Advances in forecasting harmful algal blooms using machine learning models: A case study with Planktothrix rubescens in Lake Geneva

Jonathan Derot, Hiroshi Yajima, Stéphan Jacquet

▶ **To cite this version:**

Jonathan Derot, Hiroshi Yajima, Stéphan Jacquet. Advances in forecasting harmful algal blooms using machine learning models: A case study with Planktothrix rubescens in Lake Geneva. Harmful Algae, 2020, 99, 10.1016/j.hal.2020.101906 . hal-03025770

HAL Id: hal-03025770
https://hal.science/hal-03025770

Submitted on 27 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain

Harmful Algae 99 (2020) 101906

Contents lists available at ScienceDirect

Harmful Algae

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

Advances in forecasting harmful algal blooms using machine learning models: A case study with *Planktothrix rubescens* in Lake Geneva

Jonathan Derot[a,*], Hiroshi Yajima[a], Stéphan Jacquet[b]

[a] Estuary Research Center, Shimane University, 1060 Nishikawatsu-cho, Matsue, Shimane 690-8504, Japan
[b] Université Savoie Mont Blanc, INRAE, UMR CARRTEL, 74200 Thonon-les-Bains, France

ARTICLE INFO

Keywords:
Water quality management
Planktothrix rubescens
Peri-alpine lakes
Forecast
Random forest
Individual conditional expectation plots

ABSTRACT

The development of anthropic activities during the 20th century increased the nutrient fluxes in freshwater ecosystems, leading to the eutrophication phenomenon that most often promotes harmful algal blooms (HABs). Recent years have witnessed the regular and massive development of some filamentous algae or cyanobacteria in Lake Geneva. Consequently, important blooms could result in detrimental impacts on economic issues and human health. In this study, we tried to lay the foundation of an HAB forecast model to help scientists and local stakeholders with the present and future management of this peri-alpine lake. Our forecast strategy was based on pairing two machine learning models with a long-term database built over the past 34 years. We created HAB groups via a K-means model. Then, we introduced different lag times in the input of a random forest (RF) model, using a sliding window. Finally, we used a high-frequency dataset to compare the natural mechanisms with numerical interaction using individual conditional expectation plots.

We demonstrate that some HAB events can be forecasted over a year scale. The information contained in the concentration data of the cyanobacteria was synthesized in the form of four intensity groups that directly depend on the *P. rubescens* concentration. The categorical transformation of these data allowed us to obtain a forecast with correlation coefficients that stayed above a threshold of 0.5 until one year for the counting cells and two years for the biovolume data. Moreover, we found that the RF model predicted the best *P. rubescens* abundance for water temperatures around 14°C. This result is consistent with the biological processes of the toxic cyanobacterium. In this study, we found that the coupling between K-means and RF models could help in forecasting the development of the bloom-forming *P. rubescens* in Lake Geneva. This methodology could create a numerical decision support tool, which should be a significant advantage for lake managers.

1. Introduction

The Anthropocene and the phase of great acceleration of this period are subjects of several debates in the scientific community (Chernilo, 2017; Crutzen, 2006; Steffen et al., 2007). Despite different opinions, a consensus was reached about the direct negative effects of anthropic activities on ecosystems and biodiversity (Halpern et al., 2008). Aquatic ecosystems and, more specifically, freshwaters are areas affected by human activities (Dudgeon et al., 2006; Reid et al., 2019). Typically, the decline of biodiversity in freshwater ecosystems could be more important than that in oceans or terrestrial systems (Sala et al., 2000). The increase in nutrient fluxes due, for instance, to intensive agriculture, and/or insufficient treatment of water or sewage, is one of the major causes that generate a cascade of negative effects in freshwater bodies (Roelke et al., 2011; Smith et al., 2006). This increase in nutrients causes eutrophication that, in most cases, promotes the proliferation of some phytoplanktonic populations (Camargo and Alonso, 2006; Schindler, 2006). Commonly referred to as harmful algal blooms (HABs), the intensity and size of these biological events have increased continuously over the past decades (Glibert et al., 2005; Shumway et al., 2018).

The eutrophication of freshwater ecosystems and their associated HABs are recognized global issues and are often linked to the development of filamentous or colonial cyanobacteria (Backer et al., 2015; Bartram and Chorus, 1999). These photosynthetic prokaryotes, represented by a large variety of genera (e.g. *Anabaena, Aphanizomenon, Cylindrospermopsis, Microcysis, Planktothrix* …) can produce toxins, which are likely to be transferred along the food web and cause fish mortality (Sotton et al., 2014; Vanni et al., 1997). These toxins can also directly affect both pet and human health (Briand et al., 2003;

* Corresponding author.
E-mail address: j.derot@soc.shimane-u.ac.jp (J. Derot).

https://doi.org/10.1016/j.hal.2020.101906
Received 26 May 2020; Received in revised form 15 September 2020; Accepted 21 September 2020
1568-9883/ © 2020 Elsevier B.V. All rights reserved.

Codd et al., 2005). Moreover, these biological events can detrimentally affect economic issues affecting tourism, recreational activities, agriculture, fish farms, drinking water supplies … (Carmichael and Boyer, 2016; Reynaud and Lanzanova, 2017).

This problem directly concerns Lake Geneva, which is the largest natural deep lake in Western Europe and is characterized by a mesotrophic state. One million inhabitants use this lake as a drinking water supply, and it has various linked tourism activities and fisheries linked to it (Gallina et al., 2017). Lake Geneva, like some other peri-alpine lakes (e.g., Lake Bourget, France), has been experiencing recurrent problems with the filamentous toxic cyanobacterium *Planktothrix rubescens*, for several years, which is particularly well adapted to this type of environment (Anneville et al., 2015; Gallina et al., 2017; Jacquet et al., 2005; Jacquet et al., 2014; Kerimoglu et al., 2017; Oberhaus et al., 2007). This cyanobacterium can produce different hepatotoxins that could lead to endocrine disruption and, consequently, favor the development of certain types of cancers (Catherine et al., 2017; Pearson et al., 2010). Furthermore, numerous studies have shown that global warming could increase this type of toxic cyanobacteria in the coming years (Gallina et al., 2017; Paerl and Otten, 2013; Wang et al., 2015).

HAB forecasting has become critical in environmental sciences and among the stakeholders (Pennekamp et al., 2019). Although classical hydro-ecological models can efficiently model the physical processes, a meta-analysis, which evaluated the performance of 124 such models, highlighted their difficulties in accurately reproducing the phytoplankton dynamics (Shimoda and Arhonditsis, 2016). These decreases in the predictive performance may be due to the complex interactions and nonlinear mechanisms that regulate the phytoplankton compartment, which are modeled via mathematical models based on a priori assumptions (Arhonditsis, 2009; Edwards et al., 2016; Zhao et al., 2008). In contrast to these hydro-ecological models, one of the major advantages of machine learning models is that they do not require a priori assumptions (Breiman, 2001; Tsanas and Xifara, 2012; Zhao and Zhang, 2008). In the following non-exhaustive list, it is evident that many studies use this advantage of machine learning to bypass the prediction difficulties concerning the phytoplankton: (Cho et al., 2018; Cho and Park, 2019; Du et al., 2018; Kehoe et al., 2015; Lee et al., 2016; Lee and Lee, 2018; Rivero-Calle et al., 2015; Shamshirband et al., 2019; Shin et al., 2017; Thomas et al., 2018; Yajima and Derot, 2018; Zhang et al., 2016).

Regarding restoration, many bio-assessment programs use a class system to define the healthy ecological state of an aquatic ecosystem (Poikane et al., 2019). For the European territory, the protection and restoration of water bodies are supervised by an European directive referred to as the “Water Framework Directive” (WFD), that aims to reach a good ecological state (for the ecosystems) in the coming years (Directive, 2000). This directive is based on five ecological state classes, namely poor, bad, moderate, good and high. In this context, it is important to note that these classes are mainly based on biological metrics, among which the phytoplankton (Laplace-Treyture and Feret, 2016; Le Vu et al., 2011; Phillips et al., 2013). Consequently, forecast models could help lake restoration objectives by testing scenarios with some key issues, such as eutrophication and climate change (Lehmann and Hamilton, 2018; Wang et al., 2018). Such models could also help in forecasting the HAB events and avoid financial losses by creating a decision support tool, to determine the fishing periods in lakes (Gill et al., 2018; Manning et al., 2019).

Long-term monitoring enables us to satisfy bio-assessment program requirements, such as the European Water Framework Directive (Le Vu et al., 2011). In addition, this type of monitoring generates large databases, which can be used as the input for machine learning models (McGovern et al., 2017). In lacustrine ecosystems, the pairing of long-term and or high-frequency databases and a random forest (RF) model, seems to be a good alternative for hydro-ecological models for predicting the phytoplankton biomass and bloom (Breiman, 2001; Thomas et al., 2018; Yajima and Derot, 2018). Moreover, some clustering algorithms based on machine learning models such as the K-means method, enable the creation of classes to address environmental problems (Derot et al., 2020; Hartigan and Wong, 1979; Rousseeuw et al., 2014; Solidoro et al., 2007). Furthermore, the predictive performance of an RF model can be improved by utilizing it in conjunction with a K-means model (Kwon and Park, 2016; Liu and Sun, 2019). However, this compels us to use the RF model in the classification mode; however, this is not inconsistent with bio-assessment programs, which are generally used as a system based on the ecological status in the form of categorical data.

