Environmental Science and Pollution Research (2022) 29:74967–74982
https://doi.org/10.1007/s11356-022-21168-z

RESEARCH ARTICLE

Uncovering the influence of hydrological and climate variables in chlorophyll-A concentration in tropical reservoirs with machine learning

Taís Maria Nunes Carvalho[1] · Iran Eduardo Lima Neto[1] · Francisco de Assis Souza Filho[1]

Received: 18 February 2022 / Accepted: 25 May 2022 / Published online: 1 June 2022

Abstract
Climate variability and change, associated with increasing water demands, can have significant implications for water availability. In the Brazilian semi-arid, eutrophication in reservoirs raises the risk of water scarcity. The reservoirs have also a high seasonal and annual variability of water level and volume, which can have important effects on chlorophyll-a concentration (Chla). Assessing the influence of climate and hydrological variability on phytoplankton growth can be important to find strategies to achieve water security in tropical regions with similar problems. This study explores the potential of machine learning models to predict Chla in reservoirs and to understand their relationship with hydrological and climate variables. The model is based mainly on satellite data, which makes the methodology useful for data-scarce regions. Tree-based ensemble methods had the best performances among six machine learning methods and one parametric model. This performance can be considered satisfactory as classical empirical relationships between Chla and phosphorus may not hold for tropical reservoirs. Water volume and the mix-layer depth are inversely related to Chla, while mean surface temperature, water level, and surface solar radiation have direct relationships with Chla. These findings provide insights on how seasonal climate prediction and reservoir operation might influence water quality in regions supplied by superficial reservoirs.

Keywords Chlorophyll-a · Machine learning · Tropical lakes · Water quality · Climate variability

Introduction

In most developing countries, the urbanization process is associated with an increase in water demand (UNESCO World Water Assessment Program 2018). At the same time, the availability of drinking water remains the same or even decreases (Veldkamp et al. 2017; Greve et al. 2018). Accelerated urbanization is also related to the intensification of human activity, resulting in increased nutrient loads and water quality degradation (Vörösmarty et al. 2010).

The situation is worse in regions with high climatic variability (temporal and spatial), in which the distribution of rainfall is irregular, and extreme events of droughts and floods are frequent (Easterling et al. 2000; Hirsch and Archfield 2015). This is the case in the Northeastern semi-arid region of Brazil, where multi-annual drought events are common and have severe socioeconomic and environmental impacts (Campos 2015; Pontes Filho et al. 2020). One of the management strategies historically adopted in the region to deal with this scenario is the construction of reservoirs (Gutiérrez et al. 2014), which have the important role of transferring water both temporally and spatially. Most of these reservoirs serve multiple purposes, including drinking water supply, irrigation, and fish farming. The water volume in these reservoirs can vary significantly between the dry and wet seasons and reduce drastically during drought periods (Rocha and Lima Neto 2021a).

Eutrophication, caused by the excessive increase of phosphorus and nitrogen loads, is one of the main causes of the deterioration of water quality in reservoirs (Paerl and Otten 2013). Eutrophication is associated with the proliferation of algae and cyanobacterial blooming (Yang et al. 2008), and sometimes, an increase in mortality of benthic animals and fish (Sperling 2005). Agriculture and livestock farming

Responsible Editor: Marcus Schulz

✉ Iran Eduardo Lima Neto
iran@deha.ufc.br

[1] Department of Hydraulic and Environmental Engineering, Universidade Federal Do Ceará, Campus do Pici, Bloco 713, Fortaleza, CEP 60455-760, Brazil

contribute to this process since significant loads of phosphorus and nitrogen can be carried with surface water runoff into the reservoir (Wiegand et al. 2020; Rocha and Lima Neto 2021; Lima Neto et al., 2022).

A few studies have associated phytoplankton growth rates with the volume of water stored in the reservoir (Pacheco and Lima Neto 2017; da Rocha Junior et al. 2018), but most of them relied on field studies, which are usually unavailable for a long-term horizon (more than 10 years), especially in data-scarce regions. Other researchers have related chlorophyll-a concentrations (Chla) to hydrological and/or climate variables, such as wind speed, air temperature, solar radiance, precipitation, mixing depth, and runoff (Blauw et al. 2018; Stockwell et al. 2020; Stefanidis et al. 2021), but none of them analyzed this relationship in tropical reservoirs. Past research has also shown that climate variability and future changes in frequency and intensity of drought events can increase phosphorus concentrations in tropical reservoirs (Raulino et al. 2021; Rocha and Lima Neto 2021a), hence the importance of investigating the relationship between climate variables and Chla.

The mechanisms associated with Chla fluctuations are complex and have been extensively studied (Pacheco and Lima Neto 2017; Blauw et al. 2018; Dunstan et al. 2018; Li et al. 2021), and more recently, many researchers have applied machine learning techniques for water quality assessment and to predict Chla (Liu et al. 2019; Shen et al. 2019; Najah Ahmed et al. 2019; Tong et al. 2019; Mamun et al. 2019; Nguyen et al. 2020; Yu et al. 2020). Data for most of these studies have been obtained from automated stations (Blauw et al. 2018) or long field campaigns (Liu et al. 2019; Najah Ahmed et al. 2019; Li et al. 2021), which can be expensive and time consuming. One strategy to deal with the lack of field data is using satellite data, which has been frequently used to monitor water quality and has proved to be reliable, but it has not been sufficiently explored for inland waters (Lopes et al. 2014; Gholizadeh et al. 2016; Wang and Yang 2019; Ross et al. 2019; Nguyen et al. 2020; Iiames et al. 2021).

Recent evidence suggests that reanalysis climate data can be effective in explaining the effects of climate on phytoplankton biomass (Stefanidis et al. 2021). However, to the authors' knowledge, no study has explored the predictive capacity of non-parametric models based on reanalysis climate data for semiarid climates. In these regions, Chla modeling can be challenging, as water volume has a strong interannual variability and phosphorus concentration has a weak correlation with Chla. The state-of-the art models used to explore the mechanisms for Chla variability may not be suitable for them. Machine learning models can be informative in this case, but model comparison is required, as these algorithms are mainly driven by data and their predictive capacity can be site-specific.

This study evaluates the influence of hydrological and climate variables on Chla in reservoirs located in Northeastern semi-arid Brazil. This analysis is important from the point of view of climate variability, which can significantly affect the hydrological processes of the reservoirs, and to understand the possible influence of water level and volume fluctuations on Chla. The predictive model proposed here combines climate reanalysis data, together with commonly available hydrological variables, and satellite-based predictions of Chla. The main goals of this study are (i) to explore the relationships between hydrological and climate variables and the concentration of Chla in tropical reservoirs and (ii) to evaluate the performance of nonparametric machine learning models for predicting Chla using these variables.

Materials and methods

Study area

The reservoirs analyzed in this study are located in the Northeastern region of Brazil (Fig. 1), which has a semi-arid climate and is frequently affected by multi-annual droughts. These reservoirs are part of the Jaguaribe-Metropolitano water supply system, which transfers water to Fortaleza, the capital of the State of Ceará. Castanhão is the largest reservoir for multiple uses in the country, with a capacity of 6.7 billion cubic meters. All three reservoirs are also used for irrigation. Banabuiú (capacity of 1.6 billion cubic meters) supplies the Irrigated Perimeter Morada Nova, while Orós (capacity of 2.1 billion cubic meters), the second-largest reservoir in the State of Ceará, also serves for hydroelectric use. The surface area of these reservoirs ranges between 116 and 410 km^2, and the mean water level from 90 to 192 m.

Data and variable selection

This research uses data from publicly available databases, obtained from satellite, reanalysis, and rain gauge stations. The historical series of monthly chlorophyll-a concentrations (Chla) from 2002 to 2019 were obtained from the Hidrosat portal (http://hidrosat.ana.gov.br/). The dataset obtained from Hidrosat is the result of a partnership between the Brazilian Water Agency (ANA) and the Research Institute for Development (Institut de Recherche pour le Développement, IRD). Water quality stations use data from the Terra (EOS AM) and Aqua (EOS PM) satellites.

