A New Predictive Model for Evaluating Chlorophyll-a Concentration in Tanes Reservoir by Using a Gaussian Process Regression

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

Abstract Chlorophyll-a (hereafter referred to as Chl-a) is a recognized indicator for phytoplankton abundance and biomass –hence, an effective estimation of the trophic condition– of water bodies as lakes, reservoirs and oceans. Indeed, Chl-a is the primary molecule responsible for photosynthesis. A strong and robust Bayesian nonparametric technique, termed Gaussian process regression (GPR) approach, for foretelling the dependent variable Chl-a concentration in Tanes reservoir from a dataset concerning to 268 samples is shown in this paper. Ten years (2006–2015) of monitoring water quality variables (biological and physico-chemical independent variables) in the Tanes reservoir were used to build this mathematical GPR-relied model. As an optimizer, the method known as Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGSB) iterative algorithm was used; this allows the selection of kernel optimal parameters during the GPR training phase, which greatly determines the regression precision. The results of the current investigation can be summarized in two. Firstly, the relevance of each input variable on Chl-a concentration in Tanes reservoir is determined. Secondly, the Chl-a can be successfully predicted using this hybrid LBFGSB/GPR–relied model (R^2 and r values were 0.8597 and 0.9306, respectively). The concordance between

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

observed data and the model clearly proves the high efficiency of this innovative approach.

Keywords Chlorophyll-a; Gaussian process regression (GPR); Bayesian statistics; Regression analysis; Reservoir water quality

1 Introduction

Reservoirs and lakes are large bodies of standing water and multipurpose use (drinking water storage, hydropower, irrigation or simply leisure). Chl-a is an extensively applied environmental indicator of the algae or phytoplankton biomass growth –since it is found in every single photosynthesizing organism– and of the eutrophication process in reservoirs, lakes and oceans (Latif et al. 2003).

Chlorophylls are a family of green pigments (forms a, b, c, d, e and f) found in cyanobacteria and in all those organisms that contain chloroplasts or tylocoidal membranes in their cells, which includes plants and various algae (Wetzel 2001; Schinck et al. 2020). Chl-a is the primary molecule (found in every single photosynthesizing organism) responsible for photosynthesis, a process that enables plants and algae to produce energy from sunlight.

Chlorophyll (and thus phytoplankton) can be easily quantified using its known optical properties (each form reflects slightly different ranges of green wavelengths). The optical measure of chlorophyll concentration is a simple and accurate method to

estimate the phytoplankton (microscopic algae) concentration and, indirectly, the biological activity. In this way, monitoring chlorophyll levels is a direct way of tracking phytoplankton growth and therefore, eutrophication processes in reservoirs and lakes.

Large quantities of algae in a reservoir have noteworthy consequences on its biological and physico-chemical processes. It is well known that algal blooms lead to water quality decreasing –even becoming toxic– in reservoirs and lakes. The proliferation of algal blooms, high Chl-a levels, are associated with harmful algal blooms –algal blooms containing toxins– (Pip and Bowman 2014; Yuan et al. 2014; Vilán Vilán et al. 2013; Schinck et al. 2020). For example, cyanotoxins released by some cyanobacteria in reservoir water cause serious problems if they are consumed by humans or animals, direct or indirectly, posing a threat to drinking and recreational water (Watzin et al. 2006, Kalaji et al. 2016). In this sense, when the cyanotoxin concentration is not available, the knowledge of the Chl-a concentration can be considered an alternative measure to the cyanobacteria presence in a reservoir and, even an estimate of harmful cyanobacterial blooms (HABs) (McQuaid et al. 2011; Vilán Vilán et al. 2013; Schinck et al. 2020). Hence, Chl-a concentration forecasting is crucial in water quality management to prevent this kind of contamination and avoid potential health risks (Wheeler et al. 2012; Schinck et al. 2020). Additionally, Chl-a was chosen as the parameter to indicate phytoplankton biomass required by the European Commission Water Framework Directive (Directive 2000/60/EC) and other European directives. However, Chl-a prediction in reservoirs and lakes has not been completely successful yet (Di Toro et al. 1971; Brown et al. 2000; Tufford and McKeller 1999; Reynolds 2006).

Water quality Laplacian mathematical models based on reservoir or lake internal physico-chemical processes require an enormous quantity of information that in practice is not accessible, or cannot be obtained in its entirety, and presents a difficult implementation on the computer and it is computational time consuming.

In this study, a new hybrid regressive model based on Gaussian Process Regression (GPR) technique to foretell the Chl-a concentration in Tanes reservoir (located in Asturias, Northern Spain) is applied with success (see Figs. 1a and b). Algal abnormal proliferation can be a serious environmental problem in water bodies like the Tanes reservoir. This reservoir is used to supply drinking water to the central area of Asturias. Hence, the need to avoid major risks such as the presence of toxic algae blooms using Chl-a concentration as an early alarm.

Fig. 1 a Aerial view of the Tanes reservoir; **b** a closer view of the Tanes reservoir

As a result, the implementation of the innovative technique that combines the Gaussian process regression (GPR) approximation (Rasmussen 2003; Dym and McKean 2008; Ebde 2015) with the optimization algorithm Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGSB) (Liu and Nocedal 1989; Byrd et al. 1994; Zhu et al. 1997; Fei et al. 2014) to foretell the Chl-a concentration could be an attractive methodology since, as far as authors know, it has never been tackled in prior researches about water quality in reservoirs and lakes before. Moreover, the GPR technique is a statistical learning methodology developed by statistics and Bayesian analysis, which is capable of dealing with non-linearities, including interactions among variables (Rasmussen

2003; Dym and McKean 2008; Ebden 2015). If we compare it with other classical and metaheuristic regression techniques, GPR approximation presents some benefits (Rasmussen and Williams 2005): (1) GPR has a remarkable ability to be widespread; (2) the GPR optimal parameters can be determined using heuristic optimizers; (3) the GPR results show an evident probabilistic significance; (4) the GPR works well on small datasets; (5) it makes use of the whole information available; and (6) moreover, it has the ability to provide uncertainty measurements on the predictions, which is the main characteristic that differentiates it from other regression methods. In effect, the LBFGSB optimizer has been used here satisfactorily to calculate the optimal GPR hyperparameters. In addition, former investigations indicate that GPR is a very appropriate tool in a large number of real applications as the estimation of the chlorophyll concentration in subsurface waters from remote sensing data (Pasolli et al. 2010), computational fluid dynamics (CFD) (Duan et al. 2019), geometrical characteristics of cladding tracks (Wang et al. 2020), efficiency of a solar/waste energy boosted dehumidification/regeneration cycle with a solid adsorbent bed (Akhlaghi et al. 2019), blood pressure measurement (Alghamdi et al. 2020), state of health for lithium battery (Li et al. 2020), building energy use (Zeng et al. 2020), time series analysis (Ambrogioni and Maris 2019), wind speed prediction (Cai et al. 2020), gravity field modeling (Gao and Liao 2019), sunspot cycle prediction (Gonçalves et al. 2020) and so on. However, it has never been used for evaluating Chl-a concentration from water quality certificates in reservoirs and lakes, taking into account the functions attributed to the Cantabrian Basin Authority such as the administration and control of the public hydraulic domain of its territorial scope.

The principal goal of the current research was to foretell the output Chl-a concentration in Tanes reservoir from the remaining input biological and physico-chemical parameters –measured from periodic samplings required by the Water Framework Directive (Directive 2000/60/EC) in all water bodies– employing Gaussian process regression (GPR) along with the LBFGSB optimizer. This model defines a new algorithm to analyze phytoplankton density in lakes and reservoirs, measuring the Chl-a concentration in them (Chen 1970; Smith 2006; Riegl et al. 2014). Indeed, the Chl-a concentration can be considered a fundamental indicator of excess nutrients such as total phosphorus and nitrogen in a reservoir or lake, and ultimately, of the existence of eutrophication in those bodies of water.

This paper structure is as follows: Section 2 presents the experimental arrangement, all the variables included in this research and GPR methodology; Section 3 draws up the findings gathered with this novel technique by collating the GPR results with the observed values as well as the significance ranking of the input variables; and conclusively, Section 4 concludes this study with an inventory of principal results of the investigation.

2 Materials and Methods

2.1 Study Area

Tanes reservoir is inside the Natural Park of Redes (Natural Reserve and Reservation of the Biosphere), in the south of Asturias (a region in Northern Spain), between the municipalities of Caso and Sobrescobio in the Nalon valley. The project finished in 1978, with a capacity of 33.27 hm^3, a surface of 159 ha, and a maximum depth of 95 m.

This reservoir is linked to the Rioseco reservoir in order to pump water from this one to the previous one during the night and down from Tanes to Rioseco during the day to produce energy. The Rioseco reservoir is located 3 km downstream. Its capacity is 4.3 hm^3, its surface, 63 ha and its maximum depth, 28.5 m. Both Tanes and Rioseco reservoirs supply water to almost the entire urban center of Asturias. Other uses of Tanes reservoir are hydroelectric and, recently, recreational. The reservoir has an ornithological interest due to the presence of ducks and herons, and, with a representative fauna, it is a hunting refuge.