Our goal was to lay the foundations of a numerical model to forecast a case study cyanobacterial bloom in Lake Geneva, having in mind it could be useful for both scientists and lake stakeholders. Therefore, we explored the capacity of the RF model, often considered as a black box, to determine the biological interactions that exist in this natural environment. We also studied the forecast performances of this model by varying the sliding windows over a year-scale.

2. Material and methods

2.1. Databases and sampling point

Lake Geneva, which lies at an altitude of 372 m, forms the border between France and Switzerland in the north of the French Alps. It is a 72 km long ecosystem with an area of 582 km^2 and a maximum width of 13 km. Lake Geneva is composed of two basins, namely a deep central eastern basin called the large lake (Grand Lac), whose deepest point is 309 m below the surface, and a western and more shallow basin, at small lake (Petit Lac), with a maximum depth of 74 m. It is a meromictic lake, never covered by ice, with an average temperature ranging between 4 and 22°C. It holds an approximate volume of 89 Km^3. Lake Geneva was reported as eutrophic during the 1970s and it changed to a mesotrophic state during the 1990s, following the restoration programs in response to the appropriate measures taken to reduce the phosphorus inputs to the lake (Anneville et al., 2002). In the early 1980s, the annual average total phosphorus concentration was approximately 89.5 µgP/L, while, at the present time this concentration dropped below 15 µgP/L (Salmaso et al., 2018).

The datasets used for Lake Geneva were obtained at a single sampling point referred to as SHL2, corresponding to the deepest and pelagic part of the lake (Anneville and Pelletier, 2000). Data were obtained bi-monthly, except for the winter period, for which sampling is performed once a month. The first database is a long-term dataset that starts from 1984 and ends in 2018 (Rimet et al., 2020). Notably, the phytoplankton data before 2000 corresponded to an integrated water sample from the surface to a depth of 10 m. After 2000 the sampling consisted of an extended water column down to 18 m. The phytoplankton counts were obtained using the classical Uthermol method, and the data were presented both as the number of cells per milliliter (mL) and a biovolume measured in μm^3/mL.

For biological purposes, we only used the information available for *Planktothrix rubescens*, the sum of cyanobacteria taxa and chlorophyll-a (Chl-a). The following physicochemical parameters were used: orthophosphate (PO_4), particulate phosphorus (P_{part}), total phosphorus (P_{tot}), dissolved oxygen (O_2), sulfates (SO_4), chlorides (Cl), sodium (Na), particulate organic carbon (COP), complete alkalinity titration (TAC), conductivity (Cond), reactive silica (SiO_2), pH (pH), nitrite (NO_2), ammonia (NH_4), nitrate (NO_3), particulate organic nitrogen (NOP), total nitrogen (N_{tot}) and water temperature (T°C).

The second dataset was obtained from a measuring spectro-fluorescent device, the FluoroProbe (bbe Moldaenke GmbH, Germany). This probe provides a concentration estimate for different algal classes (i.e., green algae, blue-green algae/cyanobacteria, diatoms/dinoflagellates, and cryptophytes/PE-rich groups including cyanobacteria) based on the fluorescence excitation spectra called fingerprints (Beutler et al., 2002;

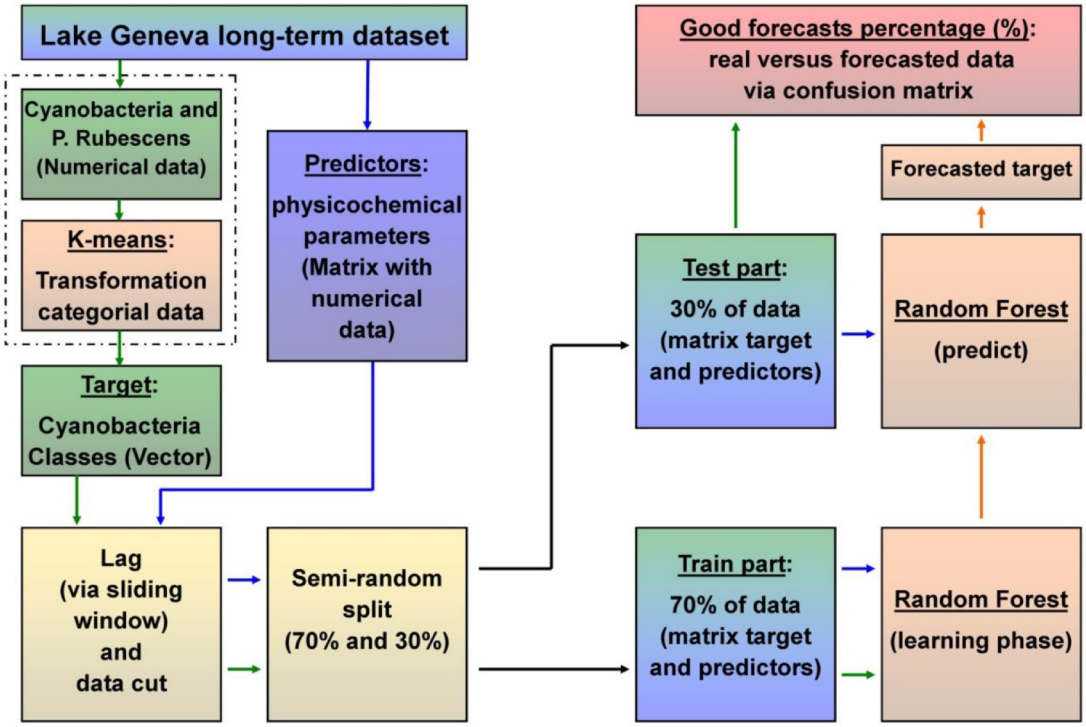

Fig. 1. Conceptual diagram presenting the methodology used to measure the forecast quality, considering all lag times and sampling frequencies.

Catherine et al., 2012; Kring et al., 2014; Leboulanger et al., 2002). The sampling began in June 2017; however, to have a complete year dataset and reduce the computation time, we only used the year 2018. Data were collected from the surface to a depth of 40 m, along with the concerned depth (m), temperature (°C), chlorophyll-a (µg/L), chlorophyll-a corrected in lab (µg/L), green algae (µg/L), blue-green algae (µg/L), diatoms (µg/L), yellow substances (µg/L), and *Planktothrix rubescens* (µg/L) after specific calibration for the cyanobacterium. For each profile obtained with the probe that descended into the water column slowly (approximately $0.1 ms^{-1}$), data were obtained every 0.1 s.

2.2. Machine learning models

A recent study, performed in another Swiss peri-alpine lake, obtained good results for the forecast of phytoplankton and cyanobacteria distribution using an RF model (Thomas et al., 2018). On a broader level, the RF model also yielded good prediction and forecast results in freshwater environments (Derot et al., 2020; Yajima and Derot, 2018). These studies indicated that the RF model is well adapted to our current problem. Furthermore, this model, which is an upgrade of the classification and regression tree (CART) model, has suitable properties regarding the studied databases that include biological components (Breiman et al., 1984). This model has no prior assumptions, is not too sensitive to missing data and is adapted to manage nonlinear processes (Breiman, 2001; Thomas et al., 2018). The term "forest" in this model is derived from the numerous CART models that are created during the learning phase. To decrease the computation time, it is critical to select an appropriate number of trees using the out-of-bag error. With respect to the learning phase for the classification mode (long-term database and Table S1), as evident in Figs. S2-S3 in the supplementary material, the out-of-bag error becomes more stable after 300 trees; therefore in this case we performed our executions with 2000 trees. Similarly, regarding the learning phase for the regression mode (Fig. S7 and Table S1), we performed all our runs with 200 trees. For the minimum number of observations in each node of the RF model, also called the minimum leaf size (min-leaf), we used the default setup proposed by MATLAB. This implies that the value of min-leaf is equal 5 and 1 in the regression and classifications modes, respective (Derot et al., 2020; Yajima and Derot, 2018).