The program MOD3R (MODIS Reflectance Retrieval over Rivers) is used to extract time series of reflectance from MODIS (sensor onboard the Terra and Aqua satellites) images of water bodies. The algorithm identifies and groups the water pixels in the image and, from the extraction of reflectance values from the visible and infrared bands,

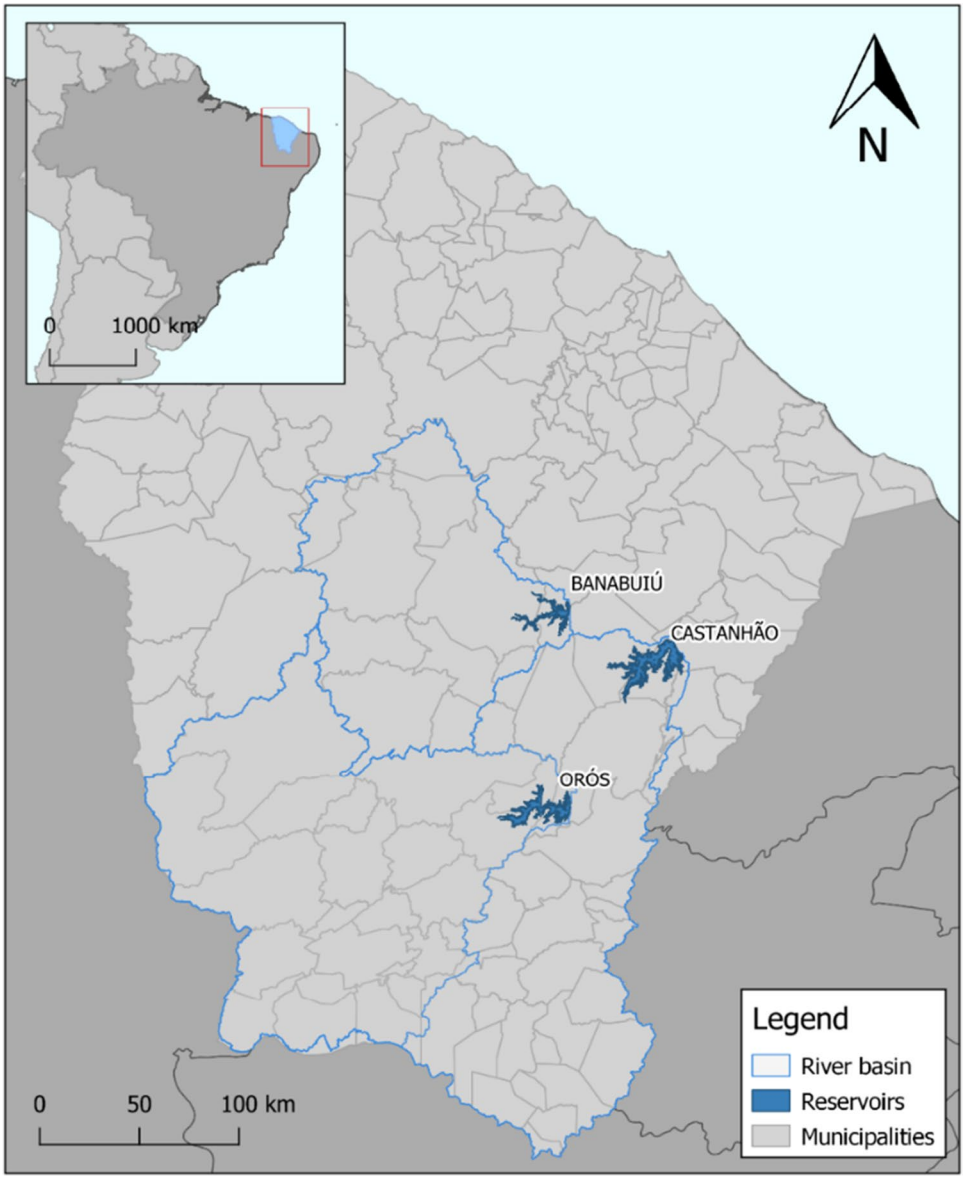

Fig. 1 Study area location. Banabuiú, Castanhão, and Orós are the main reservoirs of the State of Ceará, Brazil (highlighted in the map). Their hydrographic basins are contoured by the blue line

the water quality parameters are estimated. Mathematical models that relate reflectance data and water quality data were calibrated and validated with data collected in the field. This procedure is detailed in Lins et al. (2017).

For some months of the original series of Chla, more than one estimation was available. In these cases, the median of these values was used to represent monthly concentration. Months with missing values were filled in with the median of the historical concentration series for the corresponding month.

Hydrological and climate variables used in this research and their respective sources are described in Table 1. Precipitation data for the period between 2002 and 2019 were obtained from the spatial interpolation of the data provided by the Brazilian Water Agency, publicly available on the Hidroweb portal (http://www.snirh.gov.br/hidroweb/). Daily precipitation measured in rain gauges was interpolated using the inverse distance weighting method with exponent two into grid points with 0.05° size. This procedure was performed using the R package ipdw (Stachelek 2020). Then, the average monthly precipitation was calculated for each reservoir's hydrographic basin.

Average monthly temperature data was extracted from version 4 of the University of East Anglia's Climatic Research Unit (CRU) climate database (Harris et al. 2020). Data is publicly available in the NetCDF format, which stores multidimensional variables; for example, temperature has four dimensions: latitude, longitude, time, and temperature value. To estimate average monthly temperature over the reservoir, we extracted the pixels contained inside the limits of the reservoir and calculated its average value for each month in the time series (2002–2019).

Table 1 Explanatory variables of the regression models. Except for the variables extracted from the sources indicated with an asterisk "*" (which are available in tabular format), all other variables were obtained in NetCDF format

Variable	Unit	Description	Source	Mean	Standard deviation
Mean precipitation	mm	Average monthly precipitation on the hydrographic basin of the reservoir, calculated from rain gauge measures	Hidroweb*	61.59	74.74
Mean temperature	°C	Mean surface temperature over the reservoir calculated from CRU grid	CRU (Harris et al. 2020)	27.78	1.23
Water volume	m^3	Total water volume in the reservoir	COGERH*	1.42E + 09	1.56E + 09
Water level	m	Distance from the bottom of the reservoir to the water surface	COGERH*	137.18	43.99
Runoff	m	Monthly average of surface and subsurface runoff accumulated over one day in the hydrographic basin	ERA5	1.75E-04	3.59E-04
2 m temperature	K	Air temperature at 2 m above the reservoir	ERA5	300.90	1.29
Lake bottom temperature	K	Water temperature at the bottom of the reservoir	ERA5	299.00	1.20
Lake mix-layer depth	m	Thickness of the uppermost layer of the reservoir that is well mixed and has a near constant temperature	ERA5	5.19	1.48
Surface net solar radiation	Jm^{-2}	Amount of solar radiation that reaches the water surface, assuming cloudless conditions	ERA5	1.78E + 07	2.42E + 06
10 m u-component of wind	ms^{-1}	Horizontal wind speed of air moving towards the east, at a height of 10 m above the reservoir surface	ERA5	-2.43	0.80
Reservoir	Dummy	Represents the reservoir correspondent to the observation	-	-	-
Drought year	Binary	Indicates if the year of the observation was a drought year (1) or not (0)	-	0.37	0.48

Except for water volume and level, all other variables were extracted from the ERA5 gridded (lat-lon grid of 0.25°) reanalysis database of the European Center for Medium-Range Weather Forecasts (Hersbach et al., 2020). Data is also available online in the NetCDF format, in hourly or monthly scale, with a temporal coverage from 1979 to present. Reanalysis uses observed data from weather stations across the world and climate models to estimate a global dataset containing atmospheric, land, and oceanic climate variables.

Average runoff was calculated by averaging the monthly runoff for all pixels contained in the region delimited by each reservoir's hydrographic basin. For all other variables, the time series was extracted for the nearest pixel to the centroid of the reservoir, which was identified using the nearest-neighbor interpolation method. Water volume and level were obtained from the Water Resources Management Company of Ceará (COGERH), also available online on the Reservoir Monitoring System (https://www.ana.gov.br/sar).

Further improvements can be made by validating reanalysis data with field data and by incorporating more reservoirs into the analysis. However, this would require field campaigns and/or the implementation of automatic monitoring systems.

Variables that had a Pearson's correlation coefficient above 0.8 were removed from the dataset (temperature at 2 m and runoff; refer to Fig. S1 in the supplementary material for the correlation matrix). As the effect of hydrological variables can be site-specific, a dummy variable was included to indicate the corresponding reservoir of each observation. To account for the effect of drought on Chla, a binary variable was included to indicate if the observation was registered during a drought year, according to drought records of the area (Pontes Filho et al. 2020).

All explanatory variables were re-scaled to range between 0 and 1 using the min–max normalization:

$$x = \frac{x - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}$$

where x is the original value and x is the scaled value. The final dataset contained 679 samples from the three reservoirs analyzed in this study. All analyses were performed using R (version 4.0.5) software.

Regression models

Six nonparametric machine learning models were compared with standard linear regression and one semi-parametric algorithm to investigate the best-performing predictive model. Data were randomly split into training (80%) and testing (20%) datasets. The training dataset was used to tune model hyperparameters, and the testing dataset was used to evaluate model performance. Model tuning and performance evaluation are detailed in the "Model parameters and performance evaluation" section.

In the following topics, there is a brief explanation of the regression models used in this study. It is important to highlight an essential property of the predictive models, which is the bias-variance tradeoff. When fitting regression models, the best outcome is obtaining a model that not only provides accurate predictions (low bias) but also generalizes well to new data (low variance). The bias error is associated with a poor learning process, in which the relationship between explanatory and response variables is not properly captured (underfitting). The variance error happens when the model is sensitive to small variations during training, i.e., fits too perfectly and ends up modeling random noise (overfitting). One wants to avoid models that are either too complex or too simple and get the one that presents similar performances during training and testing.

Linear regression model

Linear regression aims to explain the relationship between a set of independent variable vectors (x) and a dependent variable (y) based on the linear function described below:

$$\widehat{Y} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

where X_j is a vector for the jth independent variable, and β_j and β_0 are unknown parameters (coefficients and an intercept, respectively). The algorithm calculates the parameters by minimizing the sum of the squares of the residuals (SSR), i.e., the difference between observed and predicted values.

Elastic-net regularized generalized linear model

While in the ordinary least squares regression, the distribution of errors is normal, in the generalized linear model (GLM), it may assume different distributions, such as binomial, Poisson, and gamma. In GLMs, the variance of the response variable can be non-constant and a linking function can be used to connect the predictor and the mean of the distribution function (Nelder & Wedderburn, 1972). In this study, the error distribution was assumed to be normal.