The research area is located in the Central Carboniferous Basin with a wide diversity of lithologies that include mainly quartzite and limestone materials. The Tanes reservoir basin is made up of quartzites, sandstones and slates. The predominant materials are basic. This section of the river, where Tanes reservoir is located, receives underground discharges from the carbonate aquifers that are generally disconnected from each other by slates with sandstone intercalations and by quartzites, both waterproof.

2.2 Experimental Dataset

The numerical outcome used for the LBFGSB/GPR analysis were acquired over 10 years (from 2006 to 2015) from 268 samples monthly (minimum frequency, sometimes bimonthly even weekly) picked up from January 16, 2006, to December 31, 2015, following the sampling protocols for lakes and reservoirs developed by *Spanish Ministry for the Ecological Transition and Demographic Challenge* in agreement with those validated by the European Union (Smith et al. 2008; World Health Organization 1998). Samples were taken with a Niskin hydrographic bottle at the reservoir maximum

depth site –established with a depth gauge (Willame et al. 2005)–. The Niskin bottle is like a tube with stoppers at each end that allows vertical sampling at desired depths (see Fig. 2a). These depths were selected at equal intervals in the euphotic zone (enough illuminated region for photosynthesis) calculated from Secchi depth –where the pattern on a Secchi disk (see Fig. 2b), is no longer visible because of the water turbidity–. Thus, five subsamples were collected (Brönmark and Hansson 2005; Quesada et al. 2006) and then homogenized to obtain a composite sample in which the Chl-a and phytoplankton content were determined. The other experimental data considered in this work were the common biological and physico–chemical parameters evaluated in limnological studies (Brönmark and Hansson 2005; Negro et al. 2000; Reynolds 2006) and were analyzed in the field (Water temperature, pH, conductivity, Dissolved Oxygen and Secchi Depth) or in an accredited laboratory.

Fig. 2 a A Niskin bottle; **b** examples of Secchi disks

2.3 Variables of the Model

The aim of this work was to find a way to foretell the Chl-a concentration (µg/L) in Tanes reservoir. Chl-a is extensively employed as algae biomass index and consequently, as a sign of reservoirs and lakes eutrophication (Latif et al. 2003). High chlorophyll content in reservoirs and lakes waters normally indicates the existence of algal blooms (the United States Environmental Protection Agency 2014) and therefore a decrease in these waters' quality. Certainly, chlorophyll is a noteworthy pigment in green organisms because it is responsible for absorbing the light energy needed in the photosynthetic process. Phytoplankton, made up of green organisms, is related to the

Chlorophyll presence (American Public Health Association 2005). Chlorophyll concentration is also accepted as a substitute measure of the presence of cyanobacteria which can generate cyanotoxins in the water body and the subsequent potential risk to public health (Wheeler et al. 2012; Schinck et al. 2020). The built model (GPR-relied model) uses the concentration of different kinds of phytoplankton, as well as some chemical and physico-chemical parameters, as independent predictors.

Input parameters:

- Biological variables:

- Cyanobacteria concentration (mm^3/L): photosynthetic bacteria in the phytoplankton community (see Fig. 3a), promoted in fertilized environments, especially in eutrophic ones. Their presence in freshwater ecosystems is a matter of concern, especially in lentic water bodies (Quesada et al. 2006; Texeira and Rosa 2006; Willame et al. 2005; Vilán Vilán et al. 2013).

- Diatoms concentration (mm^3/L): another frequent organism in the phytoplankton community (photosynthetic microorganisms that drift about in water; see Fig. 3b).

- Euglenophytes concentration (mm^3/L): all of them are primary producers (organisms that build complex organic molecules from simple inorganic compounds), that is, they can photosynthesize and they are part of phytoplankton (see Fig. 3c).

- *Dinophlagellata* concentration (mm^3/L): another kind of phytoplankton, unicellular and eukaryote –kingdom Protista– (see Fig. 3d).

- Chrysophytes concentration (mm^3/L): in essence photosynthetic even though some of them can be organotrophic (see Fig. 3e).

- Chlorophytes concentration (mm^3/L): both single and many-celled algae species in phytoplankton (see Fig. 3f).

- Chryptophytes concentration (mm^3/L): unicellular tiny flagellated kind of phytoplankton (see Fig. 3g).

Fig. 3 Microorganisms in Tanes reservoir whose concentrations were used as input variables: **a** Cyanobacteria; **b** Diatoms; **c** Euglenophytes; **d** *Dinophlagella*; **e** Chrysophytes; **f** Clorophytes; and **g** Chryptophytes

➢ Physico-chemical variables:

- Water temperature (ºC): reservoir average thermal energy measure. It affects the physical and chemical process and biological activity and growth in water ecosystems.

- Orthophosphates concentration (mg PO_4^{3-}/L): an expression of the reactive phosphates –i.e. the phosphorus form taken up by plants– density. High values have a significant impact on water ecosystems since they promote the growth of phosphate-dependent organisms –algae and phytoplankton in general, for instance–, that can lead to eutrophication decreasing water quality and causing the extinction of some oxygen-dependent organisms.

- Total phosphorus concentration (mg P/m^3). It includes the above mentioned phosphates and the other phosphorus forms in both the soluble and the sestonic water fractions. Phosphorus is mostly present in the environment as orthophosphates and can be considered a reservoir of them since all the other forms can be converted into orthophosphates depending on environmental conditions. Hence, Phosphorus is a limiting factor in biological productivity.

- Nitrite concentration (mg NO^{2-}/L). Nitrite and nitrate are ubiquitous and naturally occurring ions in the environment. Both are products of the oxidation of nitrogen. Nitrite is a potential problem in aquatic environments. It is capable of inducing methemoglobinemia in a wide range of species, including humans, and other multiple physiological effects. It can reach water bodies from waste-water disposal, from nitrate reduction and also from ammonia oxidation.

- Nitrate concentration (mg NO_3^-/L): It is one of the most abundant nitrogen forms in water –since the main source of nitrogen compounds in water are fertilizers– and the one used by plants. High consumption cause methemoglobinemia in humans and many animals since their bodies convert a portion of that nitrate into nitrite. In large quantities, it is also harmful to fish and can cause, together with a high concentration of phosphate, eutrophication processes, strongly linked to cyanobacteria blooms and their toxic metabolites, the cyanotoxins (Schinck et al. 2020).

- Ammonium concentration (mg NH_4^+/L). Together with nitrite and nitrate, ammonium is the major inorganic nitrogen compound occurring in surface waters. In oxidizing media, it can be converted into nitrate, or nitrite if the

oxidation process is incomplete. When dissolved in surface water, ammonia exists in two forms: NH_3^0 (unionized) and NH_4^+ (ionized, called ammonium). The unionized form can enter into fish, and once inside, some converts to the ionized form, which causes cellular damage.

- Dissolved oxygen concentration (mg O_2/L): a measure of oxygen diffused in water from the atmosphere and crucial for most aquatic species. Adequate dissolved oxygen concentration is necessary to minimize fish stress. Algal blooms produce much oxygen during the day but they also can cause its lowering at night or when they die and decompose.

- Iron concentration (mg Fe/L). Iron is found in the aquatic environment as a result of natural runoff, erosion of certain soils and other geological sources, besides its presence in water wastes that end up in water bodies. It is a trace element essential for life but in higher concentrations can be toxic. Its oxidized form (iron hydroxide) may precipitate out and form a slim on bottom sediment which decreases light penetration in water bodies inhibiting algal growth.

- Manganese concentration (mg Mn/L). Manganese is a crucial micronutrient in phytoplankton growth. It has been shown that manganese concentration affects phytoplankton composition (Patrick et al. 1969). Water bodies with significant iron and manganese levels can cause blooms of ferromanganese-depositing bacteria declining algal abundance (Sheldon and Skelly 1990).

- Conductivity (μS/cm): refers to the water ability to conduct electricity which in turn depends on water salinity that affects phytoplankton concentration and composition (Redden and Rukminasari 2008).

- Volume of water (hm^3). It represents the quantity of water in the reservoir at the sampling time and it varies according to the impoundment purpose (drinking water supply, hydroelectric use, etc.). This volume affects all substance concentrations in it. High volume dilutes the wastewater that reaches the reservoir reducing substances concentrations and consequently, their toxicity and the risk of eutrophication. In this way, the trophic state improves.

- pH: a measure of the water acidity/alkalinity which affects biological and chemical processes and an indicator of the eutrophication process.