The individual conditional expectation (ICE) plots, which are an improved version of the partial dependence plot (PDP), enables the interpretation of the interactions between the predictors created by the RF model during the learning phase (Friedman et al., 2001; Goldstein et al., 2015). These numerical tools allowed us to find some similarities between the learning phase interactions and biological processes (Cutler et al., 2007; Derot et al., 2020; Roubeix et al., 2016; Teichert et al., 2016). To categorize the cyanobacteria into some intensity classes, we used a clustering algorithm called K-means based on machine learning (Hartigan and Wong, 1979). We used this model at the default setting, which calculates the distance between the centroids with the squared Euclidean distance:

$$d(x, c) = (x - c)(x - c)'$$

where x is an observation and c is the centroid. Here, we used four centroids to obtain four intensity classes. All numerical analyses in this study were performed using MATLAB and its *Statistics and Machine Learning Toolbox*. Concerning the RF model we used the function *TreeBagger*, for the ICE plot, and the functions *plotPartialDependence* and *kmeans* for the clustering model. Moreover, we used the *rng* function to obtain reproducible results. This command allows us to specify the seed, in order to control the random number generation used in the K-means and RF models (process bootstrap during the learning phase), as well as the *cvparation* function (see below in Section 2.3). In other words, all functions that use a random draw will have the same result between each new run.

2.3. Forecast and prediction strategies

To apply the forecasting strategy (Fig. 1), we require a long-term dataset. Therefore, we only applied this strategy to the long-term database of Lake Geneva. The RF model can predict a categorical target with numerical predictors and *vice versa*; additionally, a mix of

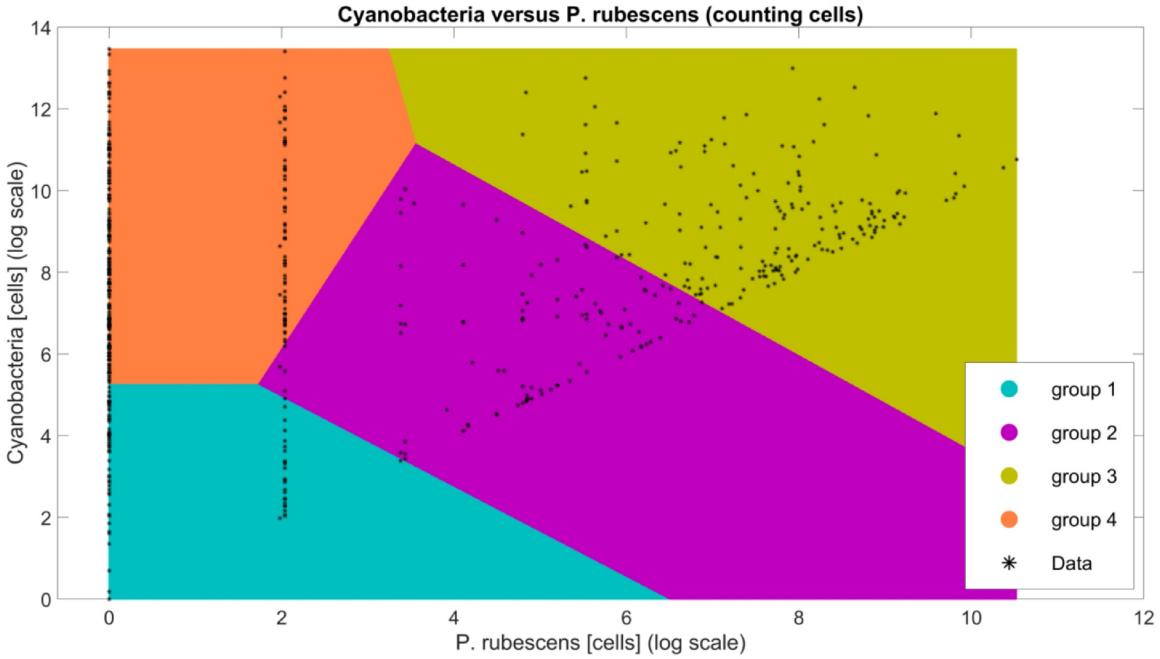

Fig. 2. Output of the K-means model with four *P.rubescens* and cyanobacteria intensity groups, based on the counting cells. Each colored area corresponds to an intensity class: cyan for low concentrations of *P.rubescens* and cyanobacteria; orange for low concentration of *P.rubescens* and high concentration of cyanobacteria; purple for middle concentrations of *P.rubescens* and cyanobacteria; and green for high concentrations of *P.rubescens* and cyanobacteria.

categorical and numerical predictors can be used. The first step of our forecast strategy involved the separation of the target from the predictors. Here, we used a categorical target in the form of four intensity classes depending on the cyanobacteria and *P. rubescens*. We explain this classification in Section 3.1 (Fig. 1). To build these classes, we used the sum of the total cyanobacteria and the *P. rubescens* data, which enabled us to create our target signal via a K-means clustering model. Notably, we created two different pools of intensity classes; one was created with the cell counting and the other with the biovolume. Regarding the predictor matrix in both cases, we used all physicochemical parameters that were presented in Section 2.1 (Fig. 1). It should be also kept in mind that the predictors were not transformed via the K-means model. Therefore, this matrix was composed of numerical data.

The second step of our strategy involved the application of a lag via a sliding window method (Herrera et al., 2010; Yajima and Derot, 2018). We performed feature engineering on the long-term dataset. In other words, we reframed this dataset as a supervised learning problem with the lagged target signal. To provide a practical example, when we applied a lag time of one year, this lag was applied to the target vector (cyanobacteria intensity classes), but not on the predictor matrix (Fig. S10). Therefore, in this case, the input of the RF model used the first value of the target signal and corresponded to the first sampling in 1985. Comparatively, the first values of all predictors correspond to the first samplings in the year 1984. To run an RF model, the length of the target vector must be the same as that of the predictor matrix. Consequently, with this sliding window strategy, we are compelled to discard some data: we discard the first data record of the target signal vector, which corresponds to the lag time; and we also discard the last data record of all predictors, which correspond to the lag time (Fig. S10). Then, we divided the data into two pools; the train part that contains 70% of the data for the learning phase; and the test part with 30% of the remaining data.

These splits were realized with the MATLAB function *cvparation* to obtain a semi-random draw. Using this function avoids, for example, to have one pool that only contains all the highest target values and another one with all the lowest values. Consequently, we obtained two pools with close intensity values. Following this split, we used the train part to perform the learning phase of the RF model. We then only used the predictors from the test part with the MATLAB function *predict* to get the forecast from the trained RF model. We compared these outputs with the real target signal from the test part (Fig. 1). As a reminder, we used the model in the classification mode, because our target signal was composed of numerical data. Therefore, to compare the forecasted target signal versus the lagged real classes, we could not use classical linear regression. Therefore, we compared these two signals with a confusion matrix (Figs 1, 3, and S4). We calculated the entire confusion matrix with the MATLAB function *plotconfusion*. Furthermore, we took the mean of the good forecasting percentage to compare the outputs of these matrixes (see the blue boxes in Figs. 3 and S4). It should be noted that we also explored the performance of the RF model during its learning phase (test part). The misclassification probabilities in Fig. S13, were extracted via the MATLAB function *oobError*. We used the same function to plot the out-of-bag error for the Figs. S2, S3, and S7.

Concerning the other dataset recorded with FluoroProbe and our pretests in the supplementary material (Table S1 and discussion part), we only applied the RF model to obtain a prediction (no lag times). We just extracted the ICE plot for the FluoroProbe database, to examine the interactions between the predictors and the target created by this model during the learning phase (see Section 2.2). In all of these analyses, we also used the *cvparation* function to split these datasets between train (70%) and test (30%) parts. The RF models are still constructed using the MATLAB function *TreeBagger*; however, for the analyses with numerical values, we tuned the model in the regression mode. In addition, when we used this mode the correlation coefficients (R^2) were calculated via the coefficient of determination as follows (Du et al., 2018; Kehoe et al., 2015; Lee et al., 2016; Lee and Lee, 2018; Shamshirband et al., 2019):

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where SS_{tot} is the total sum of squares, SS_{res} is the residual sum of squares, n is the number of observations, $\hat{y}_i$ is the predicted data, y_i is the observed data, and $\bar{y}$ is the mean of the observed data.

3. Results

3.1. Intensity classes based on cyanobacteria

Fig. 2 reveals four groups created by the K-means method from the counting cell dataset. On the x-axis, we placed the *P. rubescens* data and on the y-axis the total of cyanobacteria. There is a wide range in the data with a factor of 4×10^5, between the highest and lowest values in the biovolume of cyanobacteria. Therefore, we applied a log transformation to reduce this range. We performed the same procedure with the biovolume database (Fig. S1). It should be noted that the transformation was realized with the natural logarithm. At the bottom left in the cyan area, we can find the first group. This class contains a low concentration of cyanobacteria and *P. rubescens*. We found the same pattern with the biovolume data, except that more points are agglomerated on the zero. The second group in orange contains a low concentration of *P. rubescens* and a high concentration of cyanobacteria. We also found the same pattern with the biovolume as for the two other classes.

Consequently, the following observations for Fig. 2 are also valid for Fig. S1. The four groups created by the K-means model are not perfectly distributed in four squares. The purple area that represents the third class is composed of a middle-range concentration of *P. rubescens* and the sum of cyanobacteria. The last class in green at the top right contains high values of cyanobacteria and *P. rubescens*. In both cases, for the cell counting and biovolumes, many points are agglomerated on the y-axis, corresponding to the samplings where *P. rubescens* was absent in Lake Geneva.

