Science of the Total Environment 648 (2019) 472–480

Contents lists available at ScienceDirect

Science of the Total Environment

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

Chlorophyll *a* predictability and relative importance of factors governing lake phytoplankton at different timescales

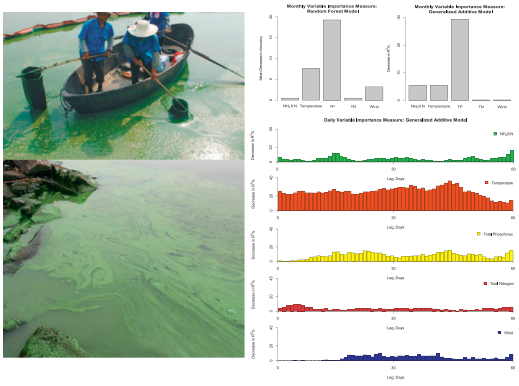
Xia Liu, Jianfeng Feng *, Yuqiu Wang *

Key Laboratory of Pollution Processes and Environmental Criteria, Ministry of Education, and Tianjin Key Laboratory of Environmental Remediation and Pollution Control, College of Environmental Science and Engineering, Nankai University, Tianjin 300071, China

HIGHLIGHTS

- Water temperature is more important predictor of daily chlorophyll *a* than nutrient.
- Nutrients are a more important predictor than water temperature at a monthly scale.
- The drivers of phytoplankton fluctuations vary at different timescales.
- Timescales have an influence on the relative role of N and P limitation in lakes.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 June 2018
Received in revised form 10 August 2018
Accepted 10 August 2018
Available online 11 August 2018

Editor: Sergi Sabater

Keywords:
Phytoplankton
Predictability
Random forest
Generalized additive model
Timescale

ABSTRACT

Assessing the key drivers of eutrophication in lakes and reservoirs has long been a challenge, and many studies have developed empirical models for predicting the relative importance of these drivers. However, the relative roles of various parameters might differ not only spatially (between regions or localities) but also at a temporal scale. In this study, the relative roles of total phosphorus, total nitrogen, ammonia, wind speed and water temperature were selected as potential drivers of phytoplankton biomass by using chlorophyll *a* as a proxy for biomass. A generalized additive model (GAM) and a random forest model (RF) were developed to assess the predictability of chlorophyll *a* and the relative importance of various predictors driving algal blooms at different timescales in a freshwater lake. The results showed that the daily datasets yielded better predictability than the monthly datasets. In addition, at a daily scale, water temperature was a more important predictor of chlorophyll *a* than nutrients, and the importance of phosphorus was comparable to that of nitrogen. In contrast, at a monthly scale, nutrients are more important predictors than water temperature and phosphorus is a better predictor than nitrogen. This study indicates that the drivers of phytoplankton fluctuations vary at different timescales and that timescale has an influence on the relative roles of nitrogen and phosphorus limitation in lakes, which suggests that the temporal scale should be considered when explaining phytoplankton fluctuations. Moreover, this study provides a reference for the monitoring of phytoplankton fluctuations and for understanding the mechanisms underlying phytoplankton fluctuations at different timescales.

© 2018 Elsevier B.V. All rights reserved.

* Corresponding authors.
E-mail addresses: fengjf@nankai.edu.cn (J. Feng), yqwang@nankai.edu.cn (Y. Wang).

https://doi.org/10.1016/j.scitotenv.2018.08.146
0048-9697/© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Algal blooms are a major problem in freshwater ecosystems throughout the world (Paerl and Huisman, 2008; Page et al., 2018). Predictive models can be useful for developing strategies to reduce bloom frequency and severity and for guiding actions to reduce bloom impacts (Cha et al., 2014). Identifying algal bloom drivers is essential for developing predictive models. Over the past five decades, studies have attempted to identify the drivers of undesirable or harmful algal blooms (HABs) (Cha et al., 2017). Although an excess of nutrients is always a key determinant of blooms, the relative roles of phosphorus (P) and nitrogen (N) in bloom dynamics remains hotly debated (Carpenter et al., 2016; Conley et al., 2009; Elser et al., 2009; Schindler, 1977; Schindler, 2014). Moreover, the impacts of temperature on blooms has received much attention in the past decade (Trolle et al., 2015), and these impacts will increase at an alarmingly fast rate in the future (O'Reilly et al., 2015). Some studies have found that global warming plays an important role in the global spread of phytoplankton blooms (Ma et al., 2016). The formation of HABs has been attributed, in part, to extra potential meteorological conditions (Scavia et al., 2016). For example, climate change will affect not only thermal regimes in lakes but also precipitation and thus runoff. Catchment-related fluxes in nutrients and water residence time might also have strong impacts on algal blooms and lake productivity, and changes in wind speed and thus mixing and turbulence will also contribute to these impacts. Thus, nutrients, temperature and meteorological conditions are likely to play significant roles in driving algal blooms (Rigosi et al., 2014; Scavia et al., 2016).