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

2.4 Gaussian Process Regression (GPR)

A Gaussian process (GP) is a stochastic process that generates samples over time $\{X_t\}_{t\in\tau}$ in such a way that it does not affect the finiteness of a linear combination X_t having (or more generally any linear function of the sample function X_t), linear combination that will normally be distributed (Rychlik et al. 1997; Bishop 2011; Daemi et al. 2019; Li et al. 2020). Let's assume that $T=\{(\mathbf{x}_i, y_i)/i=1,2,...,N\}$ describes the training collection data of the Gaussian method. When we approximate a regression problem using Gaussian processes (also termed *kriging*), the following hypothesis is made: for a Gaussian process f observed at the $\mathbf{x}$ coordinates, the vector of

values $f(\mathbf{x})$ is only a sample of a multivariate Gaussian distribution of dimension equal to the number of observed coordinates *n*. It is a well-known fact that Gaussian processes can be utterly established by their second-order statistics. Hence, supposing a Gaussian process with a zero mean, the definition of the *covariance matrix K* (positive definite kernel) will completely determine the performance of the Gaussian process. The covariance matrix permits to define Gaussian process' concepts like isotropy, stationarity, smoothness and periodicity. The following is a list of some common covariance functions used in many regression problems (Rychlik et al. 1997; Rasmussen and Williams 2005; Dym and McKean 2008; Bishop 2011; Daemi et al. 2019; Li et al. 2020):

- Linear: $K_L(\mathbf{x},\mathbf{x}') = \mathbf{x}^T\mathbf{x}'$
- Squared exponential: $K_{SE}(\mathbf{x},\mathbf{x}') = \sigma_f^2 \cdot e^{-\frac{|d|^2}{2\ell^2}}$
- Rational quadratic: $K_{RQ}(\mathbf{x},\mathbf{x}') = \left(1+|d|^2\right)^{-\alpha}$, $\alpha \geq 0$
- Periodic: $K_P(\mathbf{x},\mathbf{x}') = e^{-\frac{2\sin^2\left(\frac{d}{2}\right)}{\ell^2}}$
- Ornstein–Uhlenbeck: $K_{OU}(\mathbf{x},\mathbf{x}') = e^{-\frac{2\sin^2\left(\frac{d}{2}\right)}{\ell^2}}$

where d is equal to $d = \mathbf{x} - \mathbf{x}'$, ℓ is the characteristic length-scale of the process and σ_f^2 is the signal variance. In this study, the squared exponential kernel, also called *radial basis function* (RBF), was used due its better performance compared to other kernels with a large volume of training data. It is clear that the results of this approach relied on Gaussian processes (GPs) rely on the values of the hyperparameter θ (for

example, ℓ and σ_f^2 values) since these values determine the behavior of the model. In practice, the real experimental data are noisy observations so that:

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon \tag{1}$$

where ε depicts the white noise as an additive term in Eq. (1). In practice, it is common to consider $\varepsilon \sim N(0, \sigma_n^2)$, which means that Gaussian noise will be independent and identically distributed, being σ_n the standard deviation of this noise. To fix ideas, it is possible to take into account a finite ensemble of the noisy real experimental data as an separate Gaussian procedure expressed as (Rasmussen and Williams 2005; Bishop 2011; Daemi et al. 2019; Li et al. 2020):

$$\mathbf{y} \sim GP\left(m(\mathbf{x}), K(\mathbf{x}, \mathbf{x}') + \sigma_n^2 \delta_{ij}\right) = GP\left(0, K(\mathbf{x}, \mathbf{x}') + \sigma_n^2 \delta_{ij}\right) \tag{2}$$

where δ_{ij} is the Kronecker delta distribution and

$$\begin{aligned} m(\mathbf{x}) &= E\left[f(\mathbf{x}) \right] \\ K(\mathbf{x}, \mathbf{x}') &= Cov(\mathbf{x}, \mathbf{x}') = E\left[\left(f(\mathbf{x}) - m(\mathbf{x}) \right)\left(f(\mathbf{x}') - m(\mathbf{x}') \right)^T \right] \end{aligned} \tag{3}$$

Therefore, according to the hypothesis of a zero mean distribution, $f(\mathbf{x}) \sim N\left(0, K(\theta, \mathbf{x}, \mathbf{x}')\right)$, we have that $K(\theta, \mathbf{x}, \mathbf{x}')$ is termed *covariance matrix* for all possible pairs $(\mathbf{x}, \mathbf{x}')$ of a given set of hyperparameters θ. In this case, the logarithmic marginal probability is expressed by:

$$\log p\left(f(\mathbf{x}) \middle| \theta, \mathbf{x} \right) = -\frac{1}{2} f(\mathbf{x})^T K(\theta, \mathbf{x}, \mathbf{x}') f(\mathbf{x}') - \frac{1}{2} \log \det\left(K(\theta, \mathbf{x}, \mathbf{x}') \right) - \frac{n}{2} \log 2\pi \tag{4}$$

The calculation of the maximum of this marginal likelihood with respect to θ determines the whole requirement of the Gaussian process *f*. Note in (4) that the first term on the right-hand side appertains to a penalization term as a consequence of the unsuitability of the model to fit to the experimental values while the second term is a penalization term that grows in proportion to the intricacy of the model.

Making predictions about an unobserved value $f(\mathbf{x})$ in the $\mathbf{x}$ coordinate, after specifying θ, is just an issue of extracting samples from the forecasted distribution $p\left(y^* \left| \mathbf{x}^*, f(\mathbf{x}), \mathbf{x} \right.\right) = N\left(y^* \left| A, B \right.\right)$ where the subsequent estimation of the mean *A* is expressed by:

$$A = K\left(\theta, \mathbf{x}^*, \mathbf{x}\right) K\left(\theta, \mathbf{x}, \mathbf{x}'\right)^{-1} f(\mathbf{x}) \tag{5}$$

and the subsequent estimation of the variance is determined by:

$$B = K\left(\theta, \mathbf{x}^*, \mathbf{x}^*\right) - K\left(\theta, \mathbf{x}^*, \mathbf{x}\right) K\left(\theta, \mathbf{x}, \mathbf{x}'\right)^{-1} K\left(\theta, \mathbf{x}^*, \mathbf{x}\right)^T \tag{6}$$

where:

- $K\left(\theta, \mathbf{x}^*, \mathbf{x}^*\right)$: would be the variance matrix at the new unobserved point $\mathbf{x}^*$ for a given vector of θ hyperparameters;
- $K\left(\theta, \mathbf{x}^*, \mathbf{x}\right)$: would be the covariance matrix between a new unobserved value $\mathbf{x}^*$ and all the remaining observed values of the $\mathbf{x}$ coordinate for a given vector of θ hyperparameters;
- $K\left(\theta, \mathbf{x}, \mathbf{x}'\right)$: is the covariance matrix for all possible pairs $\left(\mathbf{x}, \mathbf{x}'\right)$, as previously defined.

It is possible to point out that the subsequent mean estimation $f\left(\mathbf{x}^*\right)$ at the new unobserved point $\mathbf{x}^*$ is expressed as a linear combination of the observed values of $f\left(\mathbf{x}\right)$. Additionally, the variance of $f\left(\mathbf{x}^*\right)$ is independent of the observed values of $f\left(\mathbf{x}\right)$.

Hence, the GPR technique is relied on a nonparametric methodology since its predictive capacity falls on the observed values $\mathbf{y}$ and on the input data. Following this procedure, the values $\theta = \left\{l, \sigma_f, \sigma_n\right\}$ are called the GPR model hyperparameters (Lantz 2019). In order to determine the optimal hyperparameters $\theta' = \arg\max_{\theta} \log p\left(\mathbf{y}\middle| X, \theta\right)$, it is possible to employ any standard optimiser after parameter initialization. In this investigation work, the optimisation method, termed LBFGSB algorithm (Liu and Nocedal 1989; Byrd et al. 1994; Zhu et al. 1997; Fei et al. 2014) described below, is successfully applied.

2.5 Approach Accuracy

This novel LBFGSB/GPR–based method was developed with twenty predictive input variables already described in subsection 2.3 above. The Chl-a concentration is, as we know, the dependent variable to be predicted. In order to accurately and reliably forecast Chl-a from the twenty remaining input variables, it is mandatory to select the best model that fits the observed dataset. Although several possible statistics can be applied to ascertain the goodness–of–fit, the rule employed in this study was the coefficient of determination R^2 (Knafl and Ding 2016; McClave and Sincich 2016). The coefficient of determination is a statistic employed in the scope of a statistical model whose principal objective is to foretell upcoming results or to check an assumption.

Next, we will call the observed values t_i versus the values predicted by the model y_i. Now we can define the following sums of squares given by (McClave and Sincich 2016):

- $SS_{tot} = \sum_{i=1}^{n} (t_i - \bar{t})^2$: is the overall sum of squares, proportional to the sample variance.
- $SS_{reg} = \sum_{i=1}^{n} (y_i - \bar{t})^2$: is the regression sum of squares, also termed the *explained sum of squares.*
- $SS_{err} = \sum_{i=1}^{n} (t_i - y_i)^2$: is the residual sum of squares.

where $\bar{t}$ is the mean of the *n* observed data:

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

Considering the former sums, the coefficient of determination is specified by the following equation:

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

Moreover, the GPR methodology relies heavily on three hyperparameters (Rasmussen 2003; Dym and McKean 2008; Ebden 2015):

- Variance (σ_f^2): this parameter refers to the variance of the signal; its purpose is to control the vertical range of the kernel function.
- Lengthscale (ℓ): this parameter provides the characteristic size of the length. It allows controlling the horizontal scale where the kernel function alters.