3.2. Forecast of intensity classes at a year-scale

Fig. 3 shows the confusion matrix results for the counting cells without lag. The green boxes show the good predictions. The upper numbers in bold indicate the number of observations that were well classified; the percentage below the bold numbers represent the validation phase that was performed from the test par dataset. For example, in class 4, 71 observations were well classified, representing 32.1% of the test part data. The red boxes use the same annotation system for the wrong classification. For instance, 9.5% of the observations in class 4 were predicted in class 3, for 21 misclassified observations in that group. The gray boxes show the average percentage for a specific class, such as class 4, where 84.5% of the data were well classified. We can find the overall average of good classified observation in green in the blue box. In that case 62% of these intensity classes were well predicted. Fig. S4 shows the confusion matrix for the biovolume without lag. There is more misclassification for class 1, which could explain the difference of 2% in the overall average of good prediction between these two cases. Regarding the learning phase, in the supplementary material (Fig S13), we extracted the misclassification probability from the train part dataset.

We extracted the percentage of the well-predicted results of the blue boxes, as shown in Fig. 3, for each lag time to create Fig. 4. Therefore, on the y-axis that represents the average of good percentage of good forecast; two values of 60.2% and 62% were observed for the no-lag cases. The x-axis shows the input lag that was introduced with the sliding window strategy. To facilitate reading, the x-axis is represented on a log scale. The red line and green lines denote the results for the biovolume and counting cell, respectively. With respect to the forecast of the biovolume, the value of R^2 starts to decrease from the two-year lag, but these coefficients always stay above a threshold of 50% of the good forecast. In the case of the counting cells, even if the no-lag coefficient is better, the lagged R^2 decreases quicker than the red line. Furthermore, these correlation coefficients start to decrease from the one-year lag. After this point the other R^2 values fall below the 0.5 threshold. However, the general evolution of these coefficients follows the same tendency in both cases. In the supplementary material (Fig. S20 for biovolume and Fig S21. for counting cells), we also extracted the percentages of good forecast from the output classes for each group.

3.3. Similitude between interactions

We did not implement the transformation via the K-means method and the sliding window strategy on the FluoroProbe data from 2018. We only predicted the concentration of *P. rubescens* using the RF model in the regression mode. Fig. 5 shows the results of the comparison between the real data from the test part and the forecast target signal. Both, standard and adjusted correlations coefficients are extremely high because they are equal to 0.90. The Kendall and Spearman coefficients that measure the dynamics of the co-evolution between the two signals were found to be 0.93 and 0.97, respectively. Thus, it can be elucidated that there is a strongly correlated dynamic between the prediction and real data. This is demonstrated in another manner in Fig. S9, which shows that the dynamic of the red line (prediction) follows the black line that represents the real concentration of *P. rubescens* from the test part. Furthermore, a disparity is evident in the predicted signal between the 1300 and 1400 points. This irregularity is displayed in Fig. 5, where the real data above 4µg/L are not well correlated.

Fig. S8 shows the results of the out-of-bag error extraction for this run. Our findings show that temperature is the most important predictor with an intensity of over 3.5. The diatoms, which are the second most important predictor, only reach an importance of 2. Fig. 6 illustrates the extraction of the ICE plot from the learning phase for the temperature. The ICE plots of other predictors are included in the supplementary material. Thus, we can observe the internal interactions created by the RF model during its learning phase between the temperature (x-axis) and concentration of *P. rubescens* (y-axis). The red line represents the PDP analysis, and the blue points were created by the ICE method. The highest values of *P. rubescens* were predicted around 14°C, and the lowest concentrations of cyanobacteria were predicted for the lowest temperature under 8°C. Notably, owing to the tree structure, the interactions created during the learning phase of the RF model are co-dependent. In other words, each child node in this structure depends on the previous child node or root node. Consequently, each prediction in the terminal nodes is dependent on several different predictors. Trivially, for example, if the temperature is above 10°C (root node), then the depth of 15m or less (first child node), then the concentration of the diatoms is above 1.2µg/L (second child node), the RF model would predict a *P. rubescens* concentration equal to 2.5 µg/L (one terminal node). In this example, the influence of the 15m depth is only valid if the water temperature is greater than 10°C, and so on. Therefore, this ICE plot is not a perfect tool, but it gives some good indications for the physicochemical thresholds (Derot et al., 2020; Roubeix et al., 2016).

4. Discussion

In our pretest on this dataset, we first expected to directly forecast the *P. rubescens* concentration as numerical data. Unfortunately, the correlation coefficient between the test part and real data was low (Table S1). Then, following the same process, we attempted to predict the chlorophyll-a and cyanobacteria; we even tried to predict the *P. rubescens* signal transformed in the categorical data (Table S1). Nevertheless, none of these tests surpassed a threshold of R^2 above 0.5, which is a general consensus where in a model begins to provide accurate prediction. Consequently, using these four groups based on the sum of cyanobacteria and P. rubescens presented in Section 3.1 (Fig. 2) enabled us to exceed this threshold for the prediction without lag time (Figs. 3 and S4). We have highlighted in the introduction that Lake Geneva is concerned by the European WFD, which uses five ecological state classes. Accordingly, we also attempted to split our data into five groups using the K-means model. However, as shown in Figs. S11-S12, it was difficult to make a clear biological distinction between some of these groups. Moreover, the average correlation coefficients which included lag times was lower. For the biovolume with four groups, an

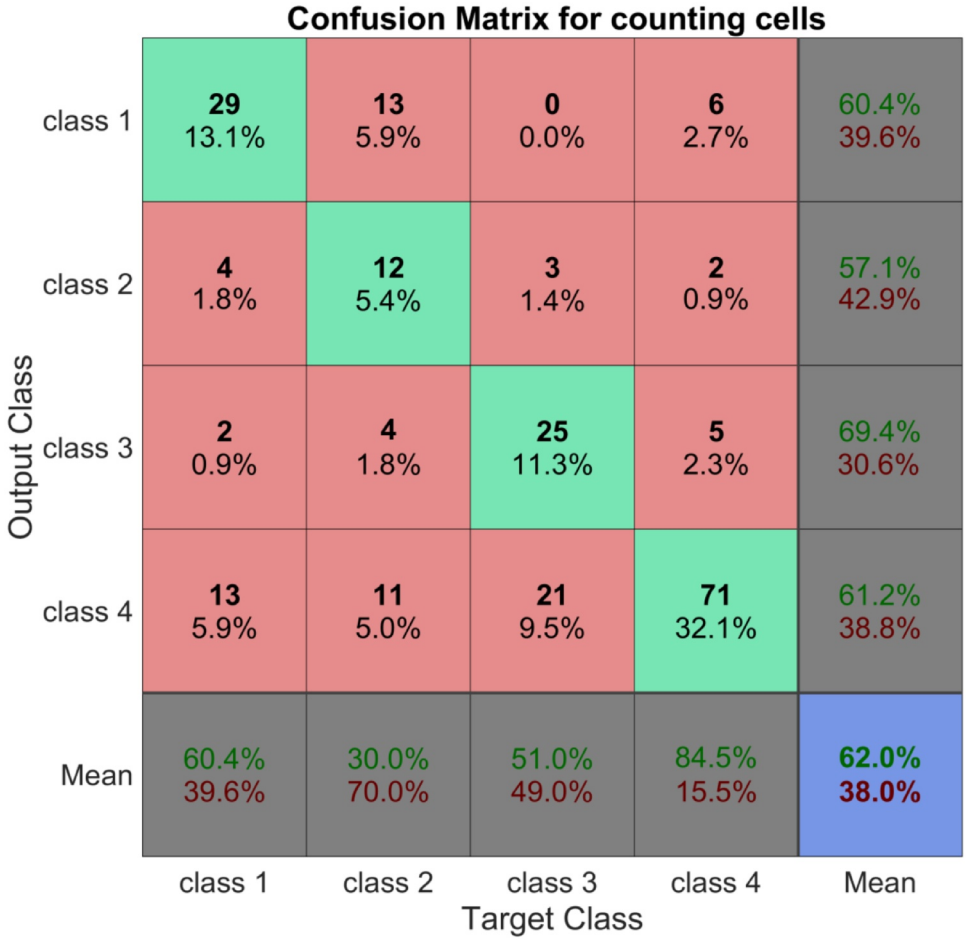

Fig. 3. Confusion matrix for counting cells with no lag time with four intensity cyanobacteria classes, between 1984 and 2018. This analysis denotes a validation phase, which was performed from the test part dataset. The gray boxes show the rate of classification for each group. The diagonal of the matrix represents the well-classified groups, and the overall rate is presented in the blue box.