The relative importance of environmental drivers might be site-specific (Rigosi et al., 2014). For example, the results of a linear regression model tested on >1000 U.S. lakes indicated that nutrient levels are more important predictors of algal blooms than temperature (Rigosi et al., 2014). In contrast, a study employing a Bayesian network model suggested that phytoplankton fluctuations are more sensitive to changes in water temperature than to changes in the concentration of total P in 20 globally distributed lakes (Rigosi et al., 2015). The relative roles of nutrients, temperature, mixing, hydrology, solar radiation and other drivers depend on both spatial and temporal resolutions (Blauw et al., 2018). At an inter-lake level or for intra-lake development over years, the nutrient levels are expected to override other drivers, whereas at the intra-lake level, both temperature and wind speed (and thus turbulence and nutrient mixing) could be important. Most studies on phytoplankton dynamics in both freshwater and marine sites have addressed drivers of phytoplankton responses at monthly or inter-annual scales.

The drivers of phytoplankton fluctuations can also differ among different temporal scales (Blauw et al., 2018). In marine waters, phytoplankton fluctuations at inter-annual and decadal scales are often impacted by climatic variation or changes in the eutrophication status (McQuatters-Gollop and Vermaat, 2011; Ottersen et al., 2001; Richardson and Schoeman, 2004), whereas at a seasonal scale, nutrients, temperature, solar irradiance, thermal stratification, and grazing are the major drivers of phytoplankton fluctuations and succession (Sharples et al., 2006; Sommer et al., 2012; Winder and Cloern, 2010). At even shorter timescales (e.g., monthly and daily), fluctuations in phytoplankton are largely affected by physical drivers, such as wind and turbulent mixing, which affect the spatial distribution of phytoplankton (Peter, 2005). For accurate predictions, it is important to separate the different impacts of various drivers at different temporal scales. Winder and Cloern (2010) conducted a comparative time series analysis of lakes and open oceans and found consistent differences in the relative importance of drivers between seasonal and inter-annual scales. Most studies that have investigated such temporal responses were based on marine environments (Cloern and Jassby, 2010). However, the roles of drivers at shorter timescales, i.e., monthly and daily scales, as determinants of algal biomass in freshwater lakes remain poorly understood.

The determination of a proper model structure is critical for developing a predictive model. The core of any forecasting model that used by management authorities or the community is to predict events at a relevant timescale. Many studies have focused on potential predictors of HABs without considering time-lag effects (Hollister et al., 2016; Segura et al., 2017), which results in the need for extra predictions for predictors. For example, explanatory variables (predictors) must be determined before the developed predictive model can be used to predict phytoplankton responses. Predictive models with a time lag between the drivers and responses can address this issue well because responses can be predicted without determining the predictors. There will always be some time lag between drivers and responses, and this time lag is related to either physical factors (e.g., turbulence and the spatial distribution of algal masses) or biological factors related to growth rate and population responses. Several studies have attempted to explore the development of forecasting models based on the time-lag effect of predictors (Kehoe et al., 2015; Xiao et al., 2017; Zhang et al., 2015b). For example, Zhang et al. (2015b) used an artificial neural network model to predict the water quality of the Yuqiao Reservoir (YQR) and found that this model is potentially useful for predicting eutrophication up to 2 weeks in advance.