- Gaussian noise variance (σ_n^2): this parameter is the variance of a Gaussian additive white noise $\varepsilon \sim N\left(0, \sigma_n^2\right)$.

It is noteworthy to consider that the GPR technique largely relied on the determination of all its three optimal hyperparameters that we have just pointed out above. In this research, a numerical optimizer, denominated Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGSB) iterative algorithm (Liu and Nocedal 1989; Byrd et al. 1994; Zhu et al. 1997), is employed here to determine these parameters because of its ability for solving unconstrained nonlinear optimization problems. Note that LBFGSB algorithm belongs to quasi-Newton methods and is therefore an extension of the secant method to tackle multidimensional problems. Furthermore, LBFGSB algorithm consists of five steps (Fei et al. 2014):

- Gradient projection;
- Generalized Cauchy point calculation;
- Subspace minimization;
- Line searching; and
- Hessian approximation.

3 Results and Discussion

All the twenty input independent variables (seven biological variables and thirteen physico-chemical variables) are shown in Tables 1 and 2, respectively. In this study, the total number of samples used here was 268, that is to say, 268 experimental samplings

were collected and processed in Tanes reservoir according to the European Water Quality Directive (Directive 2000/60/EC).

Table 1 Set of biological input independent variables used in this investigation: names, means and standard deviations

Table 2 Set of physico-chemical input independent variables used in this investigation: names, means and standard deviations

To confront this complex problem here, it is necessary to split the complete set of data into two parts: (1) a training set comprising 80% of the data; and (2) a testing set comprising the remaining 20% of the data. The key idea is to build a GPR-relied model with training data by determining the optimal parameters with the LBFGSB optimizer and then apply it to the test data to obtain predictions.

As stated above, the output variable (dependent variable) in this study is the Chl-a concentration dealt with by means of the LBFGSB/GPR–relied method. The selection of the optimal hyperparameters is a key factor in the performance of this method as we see above: (1) the lengthscale ℓ and variance σ_f^2 of the radial basis function (RBF) kernel; (2) the objective function value; and (3) Gaussian noise variance σ_n^2. The objective function used in the hyperparameter optimisation process is the –log

likelihood value (see Eq. 4). In this way, Table 3 indicates initial intervals of the three hyperparameters of the LBFGSB/GPR–relied approach fitted in this investigation.

Table 3 Search space for the three hyperparameters of the LBFGSB/GPR–relied approach fitted in this investigation

According to this methodology, Table 4 identifies the optimal parameters of the best fitted GPR–relied approach encountered with the LBFGSB optimizer.

Table 4 Optimal hyperparameters of the best fitted GPR–relied model encountered with the LBFGSB optimizer: variance σ_f^2 and lengthscale ℓ for the RBF kernel, the Gaussian noise variance σ_n^2, and the corresponding objective function value for the optimized models for the training set

The value of R^2 was determined by employing this optimized approach to the testing dataset. The unit Gpy, used to implement the Gaussian technique in python (GPy 2014; Ciaburro 2016; Stone 2016), in combination with the LBFGSB optimizer (Liu and Nocedal 1989; Byrd et al. 1994; Zhu et al. 1997; Fei et al. 2014), were employed to build the definitive regression approach.

Considering the calculations accomplished, the LBFGSB/GPR–relied technique has permitted to construct a model with high allowances to assess the Chl-a concentration

by means of the test dataset. Certainly, the value of R^2 of the best GPR approach was 0.8597 with a correlation coefficient *r* of 0.9306 for the dependent Chl-a variable.

Additionally, we have constructed models with other state of the art methods for comparison purposes such as the Multilayer Perceptron (MLP), Support Vector Regression (SVR) and Random Forest (RF) and, in all cases, the LBFGSB/GPR gets better results. The results can be seen in Table 5.

Table 5 Different error metrics (coefficient of determination (R^2), correlation coefficient (*r*), Mean Bias Error (MBE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)) for different regression methods (Gaussian Process Regression (GPR), Multilayer Perceptron (MLP), Support Vector Regression (SVR) and Random Forest (RF)) to construct the Chlorophyll model

A pictorial graph of the first-order and second-order terms that create the excellent GPR–relied approach for the Chl-a concentration is shown below in Figs. 4 and 5, respectively.

Fig. 4 First-order terms of the four more important input independent variables for the dependent Chl-a variable: **a** Chl-a vs. Chlorophytes; **b** Chl-a vs. Volume of water; **c** Chl-a vs. Secchi depth; and **d** Chl-a vs. *Dinophlagellata*

Fig. 5 Second-order terms of the four more important input independent variables for the dependent Chl-a variable: Chl-a vs. Chlorophytes and Volume of water; Chl-a vs. Chlorophytes and Secchi depth; Chl-a vs. Chlorophytes and *Dinophlagellata*; Chl-a vs. Volume of water and Secchi depth; Chl-a vs. Volume of water and *Dinophlagellata*; and Chl-a vs. Secchi depth and *Dinophlagellata*

3.1 Significance of Variables

A further relevant finding of this current study is the relevance of the input independent variables in order to foretell the Chl-a concentration for this complicated nonlinear research (see Table 6 and Fig. 6). The relevance of the input variables has been determined following the method proposed in Paananen et al. (2019). The authors argue that the traditional assessment of the relevance of the variables employing automatic relevance determination (ARD) (Seeger 2000) does not furnish a suitable technique because it automatically undervalues the importance of linear input variables in relation with nonlinear ones since they have the same significance in the calculation of the squared error (Piironen and Vehtari 2016). Instead, they suggest the determination of the posterior latent mean's variance. A big modification of the value of the latent mean if a single independent variable is altered means that this variable is relevant. We are estimating the relevance of the variable using the Variance of the Posterior Latent Mean (Paananen et al. 2019).

In this way, Chlorophytes concentration is the most relevant input variable according to LBFGSB/GPR approach in the Chl-a forecasting in Tanes reservoir. Next, the second significant input variable is volume of water followed by Secchi depth,

Dinophlagellata, Diatoms, Cyanobacteria, Chryptophytes, Euglenophytes, Total phosphorus and Nitrate concentration. The remaining input variables have a relative normalized relevance of less than 0.0001, which means that their effect is negligible.

Table 6 Relative normalized significance of the input variables in the Chl-a model following the Variance of the Posterior Latent Mean method (Note: only the variables with relevance greater than 0.0001 have been included)

Fig. 6 Relative significance of the input variables as stated in the LBFGSB/GPR–relied approach for the Chl-a concentration

Although Chlorophytes are not the main chlorophyll-containing organism in the reservoir –*is a meso-eutrophic ecosystem* (Álvarez Cobelas and Arauzo 2006) and therefore diatoms are the representative group– they have a higher biomass, higher than diatoms.

The Volume of water is the next significant input variable; obviously, taken into account that concentration is the solute mass (in this case, Chl-a mass) divided by total volume (solute+solvent volume) and, in this case, the solvent volume is the Volume of water.

Secchi depth is the third one in the ranking for Chl-a concentration forecasting. Secchi depth is a measure of the water turbidity (or transparency), so that as the chlorophyll

concentration increases, the algae concentration also increases and consequently the turbidity is higher (transparency decrease).

Dinophlagellata and Diatoms concentrations are other significant variables (the fourth and fifth ones in the ranking to predict Chl-a concentration, respectively) most due to the photosynthetic nature of those organisms, despite being heterotrophic some species of *Dinophlagellata*. Diatoms make the most representative group in this reservoir, with high cell densities.

Less important in Chl-a concentration forecasting is the Cyanobacteria concentration (sixth position in the ranking). Two reasons explain the correlation between Cyanobacteria and Chl-a concentrations (Vilán Vilán et al. 2013; Schinck et al. 2020). Firstly, because as phytoplankton biomass increase –Chl-a concentration rises–, the cyanobacteria predominance rises. Secondly, because Chl-a is present in all cyanobacteria as the main pigment for capturing sunlight and carrying out photosynthesis. High eutrophic environments are dominated by cyanobacteria but this reservoir is meso-eutrophic, and the presence of this kind of phytoplankton is low. However, it is a significant variable to predict Chl-a concentration since they are organisms containing Chl-a.

The next variable in the ranking (seventh place) is the Chryptophytes concentration. Chryptophytes, a class of freshwater or marine single-celled protists, are essentially photosynthetic. Although they can grow in short light conditions thanks to some

phycobiliprotein, they are more frequent in not high trophic states (Abirhire et al. 2015) as in Tanes reservoir case.

