Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-021-02136-4

ORIGINAL PAPER

Modeling eutrophication risks in Tanes reservoir by using a hybrid WOA optimized SVR-relied technique along with feature selection based on the MARS approximation

Paulino José García-Nieto[1] · E. García-Gonzalo[1] · José Ramón Alonso Fernández[2] · Cristina Díaz Muñiz[2]

Accepted: 7 November 2021

Abstract

Total phosphorus (from now on mentioned as TP) and chlorophyll-a (from now on mentioned as Chl-a) are recognized indicators for phytoplankton large quantity and biomass-thus, actual estimates of the eutrophic state-of water bodies (i.e., reservoirs, lakes and seas). A robust nonparametric method, called support vector regression (SVR) approach, for forecasting the output Chl-a and TP concentrations coming from 268 samples obtained in Tanes reservoir is described in this investigation. Previously, we have carried out a selection of the main features (biological and physico-chemical predictors) employing the multivariate adaptive regression splines approximation to construct reduced models for the purpose of making them easier to interpret for researchers/readers and to reduce the overfitting. As an optimizer, the heuristic technique termed as whale optimization iterative algorithm (WOA), was employed here to optimize the regression parameters with success. Two main results have been obtained. Firstly, the relative relevance of the models variables was stablished. Secondly, the Chl-a and TP can be successfully foretold employing this hybrid WOA/SVR-based approximation. The coincidence between the predicted approximation and the observed data obviously demonstrates the quality of this novel technique.

Keywords Eutrophication in reservoirs · Support vector machines (SVMs) · Whale optimization algorithm (WOA) · Multivariate adaptive regression splines (MARS) · Regression analysis

1 Introduction

In ecology, a eutrophic crisis of an aquatic ecosystem can be described as accelerated aging as a consequence of water nutrient enrichment caused by anthropogenic activities. The most widespread use refers to the contribution of inorganic nutrients containing Nitrogen and Phosphorus in an aquatic ecosystem, such as a reservoir or lake (Arauzo and Álvarez Cobelas 1994; Ansari et al. 2010). Eutrophication is a kind of water contamination, giving place to modifications such as the presence of colored waters, absence of see-through quality and poisoning by specific algae releases (Reynolds 2006; Van der Valk 2006; Howell 2017). At present, agriculture is taken into account to be a principal underlying and lasting cause of eutrophication in many basins around the world.

Water pollution, such as the eutrophication of lakes, has gradually become an urgent environmental issue worldwide (Liu et al. 2011; Lin et al. 2021). According to statistics from the Water Research Commission, South Africa, 53%, 54%, 46%, and 28% of lakes in Europe, Asia, North America, and Africa face eutrophication problems, respectively (Lin et al. 2021). The eutrophication of lakes occurs under the combined effect of natural factors and human activities, and leads to the growth of great amount of algae. The causes of eutrophication are complex, as eutrophication involves various ecological, social, economic, and other factors (Álvarez et al. 2017). The natural evolution of reservoirs and lakes from oligotrophic to

✉ Paulino José García-Nieto
pjgarcia@uniovi.es

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

[2] Cantabrian Basin Authority, Spanish Ministry for the Ecological Transition and Demographic Challenge, 33071 Oviedo, Spain

Published online: 19 November 2021

eutrophic is slow, but has been accelerated under human intervention. Therefore, adopting a scientific method for assessing water quality and identifying potential risk sources is urgent, which will help to strengthen ecological and environmental management (Reynolds 2006; García-Nieto et al. 2019; Lin et al. 2021).

The development of biomass in an ecosystem is limited by the scarcity of nitrogen or phosphorus, which primary producers need to develop. Urban effluents, or diffuse pollution from agrarian or atmospheric sources, can contribute increasing the concentration of these limiting substances. The results are important consequences on the composition, structure and dynamics of the ecosystem.

With eutrophication diversity decreases and biomass increases (Harper 1991; Reynolds 2006; Howell 2017). When cyanobacteria begin to be dominant, the potable and recreational use of lake and reservoir waters may be threatened.

The quality of the water decreases with the proliferation of algae and when it becomes explosive, it can cause the appearance of toxins, particularly when cyanobacteria are predominant; large chlorophyll-a contents, usually mean harmful algal blooms (HAB) including toxins—(Pip and Bowman 2014; Yuan et al. 2014). The cyanotoxins liberated by some cyanobacteria in water pose a menace to recreational and drinking (Watzin et al. 2006; Kalaji et al. 2016). In this sense, the knowledge of the concentration of Chl-a can be seen as an optional indicator to evaluate the possibility of blooms of harmful cyanobacteria (HABs) (Huisman et al. 2010; McQuaid et al. 2011; Shumway et al. 2018) in water bodies. Thus, it is important to predict the amount of chlorophyll-a when assessing the water quality so the pollution due to this problem and its health dangers are avoided (Wheeler et al. 2012; Kinkaid 2014; Shumway et al. 2018). Nevertheless, chlorophyll-a is still far from being adequately predicted in lakes and reservoirs (Di Toro et al. 1971; Brown et al. 2000; Tufford and McKeller 1999).

The overabundance of algae that characterizes eutrophication causes the water to become cloudy, preventing sunlight from penetrating the bottom of the ecological community. As a consequence, photosynthesis becomes impossible, while the oxygen-consuming metabolic activity of decomposers increases. In this sense, at the bottom of the ecosystem oxygen is exhausted quickly and becomes anoxic. The alteration caused by these variations makes unviable the life of most of the species of the ecological community.

The eutrophication process can end up turning a reservoir or lake into dry land. This occurs because the nutrients generate large biomass of organisms that are not totally consumed by degrading organisms. Natural eutrophication processes can be in ancient channels of the rivers that are transformed into swamps and later are covered with vegetation.

Modelling is regarded as an interesting tool (Barnes and Chu, 2010; Vinçon-Leite and Casenave 2019) since predictions and understanding of the different process stages can be made. Chl-a and phosphorus are the main indicators to assess continental water trophic state (Beiras 2018).

Chl-a is the principal compound involved in photosynthesis, and it is commonly employed as an indicator for algae growth (Reynolds 2006; Van der Valk 2006). Checking the amount of chlorophyll, using its optical characteristics, helps to control the eutrophication processes that can arise in reservoirs.

Laplacian mathematical models of water quality relied on inner physico-chemical actions in reservoirs and lakes demand a very big quantity of information that is not reachable from the practical point of view, either because it cannot be acquired in its completeness, or because it is complicated to implement or take up too much time to evaluate (Gul et al. 2020). For this reason, there are more and more research works that make use of machine learning techniques to model water quality. For example, Shamshirband et al. (2019) have developed ensemble models using the Bates–Granger approach and least square method to combine forecasts of multi-wavelet artificial neural network (ANN) models for multi-day ahead forecasting of chlorophyll a concentration in coastal waters; Tiyasha and Yaseen (2020) reports the state of the art of various artificial intelligent models implemented for river water quality simulation over the past two decades (2000–2020); Hadjisolomou et al. (2021) have used artificial neural networks (ANNs) to model freshwater eutrophication with limited limnological data; Deng et al. (2021) have implemented two different machine learning methods (artificial neural networks (ANN) and support vector machines (SVM)) to accurately forecast algal growth and eutrophication in Tolo Harbour in Hong Kong carrying out a comparative analysis with 30-year measured data.

In this investigation, a novel approximation obtained performing a regression that relies on support vector regression (SVR) in combination with whale optimization algorithm (WOA) (Mirjalili and Lewis 2016) has been used to successfully predict the phosphorus and chlorophyll concentration in a reservoir located in Principality of Asturias (an autonomous community placed in Northern Spain) known as Tanes reservoir (see Figs. 1a, b). Algal atypical rapid reproduction is a grave health matter in lakes and reservoirs like the Tanes one that provides water to the main city of the region, Oviedo, that is, it is the supply to one million inhabitants. To prevent toxic algae blooms Chl-a and TP concentrations are used as an early alarm.

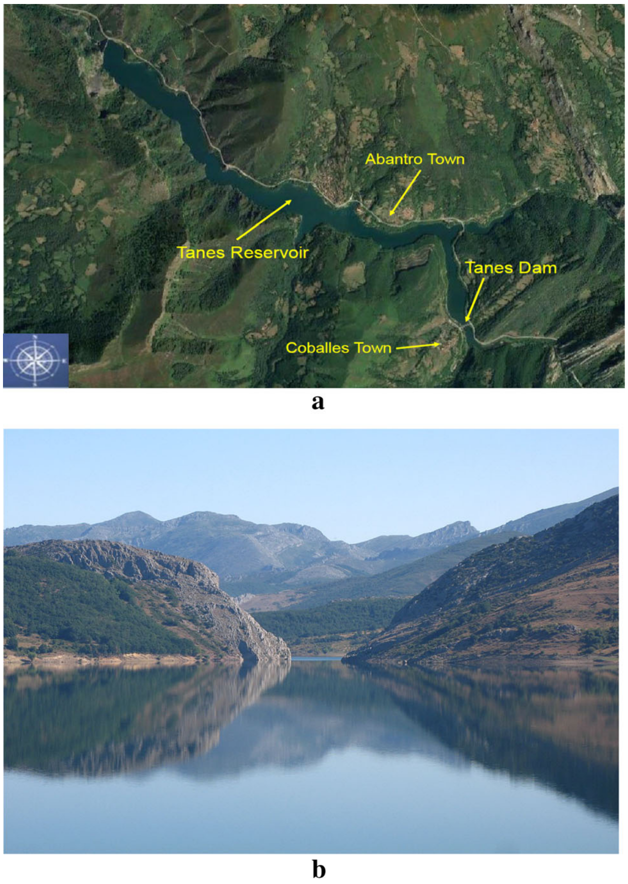

Fig. 1 Images from Tanes reservoir: **a** overview of the study area showing its location; and **b** a close-up image of this water body

This new methodology, which mixes the SVM approach (Cristianini and Shawe-Taylor 2000; Hansen and Wang 2005; Steinwart and Christmann 2008) with the whale optimization algorithm (WOA) (Mirjalili and Lewis 2016; Gharehchopogh and Gholizadeh 2019; Ebrahimgol et al. 2020), to foretell the Chl-a and TP concentrations could be an interesting procedure that has not been used so far. Moreover, the SVM method is a mathematical technique relied on the statistical machine learning which has the capacity to tackle nonlinearities as well as interactions among variables (Schölkopf et al. 2000; Steinwart and Christmann 2008; Bishop 2011). The SVM technique provides some advantages with respect to the classical regression methods (Li et al. 2008; Barnes and Chu 2010; Kuhn and Johnson 2013): (1) The SVM eludes mathematical hydraulic models of the reservoir; (2) In SVMs, the knowledge of the physico-chemical involved in the pollutants transport in the reservoir is not required; (3) SVMs allow to deal with the nonlinear relationships among the input variables of the water body; and (4) by training and testing SVR (support vector regression) enables to find nonlinear relationships between data showing an obvious significance. Certainly, the WOA optimizer has been employed well enough to determine the optimal SVM hyperparameters in this investigation. Moreover, earlier researches point out that SVM is a suitable instrument in a big number of existing applications as the foretold modelling for solar thermal energy systems (Waseem Ahmad et al. 2018), air and water quality estimation (García-Nieto et al. 2013; Xu et al. 2019), weighted multiscale SVR combined with ultraviolet–visible spectra for quantitative analysis of edible blend oil (Wu et al. 2021), prediction of the short-term electricity load employing SVR in conjunction with grey catastrophe and random forest techniques (Fan et al. 2021), analysis of SVR kernels for energy storage efficiency prediction (Ighravwe and Mashao 2020), SVR for prediction the number of Dengue infections in the capital of Indonesia (Tanawi et al. 2021), etc. Nevertheless, SVR remains a new method for assessing Chlorophyll-a and TP concentrations using biological and physico-chemical variables and thus to evaluate the quality of the water in lakes and reservoirs.