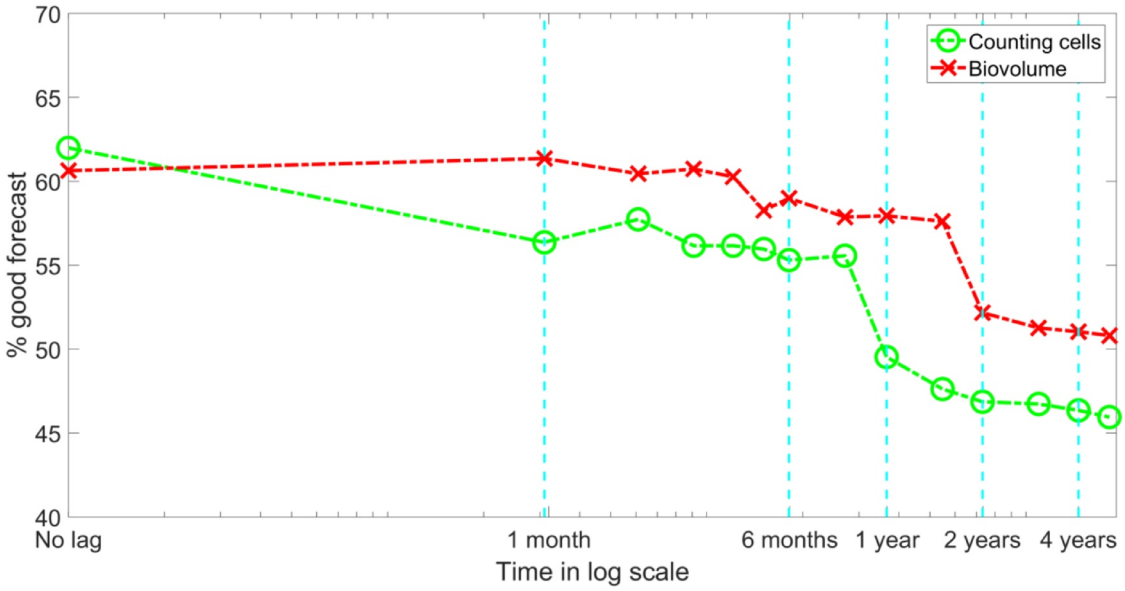

Fig. 4. Evolution of the percentage of good forecasting depending on a lag time of up to 5 years. The x-axis represents the lag times from the sliding window in the log scale. The y-axis shows the correlation matrix average of the good percentage of the forecast from an RF model, for the validation phase (test part). The red and red lines display the evolution of the biovolume dataset and counting cells, respectively.

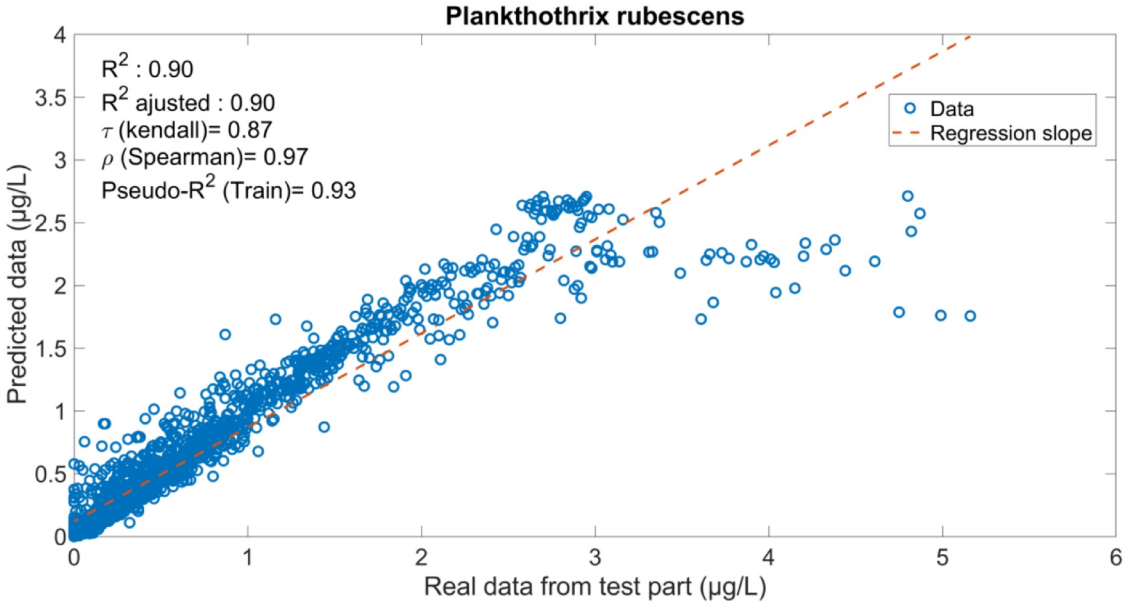

Fig. 5. Correlation between the real (x-axis) and predicted data for *P. rubescens* (y-axis) in 2018 from Lake Geneva.

average value of 56.47% was observed in comparison to that of 48.85% for five groups; for the counting cells with four groups, an average value of 51.76% was observed in comparison to that of 46.41% for five groups.

Regarding the biovolume, we obtained an R^2 of 60.6%; for the prediction based on counting cells, we obtained a correlation coefficient of 62.2%. In Table S1, these coefficients lie between 0.031 and 0.47. This corresponds to a factor improvement between 1.3 and 2000. Therefore, the use of the K-means model for transforming our target signal in the four intensity classes enabled us to overcome the starting threshold of 0.5, considering the correlation coefficient between the test part and real data. Moreover, the categorical transformation of data retains the information related to *P. rubescens*, contrary to the prediction of chlorophyll-a. This class system is not significantly from the healthy ecological state used in bio-assessment programs (Poikane et al., 2019). However, the metrics used as the basis for these ecological states in the WFD are often opaque and do not adequately represent the different management problems for various water bodies (Waylen et al., 2019). In addition, this European directive requires only 4 to 6 samples per season to determine these ecological states (Directive, 2000). This low sampling frequency can bias the determination of these states in aquatic ecosystems (Bresciani et al., 2011). In this context, the use of remote sensing data or data from automated monitoring stations at fixed points could help in preventing this type of bias (Bresciani et al., 2011; Le Vu et al., 2011). With respect to, the HAB events directly linked to the cyanobacteria in peri-alpine lakes in particular, this type of automated sampling station has already presented valuable results (Le Vu et al., 2011; Pomati et al., 2011; Thomas et al., 2018). It could, therefore, be interesting to try to develop an ecological class system, in conjunction with the Swiss and French stakeholders

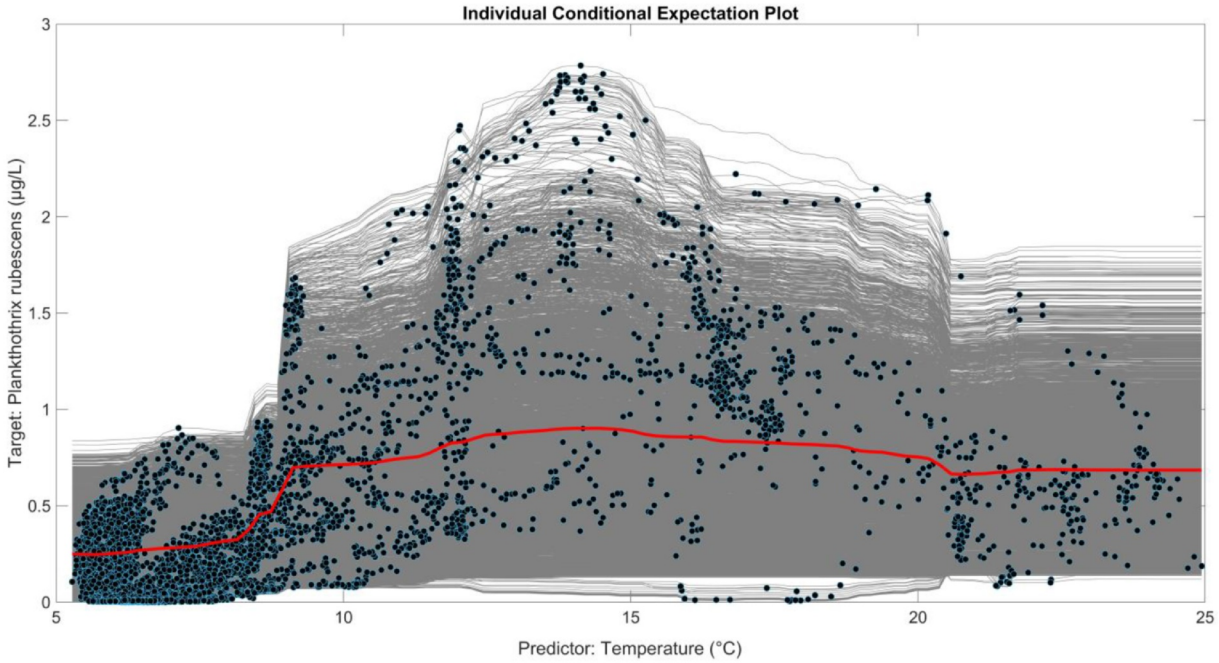

Fig. 6. ICE plots for the temperature. The red line shows the PDP; the gray lines and blue points are derived from the ICE analyses.

that would respond precisely to problems related to Lake Geneva. Subsequently, this potential new class system with the methodology presented here should be re-analyzed, to check whether the forecast performances of the RF model are still as encouraging as our current results. To summarize, the creation of these classes based on *P. rubescens* and cyanobacteria via the K-means method allowed us to drive the RF model in the classification mode for the forecast tests, while maintaining the initial R^2 greater than to 0.5.