After Chryptophytes, Euglenophytes concentration is the following relevant input variable in Chl-a concentration forecasting. The presence of this kind of phytoplankton in Tanes reservoir *is* not relevant (according to its eighth position in the ranking to predict Chl-a concentration).

Total phosphorus and Nitrate concentrations are the ninth and tenth variables, respectively, in the rank for predicting Chl-a content. Both of them correlate to Chl-a concentration because they play, as nutrients, a role in phytoplankton growth (organisms with chlorophyll). In freshwater, Total phosphorus is, generally, the growth-limiting nutrient since nitrogen can be available from the atmosphere (Fields 2004; Moura Ado et al. 2012).

To conclude, conductivity and pH seem to be no remarkable role in the Chl-a concentration in Tanes reservoir.

Overall, in this study the GPR–relied technique is shown as an accurate and satisfactory tool to indirectly assess the Chl-a concentration (dependent variable), conforming to the real observed data in this reservoir, as a function of some easy and frequent measured parameters. Certainly, Fig. 7 indicates the comparison between the experimental and foretold Chl-a concentration values employing the GPR approach for the test dataset.

Consequently, it is essential to combine the GPR methodology with the LBFGSB optimizer to overcome this nonlinear regression problem through a novel hybrid approach that is sufficiently robust and effective. Conclusively, modeled and measured Chl-a values were found to be highly correlated.

Fig. 7 Observed vs. predicted Chl-a concentration values considering the confidence interval employing the LBFGSB/GPR–relied approach for the testing dataset ($R^2 = 0.8597$ and $r = 0.9306$)

4 Conclusions

Relying on the former results, several core discoveries of this study can be drawn and indicated as follows:

- First of all, it is important to note that analytical models currently used to foretell the Chl-a concentration from the observed values are not accurate enough because they greatly simplify a highly nonlinear complex problem. Consequently, the use of machine learning methods as the novel hybrid LBFGSB/GPR–relied approach employed in this study has revealed itself as the best choice to make an accurate estimation of the Chl-a concentration from experimental samplings in Tanes reservoir.
- In the second place, the hypothesis that the identification of the Chl-a concentration can be determined accurately by means of a hybrid LBFGSB/GPR–relied approach in reservoirs and lakes has also been validated here in Tanes reservoir.

- Thirdly, the application of this GPR–relied methodology to the complete experimental dataset belonging to the Chl-a concentration resulted in a satisfactory coefficient of determination and correlation coefficient whose values were 0.8597 and 0.9306, respectively.

- In fourth place, the ranking (or order of importance) of the input variables entailed in the estimation of the Chl-a concentration from experimental samplings in Tanes reservoir was established. This is one of the principal core conclusions in this research. The Chlorophytes concentration in particular must be taken into account as the most important issue in the forecasting of Chl-a concentration. On this matter, it is also noteworthy to emphasize the principal importance in this order of the water volume, Secchi depth, *Dinophlagellata*, Diatoms, Cyanobacteria, Chryptophytes, Euglenophytes, Total phosphorus and Nitrate concentration in the obtained Chl-a concentration outcome.

- Conclusively, the principal role of the accurate hyperparameters determination in the GPR–relied methodology in relation to the regression performance carried out for Chl-a concentration is established. The calculation of these optimal hyperparameters was successfully carried out here using the LBFGSB optimizer.

To sum up, this procedure could be successfully applied to foretell the Chl-a concentration in the different kinds of reservoirs. However, it is usually mandatory to consider individual characteristics for each reservoir and experiment. Hence, it is possible to conclude that the LBFGSB/GPR–relied method is a robust useful answer to the nonlinear problem of the estimation of the Chl-a concentration from experimental samplings in reservoirs and lakes.

Acknowledgements The authors gratefully recognize the computational help supplied by the Department of Mathematics at University of Oviedo as well as the monetary help of the Research Projects PGC2018-098459-B-I00 and FC-GRUPIN-IDI/2018/000221, both partial financing from European Founds (FEDER). Likewise, it is mandatory to express gratitude to Anthony Ashworth for his revision of English grammar and spelling of this investigation paper.

References

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

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

Akhlaghi YG, Sudong Z, Shittu S, Badiei A, Cattaneo MEGV, Xiaoli M (2019) Statistical investigation of a dehumidification system performance using Gaussian process regression. Energ Buildings 202:109406

Alghamdi AS, Polat K, Alghoson A, Alshdadi AA, Abd El-Latif AA (2020) Gaussian process regression (GPR) based non-invasive continuous blood pressure prediction method from cuff oscillometric signals. Appl Acoust 164:107256

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

Ambrogioni L, Maris E (2019) Complex-valued gaussian process regression for time series analysis. Signal Process 160:215–228

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

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

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 57:1574–1583

Byrd RH, Lu P, Nocedal J, Zhu C (1994) A limited-memory algorithm for bound constrained optimization. SIAM J Sci Comp 16:1190–1208

Cai H, Jia X, Feng J, Li W, Hsu Y–M, Lee J (2020) Gaussian Process Regression for numerical wind speed prediction enhancement. Renew Energ 146:2112–2123

Chen CW (1970) Concepts and utilities of ecologic model. J Sanit Eng Div 96:1086–1097

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

Daemi A, Kodamana H, Huang B (2019) Gaussian process modelling with Gaussian mixture likelihood. J Process Contr 81:209–220

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: Non equilibrium

systems in natural water chemistry, Advances in Chemistry Series, American Chemical Society, New York, vol 106, pp 131–150

Duan Y, Cooling C, Soo Ahn J, Jackson C, Flint A, Eaton MD, Bluck MJ (2019) Using a Gaussian process regression inspired method to measure agreement between the experiment and CFD simulations. Int J Heat Fluid Fl 80:108497

Dym H, McKean HP (2008) Gaussian processes, function theory, and the inverse spectral problem. Dover, New York

Ebden M (2015) Gaussian processes: a quick introduction. https://arxiv.org/pdf/1505.02965.pdf. Accessed 29 August 2015

Fei Y, Rong G, Wang B, Wang W (2014) Technical section: parallel L-BFGS-B algorithm on GPU. Comput Graph 40:1–9

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

Gao A, Liao W (2019) Efficient gravity field modeling method for small bodies based on Gaussian process regression. Acta Astronaut 157:73–91

Gonçalves IG, Echer E, Frigo E (2020) Sunspot cycle prediction using warped Gaussian process regression. Adv Space Res 65(1):677–683

GPy (2014) A Gaussian process framework in python. http://github.com/SheffieldML/GPy. Accessed 25 January 2014

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

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

Lantz B (2019) Machine learning with R: expert techniques for predictive modeling. Packt Publishing, Birmingham, UK

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, Yuan C, Li X, Wang Z (2020) State of health estimation for Li-Ion battery using incremental capacity analysis and Gaussian process regression. Energy 190:116467

Li M, Sadoughi M, Hu Z, Hu C (2020) A hybrid Gaussian process model for system reliability analysis. Reliab Eng Syst Safe 197:106816.

Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. Math Program 45:503–528

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

Moura Ado N, do Nascimento 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

Negro AI, de Hoyos C, Vega JC (2000) Phytoplankton structure and dynamics in Lake Sanabria and Valparaíso reservoir (NW Spain). Hydrobiologia 424:25–37

Paananen T, Piironen J, Andersen MR, Vehtari A (2019) Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), Proceedings of Machine Learning Research (PMLR), Naha, Okinawa, Japan, pp 1743–1752

Pasolli L, Melgani F, Blanzieri E (2010) Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing data. IEEE Geosci Remote S 7(3):464–468

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 National Acad Sci 64(2):472–478

Piironen J, Vehtari A (2016) Projection predictive model selection for Gaussian processes. In: IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE Publisher, Vietri sul Mare, Italy, pp 1–6

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

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

Rasmussen CE (2003) Gaussian processes in machine learning: summer school on machine learning. Springer, Berlin

Rasmussen CE, Williams CKI (2005) Gaussian processes for machine learning. The MIT Press, Cambridge, MA, USA

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

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

Rychlik I, Johannesson P, Leadbetter MR (1997) Modelling and Statistical Analysis of ocean-wave data using transformed Gaussian processes. Mar Struct 10(1):13–47

Schinck M-P, L'Ecuyer-Sauvageau C, Leroux J, Kermagoret C, Dupras J (2020) Risk, drinking water and harmful algal blooms: a contingent valuation of water bans. Water Resour Manag 34:3933–3947

Seeger M (2000) Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. In: NIPS'99 Proceedings of the 12th International Conference on Neural Information Processing Systems, MIT Press Cambridge, MA, USA, vol 12, pp 603–609

Sheldon SP, Skelly DK (1990) Differential colonization and growth of algae and ferromanganese-depositing bacteria in a mountain stream. J Freshwater Ecol 5(4): 475–485

Smith VH (2006) Responses of estuarine and coastal marine phytoplankton to nitrogen and phosphorus enrichment. Limnol Oceanogr 51:377–384

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

Stone JV (2016) Bayes' rule with python: a tutorial introduction to Bayesian analysis. Sebtel Press, London