The foremost aim of the present investigation was to prognosticate the dependent Chl-a and TP concentrations using the input physico-chemical and biological variables in Tanes reservoir-obtained sampling the reservoir periodically (Directive 2000/60/EC; Spatharis and Tsirtsis 2021)—employing SVR in conjunction with WOA. This methodology describes a novel technique to study closely Chl-a and TP in water bodies (i.e., lakes and reservoirs), obtaining measurements of the Chl-a and TP concentrations in them (Smith 2006; Riegl et al. 2014). Certainly, the Chl-a concentration can be taken into account as an essential indicator of surplus nutrients such as TP concentration in a lake or reservoir, and basically, of the presence of eutrophication in these water bodies.

This research work is organized as follows: To start with, the variables and data for the investigation along with mathematical principles are detailed. Next we present the results and discoveries acquired with this new approximation by comparing the observed values with SVR results and next determining the relevance order for the parameters of the model including the discussion. Finally, conclusions derived from this work are explained in detail.

2 Materials and methods

2.1 Study area

The Tanes reservoir is located in the south of the Principality of Asturias (an autonomous community located in the north of Spain), specifically within the Redes Natural Park (this park is considered a biosphere reserve) and in the Nalón river valley. Furthermore, Tanes reservoir supplies drinking water to almost the entire urban nucleus of the

Principality of Asturias, including its capital, the city of Oviedo. Therefore, its importance is paramount from several points of view (Kerich 2020; Çadraku 2021). The Tanes reservoir has the following hydrological characteristics: (a) volume: 33.27 hm^3; (b) area: 159 ha; and (c) 95 m of depth in the deepest point. Three kilometers downstream from the Tanes reservoir is the Rioseco reservoir. The Rioseco reservoir has the following hydrological characteristics: (a) capacity of 4.3 hm^3; (b) surface area of 63 ha; and (c) maximum depth of 28.5 m. Rioseco and Tanes reservoirs provide water supply to approximately one million inhabitants. In addition, the Tanes reservoir has hydroelectric and recreational uses as well as ornithological interest.

The geological area of the Tanes reservoir is the Central Carboniferous Basin, whose lithologies are mainly quartzites and limestones. Therefore, the nature of the most common materials is basic.

2.2 Experimental dataset

The data collection in the WOA/SVR-based study were picked up during ten years. The 268 samples were obtained monthly, starting in January 2006 and ending in December 2015 (Smith et al. 2008; World Health Organization 1998). The data was obtained employing a Niskin bottle at the point in the reservoir where the depth is maximum (see Fig. 2a). The samples were taken at equally spaced depth intervals determined in relation to Secchi depth (the design is not observable on a Secchi disk as a result of the water turbidity) (see Fig. 2b). In this way, five subsamples were picked up (Brönmark and Hansson 2005; Quesada et al. 2006) and next homogenized to acquire a unique sample.

The physical–chemical parameters were analyzed by an ISO17025 accredited laboratory, following the corresponding methods in the Standard Methods for the Examination of Water and Wastewater (American Public Health Association 2005; Negro et al. 2000; Van der Valk 2006). A quality assessment program including internal laboratory control (use of standards, blanks and replicates during analysis) as well as analysis of blanks, replicates and blind samples collected in the Tanes reservoir was applied. During the sampling procedure, field blanks were also collected. A total of 10% of samples were replicated to assess variability. Some of these variables were obtained directly, like the temperature or the pH of the water, while obtaining others required a certified laboratory.

2.3 Variables of the model

The ultimate purpose of this investigation was to predict the Chl-a and TP concentrations (μg/L). Chlorophyll-a is broadly used as an indicator of the biomass of algae and therefore, as an index of eutrophication (Latif et al. 2003; Karydis 2009). Very high concentrations of chlorophyll generally point out the presence of algal blooms (United States Environmental Protection Agency 2014) and hence a diminishing in the water quality. Chlorophyll is the substance directly related with the photosynthesis. Thus, phytoplankton and Chlorophyll are mutually linked (American Public Health Association 2005). Besides, as cyanobacteria are predominant in phytoplankton for eutrophic ecosystems, Chlorophyll concentration is considered as a subrogate indicator of them and a measure of the potential risk to public health since cyanobacteria can generate cyanotoxins (Wheeler et al. 2012). The constructed model (SVR) employs the data from different kinds of phytoplankton (biological variables) together with physico-chemical variables as independent parameters, supplied by the Cantabrian Basin Authority, agency primarily responsible for the management of the hydrographic basins in the Cantabrian Sea.

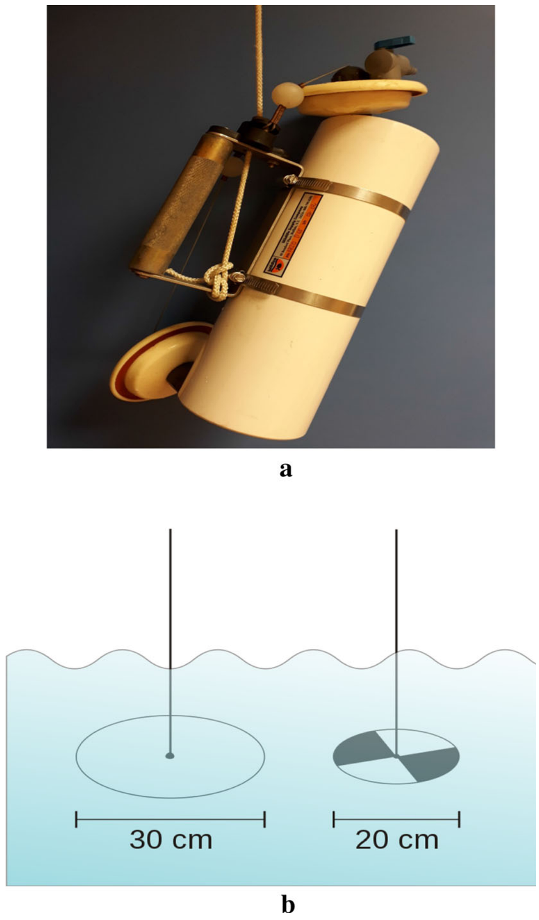

Fig. 2 **a** Niskin device; **b** Secchi disks

Independent variables:

- Biological parameters:

- Cyanobacteria concentration (mm^3/L): they have photosynthetic capacity (see Fig. 3a), advertised in environments. They should not be present is freshwater (Quesada et al. 2006; Texeira and Rosa 2006; Willame et al. 2005).
- Diatoms concentration (mm^3/L): another common organism in phytoplankton community and the major group of algae concentration (Fig. 3b).
- Euglenophytes concentration (mm^3/L): phytoplankton with photosynthetic capacity contains it (Fig. 3c).
- *Dinophlagellata* concentration (mm^3/L): it is also a type of phytoplankton (Fig. 3d).

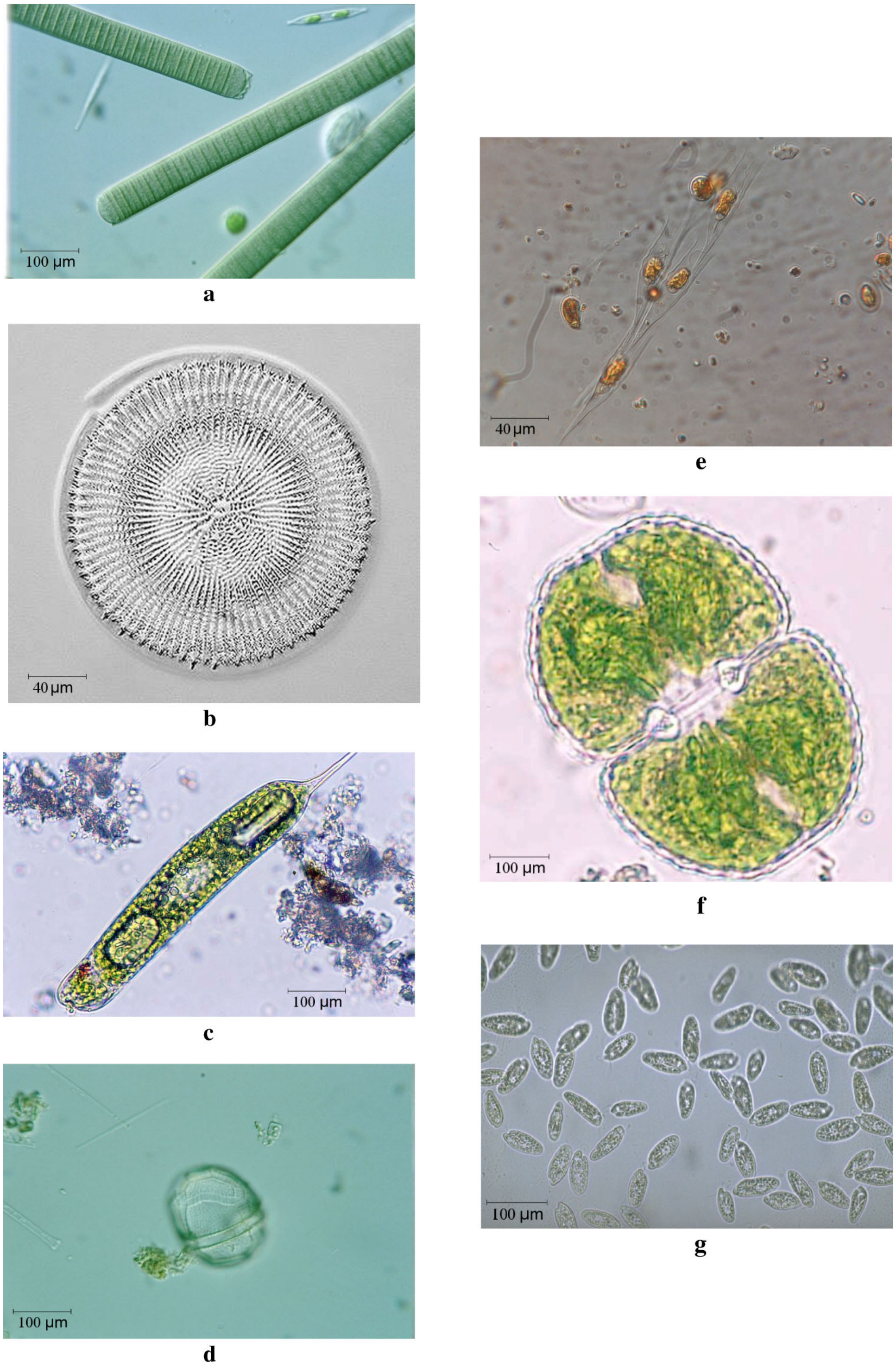

Fig. 3 Microscopic organisms used in this study: **a** Cyanobacteria; **b** Diatoms; **c** Euglenophytes; **d** *Dinophlagella*; **e** Chrysophytes; **f** Clorophytes; and **g** Chryptophytes

- Chrysophytes concentration (mm^3/L): they are fundamentally photosynthetic (Fig. 3e).
- Chlorophytes concentration (mm^3/L): they are one of the most numerous groups of algae (Fig. 3f).
- Chryptophytes concentration (mm^3/L): diminutive type of phytoplankton (Fig. 3g).