Regularization is a useful technique for learning algorithms: penalties can be added to the model to prevent overfitting issues and to deal with highly correlated explanatory variables. Ridge and Lasso regression are some of the simplest and widely used penalized models; they work by adding a penalty to the SSR. Lasso penalizes the sum of the absolute coefficients ($\ell 1$ penalty) and might lead to variable selection as it sets coefficients to zero if λ is sufficiently large. The parameter λ controls the regularization strength and might assume any positive value.

$$SSR_{\text{lasso}} = \sum_{j=1}^{n} (y_i - \widehat{y}_i)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|$$

where y_i is the observed value, $\widehat{y}_i$ is the predicted value, n is the number of samples, β is the coefficient vector, and p is the number of explanatory variables. Ridge regression penalizes the square of the magnitude of the coefficients ($\ell 2$ penalty) and shrinks the coefficients proportionally, keeping all of the variables in the model:

$$SSR_{\text{ridge}} = \sum_{j=1}^{n} (y_i - \widehat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

The linear combination of both penalties is called elastic net regularization, controlled by the parameter α, which ranges between 0 (ridge) and 1 (lasso).

Artificial neural network

An artificial neural network is composed of interconnected nodes (or neurons) arranged in layers (Hastie et al., 2009). The multilayer perceptron (MLP), a broadly used class of neural networks, consists of the input (which receives the independent vectors), output, and one or more hidden layers. These layers have weighted connections that are adjusted as training occurs and are fully connected, i.e., a neuron in one layer is connected to every neuron in the next layer. The number of neurons in the hidden layer is critical for the learning process, as they detect the characteristics present in the training data and apply a nonlinear transformation to the input data.

The training algorithm used in this study was the back-propagation of the error, in which the gradient of the error concerning the weights is calculated layer by layer. Then, the error is calculated, and all weights are updated backward through the network. The optimization algorithm used to perform this method was gradient descent.

An MLP with a single hidden layer was selected and the number of hidden nodes was adjusted in the training process (see Table 2). The number of nodes in the input layer was set to 10 (the number of explanatory variables), and the learning rate was set to 0.1.

k-nearest neighbors

The k-nearest neighbors (KNN) is a supervised algorithm (Altman, 1992) for classification and regression based on a similarity measure, such as distance functions. In this method, one finds the k observations in the training set closest to x and (i) average their responses, for regression tasks or (ii) take the majority class among its k nearest neighbors, for classification tasks. The equation for the KNN fit for Yˆ can be described as:

Table 2 Main parameters of the regression models used in this study. The values used to tune the models are indicated, and the chosen values are highlighted in bold

Model	Main parameters	Values
Linear regression model	Intercept	**True** or False
Regularized generalized linear model	Alpha	**0.10**, 0.28, 0.46, 0.64, 0.82, and 1.00
	Lambda	0.0046, 0.0173, 0.0646, 0.2409, **0.8979**, and 3.3469
Multilayer perceptron	Number of nodes in the hidden layer	**3**, 5, 10, and 20
	Decay	**0.5**, 0.1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, and 1e-7
k-nearest neighbors	Number of neighbors (k)	5, 7, **9**, 11, 13, and 15
Regression tree	Complexity parameter (cp)	**0.0274**, 0.0342, 0.0390, 0.0773, 0.1400, and 0.2066
Random forest	Number of variables for splitting (mtry)	2, 4, **6**, 8, 10, and 12
	Number of trees	**50**, 100, 250, 300
Gradient boosting	Shrinkage	0.1
	Interaction depth	1, 2, 3, 4, 5, and **6**
	Minimum observations in node	10
	Number of trees	**50**, 100, 150, 200, 250, and 300
Support vector machine	Cost parameter (C)	0.25, 0.50, 1.00, 2.00, **4.00**, and 8.00
	Sigma	0.0619
	Epsilon (ε)	0.1

$$\widehat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

where N_k is the neighborhood of x defined by the k closest points x_i in the training sample. The only parameter to be determined is the number of neighbors k.

Classification and regression tree

A decision tree provides a set of rules to express the relationship between explanatory and response variables, which are represented with a tree structure. The leaves represent class labels (classification) or estimations of the response variable (regression), and branches represent the values of the tested variable.

Regression trees predict using the average values of $\overline{y}$ within each subset, which is selected to minimize the mean square error, $MSE = \sum_i (\overline{y} - y_i)^2 / n$. To determine whether splitting should continue to be done, one can use some combination of (i) a minimum number of points in a node, (ii) purity or error threshold of a node, or (iii) maximum depth of the tree (Krzywinski & Altman, 2017). Here, the minimum number of points per node was set to 20. The complexity parameter, which corresponds to the minimum improvement in the model needed at each node, was tuned using grid search (see Table 2).

Tree-based ensemble models: random forest and gradient boosting regression

Decision trees alone can easily overfit, depending on the size of the training dataset. An ensemble of decision trees is an effective approach to build a robust model and prevent overfitting. Random forests (RF) combine shallow trees using bagging, i.e., the prediction is the average (for regression) or the majority vote (classification) of the trees in the ensemble (Breiman, 2001). The trees are constructed from bootstrap samples and a random subset of predictors (mtry) is used at each split in a tree. Together with the number of trees, these are the main parameters of random forests, which was tuned in the training process (see Table 2). The minimum number of observations per node was set to 20.

Gradient boosting (GBM) uses a different ensemble technique called boosting, where decision trees are combined in a forward stage-wise procedure. While in RF each tree is independently built, in gradient boosting, each new tree is constructed on the residuals of the previous tree to minimize the mean squared error. The maximum depth of the trees (interaction depth) was tuned between 1 to 6, while the minimum number of observations per node was set to 10. The values set for the other parameters of GBM are described in Table 2.

Support vector machine

Support vector machine (SVM) (Boser et al., 1992), although widely used for classification problems, might also be applied for regression (SVR). In SVM, the main goal is to find a hyperplane that fits the training data by minimizing the Euclidean norm of the coefficient vector. This model uses a kernel function to map input data to higher-dimensional spaces, where it can be linearly separable. In regression problems, a symmetrical "margin" is added around the estimated function, where the absolute errors should be equal or less than the maximum error ε (Awad & Khanna, 2015). SVR is an optimization problem where the objective

function minimizes the Euclidean norm of the function coefficients (w), while avoiding outliers:

$$min\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}|\xi_i|$$

Subject to:

$$|y_i - w_i x_i| \leq \varepsilon + |\xi_i|$$

where C is the cost parameter, which gives more weight to the function flatness and ξ is the slack variable and corresponds to the tolerable distance of outliers from the margin.

A radial basis function kernel was applied here, defined as:

$$K_{RBF}\left(\boldsymbol{x},\boldsymbol{x}'\right) = e^{-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|^2}$$

where x and x' are samples in the input data and γ is a parameter related to the variance of the function. This parameter was set to the inverse of the training data size.

Model parameters and performance evaluation

The tuning process of the hyperparameters of regression models is fundamental to avoiding overfitting. One of the most traditional approaches to optimize hyperparameter selection is grid search. In grid search, the modeler defines a subset of hyperparameter values and a performance metric to search for the best combination of parameters. Then, k-fold cross-validation or leave-one-out cross-validation can be used on the training set to perform the tuning process.

In this study, the RMSE was chosen to tune the model's parameters. Tuning was performed with a fivefold cross-validation. In this approach, the training dataset is split into five subsets: the predictive model is fitted for four of them and the performance metric (in this study, RMSE) is calculated for the remaining subset. This procedure is repeated five times, so that all data is used at least once to train/validate the model. Model performance is assessed by calculating the average RMSE obtained in each subset. fivefold cross-validation was applied using the R package "caret." Table 2 summarizes the main parameters of the fitted models and their correspondent values. Validation was performed for each combination of the parameters and the model with the best performance (lower RMSE) was selected.

Performance metrics

Model performance in the testing dataset was evaluated using the root mean squared error (RMSE), mean absolute error (MAE), and the R squared (R^2) measures:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\widehat{y}_i - y_i)^2}{n}}$$

$$MAE = \frac{\sum_{i=1}^{n}|\widehat{y}_i - y_i|}{n}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

where y is the observed Chla, $\widehat{y}$ is the predicted Chla, $\overline{y}$ is the mean observed Chla, and n is the number of observations in the testing dataset.

Partial dependence plots

Partial dependence plots (PDP) were introduced by Friedman (2001) to interpret complex machine learning algorithms. The PDP represents the marginal effect of independent variables on the response of a machine learning model (Friedman 2001). The partial dependence of the response on a variable x_l is represented by:

$$\widehat{f}_{x_l}(x_l) = E_{x_s}\left[\widehat{f}(x_l, x_s)\right] = \int \widehat{f}(x_l, x_s)P(x_s)dx_s$$

where x_l is the independent variable analyzed in the partial dependence plot, x_s is the subset of the other input variables of the regression model $\widehat{f}$, and $P(x_s)$ is the marginal probability density of x_s. The function shows the effect of the variable x_l on the dependent variable by marginalizing over the other explanatory variables.

Results and discussion

This section presents and compares the performance obtained with the predictive models, the relative importance of the hydrological and climate variables, and their relationships with Chla.