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

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

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

Vilán Vilán JA, Alonso Fernández JR, García Nieto PJ, Sánchez Lasheras F, de Cos Juez FJ, Díaz Muñiz C (2013) Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain). Water Resour Manag 27:3457–3476

Wang S, Zhu L, Ying Hsi Fuh J, Zhang H, Yan W (2020) Multi-physics modeling and Gaussian process regression analysis of cladding track geometry for direct energy deposition. Opt Laser Eng 127:105950

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

Wetzel RG (2011) Limnology: lake and river ecosystems. Academic Press, San Diego, USA

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

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

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

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

Zeng A, Ho H, Yu Y (2020) Prediction of building electricity usage using Gaussian Process Regression. J Build Eng 28:101054

Zhu C, Byrd RH, Lu P, Nocedal J (1997) Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM T Math Softw 23(4):550–560

Table 1 Set of biological input independent variables used in this investigation: names, means and standard deviations

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

Table 2 Set of physico-chemical input independent variables used in this investigation: names, means and standard deviations

Physico-chemical input variables	Name of the variable	Mean	Std
Water temperature (ºC)	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 Search space for the three hyperparameters of the LBFGSB/GPR–relied approach fitted in this investigation

GPS hyperparameters	Lower limit	Upper limit
RBF kernel variance σ_f^2	10^{-3}	10^5
RBF kernel lengthscale ℓ	10^{-3}	10^5
Gaussian noise variance σ_n^2	10^{-3}	10^5

Table 4 Optimal hyperparameters of the best fitted GPR–relied model encountered with the LBFGSB optimizer: variance σ_f^2 and lengthscale ℓ for the RBF kernel, the Gaussian noise variance σ_n^2, and the corresponding objective function value for the optimized models for the training set

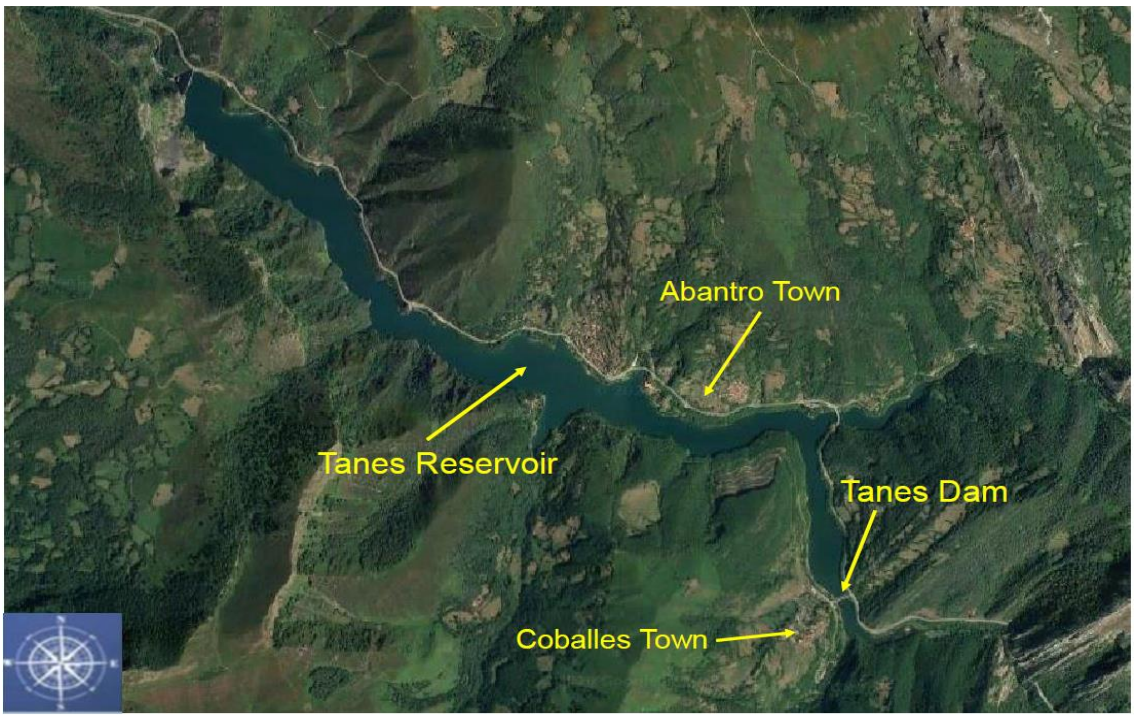
	σ_f^2	ℓ	σ_n^2	Objective fun. value
Chlorophyll	1.6212	3.2870	0.0912	197.42

Table 5 Different error metrics (coefficient of determination (R^2), correlation coefficient (r), Mean Bias Error (MBE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)) for different regression methods (Gaussian Process Regression (GPR), Multilayer Perceptron (MLP), Support Vector Regression (SVR) and Random Forest (RF)) to construct the Chlorophyll model

Model	R^2	r	MBE	MAE	RMSE
LBFGSB/GPR	0.8597	0.9306	0.1419	0.7862	1.1514
MLP	0.8239	0.9081	0.0599	0.9228	1.2896
SVR	0.7286	0.8647	0.0227	1.2388	1.6011
RF	0.4367	0.6720	0.1666	1.6504	12.128

Table 6 Relative normalized significance of the input variables in the Chl-a model following the VAR method (Note: only the variables with relevance greater than 0.0001 have been included)

Variable	Relevance
Chlorophytes	1.0000
Volume of water	0.3781
Secchi Depth	0.3580
Dinophlagellata	0.1397
Diatoms	0.1294
Cyanobacteria	0.0944
Chryptophytes	0.0625
Euglenophytes	0.0540
Total phosphorus	0.0251
Nitrate concentration	0.0203

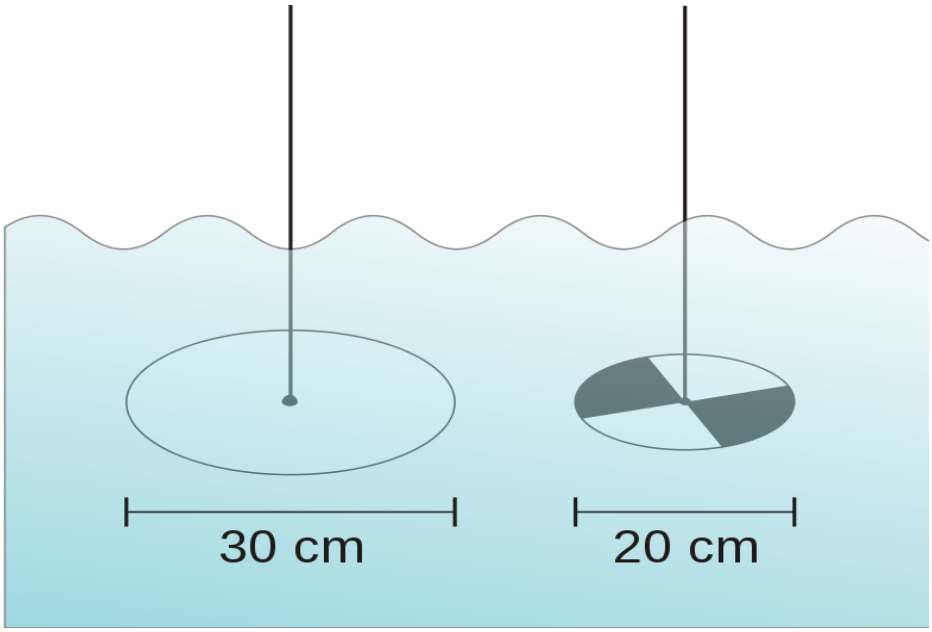


a

b

Fig. 1 a Aerial view of the Tanes reservoir; **b** a closer view of the Tanes reservoir