Physico-chemical variables (all of them referred to water column):

- Water temperature (°C): this parameter refers to the average value of measures at different depths on water column. Many chemical and biological processes that influence, among others, plant development, are sensitive to temperature.
- Orthophosphates concentration (mg PO_4^{3-}/L): this represents the phosphorus fraction assimilated by plants and, consequently, it is also related to plant development and eutrophication (it is an essential nutrient for autotrophic organisms such as phytoplankton and other chlorophyll organisms).
- Total phosphorus concentration (mg P/L): it includes all phosphorus compounds so it comprises forms not assimilated by plants but that can be converted into them when environmental conditions turn into the appropriate. Phosphorus is known as a limiting macronutrient for plant development.
- Nitrite concentration (mg NO^{2-}/L): it is one of the nitrogen compounds, an intermediate in the oxidizing process from ammonia to nitrate. It causes several effects such as, for example, methemoglobinemia in many species.
- Nitrate concentration (mg $NO_3{}^{-}$/L): nitrate is the most frequent nitrogen ion in water-and this is the nitrogen form usually taken up by plants. Nitrate is also one of the nutrients responsible for waters eutrophication.
- Ammonium concentration (mg NH_4^{+}/L): ammonium is the reduced form of nitrogen. It is a toxic excreted by aquatic organisms. In water, a high amount of this substance blocks that excretion causing damages even the death of the organism.
- Dissolved oxygen concentration (mg O_2/L): it refers to the density of oxygen dissolved in water, a key factor for many chemical and biological processes and tightly related to algae whose blooms release a lot of oxygen at daytime and gives place to its depletion at night or when bacteria decompose the dead algae.
- Iron concentration (mg Fe/L): it is a micronutrient for phytoplankton growth. Despite being a life essential element in high concentrations can be toxic. Its precipitates can cause unbalance in waterbodies.
- Manganese concentration (mg Mn/L): like iron, manganese is an essential trace element for phytoplankton. However, ferromanganese depositions constrain algal colonization and growth (Sheldon and Skelly 1990).
- Conductivity (μS/cm): it is an indirect measure of salts content in watertight related to phytoplankton composition and abundance (Redden and Rukminasari 2008).
- Volume of water (hm^3): it is an important parameter since nutrients availability for phytoplankton depends on it. High values imply the dilution of toxic substances but also the dilution of nutrients diminishing toxicity and trophic degree, respectively, and therefore, improving waterbody state.
- pH: it is an expression of acid substance concentration in water. High values indicate low acid concentration (high basic substance concentration), typically of eutrophic waterbodies.
- Secchi depth (m) or depth at which a Secchi disk immersed in water is no longer visible, measures turbidity that is mainly generated by phytoplankton so it is an indicator of its abundance.

2.4 Feature selection using multivariate adaptive regression splines (MARS)

The feature or variable selection process involves the picking of only some of the most important characteristics (known as variables or predictors) for their employ in modeling construction. Feature selection techniques are employed for five compelling reasons:

- This permits the simplification of the models for the purpose of making them easier to understand for users and researchers;
- It reduces training time: less data indicates that algorithms can be trained quicker;
- The precision is improved: less erroneous data indicates that modeling precision ameliorates.
- The curse of dimensionality can be eluded; and
- It reduces overfitting (formally, variance reduction): less duplicate data means less chance to come to decisions relied on noise.

One of the most efficient and robust techniques in the selection of variables is multivariate adaptive regression splines (MARS). The MARS tool is used here to handle complex data and to select the important features from the entire experimental dataset. Some advantages of applying the MARS method over other existing techniques in-clude (Friedman and Roosen 1995; Hastie et al. 2003): (1) it provides more flexible models than linear regression models; (2) it is easy to interpret and understand; (3) it can manage continuous and categorical data; (4) the hinge functions automatically partition the input data, so the effect of outliers is contained; (5) it does automatic variable selection (meaning that it includes important variables in the model and excludes unimportant ones); (6) it tends to have a good bias-variance trade-off; and (7) it offers an explicit mathematical expression of the dependent variable as a function of independent variables through an expansion of the base functions (hinge functions and products of several hinge functions or interactions).

Next, we will briefly describe the mathematical basics and benefits of the MARS technique. Multivariate adaptive regression splines (MARS) is a statistical approach (Friedman 1991; Sekulic and Kowalski 1992) and it is a generalization of recursive partitioning regression (RPR) that can consider complex relationships between a set of k predictor independent variables, which are denoted by $X_1, X_2, \ldots, X_k$ and a dependent variable designated by Y, and does not make starting assumptions about any type of relationship between input and output parameters. The MARS model is defined as (Friedman and Roosen 1995; Hastie et al. 2003; García-Nieto et al. 2019):

$$\hat{y} = f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) + \varepsilon \tag{1}$$

where X is a function of the independent variables and their interactions, β_0 is the intercept parameter, β is a vector of coefficients of the basis function, M is the total number of these basis functions, $h(X)$ the spline basis function in the model and ε is the fitting error.

To approximate the nonlinear relationships between the input variables X and the response parameter Y, basis functions (BF) are used. They consist of a unique spline function or the product of two or more spline functions for distinct predictors. Spline functions are piecewise linear functions, that is, truncated left-hand and right-hand functions, and take the form of hinge functions that are joined at the knots (Chou et al. 2004; Zhang et al. 2015; García-Nieto et al. 2019):

$$(x-t)_+ = \left\{ \begin{matrix} x-t & \text{if} & x > t \\ 0 & \text{if} & x \leq t \end{matrix} \right\} \tag{2}$$

$$(t-x)_+ = \left\{ \begin{matrix} x-t & \text{if} & x > t \\ 0 & \text{if} & x \leq t \end{matrix} \right\} \tag{3}$$

where t is a constant called a node that specifies the boundary between the regions that have continuity from the base functions of the regions from left to right and that are smoothly joined at the given node and adaptively selected from the data. The "+" sign refers to the positive part and sets a value equal to zero for negative values of the argument. For example, Fig. 4 indicates a pair of splines for q = 1 at the node t = 3.5.

As stated in some references (Xu et al. 2004; Cheng and Cao 2014; Zhang et al. 2015), MARS forms reflected pairs for each predictor variable with knots of each value x_j, $j \in \{1, \ldots, k\}$ with knots at each observed value x_{ij}, $i \in \{1, \ldots, n\}$ of that variable, where n is the sample size. The set of all possible pairs with the corresponding knots and the truncated linear basis functions can be expressed by the set

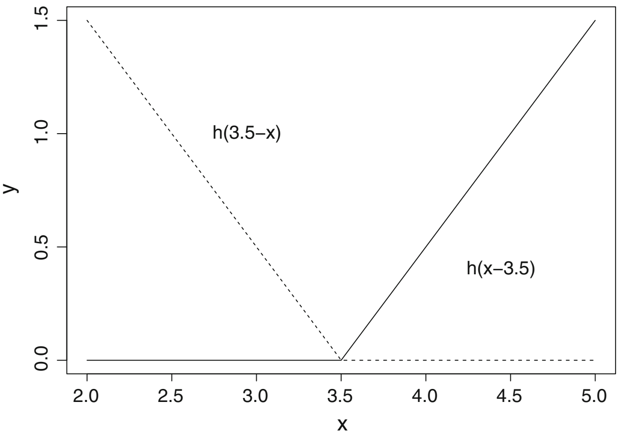

Fig. 4 A graphical representation of a spline basis function. The left spline $(x < t, -(x-t))$ is shown as a dashed line and the right spline $(x > t, +(x-t))$ as a solid line

$$D = \left\{ (x_j - t)_+, (t - x_j)_+ \middle| t \in \{x_{1j}, x_{2j}, \ldots, x_{nj}\}, j \in \{1, \ldots, k\} \right\}.$$

An adaptive regression algorithm is taken during a recursive partition strategy to automatically select the locations of the node or breakpoints, including the two-stage process: forward-stepwise regression selection and backward-stepwise elimination procedure (Cheng and Cao 2014; Zhang et al. 2015).

The first step, also called the construction phase, begins with the intersection and then, in order, adds to the approximation the predictor that further improves the fit; that is, when the maximum diminishing in the sum-of-squares residual error happens. The search for the best combination of variable and node is done iteratively. Keeping in mind a model that contains M hinge functions, the following couple will be added to the approximation in the form of

$$\beta_{M+1} h_m(X) \max(0, X_j - t) + \beta_{M+2} h_m(X) \max(0, t - X_j).$$

Given a choice for the h_m, the coefficients β_m that make up the vector β are estimated by minimizing the residual sum-of-squares, that is, by standard linear regression (Friedman and Roosen 1995; Hastie et al. 2003; Chou et al. 2004).

This procedure persists until a predetermined number of base functions ($M_{\max}$) is accomplished or the R^2 alters less than a limit (Friedman and Roosen 1995; Chou et al. 2004). A big number of BFs are summed one after another so that an overfitting model is built. Generally, the maximum number of BF is 2–4 times the number of predictor variables (Cheng and Cao 2014; Zhang et al. 2015).

The second step, also called the pruning phase, begins with the full model and simplifies it by eliminating terms by applying a backward procedure to avoid oversizing. MARS identifies the hinge functions that are less relevant for the model and removes the least significant terms sequentially. The final model is chosen employing the generalized cross-validation method (*GCV*), an adjustment of the sum-of-squares of the residuals that penalizes the complexity of the models by the number of hinge functions and the number of knots (Hastie et al. 2003; Cheng and Cao 2014; Zhang et al. 2015) and it is given by:

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f_M}(x_i)\right)^2}{\left(1 - C(M)/n\right)^2} \qquad (4)$$

where M is the number of terms in Eq. (1) (equal to the number of BFs), n is the number of data sets, y_i are the observed values and $\hat{f_M}(x_i)$ denotes the predicted values from MARS and, finally the value $C(M)$ increases with the number of basis functions used in the model. The formula for this value is (Sekulic and Kowalski 1992; Friedman and Roosen 1995; Hastie et al. 2003):

$$C(M) = (M + 1) + d\,M \qquad (5)$$

so that d is a coefficient that establishes the relevance of this parameter.

The relevance of the variables used as predictors can be assessed in different ways (Xu et al. 2004; Cheng and Cao 2014; Zhang et al. 2015): (a) using the *GCV* related to the variables: when we drop a variable from the model the *GVC* index increases. This increment is the associated value; (b) similarly, using the residual sums of squares (*RSS*); and (c) we can count the number of subsets (Nsubsets) that contains a specific variable. The more subsets, the greater its relevance.

2.5 Support vector regression (SVR)

Support vector machines (SVM) have originally emerged to address binary classification problems. In view of the situation, it was rapidly noted that the fundamental guidelines that endorse them could be employed to confront another type of issues such as the regression problems (Vapnik 1998; Pal and Goel 2007; Chen et al. 2013). In this sense, let's consider a dataset, the training set comprises the values of the output dependent variable $y_i \in \Re,\ \forall i = 1, 2, \ldots, m$ and the covariates $\mathbf{x_i} \in \Re^p,\ i = 1, 2, \ldots, m$. Thus, the method termed *support vector regression* (SVR) builds a function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b,$ so that $\mathbf{w}$ represents the perpendicular vector to the hyperplane, called *director vector* of the hyperplane, and $b/\|\mathbf{w}\|$ is the perpendicular distance from the coordinate's origin to the hyperplane. Furthermore, this approximation gives places to not more than a deviation equal to ε from y_i for all training cases $\mathbf{x}_i$, and simultaneously, it must be as flat as possible. Flatness is achieved by minimizing the Euclidean norm $\|\mathbf{w}\|_2$, while the model is fitted by penalizing the sum of deviations greater than ε. Indeed, the SVR method intends to solve the next optimization problem (Steinwart and Christmann 2008; Cristianini and Shawe–Taylor 2000; Gu et al. 2006):

$$\min_{\mathbf{w},b,\xi^+,\xi^-} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\left(\xi_i^+ + \xi_i^-\right) \qquad (6)$$

subject to

$$\left\{\begin{array}{ll} y_i - (\mathbf{w}^T\mathbf{x}_i + b) \geq \varepsilon + \xi_i^+ & i = 1, \ldots, m \\ (\mathbf{w}^T\mathbf{x}_i + b) - y_i \geq \varepsilon + \xi_i^- & i = 1, \ldots, m \\ \xi_i^+, \xi_i^- \geq 0 & i = 1, \ldots, m \end{array}\right\} \qquad (7)$$

so that C is termed the regularization constant and $\boldsymbol{\xi}^+, \boldsymbol{\xi}^- \in \Re^m$ are called slack variables. The constant C in Eq. (6) takes a positive numeric value that restrains the penalty enforced on observations that are outside the interval ε and facilitates avoiding the overfitting. This value ascertains the trade-off between the horizontality of the objective function and the complexity reduction of the model. The slack variables are presented for each training vector for the purpose of permitting deviations greater than ε, but penalizing these digressions in the objective function. The zone enclosed by $y_i \pm \varepsilon,\ \ \forall i$ is termed an ε-insensitive tube (see Fig. 5).