The aim of the current study was to evaluate the predictability of phytoplankton fluctuations and the relative importance of selected key drivers among different timescales in a freshwater lake with five potential predictors (total P, total N, ammonia N, water temperature and wind speed) as drivers and chlorophyll *a* concentrations to represent the biomass of phytoplankton. Using these five predictors, we (1) developed a generalized additive model (GAM) and a random forest (RF) model based on two datasets (monthly and daily) to assess the predictability of phytoplankton fluctuations at these two timescales, (2) assessed their importance using the GAM and RF and (3) compared the differences of the relative importance at monthly and daily scales. Although these studies were performed in a single lake, we believe the insights are relevant to other comparable freshwater bodies.

2. Materials and methods

2.1. Study area and data description

The YQR (117°34′ E and 40°02′ N) is located in the north of Tianjin City in China (Fig. 1) and serves as the largest drinking water source in Tianjin (with a population of >16 million). Two rivers (Guohe River and Linhe River) enter into the YQR. The YQR receives water from an upstream reservoir (Daheiting Reservoir) via the Guohe River, and the Linhe River is a minor contributor and often runs dry. The reservoir has a watershed area of 2060 km^2, a storage capacity of 1.559 billion m^3 and a surface area of 86.8 km^2. The YQR is considered a shallow lake with a maximum depth of 12 m and an average depth of 4.7 m. The annual precipitation in the YQR Basin over the study period was approximately 750 mm/m^2. The detailed properties of the YQR are shown in Table 1.

Because the YQR is the drinking water source in Tianjin, a project aiming to protect the watershed was implemented in 2002 to reduce or eliminate the point sources of nutrients to YQR (Zhang et al., 2015a). However, the reservoir has been mesotrophic due to increased pollution in the watershed, notably from cage aquaculture in the upstream reservoir. The water quality of YQR has gradually decreased. In fact, the YQR has experienced several HABs over the past two decades.

We used the concentration of chlorophyll *a* (Chl-a, μg/L) as a proxy for phytoplankton biomass. There certainly are variations in Chl-a to carbon ratios (Jakobsen and Markager, 2016), but one should be aware that also cell specific C may vary with species and growth conditions. When comparing direct, volume-based estimates on phytoplankton with Chl-a, there is no doubt that the Chl-a's response to light is biomass-specific, but this variation does not override the positive association between Chl-a and biomass. For example, a very strong

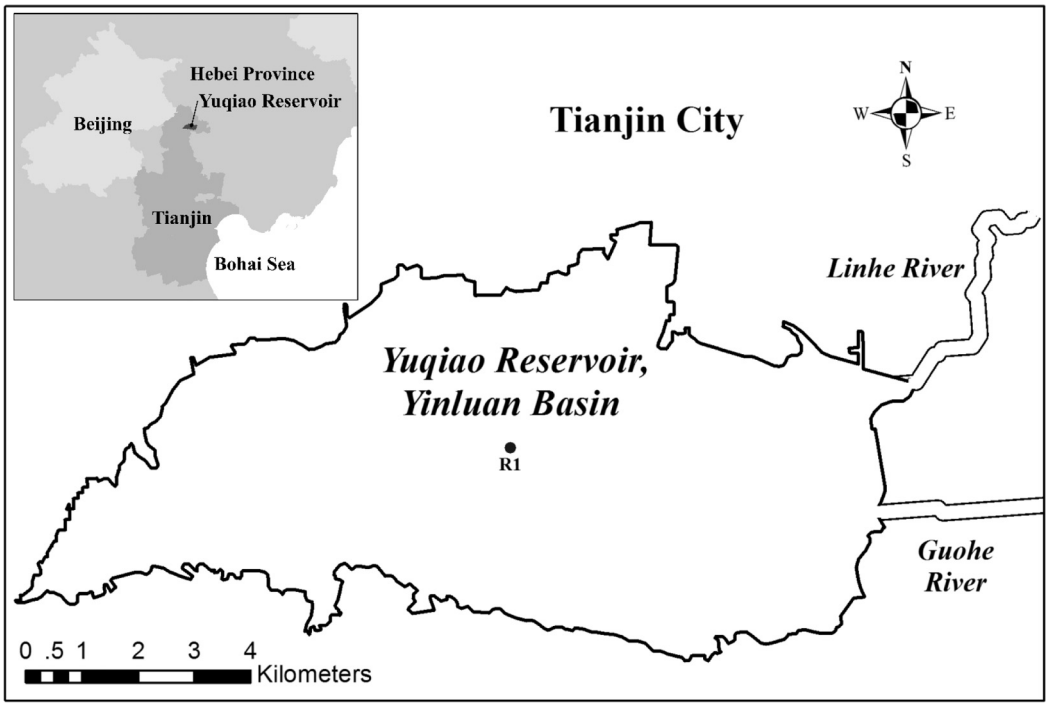


Fig. 1. Map of the Yuqiao Reservoir. The sampling sites are denoted as R1.