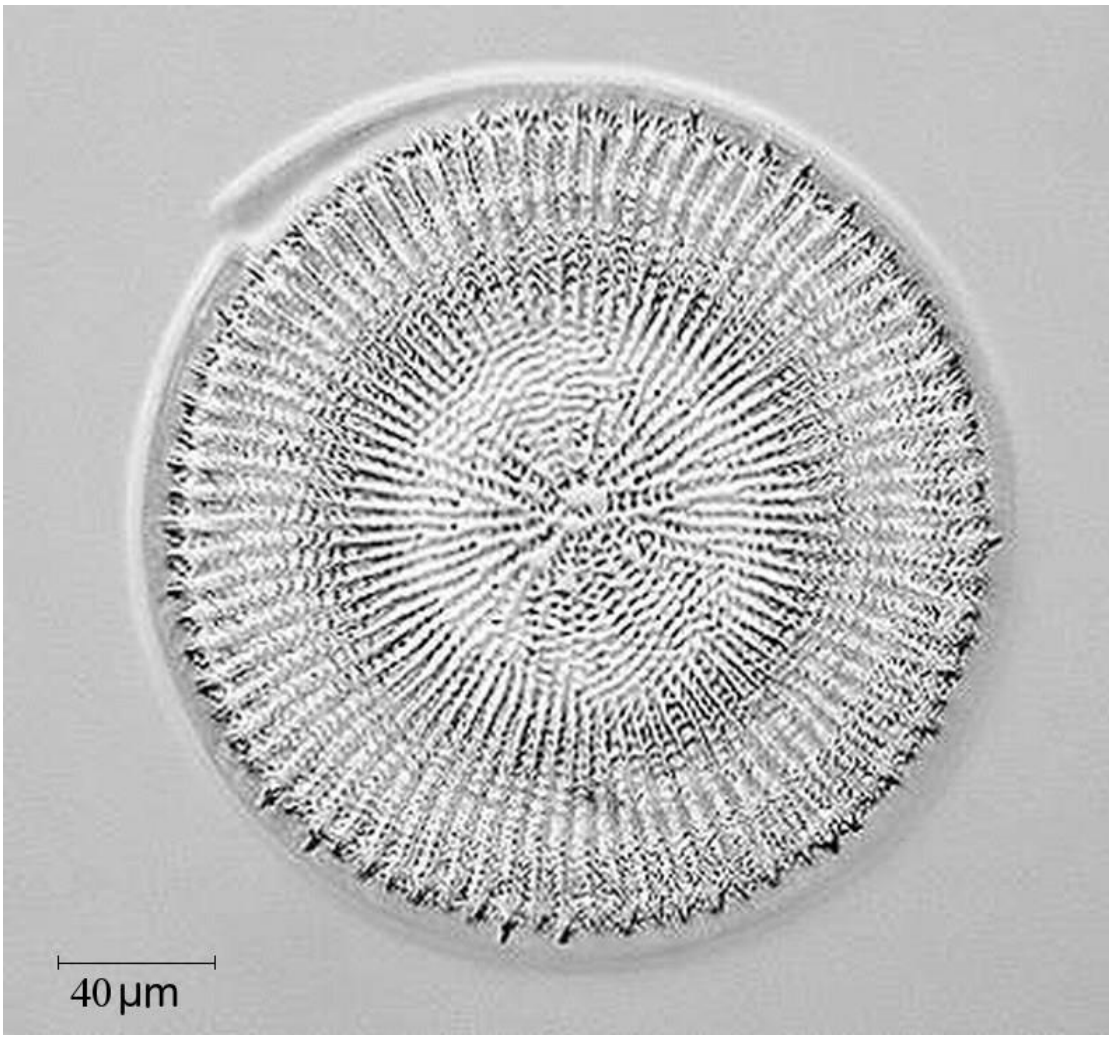
a



b

Fig. 2 a A Niskin bottle; **b** examples of Secchi disks

a

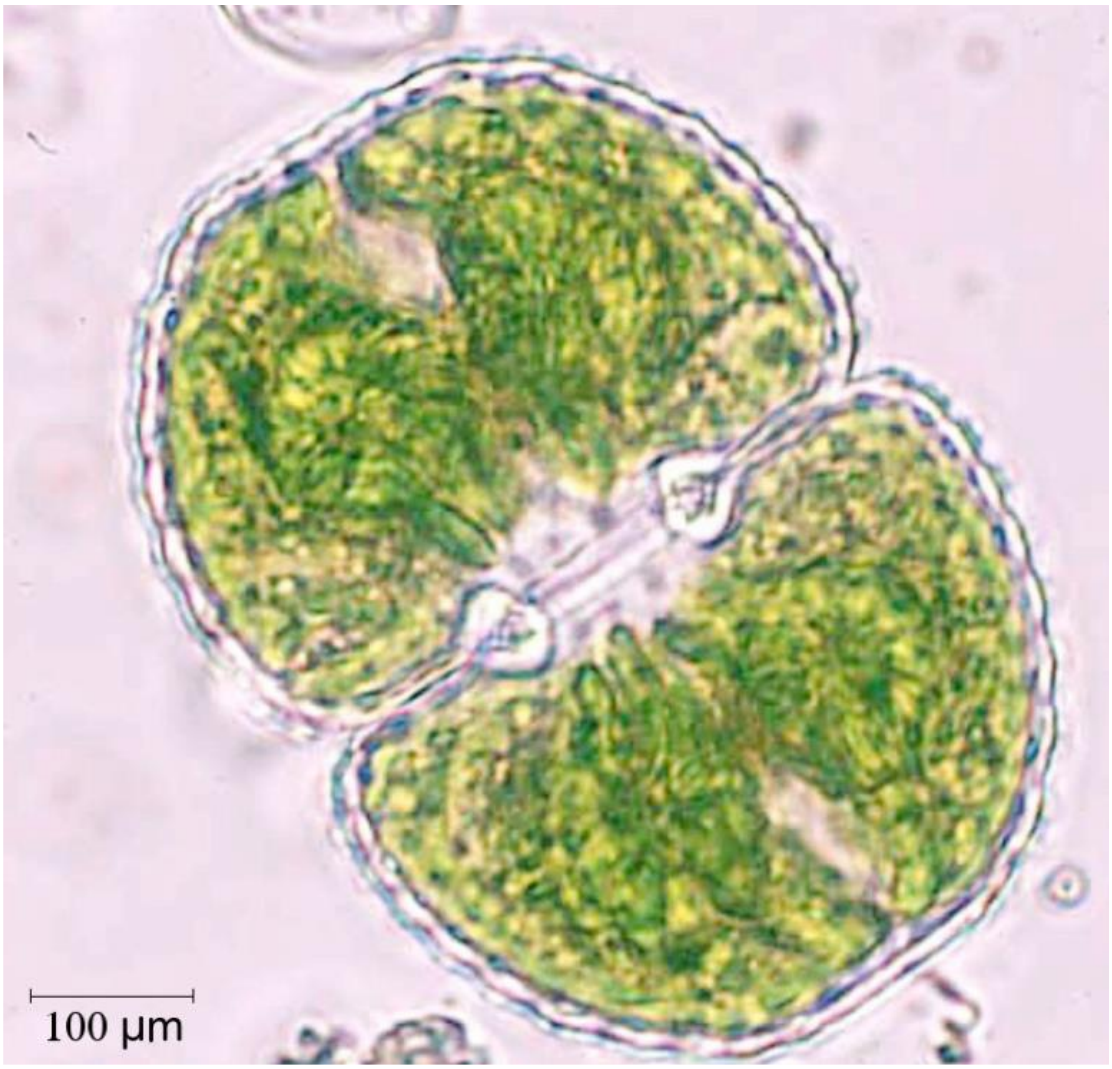


b

c

d

e



f

g

Fig. 3 Microorganisms in Tanes reservoir whose concentrations were used as input variables: **a** Cyanobacteria; **b** Diatoms; **c** Euglenophytes; **d** *Dinophlagella*; **e** Chrysophytes; **f** Clorophytes; and **g** Chryptophytes

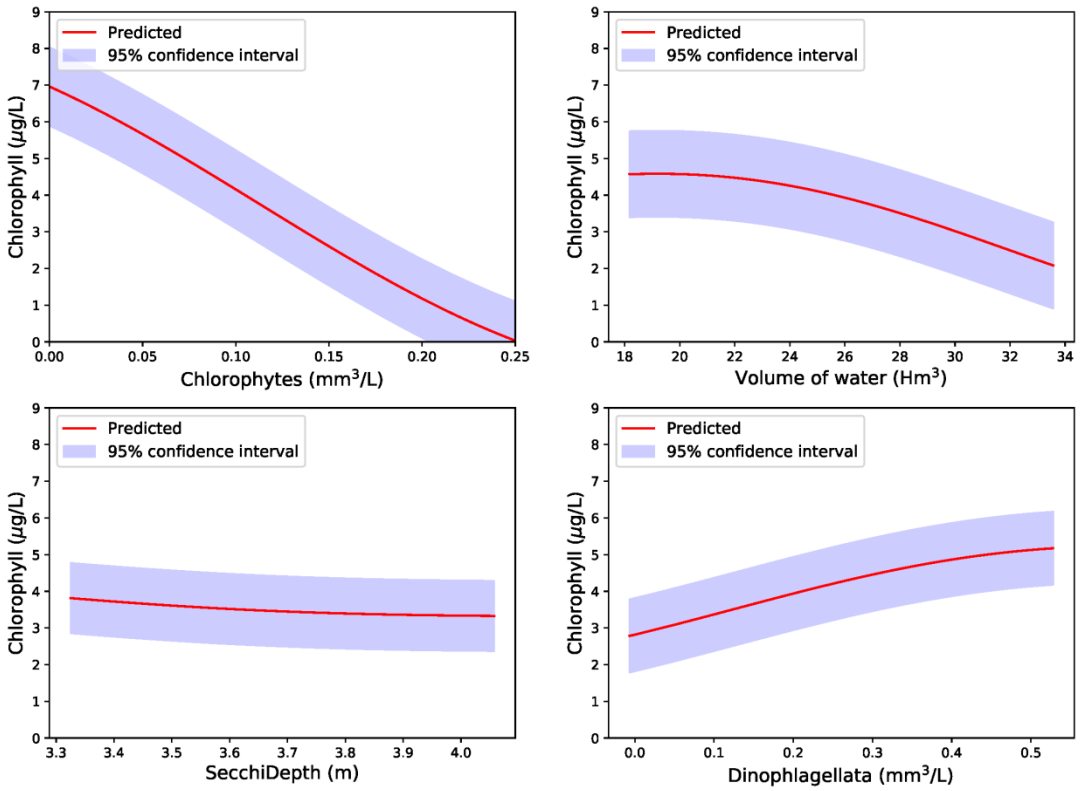


Fig. 4 First-order terms of the four more important input independent variables for the dependent Chl-a variable: **a** Chl-a vs. Chlorophytes; **b** Chl-a vs. Volume of water; **c** Chl-a vs. Secchi depth; and **d** Chl-a vs. *Dinophlagellata*