In Section 3.2, we wanted to find a way to remain above the threshold of R^2 of 0.5 for the forecast over a year-scale. Consequently, we attempted to transform the information about cyanobacteria and *P. rubescens* into categorical data. It can be observed in Figs. 3 and S4 that the average classes of intensity (see gray boxes) were well predicted. Although the correlation coefficient without lag was slightly lower for the target signal from the biovolume, its evolution over a year scale was more stable than that of the signal from the counting cells (Fig. 4). Moreover, we extracted the out-of-bag feature importance for our two cases in Figs S5. and S6. Our results show that in both cases the five more important predictors include Cl, Na, O_2, PO_4, and N_{tot}. The nutrients PO_4 and N_{tot}, which are the fundamental resources for growth and bloom development, are always 30% or 40% more important than the temperature. Furthermore, our preliminary tests allowed us to remove the light from the predictors because this parameter always had the lowest out-of-bag score. We found some similarities between our findings and the scientific literature. It has been demonstrated that ammonium chloride (NH_4Cl) and sodium nitrate ($NaNO_3$) could influence the growth rate of cyanobacteria via a mechanism linked to phycobiliprotein (Khazi et al., 2018). Furthermore, it also seems to have some links between the salinity (NaCl) and the production of microcystin (Fathabad et al., 2019). Consequently, the use of these intensity classes based on the biovolume seems to be a good alternative, and the five most influential predictors are coherent with the literature.

In another study performed in a Swiss peri-alpine lake, where the authors used an RF model to forecast the cyanobacteria (Thomas et al., 2018). Notably, they used the pseudo-R^2 to validate their forecasts (Breiman, 2001). However, several studies, including this one, know that the coefficients cannot be accessed directly as a true forecast because the pseudo-R^2 measures the goodness-of-fit during the learning phase (Large et al., 2015; Teichert et al., 2016; Thomas et al., 2018). Our results can use the R^2 from the test part instead of the learning phase. Our findings could be used as a basis for the creation of a numerical model to forecast the HAB risks. Numerical models that can forecast HAB events can be used as decision support tools for fisheries to avoid financial losses, by determining the high-risk periods (Gill et al., 2018). In the case of Lake Geneva, we could define three fishing period classes, namely: safe, moderate risk, and high risk. This kind forecast model could be useful to help the local stakeholders with the preservation and restoration of Lake Geneva. In Section 3.2, we highlighted that the use of classes based on the biovolume from a K-means model enabled us to forecast these HAB risks over a year-scale, while it holds the R^2 threshold of 0.5. Furthermore, the most important predictors could be linked to the natural mechanisms of cyanobacteria.

In Section 3.3, the database from FluoroProbe included a matrix of 6902 columns (sampling) and 9 rows. The first row corresponded to the *P. rubescens* concentration and the other 8 rows corresponded to the predictors. For the long-term dataset, we have a matrix of 754 columns (sampling) and 21 rows (target and predictors). Therefore, we have a factor of almost 10 between the length of FluoroProbe and the long-term database. Consequently, we explored whether an increase in the number of sampling could affect the prediction performance and numerical interactions created by the RF model during its learning phase. Our results demonstrate that temperature is the most influential predictor (Fig. S8). After the ICE analysis in Fig. 6, we observed that the RF model predicted the highest values of *P. rubescens* for water temperatures around 14°C. This result is consistent with the scientific literature, which revealed that *P. rubescens* has a competitive advantage for water temperatures of 15°C (Oberhaus et al., 2007) and the biomass of the cyanobacterium is likely to increase in Lake Geneva in the coming years due to global warming (Gallina et al., 2017).

We also extracted the ICE plots for the other predictors. Unfortunately, their interpretations were difficult and did not allow us to observe clear trends. As stated earlier, the interactions created between the predictors by the RF model during the learning phase are codependent. Accordingly, in many cases these interactions could be different from the known mechanisms, such as the biological processes. Furthemore, the database from FluoroProbe, which contains 10 times of sampling, enabled us to use the RF model in the classification mode and produced exceptionally good R^2 for *P. rubescens* (Fig. 5).

Our findings suggest that a high number of sampling positively affects the accuracy of the RF model to find a correlation between the cyanobacteria and temperature. Even if the RF model has no prior assumptions when used with the FluoroProbe dataset (high-frequency), it could recreate some numerical interactions that are significantly similar to the biological mechanisms. Furthermore, the positive impact of this high-frequency database indicates that the development of a sampling strategy in Lake Geneva, which is based on high-frequency sensors at a fixed point, could improve the forecast of cyanobacterial concentrations with numerical data. This analysis, based on the FluoroProbe dataset, allowed us to highlight the importance of using a high-frequency sensor to improve the performance of an RF model for biological processes. Additionally, even if the machine learning models are, in general, considered as black boxes, our results showed that this model could recreate biological interactions.

To summarize, we found that the use of a class system can enable us to forecast the intensity of harmful cyanobacteria in Lake Geneva over a year scale. Our findings also demonstrated that with a large amount of input data, even if the RF model has no prior assumption, it could find an interaction that is significantly similar to the competitive mechanism of HABs. Moreover, the correlation coefficient obtained with this database from high-frequency sensors is higher than those obtained with the long-term dataset. Consequently, our study suggests that there are two possible options for improving the forecasting of HABs in peri-alpine lakes. The first option involves the use automated system equipped with high-frequency sensors to improve the sampling strategy to detect the blooms of P. rubescens and cyanobacteria. The second option includes re-framing of these numerical data as classification problems with intensity groups, which allows us to obtain better performance with an RF model. However, these two options are not incompatible; therefore, it could be interesting to use them jointly. In this framework, it would be interesting to develop an ecological index in conjunction with the local stakeholders, to elaborate a decision support tool adapted for the preservation and restoration issues of Lake Geneva.

Authors' contributions

J.D. conceived the ideas/methodology; H.Y. contributed for the discussion of results and methodology; J.D. analyzed data and led manuscript writing; S.J helped to obtain all the data set and contributed for the discussion of results; H.Y., S.J. contributed to manuscript editing. All authors gave approval for publication.

Data availability statement

The datasets used for Lake Geneva are obtained at a single sampling point referred to as SHL2, corresponding to the deepest and pelagic part of the lake. Data were obtained bi-monthly, except for the winter period, for which sampling is performed once a month (© OLA-IS, AnaEE-France, INRAE of Thonon-les-Bains, CIPEL [Rimet and al. 10.4081/jlimnol.2020.1944]).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by a grant from the Japan Society for the Promotion of Science (JSPS). Derot J. benefited of JSPS Postdoctoral Fellowship for Research in Japan.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.hal.2020.101906.