To tackle highly nonlinear problems like this one, we will use the kernelization method. This method relies on mapping the original dataset to a larger dimensional space H termed the *feature space*. The application is carried out via a kernel function $K\left(\mathbf{x}_i, \mathbf{x}_j\right)$, which determines a scalar product in H. In order to solve the primal optimization problem given by Eqs. (6) and (7), we are going to express this problem in its dual form. The dual formulation of the optimization problem is obtained applying the Karush–

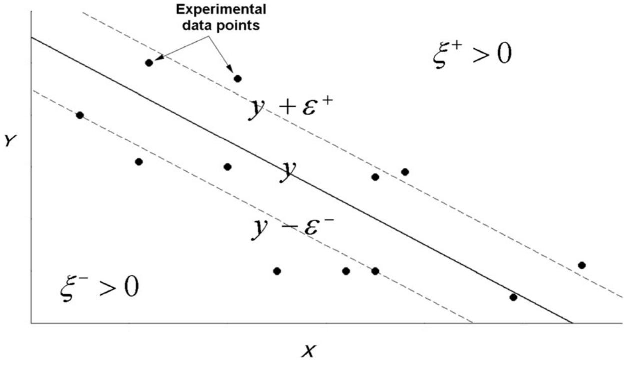

Fig. 5 Representation of the ε-insensitive tube in case of regression

Kuhn–Tucker (KKT) conditions (Li et al. 2008; Gu et al. 2006; Shawe-Taylor and Cristianini 2004):

$$\begin{aligned}\max_{\alpha^+,\alpha^-} & \sum_{i=1}^{m} y_i\left(\alpha_i^+ - \alpha_i^-\right) - \varepsilon \sum_{i=1}^{m}\left(\alpha_i^+ + \alpha_i^-\right) \\ & -\frac{1}{2}\sum_{i,j=1}^{m}\left(\alpha_i^+ - \alpha_i^-\right)\left(\alpha_j^+ - \alpha_j^-\right)K\left(\mathbf{x}_i,\mathbf{x}_j\right)\end{aligned} \tag{8}$$

subject to

$$\left\{\begin{array}{ll}\sum_{i=1}^{m}\left(\alpha_i^+ - \alpha_i^-\right) = 0, & \\ 0 \le \alpha_i^+ \le C, & i = 1,\ldots,m \\ 0 \le \alpha_i^- \le C, & i = 1,\ldots,m\end{array}\right\} \tag{9}$$

The regression estimation for a new sample $\mathbf{x}$ can be obtained with the function $f(\mathbf{x})$ given by (Steinwart and Christmann 2008; García-Nieto et al. 2013; Gu et al. 2006):

$$f(\mathbf{x}) = \sum_{i=1}^{m}\left(\alpha_i^+ - \alpha_i^-\right)K(\mathbf{x},\mathbf{x}_i) + b \tag{10}$$

Various common functions utilized as kernels in the technical bibliography are expressed as (Shawe-Taylor and Cristianini 2004; Hansen and Wang 2005; Abbaszadeh et al. 2016):

- Radial basis function termed RBF kernel:

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

- Polynomial kernel:

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

- Sigmoid kernel:

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

where a, b and σ are parameters that demonstrate the operation of the kernel.

Furthermore, typical parameters of the SVR approach can be synthesized as (Shawe–Taylor and Cristianini 2004; Steinwart and Christmann 2008; García-Nieto et al. 2013):

- Regularization constant (C): it is also called *cost function*. This constant represents the balance (or trade-off) between the margin and the slack variables. It is one of the hyperparameters of the SVR technique that must be previously determined by tuning.
- ε parameter: This value restrains the width of the allowable margin of error. The second term of the objective function (see Eq. 8) that relied on ε factor is called the *empirical error* and is determined by means of the insensitive loss function, which points out that it does not ignore errors less than ε (that is, at a distance ε of the real value).
- a, b and σ: these parameters define the mathematical expression of the distinct kernels in the final model.

Hence, it is convenient to employ some mathematical technique that determines the above hyperparameters with sufficient precision. Moreover, the whale optimizer algorithm (WOA) explained in more detail below was employed (Mirjalili and Lewis 2016; Gharehchopogh and Gholizadeh 2019) with triumph in this research work. To fix ideas, the whale optimizer algorithm (WOA) expounded in the following subsection was used (Mirjalili and Lewis 2016; Gharehchopogh and Gholizadeh 2019; Ebrahimgol et al. 2020) in this investigation with success.

2.6 Whale optimization algorithm (WOA)

The Whale Optimization Algorithm (WOA) is an optimization algorithtm first suggested by Mirjalili and Lewis (Mirjalili and Lewis 2016). It emulates the clever hunt process performed by humpback whales. The gathering performance is termed the bubble-net feeding methodology, that originate bubbles to surround their victim as they hunt. They plunge into the water at a depth of about 12 m and next originate the spiral of bubbles surrounding their victim. Then they go up pursuing the bubbles. The model that inspires the spiral bubble-net feeding performance is described below (Mirjalili and Lewis 2016; Gharehchopogh and Gholizadeh 2019; Ebrahimgol et al. 2020):

- Surrounding prey

The whales identify the position of the victim surround it. Assuming that the optimum point is not known, WOA supposes that the present best point is the prey and that this is close to the optimum. When the best scout is established, the other scouts will bring up to date their locations to the best scout. This fact is given by the following mathematical expressions:

$$\begin{aligned}\vec{D} &= \left|\vec{C}\cdot\vec{X}_p(t) - \vec{X}(t)\right| \\ \vec{X}(t+1) &= \vec{X}_p(t) - \vec{A}\cdot\vec{D}\end{aligned} \tag{14}$$

so that:

- t: This points out the present iteration;
- $\vec{A}$ and $\vec{C}$: they are called coefficient vectors;
- $\vec{X}_p$: represents the position of the prey; and
- $\vec{X}$: This represents the whale's location.

Moreover, the coefficient vectors $\vec{A}$ and $\vec{C}$ are constructed according to equations:

$$\begin{aligned}\vec{A} &= 2\vec{a}\cdot\vec{r}_1 - \vec{a} \\ \vec{C} &= 2\vec{r}_2\end{aligned} \tag{15}$$

so that components of $\vec{a}$ diminish in a linear form from 2 to 0 with advancing iterations, while $\vec{r}_1, \vec{r}_2$ are random vectors whose components lie in the interval [0,1].

- Exploitation phase: bubble-net attack procedure

The bubble-net technique is a procedure that mixes two mathematical approximations given by Mirjalili and Lewis (2016), Gharehchopogh and Gholizadeh (2019), and Ebrahimgol et al. (2020):

1. Shrinking surrounding mechanism: This procedure is accomplished by diminishing $\vec{a}$. As $\vec{A}$ takes a random value in the interval $[-a, a]$ so a diminishes from 2 to 0 with advancing iterations. Choosing values for $\vec{A}$ in $[-1, 1]$ randomly, the fresh location of a scout can be set in any point between the novel location and the position of the best point found so far.
2. Spiral upgrading location: First, the distance between the whale $\left(\vec{X}, \vec{Y}\right)$ and the prey $\left(\vec{X}^*, \vec{Y}^*\right)$ is calculated. Next, an equation called *spiral* is produced joining the whale and prey position with an helix:

$$\vec{X}(t+1) = \vec{D}' e^{bt} \cos(2\pi t) + \vec{X}^* \tag{16}$$

being:

- $\vec{D}' = \left|\vec{X}^*(t) - \vec{X}(t)\right|$ is the distance between the prey and (present best solution until now) and i-th whale;
- b is a constant that determines the form of the logarithmic spiral; and
- t takes random values that are in the interval $[-1, 1]$.

The spiraling path of the whales around their prey shrinks more and more. To implement this concurrent performance, we suppose that we choose with a 50% probability and the spiral path to update the position of the whales. This is given by the expression (Mirjalili and Lewis 2016; Gharehchopogh and Gholizadeh 2019; Ebrahimgol et al. 2020):

$$\vec{X}(t+1) = \left\{ \begin{array}{ccc} \vec{X}^*(t) - \vec{A} \cdot \vec{D} & \text{if} & p < 0.5 \\ \vec{D}' e^{bt} \cos(2\pi t) + \vec{X}^* & \text{if} & p \geq 0.5 \end{array} \right\} \tag{17}$$

so that p is a number chosen randomly that takes values in the interval $[0, 1]$. Besides the bubble-net approach, the whales look for a victim at random. The model is described below:

- Exploration phase: search for prey

The approximation that relied on the fluctuation of $\vec{A}$ can be employed to look for victims (this stage is termed exploration). Indeed, humpback whales look for at random as stated by their comparative location to each other. As a consequence, we employ $\vec{A}$ random values within the range $(-\infty, -1) \cup (1, \infty)$ to compel the scout to distance itself from a given whale. Unlike the exploitation phase, the location of a scout at this stage is upgraded by means of a chosen search scout at random. This and $\left|\vec{A}\right| > 1$ highlight exploration and allows WOA algorithm to perform an overall exploration. This is expressed (Mirjalili and Lewis 2016; Gharehchopogh and Gholizadeh 2019; Ebrahimgol et al. 2020):

$$\begin{aligned} \vec{D} &= \left|\vec{C} \cdot \vec{X}_{rand} - \vec{X}\right| \\ \vec{X}(t+1) &= \vec{X}_{rand} - \vec{A} \cdot \vec{D} \end{aligned} \tag{18}$$

so that $\vec{X}_{rand}$ gives the random position of the whale (it is called a random whale).

WOA begins with a collection of random possible solutions. Hence, according to this methodology, search agents upgrade their locations taking into account a selected search agent at random or the best solution acquired up until now at each iteration. The parameter a diminishes from 2 to 0 for the purpose of supplying both exploration and exploitation. A search scout at random is selected when $\left|\vec{A}\right| > 1$, but if $\left|\vec{A}\right| < 1$, the best solution is found upgrading the location of the search agents. Finally, WOA ends when a certain stopping criterion is fulfilled.

2.7 Approach accuracy

Twenty independent variables already specified earlier in subsection 2.3 were used in this investigation to create this innovative WOA/SVR-relied method. As is well known, the concentration of Chlorophyll-a is the predicted parameter. For the purpose of foretelling Chl-a from the twenty independent variables with enough assurance, we must choose a good approach to the experimental data. There are some indexes frequently employed to determine the goodness-of-fit in a regression problem, but the norm used in this investigation was the coefficient of determination R^2 (Freedman et al. 2007; Knafl and Ding 2016; McClave and Sincich 2016). To fix ideas, we will term the experimental values t_i and the predicted values y_i. Hence, it is feasible to specify the following additions as follows (Freedman et al. 2007; Knafl and Ding 2016; McClave and Sincich 2016):

- $SS_{tot} = \sum_{i=1}^{n} (t_i - \bar{t})^2$: is called the total sum of squares, and it is directly related to the sample variance.
- $SS_{reg} = \sum_{i=1}^{n} (y_i - \bar{t})^2$: it is called the regression sum of squares, or the explained sum of squares.

- $SS_{err} = \sum_{i=1}^{n} (t_i - y_i)^2$: it is called the residual sum of squares.

so that $\bar{t}$ is the average of the n experimental data:

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i \tag{19}$$

Thus, the coefficient of determination is defined by the mathematical expression:

$$R^2 \equiv 1 - \frac{SS_{err}}{SS_{tot}} \tag{20}$$

The closer the R^2 statistic is to the value 1.0, the smaller the disagreement between the experimental and foretold data. Similarly, the mathematical expressions for the other two statistics used in this study (*RMSE* and *MAE*) are as follows (Freedman et al. 2007; Knafl and Ding 2016):

$$RMSE \equiv \sqrt{\frac{1}{n} \sum_{i=1}^{n} (t_i - y_i)^2} \tag{21}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |t_i - y_i| \tag{22}$$

Higher values of R^2 are preferred, i.e. closer to 1 means better model performance and regression line fits the data well. Conversely, the lower the *RMSE* and *MAE* values are, the better the model performs.