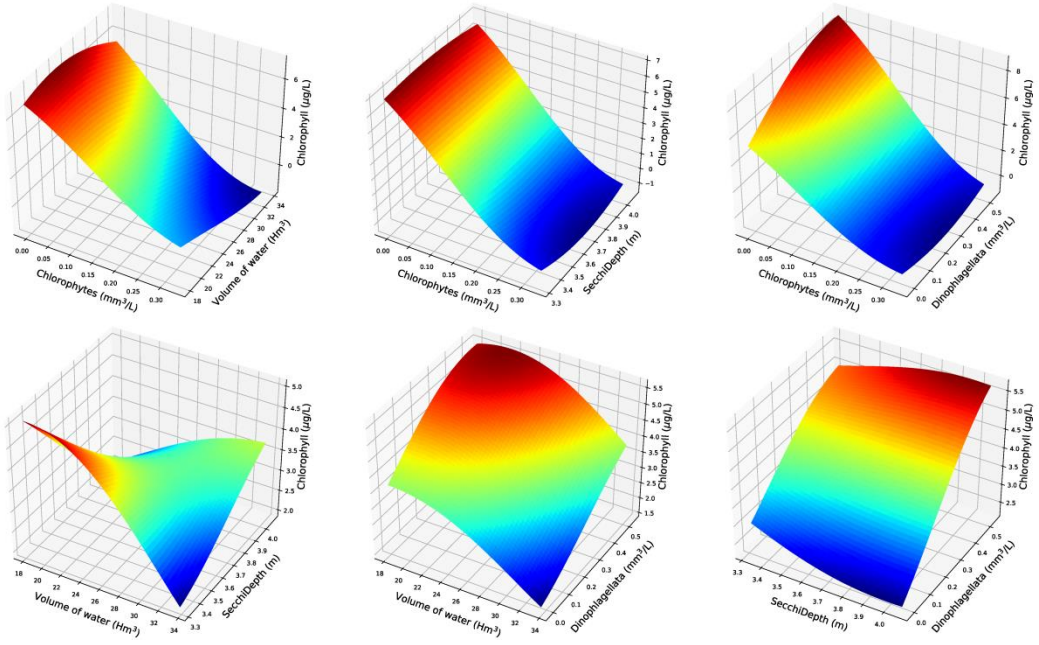


Fig. 5 Second-order terms of the four more important input independent variables for the dependent Chl-a variable: Chl-a vs. Chlorophytes and Volume of water; Chl-a vs. Chlorophytes and Secchi depth; Chl-a vs. Chlorophytes and *Dinophlagellata*; Chl-a vs. Volume of water and Secchi depth; Chl-a vs. Volume of water and *Dinophlagellata*; and Chl-a vs. Secchi depth and *Dinophlagellata*

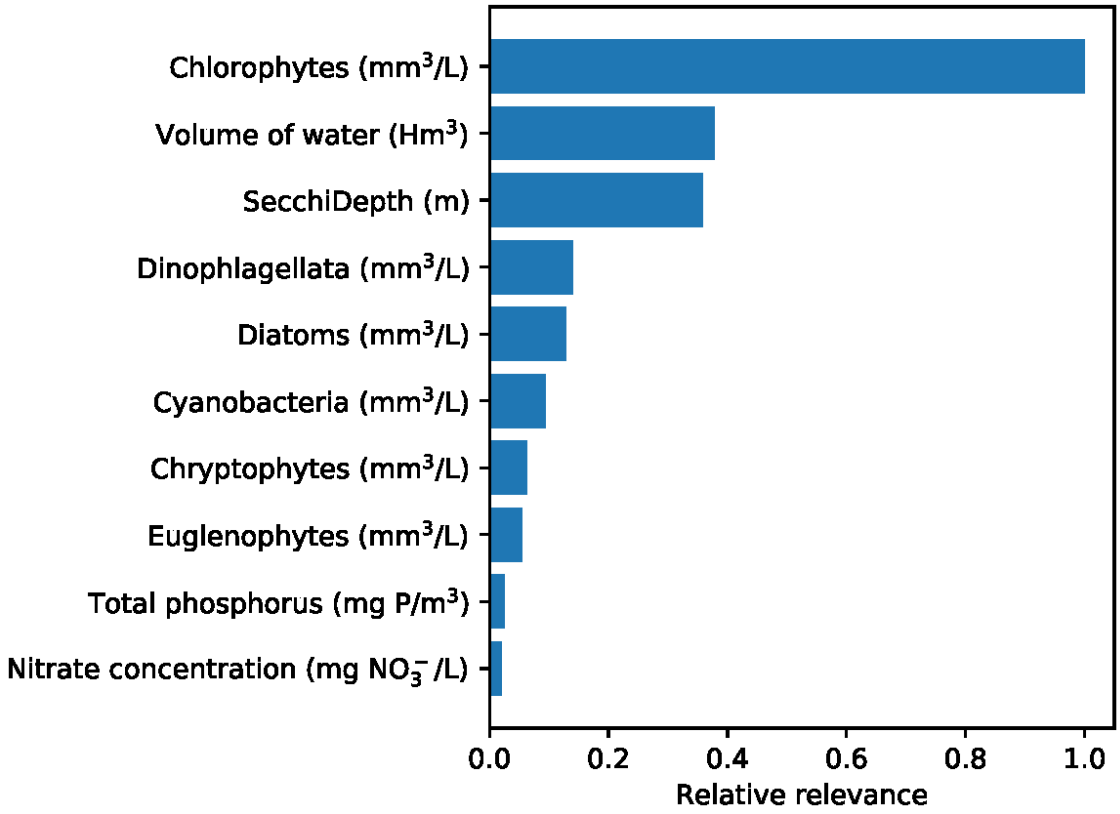


Fig. 6 Relative significance of the input variables as stated in the LBFGSB/GPR–relied approach for the Chl-a concentration

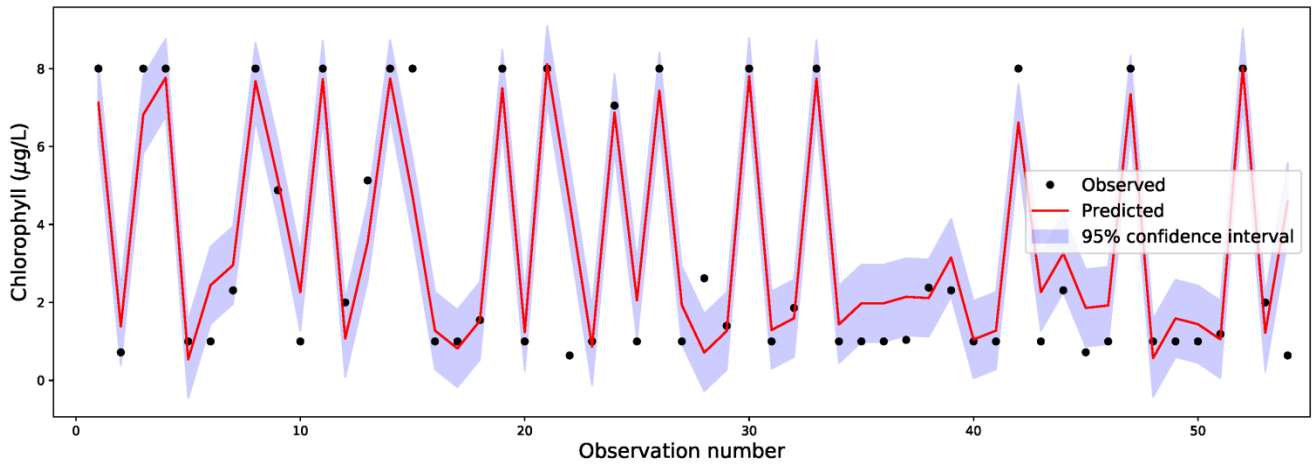


Fig. 7 Observed vs. predicted Chl-a concentration values considering the confidence interval employing the LBFGSB/GPR–relied approach for the testing dataset $\left(R^2 = 0.8597 \text{ and } r = 0.9306\right)$