correlation between Chl-a and algal biomass (microscopic estimates) was found from 400 lakes with highly differing productivity, phytoplankton composition and light attenuation (Hessen et al., 2006). The Chl-a in YQR ranged from 2.6 μg/L to 36.4 μg/L and exhibited pronounced variability. The monthly and daily fluctuations in the Chl-a are shown in Fig. 2. Five variables were used as potential environmental factors driving phytoplankton fluctuations: total P (TP, mg/L), total N (TN, mg/L), ammonia N (NH_4^+-N, mg/L), wind speed (wind, m/s) and water temperature (T, °C). Temperature and nutrients are often considered the most significant determinants of phytoplankton biomass (Cha et al., 2017; Rigosi et al., 2014). Clearly, the factors analyzed in this study do not represent an exhaustive list of potential determinants of phytoplankton development, but for obvious reasons, grazing and other losses in biomass cannot be automatically monitored at short timescales. In addition, P was represented by the total P (not orthophosphate). N was represented by both total N and NH_4^+, and NO_3^- was not analyzed separately. In our study, ammonia was poorly correlated with total N at daily and monthly timescales (daily: $p = 0.23$; monthly: $p = 0.48$, Appendix S1). Therefore, both TN and NH_4^+ were selected as potential explanatory variables. Although other variables might impact algal biomass, those included in our survey cover the main potential drivers, i.e., nutrients (N and P), temperature and meteorology.

From January 2003 to November 2017, the concentrations of TP, TN, and NH_4^+-N, and T were sampled monthly at the center of the reservoir (site R1, Fig. 1). Beginning in January 1, 2017, a buoy was also established at site R1 to monitor the in situ parameters online. The buoy had multiple probes for detecting TP, TN, NH_4^+-N, and water temperature. The monitoring data from the buoy were transferred to the data center every 4 h. The depth of the sampling for both the monthly monitoring and the buoy data were the same (0.5 m), and the dataset from the buoy was corrected with the monthly monitoring data to maintain consistency between the two datasets. A detailed description of the buoy is provided in Appendix S2. The wind speed datasets were obtained from the China Meteorological Data Service Center (http://data.cma.cn/). The meteorological station is adjacent to the YQR, and as a result, its datasets represented the meteorological conditions of the reservoir monitoring site well.

Table 1
Key properties of the Yuqiao Reservoir.

	Range	Mean	Median	Coefficient of variation
Monthly dataset of total phosphorus (2003–2017): mg/L	0.01–0.019	0.04	0.03	0.71
Monthly dataset of total nitrogen (2003–2017): mg/L	0.15–4.81	1.76	1.59	0.61
Monthly dataset of ammonia nitrogen (2003–2017): mg/L	0.01–0.71	0.14	0.11	0.87
Monthly dataset of water temperature (2003–2017): °C	1.1–31.4	14.42	14.20	0.69
Average depth: m	4.47			
Area: km^2	86.8			
Longitude: E	117° 34′			
Latitude: N	40° 02′			
Nutrient level	lightly eutrophic			

2.2. Candidate models

In this study, two candidate models (GAM and RF) were selected to analyze the predictability of Chl-a and assess the relative importance of various factors in driving phytoplankton fluctuations at different timescales in the YQR. Considering the complexity of the relationships between environmental drivers and phytoplankton biomass and that the causality and dynamics of phytoplankton are not well understood, the RF methodology was selected to quantify the relationship between phytoplankton fluctuations and drivers, assess the Chl-a predictability and analyze the relative importance of the drivers. RF use an ensemble machine learning method that develops nonlinear functions based on the mean response of an ensemble of simpler decision tree models (Breiman, 2001), and this approach is increasingly utilized to model and understand environmental systems (Kehoe et al., 2012). The RF technique represents an application of bagging to decisions trees, and bagging is an ensemble modeling method designed to avoid the overfitting of models (Breiman, 1996). A large number of simple models are constructed with random subsamples of a dataset and are then aggregated in some way, usually by averaging in the case of regression