References

Anneville, O., Domaizon, I., Kerimoglu, O., Rimet, F., Jacquet, S., 2015. Blue-green algae in a “Greenhouse Century”? New insights from field data on climate change impacts on cyanobacteria abundance. Ecosystems 18 (3), 441–458.
Anneville, O., Pelletier, J.P., 2000. Recovery of Lake Geneva from eutrophication: quantitative response of phytoplankton. Archiv für hydrobiologie 148 (4), 607–624.
Anneville, O., Souissi, S., Ibanez, F., Ginot, V., Druart, J.C., Angeli, N., 2002. Temporal mapping of phytoplankton assemblages in Lake Geneva: annual and interannual changes in their patterns of succession. Limnol. Oceanogr. 47 (5), 1355–1366.
Arhonditsis, G.B., 2009. Useless arithmetic? Lessons learnt from aquatic biogeochemical modeling. Modelling of Pollutants in Complex Environmental Systems 1. pp. 1.
Backer, L.C., Manassaram-Baptiste, D., LePrell, R., Bolton, B., 2015. Cyanobacteria and algae blooms: review of health and environmental data from the harmful algal bloom-related illness surveillance system (HABISS) 2007–2011. Toxins 7 (4), 1048–1064.
Bartram, J., Chorus, I., 1999. Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management. CRC Press.
Beutler, M., Wiltshire, K.H., Meyer, B., Moldaenke, C., Lüring, C., Meyerhöfer, M., Hansen, U.-P., Dau, H., 2002. A fluorometric method for the differentiation of algal populations in vivo and in situ. Photosynthesis Res. 72 (1), 39–53.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Wadsworth Int. Group 37 (15), 237–251.
Bresciani, M., Stroppiana, D., Odermatt, D., Morabito, G., Giardino, C., 2011. Assessing remotely sensed chlorophyll-a for the implementation of the Water Framework Directive in European perialpine lakes. Sci. Total Environ. 409 (17), 3083–3091.
Briand, J.-F., Jacquet, S., Bernard, C., Humbert, J.-F., 2003. Health hazards for terrestrial vertebrates from toxic cyanobacteria in surface water ecosystems. Veterinary Res. 34 (4), 361–377.
Camargo, J.A., Alonso, Á, 2006. Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: a global assessment. Environ. Int. 32 (6), 831–849.
Carmichael, W.W., Boyer, G.L., 2016. Health impacts from cyanobacteria harmful algae blooms: implications for the North American Great Lakes. Harmful Algae 54, 194–212.
Catherine, A., Bernard, C., Spoof, L., Bruno, M., 2017. Microcystins and nodularins. Handbook of Cyanobacterial Monitoring and Cyanotoxin Analysis. Wiley, pp. 109–126.
Catherine, A., Escoffier, N., Belhocine, A., Nasri, A., Hamlaoui, S., Yéprémian, C., Bernard, C., Troussellier, M., 2012. On the use of the FluoroProbe®, a phytoplankton quantification method based on fluorescence excitation spectra for large-scale surveys of lakes and reservoirs. Water Res. 46 (6), 1771–1784.
Chernilo, D., 2017. The question of the human in the Anthropocene debate. Eur. J. Soc. Theory 20 (1), 44–60.
Cho, H., Choi, U., Park, H., 2018. Deep learning application to time-series prediction of daily chlorophyll-a concentration. WIT Trans. Ecol. Environ. 215, 157–163.
Cho, H., Park, H., 2019. Merged-LSTM and multistep prediction of daily chlorophyll-a concentration for algal bloom forecast. IOP Conference Series: Earth and Environmental Science. IOP Publishing, 012020.
Codd, G.A., Morrison, L.F., Metcalf, J.S., 2005. Cyanobacterial toxins: risk management for health protection. Toxicol. Appl. Pharmacol. 203 (3), 264–272.
Crutzen, P.J., 2006. The “anthropocene”, Earth System Science in the Anthropocene. Springer, pp. 13–18.
Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.
Derot, J., Jamoneau, A., Teichert, N., Rosebery, J., Morin, S., Laplace-Treyture, C., 2020. Response of phytoplankton traits to environmental variables in French lakes: New perspectives for bioindication. Ecol. Indicators 108, 105659.
Directive, W.F., 2000. 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. Off. J. Eur. Comm. 22 (12), 2000.
Du, Z., Qin, M., Zhang, F., Liu, R., 2018. Multistep-ahead forecasting of chlorophyll a using a wavelet nonlinear autoregressive network. Knowl.-Based Syst. 160, 61–70.
Dudgeon, D., Arthington, A.H., Gessner, M.O., Kawabata, Z.-I., Knowler, D.J., Lévêque, C., Naiman, R.J., Prieur-Richard, A.-H., Soto, D., Stiassny, M.L., 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. Biol. Rev. 81 (2), 163–182.
Edwards, K.F., Thomas, M.K., Klausmeier, C.A., Litchman, E., 2016. Phytoplankton growth and the interaction of light and temperature: a synthesis at the species and community level. Limnol. Oceanogr. 61 (4), 1232–1244.
Fathabad, S.G., Tabatabai, B., Jafar, S., Sitther, V., 2019. Microcystin levels in selected cyanobacteria exposed to varying salinity. J. Water Resource Protect. 11 (4), 395–403.
Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. Springer Series in Statistics New York.
Gallina, N., Beniston, M., Jacquet, S., 2017. Estimating future cyanobacterial occurrence and importance in lakes: a case study with Planktothrix rubescens in Lake Geneva. Aquatic Sci. 79 (2), 249–263.
Gill, D., Rowe, M., Joshi, S.J., 2018. Fishing in greener waters: understanding the impact of harmful algal blooms on Lake Erie anglers and the potential for adoption of a forecast model. J. Environ. Manag. 227, 248–255.
Glibert, P.M., Anderson, D.M., Gentien, P., Granéli, E., Sellner, K.G., 2005. The global, complex phenomena of harmful algal blooms.
Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Gr. Stat. 24 (1), 44–65.
Halpern, B.S., Walbridge, S., Selkoe, K.A., Kappel, C.V., Micheli, F., D'Agrosa, C., Bruno, J.F., Casey, K.S., Ebert, C., Fox, H.E., 2008. A global map of human impact on marine ecosystems. Science 319 (5865), 948–952.
Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: a k-means clustering algorithm. J. R. Stat. Soc. Series C (Appl. Stat.) 28 (1), 100–108.
Herrera, M., Torgo, L., Izquierdo, J., Pérez-García, R., 2010. Predictive models for forecasting hourly urban water demand. J. Hydrol. 387 (1-2), 141–150.
Jacquet, S., Briand, J.-F., Leboulanger, C., Avois-Jacquet, C., Oberhaus, L., Tassin, B., Vinçon-Leite, B., Paolini, G., Druart, J.-C., Anneville, O., 2005. The proliferation of the toxic cyanobacterium Planktothrix rubescens following restoration of the largest natural French lake (Lac du Bourget). Harmful Algae 4 (4), 651–672.
Jacquet, S., Kerimoglu, O., Rimet, F., Paolini, G., Anneville, O., 2014. Cyanobacterial bloom termination: the disappearance of Planktothrix rubescens from Lake Bourget (France) after restoration. Freshwater Biol. 59 (12), 2472–2487.
Kehoe, M.J., Chun, K.P., Baulch, H.M., 2015. Who smells? Forecasting taste and odor in a drinking water reservoir. Environ. Sci. Technol. 49 (18), 10984–10992.
Kerimoglu, O., Jacquet, S., Vinçon-Leite, B., Lemaire, B.J., Rimet, F., Soulignac, F., Trevisan, D., Anneville, O., 2017. Modelling the plankton groups of the deep, perialpine Lake Bourget. Ecol. Model. 359, 415–433.
Khazi, M.I., Demirel, Z., Dalay, M.C., 2018. Evaluation of growth and phycobiliprotein composition of cyanobacteria isolates cultivated in different nitrogen sources. J. Appl. Phycol. 30 (3), 1513–1523.
Kring, S.A., Figary, S.E., Boyer, G.L., Watson, S.B., Twiss, M.R., 2014. Rapid in situ measures of phytoplankton communities using the bbe FluoroProbe: evaluation of spectral calibration, instrument intercompatibility, and performance range. Canad. J. Fisheries Aquatic Sci. 71 (7), 1087–1095.
Kwon, O.H., Park, S.H., 2016. Identification of influential weather factors on traffic safety using k-means clustering and random forest. Advanced Multimedia and Ubiquitous Engineering. Springer, pp. 593–599.
Laplace-Treyture, C., Feret, T., 2016. Performance of the Phytoplankton Index for Lakes (IPLAC): a multimetric phytoplankton index to assess the ecological status of water bodies in France. Ecol. Indicat. 69, 686–698.
Large, S.I., Fay, G., Friedland, K.D., Link, J.S., 2015. Quantifying patterns of change in marine ecosystem response to multiple pressures. PLoS One 10 (3).
Le Vu, B., Vinçon-Leite, B., Lemaire, B.J., Bensoussan, N., Calzas, M., Drezen, C., Deroubaix, J.-F., Escoffier, N., Degres, Y., Freissinet, C., 2011. High-frequency monitoring of phytoplankton dynamics within the European water framework directive: application to metalimnetic cyanobacteria. Biogeochemistry 106 (2), 229–242.
Leboulanger, C., Dorigo, U., Jacquet, S., Le Berre, B., Paolini, G., Humbert, J.-F., 2002. Application of a submersible spectrofluorometer for rapid monitoring of freshwater cyanobacterial blooms: a case study. Aquatic Microbial Ecol. 30 (1), 83–89.
Lee, G., Bae, J., Lee, S., Jang, M., Park, H., 2016. Monthly chlorophyll-a prediction using neuro-genetic algorithm for water quality management in Lakes. Desalination Water Treat. 57 (55), 26783–26791.
Lee, S., Lee, D., 2018. Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models. Int. J. Environ. Res. Public Health 15 (7), 1322.
Lehmann, M.K., Hamilton, D.P., 2018. Modelling water quality to support lake restoration. Lake Restoration Handbook. Springer, pp. 67–105.
Liu, D., Sun, K., 2019. Random forest solar power forecast based on classification optimization. Energy 187, 115940.
Manning, N.F., Wang, Y.-C., Long, C.M., Bertani, I., Sayers, M.J., Bosse, K.R., Shuchman, R.A., Scavia, D., 2019. Extending the forecast model: predicting Western Lake Erie harmful algal blooms at multiple spatial scales. J. Great Lakes Res. 45 (3), 587–595.
McGovern, A., Elmore, K.L., Gagne, D.J., Haupt, S.E., Karstens, C.D., Lagerquist, R., Smith, T., Williams, J.K., 2017. Using artificial intelligence to improve real-time decision-making for high-impact weather. Bull. Am. Meteorol. Soc. 98 (10), 2073–2090.
Oberhaus, L., Briand, J., Leboulanger, C., Jacquet, S., Humbert, J., 2007. Comparative effects of the quality and quantity of light and temperature on the growth of Planktothrix agardhii and P. rubescens 1. J. Phycol. 43 (6), 1191–1199.
Paerl, H.W., Otten, T.G., 2013. Blooms bite the hand that feeds them. Science 342 (6157),

