	<0.0001
L-SHADE	0.9117	0.9200	**0.9253**	0.8569	<0.0001

Table 6 *p*-values of the Mann-Whitney rank test comparing the 20 coefficients of determination (R^2) obtained for each of the different hybrid models for the Chl-a using different parameter tuning optimization algorithms. The algorithms are ranked from worst to best

	iL-SHADE	JSO	DE	SaDE	SHADE	MPEDE	JADE	L-SHADE
iL-SHADE		0.2930	0.0297	0.0064	0.0005	0.0264	0.0001	<0.0001
JSO	0.2930		0.1444	0.0483	0.0054	0.1820	0.0010	0.0001
DE	0.0297	0.1444		0.2235	0.0416	0.2163	0.0029	<0.0001
SaDE	0.0064	0.0483	0.2235		0.2005	0.2322	0.0578	0.0099
SHADE	0.0005	0.0054	0.0416	0.2005		0.0257	0.0970	0.0096
MPEDE	0.0264	0.1820	0.2163	0.2322	0.0257		0.0007	<0.0001
JADE	0.0001	0.0010	0.0029	0.0578	0.0970	0.0007		0.3882
L-SHADE	<0.0001	0.0001	<0.0001	0.0099	0.0096	<0.0001	0.3882	

Table 7 *p*-values of the Mann-Whitney rank test comparing the 20 coefficients of determination (R^2) obtained for each of the different hybrid models for the Phosphorus concentration using different parameter tuning optimization algorithms. The algorithms are ranked from worst to best

	SHADE	JSO	MPEDE	iL-SHADE	JADE	DE	SaDE	L-SHADE
SHADE		0.3555	0.1538	0.0376	0.0146	0.0132	0.0002	<0.0001
JSO	0.3555		0.1962	0.0411	0.0081	0.0131	0.0002	<0.0001
MPEDE	0.1538	0.1962		0.2361	0.1042	0.0010	0.0002	<0.0001
iL-SHADE	0.0376	0.0411	0.2361		0.4460	0.4728	0.0890	0.0020
JADE	0.0146	0.0081	0.1042	0.4460		0.2894	0.0162	0.0002
DE	0.0132	0.0131	0.0010	0.4728	0.2894		0.0045	0.0002
SaDE	0.0002	0.0002	0.0002	0.0890	0.0162	0.0045		0.0029
L-SHADE	<0.0001	<0.0001	<0.0001	0.0020	0.0002	0.0002	0.0029	

Table 8 Coefficients of determination (R^2) and correlation coefficients (*r*) for the hybrid L-SHADE/GBRT–based models fitted in this study for the Chl-a and Total P concentrations

Model	R^2 /*r* Chl-a	R^2 /*r* Total P
L-SHADE/GBRT	0.9051/ 0.9514	0.9253/ 0.9619

Table 9 Optimal hyperparameters of the best-fitted GBRT model for Chl-a concentration found with the L-SHADE technique

GBRT hyperparameters	Optimal values
Rounds	61
η	0.22
γ	2.0076
Minimum child weight (*MCW*)	1.36
Maximum Δ step (*MDS*)	0.0000
Subsample ratio (*SR*)	0.0.89

Table 10 Optimal hyperparameters of the best-fitted GBRT model for Phosphorous concentration found with the L-SHADE technique

GBRT hyperparameters	Optimal values
Rounds	99
η	0.11
γ	0.0000
Minimum child weight (*MCW*)	1.09
Maximum Δ step (*MDS*)	18
Subsample ratio (*SR*)	0.73

Table 11 Relative importance of the variables in the optimal L-SHADE/GBRT model for the Chl-a concentration estimation using the criteria Gain, Cover and Frequency

Feature	Gain	Cover	Frequency
Cyanobacteria	0.577332733	0.18174868	0.10872894
Temperature	0.080102681	0.06286236	0.08269525
Turbidity	0.065804007	0.08038312	0.06125574
SecchiDepth	0.046994179	0.04969524	0.05206738
Conductivity	0.041272094	0.05181897	0.06125574
Phosphorus	0.038038795	0.04355767	0.05053599
Dinophlagellata	0.037926841	0.09089558	0.10107198
Chlorophytes	0.021094743	0.05646994	0.07810107
Oxygen	0.019202011	0.05472848	0.06738132
Nitrate	0.016524269	0.01399537	0.02297090
Diatoms	0.016413543	0.09238219	0.09035222
Euglenophytes	0.014697541	0.07696392	0.07503828
Chryptophytes	0.009566865	0.05232867	0.06125574
Chrysophytes	0.007419592	0.04755028	0.04900459
Ammonium	0.006253431	0.02176822	0.02450230
pH	0.001356676	0.02285132	0.01378254

Table 12 Relative importance of the variables in the optimal L-SHADE/GBRT model for the Phosphorous concentration estimation using the criteria Gain, Cover and Frequency

Feature	Gain	Cover	Frequency
Turbidity	0.648768410	0.11153217	0.06766293
SecchiDepth	0.129311756	0.10307938	0.07430469
Chlorophyll	0.029348554	0.05939307	0.08966376
Conductivity	0.029214099	0.05898668	0.06600249
Temperature	0.028743796	0.08398947	0.06849315
Oxygen	0.019821412	0.04175599	0.04275633
Ammonium	0.019680094	0.04229445	0.02532171
Cyanobacteria	0.017097099	0.10495890	0.09506019
Diatoms	0.015334297	0.08518831	0.08551266
Chlorophytes	0.013728903	0.05138729	0.06185139
pH	0.011759012	0.02297087	0.01784973
Nitrate	0.011573934	0.03387213	0.02698215
Chrysophytes	0.009507991	0.05115362	0.06641760
Dinophlagellata	0.006361521	0.05939307	0.07555002
Chryptophytes	0.005490079	0.04963984	0.06434205
Euglenophytes	0.004259044	0.04040476	0.07222914