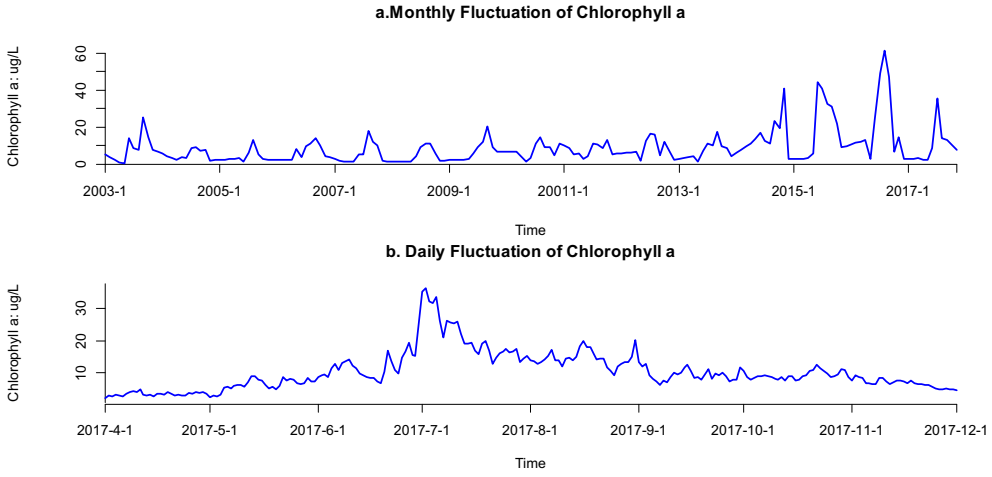


Fig. 2. Time series of chlorophyll *a* concentration (mg/mL) in the Yuqiao Reservoir at (a) monthly and (b) daily resolutions.

and by mode in the case of decision making (Kehoe et al., 2015). Bagging can be applied to any model type, such as linear regression models, and is denoted as a random forest when applied to decision trees (Breiman, 2001). The construction of a RF proceeds as follows: first, a random subset of the whole dataset is selected, and a decision tree is constructed. This model construction process is then repeated until an ensemble of decision trees is obtained. Each member of the RF ensemble is a simple decision tree biased toward predicting their own particular training data. When the mean prediction of a large number of these randomly constructed simple decisions trees (forest) is calculated, they produce low variance and unbiased predictions (Breiman, 2001).

RF were developed in this study using the "Party" package (Strobl et al., 2009) in R 3.1.1. Briefly, 500 trees were constructed for each ensemble, and leave-one-out cross-validation (CV) was conducted to test the prediction power of the RF. The model performance for the calibration and validation stages was quantified using the coefficient of determination (R^2).

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

Given the potential pitfalls of any predictive model, we also applied a GAM for the comparison and validation of the RF predictions. GAMs are widely used in environmental studies (Beale et al., 2010; Bininda-Emonds and Purvis, 2012; Sun et al., 2018). In brief, a GAM is a generalized linear model in which the linear predictor is specified as the sum of the smooth functions of some or all of the covariates (N. Wood, 2006).

The general formula of a GAM is

$$g(\mu_i) = \beta + \sum_{j=1}^{n} f_j(X_i) + \varepsilon_i \quad (2)$$

where $g(\mu_i)$ is a monotonous link function relating the response variable to the given explanatory variables, β is any strictly parametric component in the model, such as the intercept, $f_j(X_i)$ is the variable explained by the nonparametric smoothing function, and ε_i is identically and independently distributed as a normal random variable (Wood, 2004; Wood and Augustin, 2002).