Performance of the regression models

Figure 2 presents the scatterplots of predicted and observed values for all the models tested in this study. From the plots, one can notice that linear regression, regularized GLM, and the regression tree underestimate Chla. These models have strong assumptions about error distribution: homoscedasticity, normal distribution, and no autocorrelation. Although the variables with an elevated correlation have been removed, there was still some multi-collinearity between the predictors, which could be a problem for the prediction.

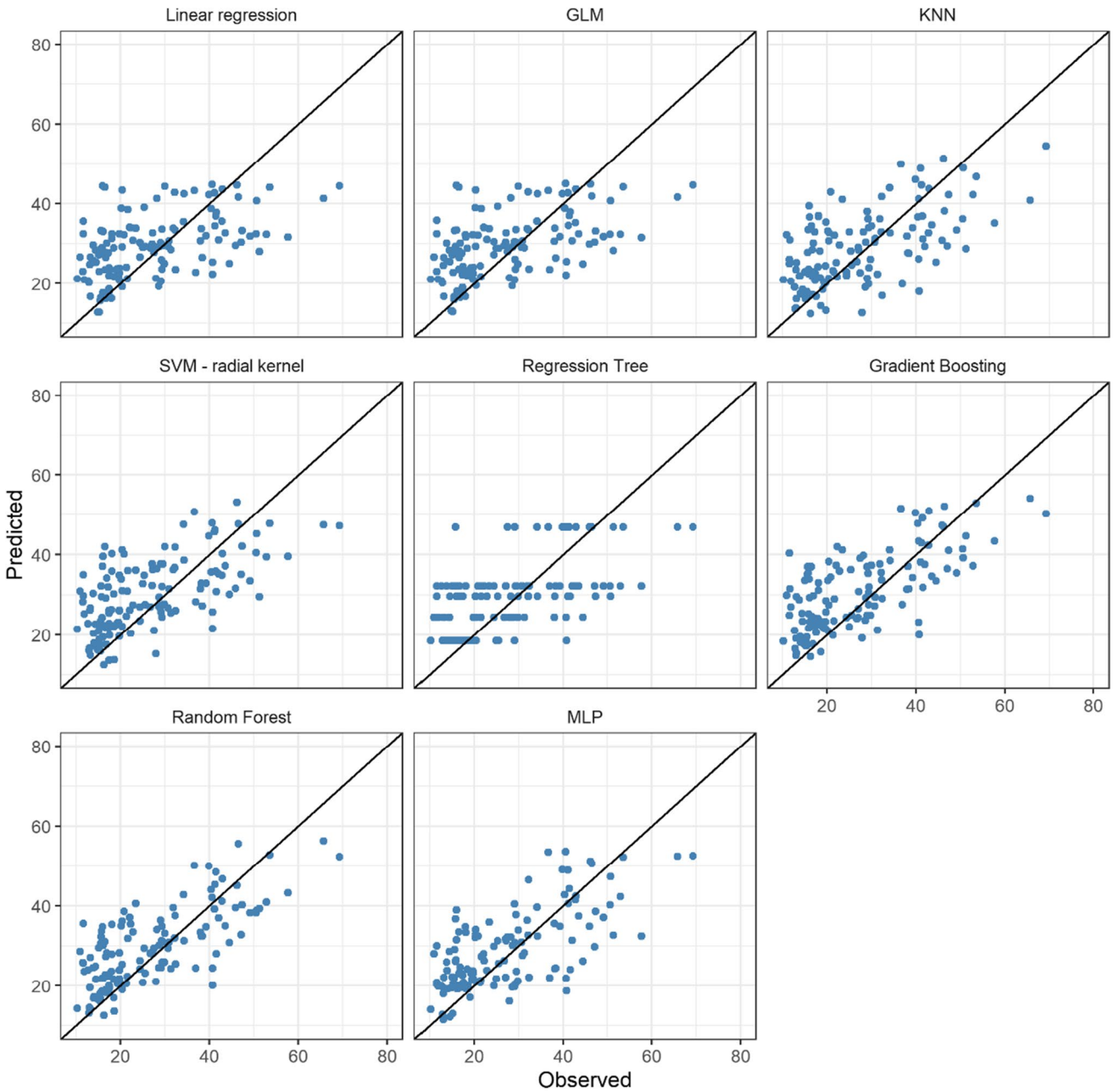

Fig. 2 Scatterplots for the predictive models tested in this study. The diagonal line represents the perfect fit between observed and predicted values

Predictors of water quality indicators will frequently be correlated (both temporally and spatially) since the mechanisms associated with their increase or decrease are interrelated (Su et al. 2012; Liu et al. 2019; Mesquita et al. 2020). It is important to keep in mind that highly correlated variables can present complementary information when combined (Guyon and Elisseeff 2003), which reinforces the need for integrating correlation analysis with model-based variable importance.

RF, GBM, and MLP provided the best predictions (Table 3). These models are designed to capture nonlinear relationships between variables, which is likely to be the case here. RF and GBM can reduce the variance of the predicted values by employing ensemble techniques (boosting and bagging, respectively), outperforming the regression tree (Hastie et al. 2009). The SVM model with a radial kernel is also able to detect nonlinearity, as it transforms data to a dimensional space where they can be linearly separable

Table 3 Performance metrics for the fitted models

Model	R^2	*RMSE*	*MAE*
RF	**0.52**	**9.32**	**7.15**
GBM	0.46	10.26	8.01
MLP	0.45	9.74	7.66
SVM	0.36	10.92	8.77
KNN	0.35	10.67	8.22
Regression tree	0.32	10.77	8.21
Linear regression	0.26	11.48	9.10
Regularized GLM	0.26	11.48	9.08

(Awad and Khanna 2015). However, SVM had a slightly worse performance than GBM, RF, and MLP.

As expected, the predictive models were able to explain only part of Chla, since the best performing model had an R^2 of 0.52 (Table 3) This performance can be considered satisfactory for a watershed-scale model, as a reference value to evaluate phosphorus (P) prediction (which can be easier to predict than Chla) is an $R^2 > 0.5$ (Moriasi et al. 2015).

This result also suggests that hydrological and climate factors alone are not enough to predict Chla and additional variables might be necessary, such as water quality indicators (Rocha et al. 2020). However, it must be emphasized that the relationship between P and Chla in tropical lakes is not comparable to that in temperate ones, where empirically estimated relationships between P and Chla provide reliable models to calculate Chla levels (Sakamoto 1966; Dillon and Rigler 1974; Jones and Bachmann 1976). A correlation analysis between measured total phosphorus concentration, obtained from COGERH database (http://www.hidro.ce.gov.br/), and estimated Chla reveals that nutrient enrichment may not be the only influencing factor on eutrophication in tropical reservoirs (Fig. 3). Although correlation between nitrogen and Chla was not analyzed here (since limited data was available), this can also be a limiting nutrient for eutrophication in reservoirs (Wiegand et al. 2020; Qin et al. 2020).

Although past studies have obtained better predictive performances (Stefanidis et al. 2021), Chla can be harder to predict in the semiarid, due to the significant water level variability (which implies more complex mechanisms behind eutrophication) and the usually higher trophic levels (Wiegand et al. 2021). There are, however, other possible explanations. The Chla time series were derived from satellite data, which has high estimation accuracy (Lins et al., 2017), but might contain noise or components that cannot be explained with known variables. Also, past studies have indicated that the drivers of Chla can vary with the temporal resolution (Blauw et al. 2018; Liu et al. 2019). For example, on a monthly scale, water temperature is less important to predict Chla than nutrient loadings (Liu et al. 2019), which

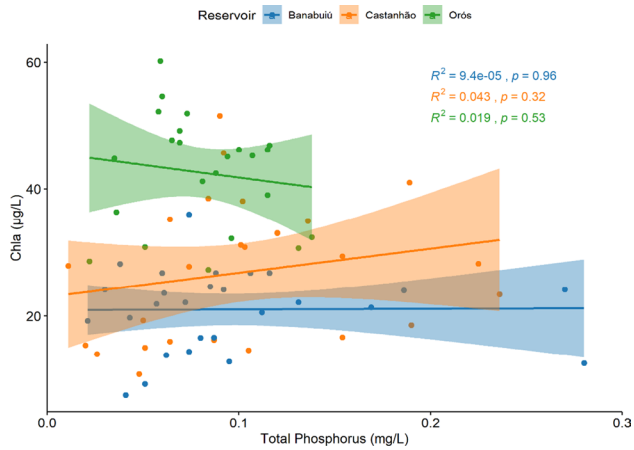

Fig. 3 Correlation between total phosphorus and Chla in the reservoirs analyzed in our study. The dark, bold line represents the fitted regression line, and the shadow area is the confidence interval. Phosphorus measurements are taken each three months and were available for a shorter period than Chla estimations (05/2008 to 11/2019)

means that part of the explanatory variables could not be able to explain Chla in our model.

Variable importance

To measure the relative influence of the model's explanatory variables, the importance measure attributed by each predictive model was extracted and scaled using min–max normalization (Fig. 4). This approach has been widely used to make machine learning models more interpretable (Hastie et al. 2009) and can be more accurate than looking only at the correlation between explanatory and dependent variables. Correlation criteria or the goodness of fitness of a linear model are simple and direct strategies to obtain information about a set of variables, but it ignores multicollinearity and interactions between them. Although this study was not intended to perform variable selection, some of the models used here have built-in processes to select the most relevant predictions, such as RF and regularized GLM, the so-called embedded methods (Guyon and Elisseeff 2003).