3 Results and discussion

Tables 1 and 2 show the input variables in this study. Seven variables are biological (see Table 1) and the remaining thirteen are physic-chemical (see Table 2). The dataset consists of 268 samples from Tanes reservoir (Directive 2000/60/EC).

In this study, firstly we have performed a choice of the principal characteristics or *feature selection* (input variables or biological and physicochemical predictors) for the two eutrophication indicators (Chl-a and TP) in bodies of water (reservoirs, lakes, etc.) using the MARS technique (ARESLab package) (Jekabsons 2016; Ciaburro 2017). This allowed us to build two simplified models (one model for Chl-a and another model for TP) in order to facilitate their interpretation for researchers and reduce overfitting. The feature selection for Chl-a and TP eutrophication indicators according to the MARS technique are shown in Tables 3 and 4, respectively. Specifically, the reduced model for Chl-a consists of thirteen input variables or predictors while the reduced model for TP consists of eleven input variables or predictors.

Table 1 Ensemble of biological independent parameters employed in this work

Biological input variables	Name of the variable	Mean	Std
Cyanobacteria (mm^3/L)	*Cyanobacteria*	0.0083	0.0074
Diatoms (mm^3/L)	Diatoms	0.5965	0.1397
Euglenophytes (mm^3/L)	Euglenophytes	0.0274	0.0133
Dinophlagellata (mm^3/L)	*Dinophlagellata*	0.1755	0.1583
Chrysophytes (mm^3/L)	Chrysophytes	0.0118	0.0102
Chlorophytes (mm^3/L)	Chlorophytes	0.1153	0.0790
Chryptophytes (mm^3/L)	Chryptophytes	0.2973	0.1279

The dataset was split into a training set (80% of the data) and a testing set (20%). The model is built with the training data using the SVR model. Previously, the metaheuristic WOA has been used to optimize the hyperparameters using a five-fold cross-validation scheme with the training dataset. The flowchart in Fig. 6 illustrates this stage. Once the parameters have been chosen, and the model obtained, this is tested with the testing dataset and predictions for these values are obtained.

As we have previously indicated, the two output variables (dependent variables) in this study are the Chl-a concentration and TP concentration both treated with the WOA/SVR-relied method. A most important issue in the efficiency of this technique is the selection of the optimal hyperparameters noted above: (1) the constant C of regularization; (2) ε the insensitive tube width; and (3) parameter σ, which condition the shape of the RBF (radial basis function) kernel in the ultimate model. The grid search method used by most computational codes is a brute force method, and as such, almost any optimization method improves its efficiency. The grid search is a very simple method that promotes an extensive searching within a predetermined grid. It can be easily improved with smarter searching methods such as the one we have chosen for this paper, WOA optimization, which is more efficient while maintaining the simplicity. Indeed, it has been applied to tune the SVR parameters with success in this study. Table 5 shows the intervals where the three parameters of SVR are searched by WOA for optimal performance of the model.

Following this process, we get the optimal parameters for the RBF-SVR model with the WOA optimizer, which are shown in Table 6.

The value of R^2 was obtained with this model and the testing dataset. The library for support vector machines, termed LIBSVM, was used here to implement the MARS technique (Chang and Lin 2011), in combination with the WOA optimizer (Mirjalili and Lewis 2016).

Taking into account these calculations, the WOA/RBF-SVR-relied method has permitted to build of a novel hybrid model that is able to predict the Chl-a and TP

Table 2 Ensemble of physico-chemical independent parameters employed in this work

Physico-chemical input variables	Name of the variable	Mean	Std
Water temperature (°C)	Water_temp	11.5146	4.9928
Phosphates concentration (mg PO_4^{3-}/L)	Phosphates	0.0209	0.0136
Total phosphorus (mg P/m^3)	Phosphorus	6.4585	2.8533
Nitrite concentration (mg NO^{2-}/L)	Nitrite	0.0019	0.0061
Nitrate concentration (mg NO^{3-}/L)	Nitrate	2.6657	0.9934
Ammonium concentration (mg NH_4^+/L)	Ammonium	0.0500	2.02×10^{-16}
Dissolved oxygen concentration (mg O_2/L)	DO	9.3335	1.1257
Iron concentration (mg Fe/L)	Iron	0.0230	0.0184
Manganese concentration (mg Mn/L)	Manganese	0.0197	0.0113
Conductivity (μS/cm)	Conductivity	183.0037	9.5099
Volume of water (hm^3)	Vol_water	26.8963	4.8194
pH values	pH_values	7.8800	0.4060
Secchi depth (m)	SD	3.6403	0.5318

Table 3 Feature selection for the eutrophication indicator Chl-a according to MARS technique

Variable	*GCV*	Nsubsets	*RSS*
Euglenophytes	16.742	28	69.814
Dinophlagellata	41.701	35	31.427
Cyanobacteria	26.112	19	69.814
Diatoms	55.792	35	0.000
Chlorophytes	100.000	37	33.921
Chryptophytes	38.665	20	100.000
Water temperature	85.260	37	83.278
Secchi depth	52.199	36	64.365
Phosphates	62.741	34	36.476
Total phosphorus	24.417	22	0.000
Manganese	30.005	23	46.761
Conductivity	16.893	28	83.278
Volume	84.615	36	0.000

Table 4 Feature selection for the eutrophication indicator TP according to MARS technique

Variable	*GCV*	Nsubsets	*RSS*
Chlorophyll	10.588	9	9.596
Euglenophytes	69.364	25	48.088
Dinophlagellata	100.000	33	100.000
Cyanobacteria	33.488	29	69.481
Diatoms	42.795	22	40.531
Chlorophytes	4.168	7	6.543
Chryptophytes	91.475	30	76.789
Phosphates	65.208	22	40.531
Nitrate	64.952	28	62.331
Iron	9.940	9	9.596
Volume	95.085	33	100.000

concentrations using the test dataset. Moreover, in Table 7 we can see the different metrics for the evaluation of the performance of this WOA/SVR model with different types of kernels and three additional models adjusted for the Chl-a and TP concentrations.

The R^2 value for the optimal SVR model was 0.8582 for the variable Chl-a predicted with RBF kernel and 0.9750 for the variable TP also predicted with RBF kernel. Similar works obtain worse results (Jimeno-Sáez et al. 2020; Liao et al. 2021).

3.1 Importance of the variables

A significant outcome of the actual research is the relative importance of the parameters used as predictors in the models that predict Chl-a and TP concentrations. Table 8 shows the weights of the thirteen variables in the WOA/SVR-RBF model for the Chl-a forecast. Similarly, Table 9 shows the weights for the eleven variables WOA/SVR-RBF model for the TP prediction. Taken in absolute value, these weights illustrate the relative relevance of the variables in this methodology. The greater the absolute value of the weight, the greater the relative relevance of the variable within the model.

In this sense, Chlorophytes concentration is the weightiest variable in WOA/SVR approach for Chl-a prediction followed, far away, by Cyanobacteria concentration, volume of water, Euglenophytes, Chryptophytes, *Dinophlagellata*, Manganese, Water temperature, Diatoms, Total phosphorus, Secchi depth, Phosphates and Conductivity (Table 8 and Fig. 7).

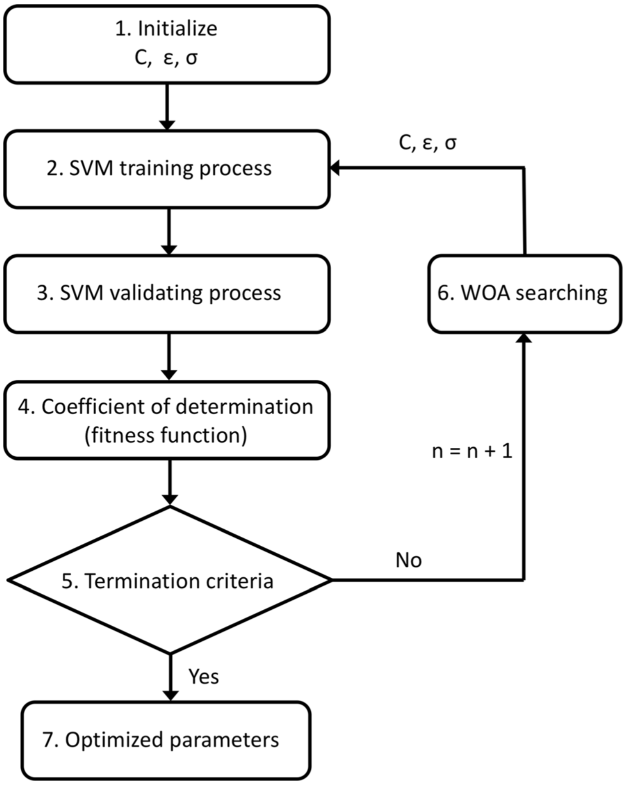

Fig. 6 Flowchart of the construction of the WOA/SVR model

Table 5 Search intervals for the RBF–SVR hyperparameters fitted by using the WOA optimizer during the tuning process

SVR hyperparameters	Lower limit	Upper limit
C	10^{-2}	10^{2}
ε	10^{-6}	10^{0}
σ	10^{-2}	10^{2}

Cyanobacteria (0.4799), Nitrate (0.3303) and Phosphates with only a thirtieth of the Iron weight (0.0581 versus 1.4933 for Iron).

As Fig. 7 shows, Chlorophytes influence in Chl-a is nearly three times all the others. Consequently, in Tanes reservoir Chl-a concentration can be predicted from Chlorophytes concentration with remarkably precision since this water body is mesotrophic. One of the reasons why chlorophytes outcompete cyanobacteria at high nutrient levels may be the balance between the rates of cellular growth and losses (Reynolds 2006; Ansari et al. 2010). Chlorophytes have a high demand for nutrients as reflected in their high growth rates.

Cyanobacteria is the second most important input variable in the prediction of Chl-a. Indeed, cyanobacteria include bacteria capable of oxygenic photosynthesis. They are the only prokaryotes that carry out this type of photosynthesis, which is why they are also called oxyphotobacteria. Cyanobacteria have also been known by the names of blue-green or chloroxybacteria algae, due both to the presence of chlorophyll pigments that give it that characteristic tone, and to its similarity with the morphology and functioning of algae.

Among the non-biological variables, Water Volume is the most important one in Chl-a forecast and the third in the general ranking. The relationship between Water Volume and phytoplankton growth, or Chl-a concentration, was pointed out by other authors (Brasil et al. 2016; Costa et al. 2016) who conclude that in a deep reservoir the reduction in its Water Volume favours phytoplankton growth and consequently, Chl-a concentration.