As stated previously, Chl-a was the response variable, and TP, TN, NH_4^+-N, wind speed (Wind) and T were the independent variables, which yielded the following:

$$g(\text{Chl a}) = \beta + f_1(TP) + f_2(TN) + f_3(NH_4^+) + f_4(T) + f_5(Wind) + \varepsilon_i \quad (3)$$

In this study, the GAM was fitted using the 'mgcv' package in R 3.1.1. Additionally, leave-one-out CV was performed to test the predictive power of the GAM. For each fitting and CV, the R^2 was calculated as an indicator of goodness-of-fit and CV.

2.3. Predictability of Chl-a

We compared the predictive performances of models with different time lags at monthly and daily scales. For daily predictions, we tested 60 different forecasting time lags ranging from 0 to 60 days. The monthly predictive models were calibrated for three different forecasting time lags (1, 2, 3 months in advance). More formally, both the RF and GAM were calibrated to predict the concentration of Chl-a (y_t) for the predictor variables x_{t-n},

$$y_t = M(x_{t-n}) \quad (4)$$

where M is the specific model (RF or GAM), and n is a range of different time lags (daily scale: $n = 0,1,2\ldots26$, monthly scale: $n = 0,1,2,3$).

2.4. Relative importance of drivers

To assess the relative importance of the five selected parameters in driving phytoplankton biomass, we analyzed the loss of predictive power by excluding drivers from the model. For the RF, this was calculated as the mean reduction in the mean square error (MSE), and for the GAM, this was calculated as the reduction in R^2. Many studies have implemented these methods to assess the relative importance of variables (Hu et al., 2017; Zhang et al., 2017b). We analyzed the variable importance with no lag between the five drivers and Chl-a. Besides, we quantified the relative importance of five potential environmental factors with the time-lag effects in driving phytoplankton fluctuations.

3. Results

The results of the assessment of the predictability of Chl-a at a monthly scale for the calibration and CV of the RF for time lags ranging from 1 to 3 months are shown in Fig. 3a. Higher R^2 values were obtained for calibration and CV of the RF when there was no time lag, but R^2 values were still high for the calibration of the RF (R^2 values ranging from 88% to 76%) with the time lag of 1 to 3 months However, the R^2 values for the CV decreased with increases in the time lag (R^2 from 60% to 17%) and reflected poor performance ($R^2 < 20\%$) with a time lag of 3 months. The results of the assessment of the predictability of Chl-a at a monthly scale using the GAM for time lags ranging from 1 to 3 months are shown in Fig. 3b. Similar to the RF results, the R^2 values for both the validation and CV of the GAM were higher when there was no time lag. However, unsatisfactory values of R^2 were obtained for the

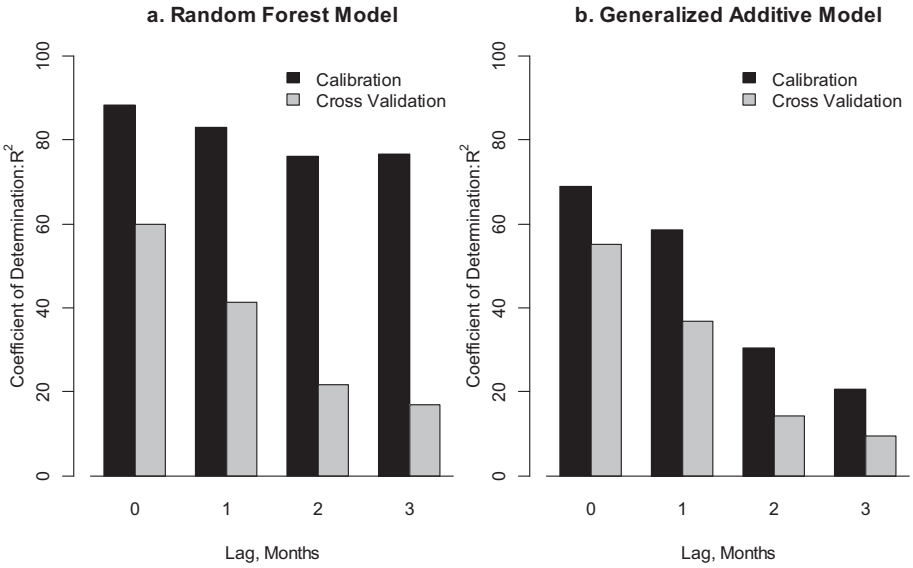


Fig. 3. Predictability of the monthly chlorophyll *a* concentration by (a) a generalized additive model and (b) a random forest model.

calibration and CV of the GAM (calibration: R^2 values ranging from 69% to 21%; and CV: R^2 values ranging from 55% to 9%) with increases in the time lag and reflected poor performance ($R^2 < 10\%$) with a time lag of 3 months. In general, the GAM performed more poorly than the RF, and these findings were obtained for both the calibration and CV (Fig. 3a–b).