Radial SVM and KNN models were excluded from this analysis since they do not have a direct importance measure. For RF, GBM, and the regression tree models, the importance corresponds to the reduction in predictive performance obtained by removing the variable from the model. In GLM and MLP, the importance is associated with the weights attributed to each variable.

The boxplots in Fig. 4 reveal that water volume was considered the most important predictive variable in all models. The models do not agree regarding the mix-layer depth and bottom temperature importance, as these presented a high variation among them. The dummy variables related to the spatial location of the reservoirs (Castanhão, Orós

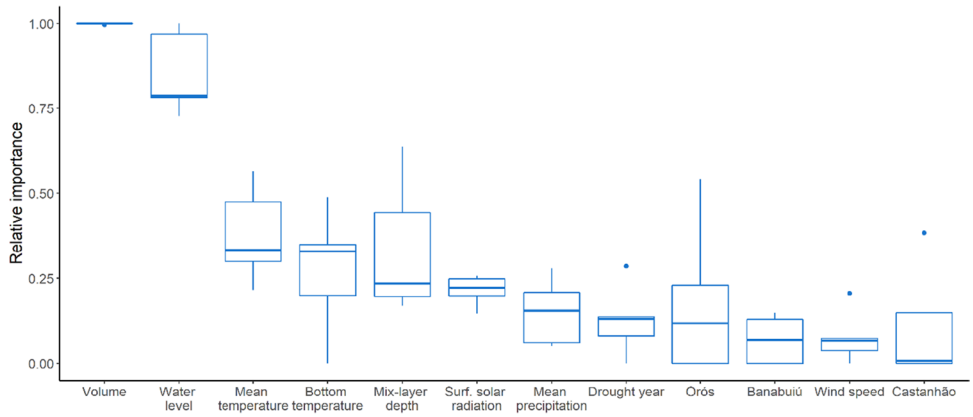

Fig. 4 The relative importance of explanatory variables considering the importance measures of each predictive model, ordered by the median value. Relative importance was scaled between 0 and 1

e Banabuiú) did not seem to significantly influence Chla, indicating that spatial variability could be less important than climate variability, or yet, that the relationships between explanatory variables and Chla are similar for all three reservoirs.

The relative influence of the variables depends on the interactions identified by each model and the procedure used to do it. For example, decision trees choose the optimal variable in each split based on the information gained by adding it to the tree. The regression tree constructed to predict Chla had only the mix-layer depth and water volume as predictors (Supplementary material, Fig. S2). This means that these two variables provide enough information to give us an approximate estimation of Chla. The regression tree alone can be considered a weak predictor, as it is very sensitive to small changes in the dataset and can easily overfit. Since they assume all variables have some interaction between them, it suits well our problem, but it fails to provide accurate estimations of Chla (here, it presented an R^2 of only 0.32). However, it can still give us interesting information on variable importance.

GBM and RF, as explained in the "Methods" section, combine several regression trees to provide stronger predictive models. RF performs variable selection during its model building process, as the variables used to construct each tree in the ensemble are selected from a random subset of the explanatory variables. The trees are fitted to bootstrap samples of the data, and the importance measure is calculated on the left-out observations (out-of-bag set). The advantage of RF's strategy to calculate variable importance is that it considers both the individual effect and the interactions between the variables (Strobl et al. 2007). GBM, on the other hand, calculates importance on the entire training set instead of using the out-of-bag sets.

To verify the effect of the season on the relationships between the explanatory variables and Chla, all the models were run again for the wet season (observations registered between February and May), and the dry season (observations from the remaining months). Variable importance was extracted for each model and normalized so one could visualize their relative influence on Chla prediction (Fig. 5).

Water volume and water level continue to be the most relevant indicators of Chla in both scenarios. However, mix-layer depth and mean temperature seem to be more important in the wet season. It is important to keep in mind that the dry season model has a smaller dataset than the wet season, as it corresponds to the observations of 4 months only. For this reason, the model can be biased, and more data could be necessary to provide reliable predictions.

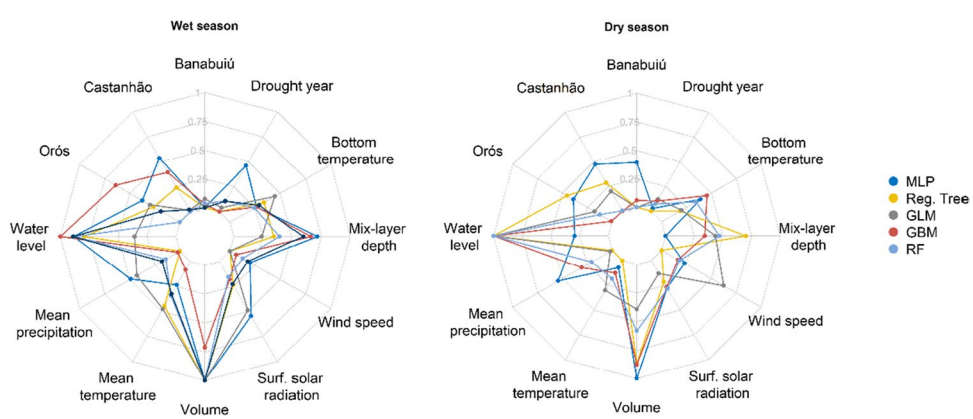

Fig. 5 Relative importance of explanatory variables considering separated models for the wet season and dry season

Relative influence of hydrological and climate variables on Chla

The PDPs in Fig. 6 illustrate the relationships between hydrological and climate variables and Chla. The RF model was selected for this analysis, as it presented the best performance according to all the metrics evaluated. These plots, however, should be interpreted with caution, as they may not display all interactions of the explanatory variables.

Confirming the findings of previous studies, Chla tends to increase as water volume reduces (da Rocha Junior et al. 2018; Wiegand et al. 2021). The decrease in water volume due to evaporation loss, water withdrawals, and extended drought periods are usually associated with higher phosphorus loads in tropical reservoirs (Raulino et al. 2021; Rocha and Lima Neto 2021a). During the dry period, sediment release and nutrient resuspension are important mechanisms associated with Chla in these reservoirs. Although the effect of internal loading has been pointed as more significant in shallow reservoirs, in the semiarid, precipitation levels come close to zero and inflow decreases drastically during the dry season, so that deep reservoirs reach very low volumes and almost no external loads are carried to them (Delmiro Rocha and Lima Neto 2021; Lima Neto et al., 2022).

Wind speed did not seem to play an important role in Chla levels, which might be due to reservoirs' morphology and the temporal scale considered here. In deep reservoirs, wind speed is indeed unimportant to Chla, as it is not a relevant driver of water column mixing. Shallow reservoirs, on the other hand, present a significant correlation with nutrient

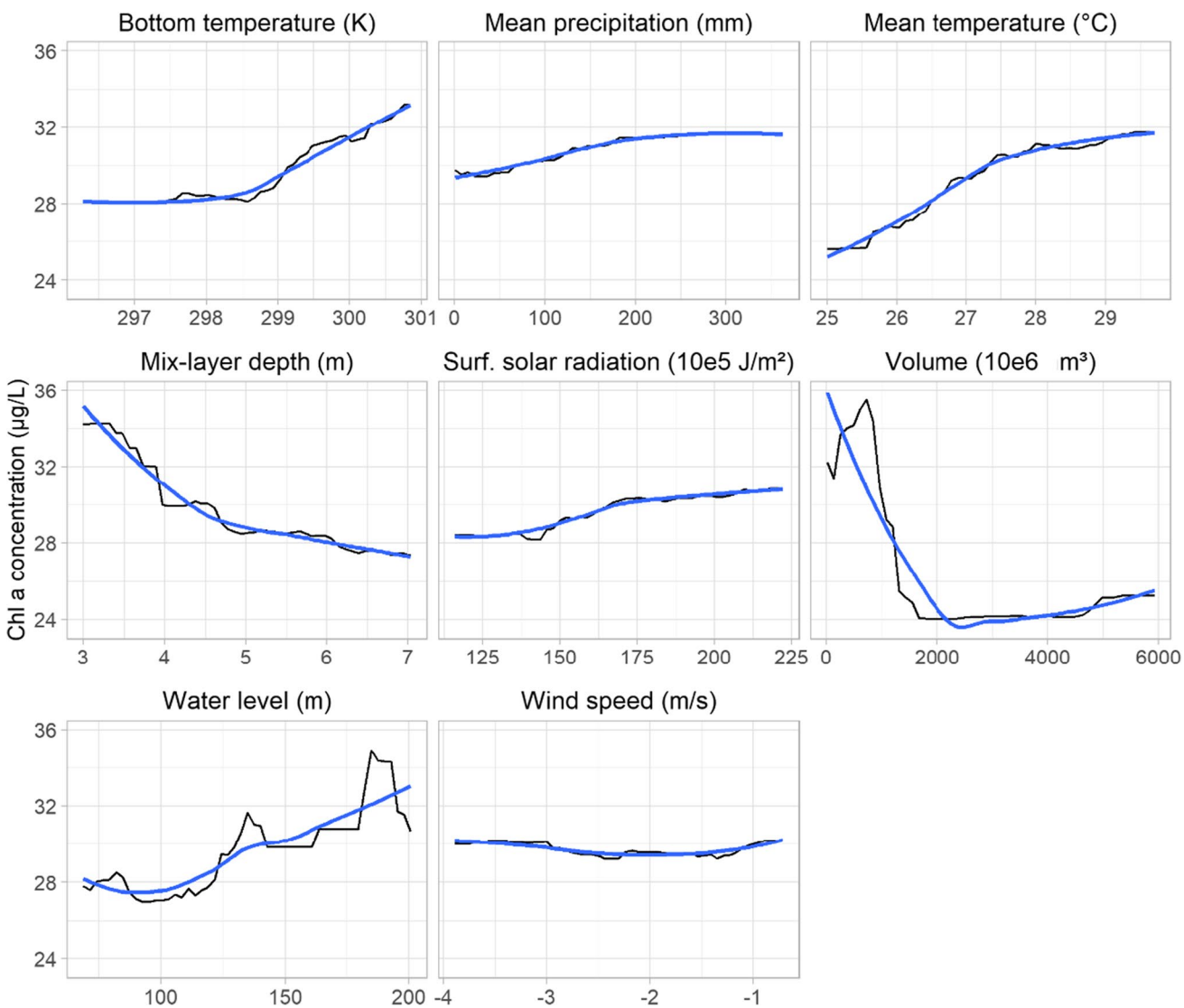

Fig. 6 PDPs for predictors of the RF model. The blue smooth line was produced using LOESS (locally weighted smoothing) to better visualize the relationship between the explanatory and response variables