Euglenophytes, Chryptophytes and Dinophlagellata concentrations are less relevant than other kinds of phytoplankton in Chl-a concentration forecasting according to their fourth, fifth and sixth position, respectively, in the ranking and their weight, nearly four times lower than Chlorophytes concentration weight and also lower than Cyanobacteria concentration weight. The concentration of Euglenophytes is the fourth most significant variable in the prediction of Chl-a (output variable) because dammed waters are usually rich in Euglenophytes. The concentration of Chryptophytes is the fifth most important variable in

Table 6 Optimal parameters for the RBF–SVR–relied models for the training set encountered with the WOA optimizer for the Chl-a and TP eutrophication indicators

RBF Kernel	Values of optimal hyperparameters
Chlorophyll (Chl-a)	Regularization factor $C = 1.3327 \times 10^{0}$, $\varepsilon = 1.4990 \times 10^{-5}$, $\sigma = 1.3499 \times 10^{0}$
Total phosphorus (TP)	Regularization factor $C = 7.4222 \times 10^{0}$, $\varepsilon = 1.4990 \times 10^{-5}$, $\sigma = 7.8960 \times 10^{-1}$

However, in the TP forecasting using WOA/SVR approach Iron concentration is the weightiest variable. The second one, Chryptophytes concentration, has similar significance. The remaining input variables have much less weight (see Table 9 and Fig. 8): Chlorophytes (0.8964), Chl-a (0.85939), *Dinophlagellata* (0.7661), Volume of water (0.7285), Diatoms (0.6987), Euglenophytes (0.6869),

Table 7 Coefficient of determination (R^2), correlation coefficient (r), root-mean-square error (*RMSE*) and mean absolute error (*MAE*) for the novel WOA/SVR-relied model with different types of kernels and linear regression (*LR*), multilayer perceptron (*MLP*) and random forest (*RF*) models for the Chl-a and TP concentrations for the test dataset

Model	Chlorophyll (Chl-a)				Total phosphorus (TP)			
	RMSE	MAE	r	R^2	RMSE	MAE	r	R^2
SVM-RBF	1.2288	0.7610	0.9278	0.8582	0.3163	0.1587	0.9877	0.9750
SVM-Linear	2.6015	2.3410	0.6511	0.3645	1.8449	1.5291	0.3993	0.1503
SVM-Polyn	1.4695	1.1059	0.8981	0.7972	0.3493	0.2801	0.9855	0.9695
SVM-Sigmoid	2.6631	2.3853	0.6275	0.3341	1.8682	1.5094	0.3609	0.1287
LR	2.7865	2.4318	0.5389	0.2710	1.8566	1.5102	0.4031	0.1395
MLP	2.1056	1.5021	0.8223	0.5837	1.7510	1.2717	0.7249	0.2267
RF	2.0672	1.7955	0.8525	0.5988	1.2574	1.0337	0.8336	0.6053

Table 8 Relevance of the variables used as predictors in the WOA/SVR-RBF model for the Chl-a forecast

Input variable	Weight
Chlorophytes	– 2.3263
Cyanobacteria	– 0.8200
Volume	0.6386
Euglenophytes	0.6370
Chryptophytes	0.6262
Dinophlagellata	0.5778
Manganese	– 0.4954
Water temperature	0.4894
Diatoms	0.3717
Total phosphorus	– 0.2453
Secchi depth	0.0268
Phosphates	– 0.0231

Table 9 Relevance of the variables used as predictors in the WOA/SVR–RBF model for the TP forecast

Input variable	Weight
Iron	1.4933
Chryptophytes	1.4570
Chlorophytes	– 0.8964
Chlorophyll	0.8593
Dinophlagellata	0.7661
Volume	– 0.7285
Diatoms	0.6987
Euglenophytes	0.6869
Cyanobacteria	0.4799
Nitrate	– 0.3303
Phosphates	0.0581

Chl-a concentration foretelling. They are important members of phytoplankton and can be found in stagnant waters, withstanding moderate levels of contamination. They are especially abundant in cold waters such as high mountain reservoirs and lakes. In general, cryptophytes are mixotrophic, that is, capable of both photosynthesis and

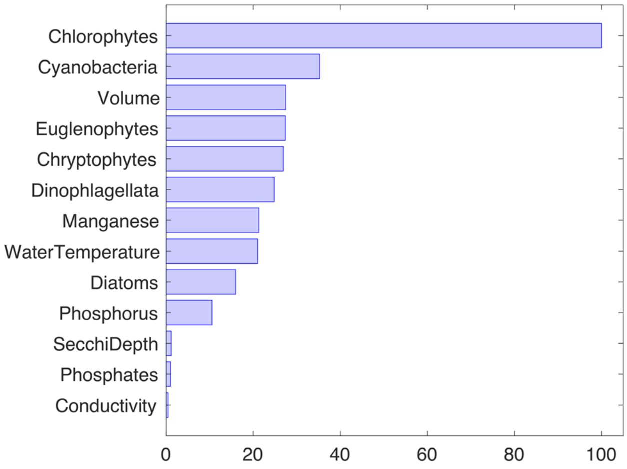

Fig. 7 Relevance ranking of the variables used as predictors for the WOA/SVR-relied approach to forecast the Chl-a concentration

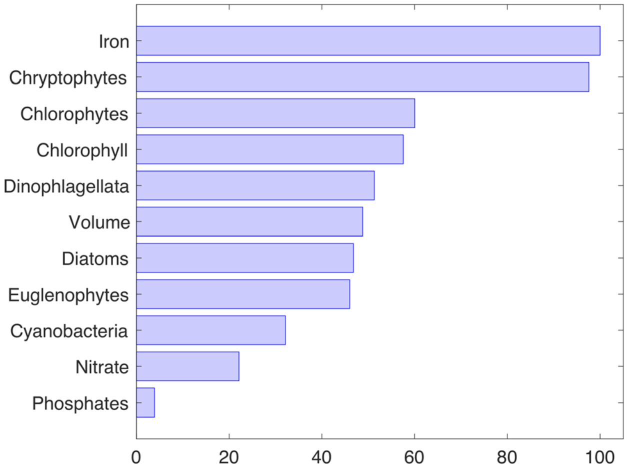

Fig. 8 Relevance ranking of the variables used as predictors for the WOA/SVR-relied approach to forecast the TP concentration

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

The less important phytoplanktonic predictor of Chl-a is Diatoms concentration, the ninth input variable in relevance, despite dominating the phytoplankton groups in Tanes reservoir. They are important producers within the food chain (Reynolds 2006; Van der Valk 2006).

Manganese concentration is an essential micronutrient in phytoplankton growth and appears in position seven in the ranking of the Chl-a ranking. It is well known for quite some time that manganese concentration concerns phytoplankton structure (Patrick et al. 1969).

Water temperature is the most relevant physico chemical variable and it is the eighth one in importance for Chl-a concentration prediction. Increasing temperature favor phytoplankton growth. In fact, Climate Change is a matter of concern in water eutrophication (Moss et al. 2011; Havens 2019).

Much less important is Total phosphorus, the tenth input variable in significance to predict Chl-a density. Total phosphorus is correlated to Chl-a because it is a nutrient for phytoplankton. The other essential nutrient to grow phytoplankton and, as a consequence, to increase Chl-a, is Nitrogen. However, this element is not relevant in Chl-a forecast perhaps because it is not a limiting factor for phytoplankton since it can take the Nitrogen up from the atmosphere (Fields 2004; Moura Ado et al. 2012).

The other input variables (Secchi depth, Phosphates concentration and Conductivity) hardly are influential on Chl-a prediction.

The higher the Secchi depth value the lower the turbidity value and therefore more light availability and more phytoplankton biomass (Costa et al. 2016) or, in other words, more Chl-a.

Phosphates are the highly biological available form of phosphorus. They are the soluble fraction of total phosphorus, the part easily used by phytoplankton to grow. They foster the biomass increase of phytoplankton reducing water quality and unbalancing the ecosystem leading to some species disappearance. However, they barely have influence as predictor of Chl-a (weight 0.0231 and twelfth position in the ranking).

Conductivity seems to have an irrelevant contribution predicting Chl-a density in the studied reservoir (the thirteenth input variable in importance and it has even less weight than phosphates).

In the case of TP content forecasting, as Fig. 8 shows, Iron and Chryptophytes concentrations are the most important input variables. An expected result considering the strong affinity Iron has for phosphorus (Koopmans et al. 2020) and the significance of Chryptophytes in phytoplankton. They can be found in stagnant waters, supporting moderate levels of contamination, which correspond to contributions of nutrients rich in phosphorus, which can be the reason for the importance of this input variable in the TP foretelling (Abirhire et al. 2015).

The rest of the input variables are much less relevant in PT prediction-about 50% of the iron or Chryptophytes weights, or less, depending on the variable considered. Most of them, except Volumen (the sixth relevant variable in the foretelling), are phytoplankton species or related to them as Chl-a concentration.

An explanation of the phytoplankton relevance in total phosphorus forecast could be that this latter is a nutrient that stimulates the growth for all these species since the phosphorus is the limiting nutrient in plant growth, particularly in lakes and reservoirs as is the case at hand. In fact, most Chlorophytes, one of the most numerous groups of algae and the most relevant phytoplanktonic predictor excluding Chryptophytes, have a wide distribution and many are cosmopolitan, hence their presence in waters contaminated with organic material, linked to contributions rich in phosphorus (Arauzo and Álvarez Cobelas 1994; Reynolds 2006). Furthermore, Dinophlagellata, the fifth significant variable—third of phytoplankton species—to predict PT density, proliferates as the amount of phosphorus increases in the water body causing the decrease in phytoplankton diversity and productivity.

Among all the phytoplankton species, cyanobacteria are the less relevant predictor of PT content according to the fact that this kind of organism only dominates environments with high trophic degree (Quesada et al. 2004), but Tanes reservoir has a fair trophic state. Cyanobacteria are a group of photosynthetic bacteria, some of which fix nitrogen, living in a wide variety of moist soils and water freely or in a symbiotic relationship with lichen-forming plants or fungi.

Chl-a concentration is related to the sum of all kinds of phytoplankton species measured since all of them are photosynthetic. Chl-a concentration is the fourth important variable for TP content.

Volume of water is another relevant variable (the sixth one according to the proposed model) to predict TP concentration; an expected result since concentration depends on the water volume.

After phytoplankton species, nitrate concentration is the next input variable ¬the tenth one in the rank for predicting TP content, with an influence nearly 70% of cyanobacteria influence in the prediction—it is another nutrient and it is usually together with phosphorus in waste water inputs. This predictor is also connected with the Chl-a concentration since it is a necessary nutrient for the growth of the phytoplankton, which contains chlorophyll. Excessive nitrate concentrations in reservoirs and lakes can cause accelerated eutrophication and loss of dissolved oxygen.

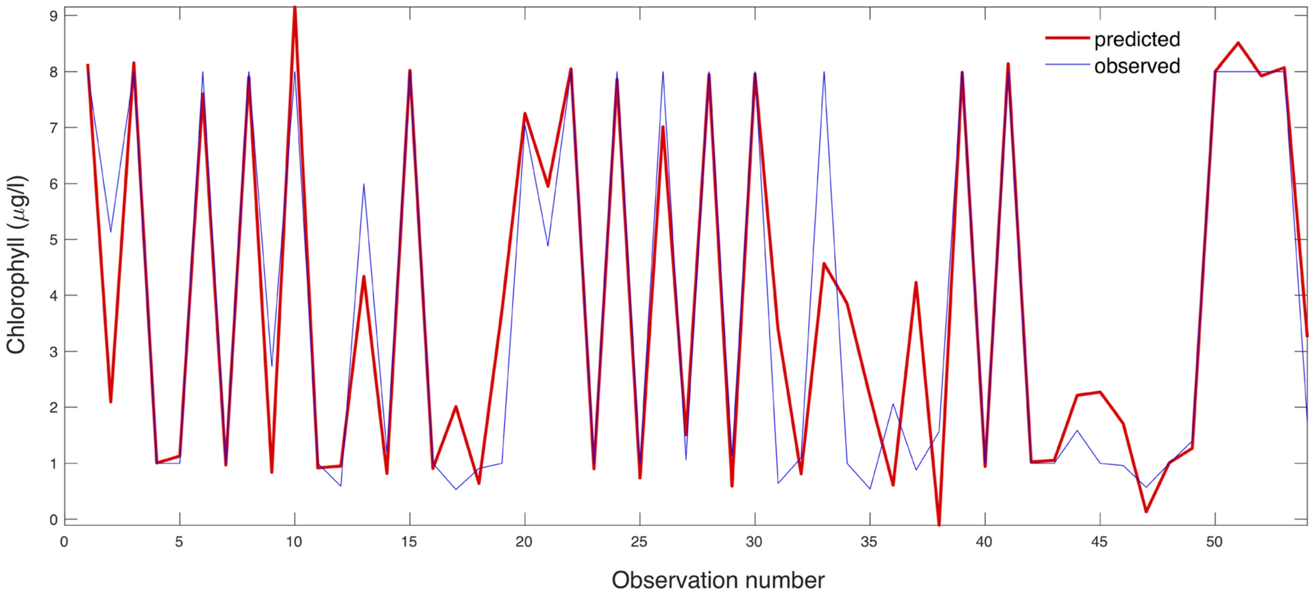

Fig. 9 Observed versus predicted Chl-a concentrations employing the WOA/SVR-RBF for model the testing dataset ($R^2 = 0.8582$)