433–434.
Pearson, L., Mihali, T., Moffitt, M., Kellmann, R., Neilan, B., 2010. On the chemistry, toxicology and genetics of the cyanobacterial toxins, microcystin, nodularin, saxitoxin and cylindrospermopsin. Marine Drugs 8 (5), 1650–1680.
Pennekamp, F., Iles, A.C., Garland, J., Brennan, G., Brose, U., Gaedke, U., Jacob, U., Kratina, P., Matthews, B., Munch, S., 2019. The intrinsic predictability of ecological time series and its potential to guide forecasting. Ecol. Monographs 89 (2), e01359.
Phillips, G., Lyche-Solheim, A., Skjelbred, B., Mischke, U., Drakare, S., Free, G., Järvinen, M., de Hoyos, C., Morabito, G., Poikane, S., 2013. A phytoplankton trophic index to assess the status of lakes for the Water Framework Directive. Hydrobiologia 704 (1), 75–95.
Poikane, S., Kelly, M.G., Herrero, F.S., Pitt, J.-A., Jarvie, H.P., Claussen, U., Leujak, W., Solheim, A.L., Teixeira, H., Phillips, G., 2019. Nutrient criteria for surface waters under the European Water Framework Directive: current state-of-the-art, challenges and future outlook. Sci. Total Environ. 695, 133888.
Pomati, F., Jokela, J., Simona, M., Veronesi, M., Ibelings, B.W., 2011. An automated platform for phytoplankton ecology and aquatic ecosystem monitoring. Environ. Sci. Technol. 45 (22), 9658–9665.
Reid, A.J., Carlson, A.K., Creed, I.F., Eliason, E.J., Gell, P.A., Johnson, P.T., Kidd, K.A., MacCormack, T.J., Olden, J.D., Ormerod, S.J., 2019. Emerging threats and persistent conservation challenges for freshwater biodiversity. Biol. Rev. 94 (3), 849–873.
Reynaud, A., Lanzanova, D., 2017. A global meta-analysis of the value of ecosystem services provided by lakes. Ecol. Econ. 137, 184–194.
[dataset]Rimet, F., Anneville, O., Barbet, D., Chardon, C., Crépin, L., Domaizon, I., Dorioz, J.-M., Espinat, L., Frossard, V., Guillard, J., Goulon, C., Hamelet, V., Hustache, J.-C., Jacquet, S., Lainé, L., Montuelle, B., Perney, P., Quetin, P., Rasconi, S., Schellenberger, A., Tran-Khac, V., Monet, G., 2020. The Observatory on LAkes (OLA) database: sixty years of environmental data accessible to the public. J. Limnol. 10.4081/jlimnol.2020.1944.
Rivero-Calle, S., Gnanadesikan, A., Del Castillo, C.E., Balch, W.M., Guikema, S.D., 2015. Multidecadal increase in North Atlantic coccolithophores and the potential role of rising CO_2. Science 350 (6267), 1533–1537.
Roelke, D.L., Grover, J.P., Brooks, B.W., Glass, J., Buzan, D., Southard, G.M., Fries, L., Gable, G.M., Schwierzke-Wade, L., Byrd, M., 2011. A decade of fish-killing Prymnesium parvum blooms in Texas: roles of inflow and salinity. J. Plankton Res. 33 (2), 243–253.
Roubeix, V., Danis, P.-A., Feret, T., Baudoin, J.-M., 2016. Identification of ecological thresholds from variations in phytoplankton communities among lakes: contribution to the definition of environmental standards. Environ. Monit. Assess. 188 (4), 246.
Rousseeuw, K., Caillault, E.P., Lefebvre, A., Hamad, D., 2014. Hybrid hidden Markov model for marine environment monitoring. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 8 (1), 204–213.
Sala, O.E., Chapin, F.S., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L.F., Jackson, R.B., Kinzig, A., 2000. Global biodiversity scenarios for the year 2100. Science 287 (5459), 1770–1774.
Salmaso, N., Anneville, O., Straile, D., Viaroli, P., 2018. European large perialpine lakes under anthropogenic pressures and climate change: present status, research gaps and future challenges. Hydrobiologia 824 (1), 1–32.
Schindler, D.W., 2006. Recent advances in the understanding and management of eutrophication. Limnol. Oceanogr. 51 (1part2), 356–363.
Shamshirband, S., Jafari Nodoushan, E., Adolf, J.E., Abdul Manaf, A., Mosavi, A., Chau, K.-w., 2019. Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters. Eng. Appl. Comput. Fluid Mech. 13 (1), 91–101.
Shimoda, Y., Arhonditsis, G.B., 2016. Phytoplankton functional type modelling: running before we can walk? A critical evaluation of the current state of knowledge. Ecol. Modell. 320, 29–43.
Shin, J., Yoon, S., Cha, Y., 2017. Prediction of cyanobacteria blooms in the lower Han River (South Korea) using ensemble learning algorithms. Desalinat. Water Treat. 84, 31–39.
Shumway, S.E., Burkholder, J.M., Morton, S.L., 2018. Harmful Algal Blooms: A Compendium Desk Reference. John Wiley & Sons.
Smith, V.H., Joye, S.B., Howarth, R.W., 2006. Eutrophication of freshwater and marine ecosystems. Limnol. Oceanogr. 51 (1part2), 351–355.
Solidoro, C., Bandelj, V., Barbieri, P., Cossarini, G., Umani, S.F., 2007. Understanding dynamic of biogeochemical properties in the northern Adriatic Sea by using self--organizing maps and k-means clustering. J. Geophys. Res.: Oceans 112 (C7).
Sotton, B., Guillard, J., Anneville, O., Maréchal, M., Savichtcheva, O., Domaizon, I., 2014. Trophic transfer of microcystins through the lake pelagic food web: evidence for the role of zooplankton as a vector in fish contamination. Sci. Total Environ. 466, 152–163.
Steffen, W., Crutzen, P.J., McNeill, J.R., 2007. The Anthropocene: are humans now overwhelming the great forces of nature. AMBIO: J. Hum. Environ. 36 (8), 614–621.
Teichert, N., Borja, A., Chust, G., Uriarte, A., Lepage, M., 2016. Restoring fish ecological quality in estuaries: implication of interactive and cumulative effects among anthropogenic stressors. Sci. Total Environ. 542, 383–393.
Thomas, M.K., Fontana, S., Reyes, M., Kehoe, M., Pomati, F., 2018. The predictability of a lake phytoplankton community, over time-scales of hours to years. Ecol. Lett. 21 (5), 619–628.
Tsanas, A., Xifara, A., 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy Build. 49, 560–567.
Vanni, M.J., Layne, C.D., Arnott, S.E., 1997. "Top–down" trophic interactions in lakes: effects of fish on nutrient dynamics. Ecology 78 (1), 1–20.
Wang, D., Gouhier, T.C., Menge, B.A., Ganguly, A.R., 2015. Intensification and spatial homogenization of coastal upwelling under climate change. Nature 518 (7539), 390–394.
Wang, Y., Hu, W., Peng, Z., Zeng, Y., Rinke, K., 2018. Predicting lake eutrophication responses to multiple scenarios of lake restoration: a three-dimensional modeling approach. Water 10 (8), 994.
Waylen, K.A., Blackstock, K.L., Van Hulst, F.J., Damian, C., Horváth, F., Johnson, R.K., Kanka, R., Külvik, M., Macleod, C.J., Meissner, K., 2019. Policy-driven monitoring and evaluation: Does it support adaptive management of socio-ecological systems? Sci. Total Environ. 662, 373–384.
Yajima, H., Derot, J., 2018. Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. J. Hydroinformat. 20 (1), 206–220.
Zhang, F., Wang, Y., Cao, M., Sun, X., Du, Z., Liu, R., Ye, X., 2016. Deep-learning-based approach for prediction of algal blooms. Sustainability 8 (10), 1060.
Zhao, J., Ramin, M., Cheng, V., Arhonditsis, G.B., 2008. Competition patterns among phytoplankton functional groups: how useful are the complex mathematical models? Acta Oecologica 33 (3), 324–344.
Zhao, Y., Zhang, Y., 2008. Comparison of decision tree methods for finding active objects. Adv. Space Res. 41 (12), 1955–1959.