resuspension (Araújo et al. 2019; Mesquita et al. 2020). Past research has indicated that although wind speed affects the dynamics of algal growth and eutrophication, there is a loss of information on wind dynamics on a monthly scale (Stefanidis et al. 2021).

Mix-layer depth has an inverse relationship with Chla, which is consistent with previous findings (Stockwell et al. 2020; Stefanidis et al. 2021). There are several factors to consider when interpreting this relationship, such as water temperature, reservoir morphology, and the ratio between the mix-layer depth and thermocline depth. In deep reservoirs, stratification is more likely to occur and lake stability tends to increase, with a higher possibility of solute accumulation in the hypolimnion, dissolved oxygen depletion, and phosphorus release from sediments (Butcher et al. 2015; Kraemer et al. 2015; Moura et al. 2020a). But an increase in mix-layer depth also results in a reduction of the light available to phytoplankton (Stockwell et al. 2020) and in lower water temperatures, which could inhibit Chla growth (Zhao et al. 2020).

Bottom temperature, mean temperature, solar radiation, and water level have direct relationships with Chla. The first three variables are directly related to each other, and their increase usually enhances phytoplankton productivity (Liu et al. 2019). The direct influence of water level on Chla is surprising, as previous studies have reported the opposite relationship (Medeiros et al. 2015; Wiegand et al. 2020; Braga and Becker 2020). These studies, however, were performed for small reservoirs, where the relationship between P and Chla is stronger than that for larger reservoirs, i.e., the mechanisms associated with Chla growth are less complex.

The effect of increasing water levels on Chla depends on the quality of the inflow, whether it is related or not to a reduction in the outflow (Bakker and Hilt 2015), the depth, and the trophic state of the reservoir (Costa et al. 2015). When precipitation occurs (and water levels start to rise), external loads from rivers and surface runoff add up to internal loads due to thermal stratification and phosphorus release from sediment, which is highly correlated with Chla growth (Moura et al. 2020a). Agriculture and cattle raising are important activities in all reservoirs analyzed here and are the main cause of nonpoint source pollution that increases external total phosphorus loading (Rocha and Lima Neto 2021; Lima Neto et al., 2022).

Although volume and water level are directly related, they have a nonlinear relationship, which can be approximated as a logarithmic curve. Hence, for a certain range, water level fluctuations have little effect on water volume. In this case, Chla growth could be related to some of the factors mentioned above (e.g., the quality of external loads). Reservoir's morphology should also be considered, as the storage depends on the water height-area relationship. Hence, the effect of water level on Chla might depend on how much water is already stored in the reservoir (i.e., at which position in the water height-area-volume curve the reservoir is), the reservoir's morphology, and the quality of external loads.

The PDPs for the dry and wet season models were also examined. Except for mean precipitation and wind speed, all variables maintained the patterns observed in the general model. Figure 7 presents the variables with opposing behaviors. While precipitation has a positive effect in the dry season, it presents a negative and almost insignificant effect during the wet season.

One explanation for this behavior is that water volumes tend to be reduced over the dry season. Hence, precipitation can increase nutrient loadings (Jeppesen et al. 2015; da Rocha Junior et al. 2018) but not have a significant effect on water volume. During the wet season, increased precipitation might induce greater flushing and lower Chla (Reichwaldt and Ghadouani 2012). Because the reservoirs have higher water volumes during this season, as the precipitation volume increases, water volume grows exponentially with respect to water level, and Chla might decrease because of mixing and flushing. This effect, however, seems to be not very relevant as produces a little variation on Chla.

The extent of precipitation influence on Chla is difficult to generalize, as it depends on the intensity and frequency of rainfall events (Reichwaldt and Ghadouani 2012; Ho and Michalak 2020) and the initial conditions of the reservoir (water volume, trophic state, etc.). The reduced stratification during the wet season (Lima Neto 2019) can also explain the reduction in Chla during this season, while stronger winds during the dry season can lead to higher Chla concentrations. Hence, precipitation alone is not the only factor to explain Chla fluctuations in both seasons, as its mechanisms are complex.

During the wet season, stronger winds seem to result in a slight decrease of Chla (up to 3 μg/L), while in the dry season, it has the opposing effect. The influence of wind speed on Chla can differ according to the water depth, and the sign of this relationship needs further investigation. Previous studies have indicated that increased wind speed can result in greater mixing of the upper layer, thus reducing Chla (Stockwell et al. 2020); however, under oligotrophic conditions, stronger winds can carry nutrients to the bottom layer and increase Chla (Kahru et al. 2010; Kim et al. 2014). This mechanism also depends on the reservoirs' morphology and water level, hence for shallow reservoirs (or for reduced water levels in the dry season), stronger winds can induce resuspension and increase internal nutrient loads (Araújo et al. 2019; Rocha and Lima Neto 2022). In the wet season, wind-induced resuspension is less significant, as external sources of nutrients play a more important role in Chla fluctuations (Rocha and Lima Neto 2021b).

The relationship between wind speed and internal phosphorus loading has been explored for artificial reservoirs

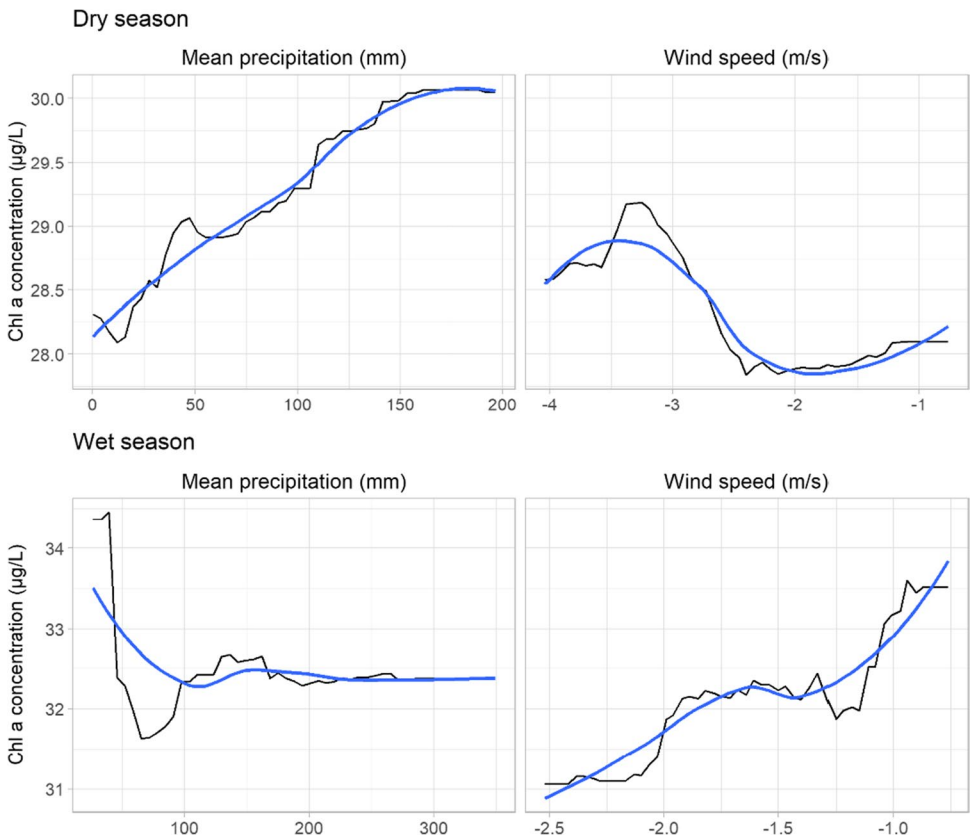

Fig. 7 PDPs for precipitation and wind speed for two separate models, one considering the months in the dry season, and the other, the months in the wet season

in Ceará, including the ones analyzed here (Rocha and Lima Neto 2022). In this study, the authors found that P release increases with stronger winds (with a threshold value of 3.5 m/s) and the trophic state of the reservoir. As internal loading can increase the risk of eutrophication, wind speed is very likely to be related to Chla in the dry season, when reservoirs become shallower.