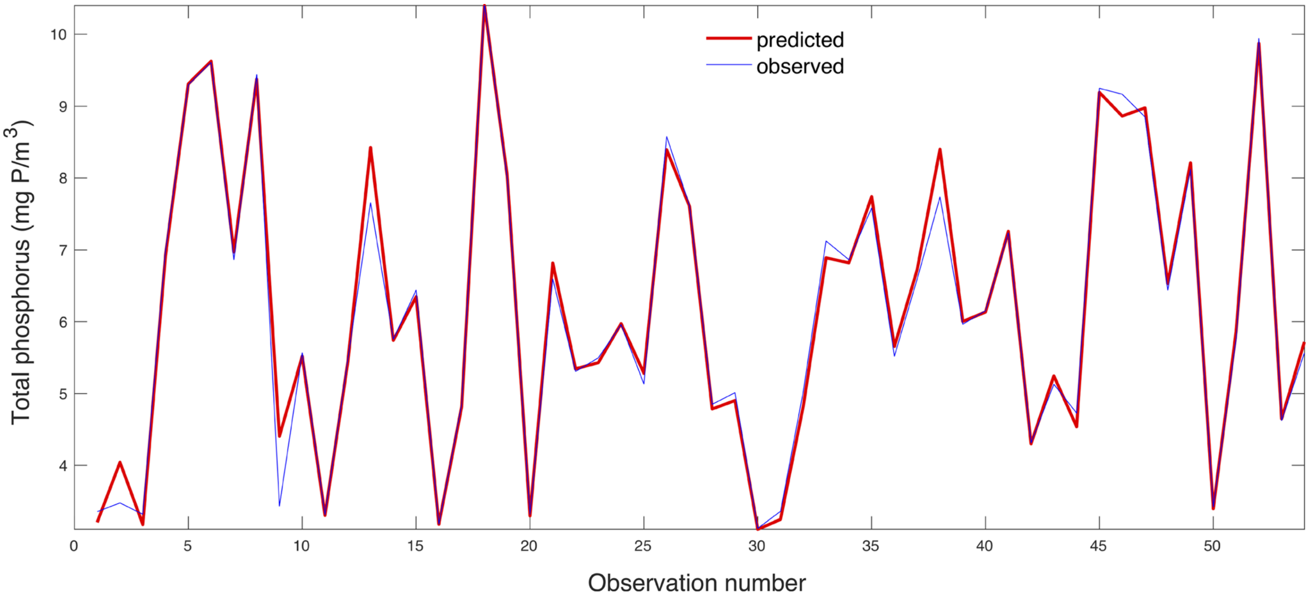

Fig. 10 Observed versus predicted TP concentrations employing the WOA/SVR-RBF model for the testing dataset ($R^2 = 0.9750$)

The last input variable in importance to forecast TP content in Tanes reservoir, and with much less influence, is Phosphates concentration. Obviously, there is a relationship between the amount of phosphates and the total phosphorus given that phosphates are a fraction of total phosphorus. However, the relationship is not so tight as with other variables as iron or the measured phytoplankton species.

On the whole, the SVR–RBF method is an accurate tool to predict the concentration of Chl-a and TP (output variables or eutrophication indicators), taking as input parameters that can be measured easily and frequently. Certainly, Figs. 9 and 10 compares the observed and predicted concentrations of Chl-a and TP using the SVR-RBF technique over the test dataset, respectively. In this method, it is important to optimize the hyperparameters of the SVR effectively and robustly. This function is performed by the metaheuristic optimizer WOA. Conclusively, the observed and predicted Chl-a and TP concentrations obtained with these models were correlated to a high degree.

4 Conclusions

According to the earlier results, several key findings from this study can be deduced and reported as follows: (1) analytical (or Laplacian) models for predicting Chl-a and TP concentrations from experimental parameters do not give good enough results due to the nonlinear character of the problem and the required simplifications. Hence, the necessity of a machine learning method such as the WOA/SVR is the verification that Chl-a and TP concentrations can be calculated precisely using this hybrid approximation based on WOA/SVR. Indeed, coefficients of determination of values 0.8582 and 0.9750 were obtained for the concentrations of Chl-a and TP, respectively; (2) Moreover, the relative importance of the predictors in the models was established. This finding can be considered one of the main ones of this investigation. In particular, Chlorophytes concentration must be kept in mind as the most noteworthy input variable in the foretelling of Chl-a concentration. Similarly, Iron appears as the most important variable in the prediction of TP; (3) conclusively, the importance of the hyperparameters precise tuning in the SVR-based approximation concerning the regression performance accomplished for Chl-a and TP concentrations was determined. The calculation of the optimal hyperparameters requires solving an optimization problem with inequality constraints. Here we have used the WOA optimizer with success.

Moreover, further application to other aquatic environments with similar characteristics such as ponds in gardens or rivers in zones with low speed where the assessment of the eutrophic state is of first importance would be desirable. Also, more experimentation is needed to take advantage of this study, modifying the relevant parameters obtained in this study to improve the prediction of the eutrophic state. For instance, an additional future line of research will be to build other novel hybrid mathematical models to address new challenges.

Acknowledgements The authors express their gratitude to the Department of Mathematics at University of Oviedo for the computational help. They also thank Anthony Ashworth for his English revision of the paper and spelling of this investigation paper.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abbaszadeh M, Hezarkhani A, Soltani-Mohammadi S (2016) Proposing drilling locations based on the 3D modeling results of fluid inclusion data using the support vector regression method. J Geochem Explor 165:23–34. https://doi.org/10.1016/j.gexplo.2016.02.005

Abirhire O, North RL, Hunter K, Vandergucht DM, Sereda J, Hudson JJ (2015) Environmental factors influencing phytoplankton communities in Lake Diefenbaker, Saskatchewan, Canada. J Great Lakes Res 41:118–128. https://doi.org/10.1016/j.jglr.2015.07.002

Álvarez X, Valero E, Santos RMB, Varandas SGP, Sanches Fernandes LS, Pacheco FAL (2017) Anthropogenic nutrients and eutrophication in multiple land use watersheds: best management practices and policies for the protection of water resources. Land Use Policy 69:1–11. https://doi.org/10.1016/j.landusepol.2017.08.028

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

Ansari AA, Gill SS, Lanza GR, Rast W (2010) Eutrophication: causes, consequences and control. Springer, New York

Arauzo M, Álvarez Cobelas M (1994) Phytoplankton strategies and time scales in a eutrophic reservoir. Hydrobiologia 291:1–9. https://doi.org/10.1007/BF00024234

Barnes DJ, Chu D (2010) Introduction to modeling for biosciences. Springer, New York

Beiras R (2018) Marine pollution: sources, fate and effects of pollutants in coastal ecosystems. Elsevier, Amsterdam

Bishop CM (2011) Pattern recognition and machine learning. Springer, New York

Brasil J, Attayde JL, Vasconcelos FR, Dantas DDF, Huszar VLM (2016) Drought-induced water-level reduction favors cyanobacteria blooms in tropical shallow lakes. Hydrobiologia 770(1):145–164. https://doi.org/10.1007/s10750-015-2578-5

Brönmark C, Hansson L-A (2005) The biology of lakes and ponds. Oxford University Press, New York

Brown CD, Hoyer MV, Bachmann RW, Canfield DE Jr (2000) Nutrient-chlorophyll relationships: an evaluation of empirical nutrient-chlorophyll models using Florida and northern temperate lake data. Can J Fish Aquat Sci 57(8):1574–1583. https://doi.org/10.1139/cjfas-57-8-1574

Çadraku HS (2021) Groundwater quality assessment for irrigation: case study in the Blinaja river basin, Kosovo. Civil Eng J 7(9):1515–1528. https://doi.org/10.28991/cej-2021-03091740

Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:1–27. https://doi.org/10.1145/1961189.1961199

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. Energy Convers Manag 75:311–318. https://doi.org/10.1016/j.enconman.2013.06.034

Cheng M-Y, Cao M-T (2014) Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. Appl Soft Comput 22:178–188. https://doi.org/10.1016/j.asoc.2014.05.015

Chou S-M, Lee S-M, Shao YE, Chen I-F (2004) Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Syst Appl 27(1):133–142. https://doi.org/10.1016/j.eswa.2003.12.013

Ciaburro G (2017) MATLAB for machine learning. Packt Publishing, Birmingham

Costa MRA, Attayde JL, Becker V (2016) Effects of water level reduction on the dynamics of phytoplankton functional groups in tropical semi-arid shallow lakes. Hydrobiologia 778(1):75–89. https://doi.org/10.1007/s10750-015-2593-6

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge

Deng T, Chau K-W, Duan H-F (2021) Machine learning based marine water quality prediction for coastal hydro-environment management. J Environ Manag 284:112051. https://doi.org/10.1016/j.jenvman.2021.112051

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

Di Toro DM, O'Connor DJ, Thomann RV (1971) A dynamic model of the phytoplankton population in the Sacramento-San Joaquin Delta. In: Advances in chemistry series, non equilibrium systems in natural water chemistry, vol 106. American Chemical Society, New York, pp 131–150

Ebrahimgol H, Aghaie M, Zolfaghari A, Naserbegi A (2020) A novel approach in exergy optimization of a WWER1000 nuclear power plant using whale optimization algorithm. Ann Nucl Energy 145:107540. https://doi.org/10.1016/j.anucene.2020.107540

Fan G-F, Yu M, Dong S-Q, Yeh Y-H, Hong W-C (2021) Forecasting short-term electricity load using hybrid support vector regression with grey catastrophe and random forest modeling. Util Policy 73:101294. https://doi.org/10.1016/j.jup.2021.101294

Fields S (2004) Global nitrogen: cycling out of control. Environ Health Perspect 112(10):A556–A563. https://doi.org/10.1289/ehp.112-a556

Freedman D, Pisani R, Purves R (2007) Statistics. WW Norton & Company, New York

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–141. https://doi.org/10.1214/aos/1176347963

Friedman JH, Roosen CB (1995) An introduction to multivariate adaptive regression splines. Stat Methods Med Res 4:197–217. https://doi.org/10.1177/096228029500400303

García-Nieto PJ, Combarro EF, del Coz Díaz JJ, 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. https://doi.org/10.1016/j.amc.2013.03.018

García-Nieto PJ, García-Gonzalo E, Alonso Fernández JR, Díaz Muñiz C (2019) Modeling algal atypical proliferation using the hybrid DE-MARS-based approach and M5 model tree in La Barca reservoir: a case study in northern Spain. Ecol Eng 130:198–212. https://doi.org/10.1016/j.ecoleng.2019.02.020

Gharehchopogh FS, Gholizadeh H (2019) A comprehensive survey: whale optimization algorithm and its applications. Swarm Evol Comput 48:1–24. https://doi.org/10.1016/j.swevo.2019.03.004

Gu T, Lu W, Bao X, Chen N (2006) Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors. Solid State Sci 8(2):129–136. https://doi.org/10.1016/j.solidstatesciences.2005.10.01

Gul A, Shahzada K, Alam B, Badrashi YI, Khan SW, Khan FA, Ali A, Rehman ZU (2020) Experimental study on the structural behavior of cast in-situ hollow core concrete slabs. Civil Eng J 6(10):1983–1991. https://doi.org/10.28991/cej-2020-03091597

Hadjisolomou E, Stefanidis K, Herodotou H, Michaelides M, Papatheodorou G, Papastergiadou E (2021) Modelling freshwater eutrophication with limited limnological data using artificial neural networks. Water 13(11):1590. https://doi.org/10.3390/w13111590

Hansen T, Wang CJ (2005) Support vector based battery state of charge estimator. J Power Sources 141:351–358. https://doi.org/10.1016/j.jpowsour.2004.09.020

Harper D (1991) Eutrophication of freshwaters: principles, problems and restoration. Springer, New York

Hastie T, Tibshirani R, Friedman JH (2003) The elements of statistical learning. Springer, New York

Havens K (2019) Effects of climate change on the eutrophication of lakes and estuaries. SGEF-189, one of a series of the Sea Grant Department, UF/IFAS Extension. University of Florida. https://edis.ifas.ufl.edu/pdf%5CSG%5CSG12700.pdf

Howell F (2017) Eutrophication: causes, mechanisms and ecological effects. Nova Science Publishers, New York

Huisman J, Matthijs HCP, Visser PM (2010) Harmful cyanobacteria. Springer, New York

Ighravwe DE, Mashao D (2020) Analysis of support vector regression kernels for energy storage efficiency prediction. Energ Rep 6(9):634–639. https://doi.org/10.1016/j.egyr.2020.11.171

Jekabsons G (2016) ARESLab: adaptive regression splines toolbox for Matlab/Octave. http://www.cs.rtu.lv/jekabsons/regression.html

Jimeno-Sáez P, Senent-Aparicio J, Cecilia JM, Pérez-Sánchez J (2020) Using machine-learning algorithms for eutrophication modeling: case study of Mar Menor lagoon (Spain). Int J Environ Res Public Health 17(4):1189. https://doi.org/10.3390/ijerph17041189

Kalaji HM, Sytar O, Brestic M, Samborska IA, Cetner MD, Carpentier C (2016) Risk assessment of urban lake water quality based on in-situ cyanobacterial and total Chl-a monitoring. Pol J Environ Stud 25:45–56. https://doi.org/10.15244/pjoes/60895

Karydis M (2009) Eutrophication assessment of coastal waters based on indicators: a literature review. Glob NEST J 11(4):373–390. https://doi.org/10.30955/gnj.000626

Kerich EC (2020) Households drinking water sources and treatment methods options in a regional irrigation scheme. J Hum Earth Future 1(1):10–19. https://doi.org/10.28991/HEF-2020-01-01-02

Kinkaid C (2014) Toxic algae: how to treat and prevent harmful algal blooms in ponds, lakes, rivers and reservoirs. Solardyne, Portland

Knafl GJ, Ding K (2016) Adaptive regression for modeling nonlinear relationships. Springer, Berlin

Koopmans GF, Hiemstra T, Vaseur C, Chardon WJ, Voegelin A, Groenenberg JE (2020) Use of iron oxide nanoparticles for immobilizing phosphorusin-situ: increase in soil reactive surface area and effect on soluble phosphorus. Sci Total Environ 711:135220. https://doi.org/10.1016/j.scitotenv.2019.135220

Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York

Latif Z, Tasneem MA, Javed T, Butt S, Fazil M, Ali M, Sajjad MI (2003) Evaluation of water-quality by chlorophyll and dissolved oxygen. In: Water resources in the south: present scenario and future prospects, commission on science and technology for sustainable development in the South, Islamabad, Pakistan, pp 122–135

Li X, Lord D, Zhang Y, Xie Y (2008) Predicting motor vehicle crashes using support vector machine models. Accid Anal Prev 40:1611–1618. https://doi.org/10.1016/j.aap.2008.04.010

Liao Z, Zang N, Wang X, Li C, Liu Q (2021) Machine learning-based prediction of chlorophyll-a variations in receiving reservoir of world's largest water transfer project—a case study in the Miyun reservoir. North China Water 13(17):2406. https://doi.org/10.3390/w13172406

Lin S-S, Shen S-L, Zhou A, Lyu H-M (2021) Assessment and management of lake eutrophication: a case study in Lake Erhai, China. Sci Total Environ 751:141618. https://doi.org/10.1016/j.scitotenv.2020.141618

Liu XJ, Duan L, Mo JM, Du E, Shen J, Lu X, Zhang Y, Zhou X, He C, Zhang F (2011) Nitrogen deposition and its ecological impact in China: an overview. Environ Pollut 159(10):2251–2264. https://doi.org/10.1016/j.envpol.2010.08.002

McClave JT, Sincich TT (2016) Statistics. Pearson, New York

McQuaid N, Zamyadi A, Prevost M, Bird DF, Dorner S (2011) Use of in vivo phycocyanin fluorescence to monitor potential microcystin-producing cyanobacterial biovolume in a drinking water source. J Environ Monit 13:455–463. https://doi.org/10.1039/c0em00163e

Mirjalili S, Lewis A (2016) The whale optimization algorithm. Adv Eng Softw 95:51–67. https://doi.org/10.1016/j.advengsoft.2016.01.008

Moss B, Kosten S, Meerhoff M, Battarbee RW, Jeppesen E, Mazzeo N, Havens K, Lacerot G, Liu Z, De Meester L, Paerl H, Scheffer M (2011) Allied attack: climate change and eutrophication. Inland Waters 1(2):101–105. https://doi.org/10.5268/IW-1.2.359

Moura Ado N, DoNascimento EC, Dantas EW (2012) Temporal and spatial dynamics of phytoplankton near farm fish in eutrophic reservoir in Pernambuco Brazil. Rev Biol Trop 60(2):581–597. https://doi.org/10.15517/rbt.v60i2.3939

Negro AI, de Hoyos C, Vega JC (2000) Phytoplankton structure and dynamics in Lake Sanabria and Valparaíso reservoir (NW Spain). Hydrobiologia 424(1):25–37. https://doi.org/10.1023/A:1003940625437

Pal M, Goel A (2007) Estimation of discharge and end depth in trapezoidal channel by support vector machines. Water Resour Res 21(10):1763–1780. https://doi.org/10.1007/s11269-006-9126-z

Patrick R, Crum B, Coles J (1969) Temperature and manganese as determining factors in the presence of diatom or blue-green algal floras in streams. Proc Natl Acad Sci 64(2):472–478. https://doi.org/10.1073/pnas.64.2.472

Pip E, Bowman L (2014) Microcystin and algal chlorophyll in relation to nearshore nutrient concentrations in Lake Winnipeg. Canada Environ Pollut 3(2):36–47. https://doi.org/10.5539/ep.v3n2p36

Quesada A, Sanchis D, Carrasco D (2004) Cyanobacteria in Spanish reservoirs. How frequently are they toxic? Limnetica 23:109–118. https://doi.org/10.23818/limn.23.09

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. https://doi.org/10.1080/09670260500480926

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. https://doi.org/10.1007/s10750-008-9376-2

Reynolds CS (2006) Ecology of phytoplankton. Cambridge University Press, New York

Riegl B, Glynn PW, Wieters E, Purkis S, d'Angelo C, Wiedenmann J (2014) Water column productivity and temperature predict coral reef regeneration across the Indo-Pacific. Sci Rep 5:8273–8279. https://doi.org/10.1038/srep08273

Schölkopf B, Smola AJ, Williamson R, Bartlett P (2000) New support vector algorithms. Neural Comput 12(5):1207–1245. https://doi.org/10.1162/089976600300015565

Sekulic SS, Kowalski BR (1992) MARS: a tutorial. J Chemom 6:199–216. https://doi.org/10.1002/cem.1180060405

Shamshirband S, Nodoushan EJ, Adolf JE, Manaf AA, Mosavi A, Chau K-W (2019) Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters. Eng Appl Comput Fluid 13(1):91–101. https://doi.org/10.1080/19942060.2018.1553742

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge

Sheldon SP, Skelly DK (1990) Differential colonization and growth of algae and ferromanganese-depositing bacteria in a mountain stream. J Freshw Ecol 5(4):475–485. https://doi.org/10.1080/02705060.1990.9665264

Shumway SE, Burkholder JM, Morton SL (2018) Harmful algal blooms: a compendium desk reference. Wiley-Blackwell, New York

Smith VH (2006) Responses of estuarine and coastal marine phytoplankton to nitrogen and phosphorus enrichment. Limnol Oceanogr 51:377–384. https://doi.org/10.4319/lo.2006.51.1_part_2.0377

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. https://doi.org/10.1002/tox.20356

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. https://doi.org/10.1016/j.ecolind.2010.01.005

Steinwart I, Christmann A (2008) Support vector machines. Springer, New York

Tanawi IN, Vito V, Sarwinda D, Tasman H, Hertono GF (2021) Support vector regression for predicting the number of Dengue incidents in DKI Jakarta. Procedia Comput Sci 179:747–753. https://doi.org/10.1016/j.procs.2021.01.063

Texeira MR, Rosa MJ (2006) Comparing dissolved air flotation and conventional sedimentation to remove cyanobacterial cells of *Microcystis aeruginosa*: part I: the key operating conditions. Sep Purif Technol 52:84–94. https://doi.org/10.1016/j.seppur.2006.03.017

Tiyasha TTM, Yaseen ZM (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. J Hydrol 595:124670. https://doi.org/10.1016/j.jhydrol.2020.124670

Tufford DL, McKeller HN (1999) Spatial and temporal hydrodynamic and water quality modeling analysis of a large reservoir on the South Carolina (USA) coastal plain. Ecol Model 114:137–173. https://doi.org/10.1016/S0304-3800(98)00122-7

United States Environmental Protection Agency (2014) Chapter 4: Eutrophication. http://www.epa.gov/emap2/maia/html/docs/Est4.pdf. Accessed 24 Aug 2014

Van der Valk AG (2006) The biology of freshwaters wetlands. Oxford University Press, New York

Vapnik V (1998) Statistical learning theory. Wiley-Interscience, New York

Vinçon-Leite B, Casenave C (2019) Modelling eutrophication in lake ecosystems: a review. Sci Total Environ 651:2985–3001. https://doi.org/10.1016/j.scitotenv.2018.09.320

Waseem Ahmad M, Reynolds J, Rezgui Y (2018) Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. J Clean Prod 203:810–821. https://doi.org/10.1016/j.jclepro.2018.08.207

Watzin MC, Miller EB, Shambaugh AD, Kreider MA (2006) Application of the WHO alert level framework to cyanobacterial monitoring of Lake Champlain, Vermont. Environ Toxicol 21:278–288. https://doi.org/10.1002/tox.20181

Wheeler SM, Morrissey LA, Levine SN, Livingston GP, Vincent WF (2012) Mapping cyanobacterial blooms in Lake Champlain's Missisquoi Bay using Quick Bird and MERIS satellite data. J Great Lakes Res 38(1):68–75. https://doi.org/10.1016/j.jglr.2011.06.009

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. https://doi.org/10.1007/s10750-005-4453-2

World Health Organization (1998) Guidelines for drinking-water quality: health criteria and other supporting information, vol 2, World Health 408 Organization, Geneva

Wu X, Bian X, Lin E, Wang H, Guo Y, Tan X (2021) Weighted multiscale support vector regression for fast quantification of vegetable oils in edible blend oil by ultraviolet-visible spectroscopy. Food Chem 342:128245. https://doi.org/10.1016/j.foodchem.2020.128245

Xu QS, Dazykowski M, Walczak B, Daeyaert F, de Jonge MR, Heeres J, Koymans LMH, Lewi PJ, Vinkers HM, Janssen PA, Massart DL (2004) Multivariate adaptive regression splines—studies of HIV reverse transcriptase inhibitors. Chemom Intell Lab 72(1):27–34. https://doi.org/10.1016/j.chemolab.2004.02.007

Xu X, Liu Y, Liu S, Li J, Guo G, Smith K (2019) Real-time detection of potable-reclaimed water pipe cross-connection events by conventional water quality sensors using machine learning methods. J Environ Manag 238:201–209. https://doi.org/10.1016/j.jenvman.2019.02.110

Yuan LL, Pollard AI, Pather S, Oliver JL, D'Anglada L (2014) Managing microcystin: Identifying national-scale thresholds for total nitrogen and chlorophyll a. Freshw Biol 59(9):1970–1981. https://doi.org/10.1111/fwb.12400

Zhang W, Goh ATC, Zhang Y, Chen Y, Xiao Y (2015) Assessment of soil liquefaction based on capacity energy concept and multivariate adaptive regression splines. Eng Geol 188:29–37. https://doi.org/10.1016/j.enggeo.2015.01.009

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