PDPs can also be plotted for two variables at the same time (Supplementary material, Fig. S3). Again, one must be careful when interpreting these plots, as they can show correlations between variables rather than a causal relationship. When considering higher values of solar radiation, wind speed presents an inverse relationship with Chla. Whether the mix-layer is shallow or deep, when solar radiation is higher, Chla tends to increase, a relationship that is confirmed by previous research (Berger et al. 2006). One can also notice that mix-layer depth seems to have a stronger effect on Chla only up to a certain point.

Wind speed had little effect on Chla when the water volume was constant. Again, this might be related to the size of the reservoirs analyzed here and does not necessarily mean that wind speed does not influence Chla. Previous studies have indicated that wind speed can be an important driver of internal phosphorus loadings in the dry period (Rocha and Lima Neto 2022), thus, this variable should not be neglected.

Precipitation can have distinct effects on nutrient concentrations (Ho and Michalak 2020). Our analysis indicates that when the water volume is high, increased precipitation levels mean higher Chla (Wiegand et al. 2020), while for low water volumes, increased precipitation levels mean lower Chla. This, again, can be related to the climate season, as previously discussed. Although there might have been some information loss due to the temporal resolution of the analysis presented here, the results are consistent with the findings of other studies performed for the semiarid region (Moura et al. 2020b; Mesquita et al. 2020; Rocha and Lima Neto 2021a, 2022). Rather than providing accurate predictions of Chla, the predictive models explored in this study can indicate the magnitude and the overall direction of the relationship between hydro-climatic variables and Chla.

Conclusions

In the semiarid region, complex mechanisms regulate phytoplankton growth, so that estimates of P may not result in reliable predictions of Chla. This study revealed that a combination of hydrological and climate factors can provide insightful information on Chla fluctuations on a monthly scale. To do that, RF and GBM are the most suitable models, with satisfactory predictive performance.

Looking at the interaction between variables, increasing solar radiation and reducing wind speed result in higher Chla, while for a deeper mix-layer, the increase of solar radiation has a positive effect on Chla. Another interesting finding was that precipitation and wind speed present opposing effects on Chla depending on the season. Water level and volume have opposite relationships with Chla: the underlying mechanism associated with Chla is reverted after the dry season (when the internal load is more significant).

These results suggest that climate and hydrological variables have nonlinear relationships with Chla, with an exploratory potential that should not be ignored. Machine learning models can provide important insight on the mechanisms related to Chla increase or decrease in reservoirs, especially when using interpretation methods such as PDPs. By understanding some of the mechanisms associated with hydrological and climatic variability and Chla, policymakers can design more specific strategies to mitigate eutrophication.

There are, however, a few drawbacks of this study, such as the temporal-spatial resolution of the time series, which can hide some of the mechanisms associated with Chla fluctuations. However, extensive field data collection would be needed to overcome this limitation. An interesting approach to be investigated in future studies is the combination of mechanistic water quality modeling and machine learning methods (the so-called scientific machine learning) to assess eutrophication mechanisms. Within this framework, physical and chemical relationships can be incorporated into machine learning modeling, facilitating uncertainty quantification and interpretability.

Supplementary information.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11356-022-21168-z.

Author contribution Conceptualization: TMNC, IELN, and FASF. Methodology: TMNC, IELN, and FASF. Formal analysis and investigation: TMNC, IELN, and FASF. Writing — original draft preparation: TMNC. Writing — review and editing: TMNC, IELN, and FASF.

Funding The present study was supported through by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (88887.123932/2015–00 and 88882.344015/2019–01); the Brazilian Council for Scientific and Technological Development (441457/2017–7); and the Cearense Foundation for Scientific and Technological Support (Cientista-chefe program).

Data availability The data and code used in the study are available at Github via https://github.com/taiscarvalho/chla-prediction-ce.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

Araújo GM, Lima Neto IE, Becker H (2019) Phosphorus dynamics in a highly polluted urban drainage channelshallow reservoir system in the Brazilian semiarid. An Acad Bras Cienc. https://doi.org/10.1590/0001-3765201920180441

Awad M, Khanna R (2015) Support vector regression. Efficient Learning Machines. Apress, Berkeley, CA, pp 67–80

Bakker ES (2015) Hilt S (2015) Impact of water-level fluctuations on cyanobacterial blooms: options for management. Aquat Ecol 503(50):485–498. https://doi.org/10.1007/S10452-015-9556-X

Berger SA, Diehl S, Stibor H et al (2006) (2006) Water temperature and mixing depth affect timing and magnitude of events during spring succession of the plankton. Oecologia 1504(150):643–654. https://doi.org/10.1007/S00442-006-0550-9

Blauw AN, Benincà E, Laane RWPM et al (2018) Predictability and environmental drivers of chlorophyll fluctuations vary across different time scales and regions of the North Sea. Prog Oceanogr 161:1–18. https://doi.org/10.1016/J.POCEAN.2018.01.005

Braga GG, Becker V (2020) Influence of water volume reduction on the phytoplankton dynamics in a semi-arid man-made lake: a comparison of two morphofunctional approaches. An Acad Bras Cienc 92:20181102. https://doi.org/10.1590/0001-3765202020181102

Butcher JB, Nover D, Johnson TE (2015) Clark CM (2015) Sensitivity of lake thermal and mixing dynamics to climate change. Clim Chang 1291(129):295–305. https://doi.org/10.1007/S10584-015-1326-1

Campos JNB (2015) Paradigms and public policies on drought in Northeast Brazil: a historical perspective. Environ Manage 55:1052–1063. https://doi.org/10.1007/s00267-015-0444-x

da Costa MRA, Attayde JL (2015) Becker V (2015) Effects of water level reduction on the dynamics of phytoplankton functional groups in tropical semi-arid shallow lakes. Hydrobiol 7781(778):75–89. https://doi.org/10.1007/S10750-015-2593-6

da Rocha Junior CAN, da Costa MRA, Menezes RF et al (2018) A redução do volume intensifica o risco a eutrofização em reservatórios do semiárido tropical. Acta Limnol Bras. https://doi.org/10.1590/s2179-975x2117

Dillon PJ, Rigler FH (1974) The phosphorus-chlorophyll relationship in lakes. Limnol Oceanogr 19:767–773. https://doi.org/10.4319/LO.1974.19.5.0767

Dunstan PK, Foster SD, King E et al (2018) Global patterns of change and variation in sea surface temperature and chlorophyll a. Sci Rep 8:1–9. https://doi.org/10.1038/s41598-018-33057-y

Easterling DR, Meehl GA, Parmesan C, et al (2000) Climate extremes: observations, modeling, and impacts. Science (80-) 289:2068–2074. https://doi.org/10.1126/science.289.5487.2068

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232. https://doi.org/10.1214/aos/1013203451

Gholizadeh M, Melesse A, Reddi L (2016) A comprehensive review on water quality parameters estimation using remote sensing techniques. Sensors 16:1298. https://doi.org/10.3390/s16081298

Greve P, Kahil T, Mochizuki J et al (2018) Global assessment of water challenges under uncertainty in water scarcity projections. Nat Sustain 1:486–494. https://doi.org/10.1038/s41893-018-0134-9

Gutiérrez APA, Engle NL, De Nys E et al (2014) Drought preparedness in Brazil. Weather Clim Extrem 3:95–106. https://doi.org/10.1016/j.wace.2013.12.001

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182. https://doi.org/10.5555/944919.944968

Harris I, Osborn TJ, Jones P, Lister D (2020) Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Sci Data 7:1–18. https://doi.org/10.1038/s41597-020-0453-3

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York, New York, NY

Hirsch RM, Archfield SA (2015) Flood trends: not higher but more often. Nat Clim Chang 5:198–199

Ho JC, Michalak AM (2020) Exploring temperature and precipitation impacts on harmful algal blooms across continental U.S. lakes. Limnol Oceanogr 65:992–1009. https://doi.org/10.1002/LNO.11365

Iiames JS, Salls WB, Mehaffey MH, et al (2021) Modeling anthropogenic and environmental influences on freshwater harmful algal bloom development detected by MERIS over the Central United States. Water Resour Res 57:e2020WR028946. https://doi.org/10.1029/2020WR028946

Jeppesen E, Brucet S, Naselli-Flores L et al (2015) (2015) Ecological impacts of global warming and water abstraction on lakes and reservoirs due to changes in water level and related changes in salinity. Hydrobiol 7501(750):201–227. https://doi.org/10.1007/S10750-014-2169-X

Jones JR, Bachmann RW (1976) Prediction of phosphorus and chlorophyll levels in lakes. Water Pollut Control Fed 2176–2182

Kahru M, Gille ST, Murtugudde R et al (2010) Global correlations between winds and ocean chlorophyll. J Geophys Res Ocean 115:12040. https://doi.org/10.1029/2010JC006500

Kim T-W, Najjar RG, Lee K (2014) Influence of precipitation events on phytoplankton biomass in coastal waters of the eastern United States. Global Biogeochem Cycles 28:1–13. https://doi.org/10.1002/2013GB004712

Kraemer BM, Anneville O, Chandra S et al (2015) Morphometry and average temperature affect lake stratification responses to climate change. Geophys Res Lett 42:4981–4988. https://doi.org/10.1002/2015GL064097

Li T, Zhang Y, He B et al (2021) Periodically hydrologic alterations decouple the relationships between physicochemical variables and chlorophyll-a in a dam-induced urban lake. J Environ Sci (china) 99:187–195. https://doi.org/10.1016/j.jes.2020.06.014

Lima Neto IE (2019) Impact of artificial destratification on water availability of reservoirs in the Brazilian semiarid. An Acad Bras Cienc. https://doi.org/10.1590/0001-3765201920171022

Lins R, Martinez J-M, Motta Marques D et al (2017) Assessment of chlorophyll-a remote sensing algorithms in a productive tropical estuarine-lagoon system. Remote Sens 9:516. https://doi.org/10.3390/rs9060516

Liu X, Feng J, Wang Y (2019) Chlorophyll a predictability and relative importance of factors governing lake phytoplankton at different timescales. Sci Total Environ 648:472–480. https://doi.org/10.1016/j.scitotenv.2018.08.146

Lopes FB, Barbosa CCF, de Novo EML, M, et al (2014) Modelagem da qualidade das águas a partir de sensoriamento remoto hiperespectral. Rev Bras Eng Agrícola e Ambient 18:13–19. https://doi.org/10.1590/1807-1929/agriambi.v18nsupps13-s19

Mamun M, Kim J-J, Alam MA, An K-G (2019) Prediction of algal chlorophyll-a and water clarity in monsoon-region reservoir using machine learning approaches. Water 12:30. https://doi.org/10.3390/w12010030

de Medeiros L (2015) C, Mattos A, Lürling M, Becker V (2015) Is the future blue-green or brown? The effects of extreme events on phytoplankton dynamics in a semi-arid man-made lake. Aquat Ecol 493(49):293–307. https://doi.org/10.1007/S10452-015-9524-5

de Mesquita JB, F, Lima Neto IE, Raabe A, de Araújo JC, (2020) The influence of hydroclimatic conditions and water quality on evaporation rates of a tropical lake. J Hydrol 590:125456. https://doi.org/10.1016/J.JHYDROL.2020.125456

Moriasi DN, Gitau MW, Pai N, Daggupati P (2015) Hydrologic and water quality models: performance measures and evaluation criteria. Trans ASABE 58:1763–1785. https://doi.org/10.13031/TRANS.58.10715

Moura DS, Lima Neto IE, Clemente A et al (2020a) Modeling phosphorus exchange between bottom sediment and water in tropical semiarid reservoirs. Chemosphere 246:125686. https://doi.org/10.1016/J.CHEMOSPHERE.2019.125686

Najah Ahmed A, Binti Othman F, Abdulmohsin Afan H et al (2019) Machine learning methods for better water quality prediction. J Hydrol 578:124084. https://doi.org/10.1016/j.jhydrol.2019.124084

Nguyen H-Q, Ha N-T (2020) Pham T-L (2020) Inland harmful cyanobacterial bloom prediction in the eutrophic Tri An Reservoir using satellite band ratio and machine learning approaches. Environ Sci Pollut Res 279(27):9135–9151. https://doi.org/10.1007/S11356-019-07519-3

Pacheco CHA, Lima Neto IE (2017) Effect of artificial circulation on the removal kinetics of cyanobacteria in a hypereutrophic shallow lake. J Environ Eng (united States) 143:1–8. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001289

Paerl HW, Otten TG (2013) Harmful cyanobacterial blooms: causes, consequences, and controls. Microb Ecol 65:995–1010. https://doi.org/10.1007/s00248-012-0159-y

Pontes Filho JD, de Souza Filho F, A, Martins ESPR, Studart TM de C, (2020) Copula-based multivariate frequency analysis of the 2012–2018 drought in Northeast Brazil. Water 12:834. https://doi.org/10.3390/w12030834

Qin B, Zhou J, Elser JJ et al (2020) Water depth underpins the relative roles and fates of nitrogen and phosphorus in lakes. Environ Sci Technol 54:3191–3198. https://doi.org/10.1021/ACS.EST.9B05858/SUPPL_FILE/ES9B05858_SI_002.XLSX

Raulino JBS, Silveira CS, Lima Neto IE (2021) Assessment of climate change impacts on hydrology and water quality of large semi-arid reservoirs in Brazil. Hydrol Sci J 66:1321–1336. https://doi.org/10.1080/02626667.2021.1933491

Reichwaldt ES, Ghadouani A (2012) Effects of rainfall patterns on toxic cyanobacterial blooms in a changing climate: between simplistic scenarios and complex dynamics. Water Res 46:1372–1393. https://doi.org/10.1016/J.WATRES.2011.11.052

de Rocha M, JD, Lima Neto IE, (2021a) Phosphorus mass balance and input load estimation from the wet and dry periods in tropical semiarid reservoirs. Environ Sci Pollut Res 2021:1–20. https://doi.org/10.1007/S11356-021-16251-W

de Rocha M, JD, Lima Neto IE, (2021b) Modeling flow-related phosphorus inputs to tropical semiarid reservoirs. J Environ Manage 295:113123. https://doi.org/10.1016/J.JENVMAN.2021.113123

de Rocha M, JD, Lima Neto IE, (2022) Internal phosphorus loading and its driving factors in the dry period of Brazilian semiarid reservoirs. J Environ Manage 312:114983. https://doi.org/10.1016/J.JENVMAN.2022.114983

Rocha SMG, de Mesquita JB, F, Lima Neto IE, (2020) Análise E Modelagem Das Relações Entre Nutrientes E Fitoplâncton Em Reservatórios Do Ceará. Rev Bras Ciências Ambient. https://doi.org/10.5327/z2176-947820190536

Ross MRV, Topp SN, Appling AP et al (2019) AquaSat: a data set to enable remote sensing of water quality for inland waters. Water Resour Res 55:10012–10025. https://doi.org/10.1029/2019WR024883

Sakamoto M (1966) Primary production by phytoplankton community in some Japanese lakes and its dependence on lake depth. Arch Hydrobiol 62:1–28

Shen J, Qin Q, Wang Y, Sisson M (2019) A data-driven modeling approach for simulating algal blooms in the tidal freshwater of

James River in response to riverine nutrient loading. Ecol Modell 398:44–54. https://doi.org/10.1016/J.ECOLMODEL.2019.02.005
Sperling M von (2005) Introdução à Qualidade das Águas e ao Tratamento de Esgotos. DESA-UFMG, Belo Horizonte
Stachelek J (2020) Spatial interpolation by inverse path distance weighting
Stefanidis K, Varlas G, Vourka A et al (2021) Delineating the relative contribution of climate related variables to chlorophyll-a and phytoplankton biomass in lakes using the ERA5-Land climate reanalysis data. Water Res 196:117053. https://doi.org/10.1016/J.WATRES.2021.117053
Stockwell JD, Doubek JP, Adrian R et al (2020) Storm impacts on phytoplankton community dynamics in lakes. Glob Chang Biol 26:2756–2784. https://doi.org/10.1111/GCB.15033
Strobl C, Boulesteix A-L, Zeileis A (2007) Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinforma 81(8):1–21. https://doi.org/10.1186/1471-2105-8-25
Su S, Xiao R, Xu X et al (2012) (2012) Multi-scale spatial determinants of dissolved oxygen and nutrients in Qiantang River. China Reg Environ Chang 131(13):77–89. https://doi.org/10.1007/S10113-012-0313-6
Tong Y, Xu X, Zhang S et al (2019) Establishment of season-specific nutrient thresholds and analyses of the effects of nutrient management in eutrophic lakes through statistical machine learning. J Hydrol 578:124079. https://doi.org/10.1016/j.jhydrol.2019.124079
UNESCO World Water Assessment Programme (2018) The United Nations world water development report 2018: nature-based solutions for water
Veldkamp TIE, Wada Y, Aerts JCJH et al (2017) Water scarcity hotspots travel downstream due to human interventions in the 20th and 21st century. Nat Commun 8:1–12. https://doi.org/10.1038/ncomms15697
Vörösmarty CJ, McIntyre PB, Gessner MO et al (2010) Global threats to human water security and river biodiversity. Nature 467:555–561. https://doi.org/10.1038/nature09440
Wang X, Yang W (2019) Water quality monitoring and evaluation using remote-sensing techniques in China: a systematic review. Ecosyst Heal Sustain 5:47–56
Wiegand MC, do Nascimento ATP, Costa AC, Lima Neto IE, (2021) Trophic state changes of semi-arid reservoirs as a function of the hydro-climatic variability. J Arid Environ 184:104321. https://doi.org/10.1016/J.JARIDENV.2020.104321
Wiegand MC, Nascimento ATP, do, Costa AC, Lima IE, (2020) Avaliação de nutriente limitante da produção algal em reservatórios do semiárido brasileiro. Rev Bras Ciências Ambient 55:456–478. https://doi.org/10.5327/Z2176-947820200681
Yang XE, Wu X, Hao HL, He ZL (2008) Mechanisms and assessment of water eutrophication. J Zhejiang Univ Sci B 9:197–209
Yu X, Shen J, Du J (2020) A machine-learning-based model for water quality in coastal waters, taking dissolved oxygen and hypoxia in Chesapeake Bay as an example. Water Resour Res 56:e2020WR027227. https://doi.org/10.1029/2020WR027227
Zhao Y, Han Q, Ding C et al (2020) Effect of low temperature on chlorophyll biosynthesis and chloroplast biogenesis of rice seedlings during greening. Int J Mol Sci. https://doi.org/10.3390/IJMS21041390