The calibration and CV of the RF model for daily predictions involving time lags ranging from 1 to 3 months are shown in Fig. 4a. The RF showed satisfactory performance, irrespective of the different time lags. As shown in Fig. 4a, the R^2 values for the CV fluctuated between 60% and 80% with different time lags, suggesting that daily datasets have high predictive power for the prediction of algal blooms. In general, the results showed that the daily predictor datasets demonstrated better predictability than the monthly predictor datasets (Figs. 3 and 4). The GAM performance was worse than that of the RF, and this finding was obtained for both the calibration and CV (Fig. 4a–b). A decreasing trend was obtained for increase in the time lag from 0 to 8 days, and the R^2 values for the CV fluctuated between approximately 40% and 60% as the time lag increased from 10 to 60 days in the GAM. Furthermore, a comparison between the observations and the modeled values obtained with the calibration and CV is shown in Fig. 5 for time lag of 30 and 60 days. Both the RF and GAM showed better performance at daily than at monthly timescales, but irrespective of the timescale, the RF generally performed better than the GAM (Fig. 5a–d and e–h).

To reveal the mechanistic causes for the predictions, we assessed the relative importance of the five selected drivers. Based on both the RF and GAM predictions at a monthly scale, TP was identified as the most important driver, followed by T and NH_4^+-N (Fig. 6). The results of the GAM and RF indicated that TP and T are the two most important predictors. The GAM indicated that NH_4^+-N and T had nearly identical importance, whereas the RF showed that the wind speed was the third most important driver. The results of the GAM showed that TN and wind had the lowest importance of all the predictors, whereas the RF results showed that TN and NH_4^+-N exhibited the lowest importance of all the predictors.

For daily predictions, both models indicated that temperature was the most important predictor driving phytoplankton fluctuations when there was no time lag (Fig. 7). The relative importance of the other drivers corresponded to the above-described results: NH_4^+-N was the second most important driver, and wind had the least

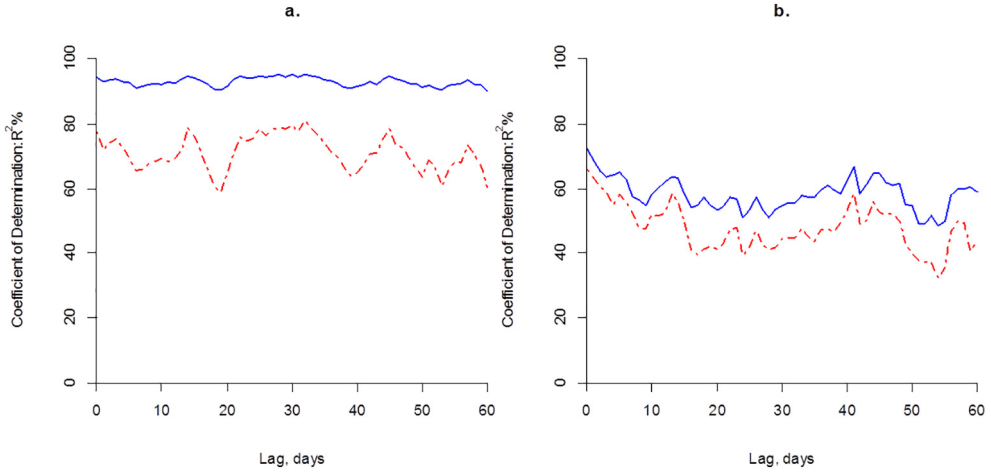


Fig. 4. Predictability of the daily chlorophyll *a* concentration by (a) a generalized additive model and (b) a random forest model. The blue lines represent the R^2 of the calibration, and the red lines represent the R^2 of the cross-validation.

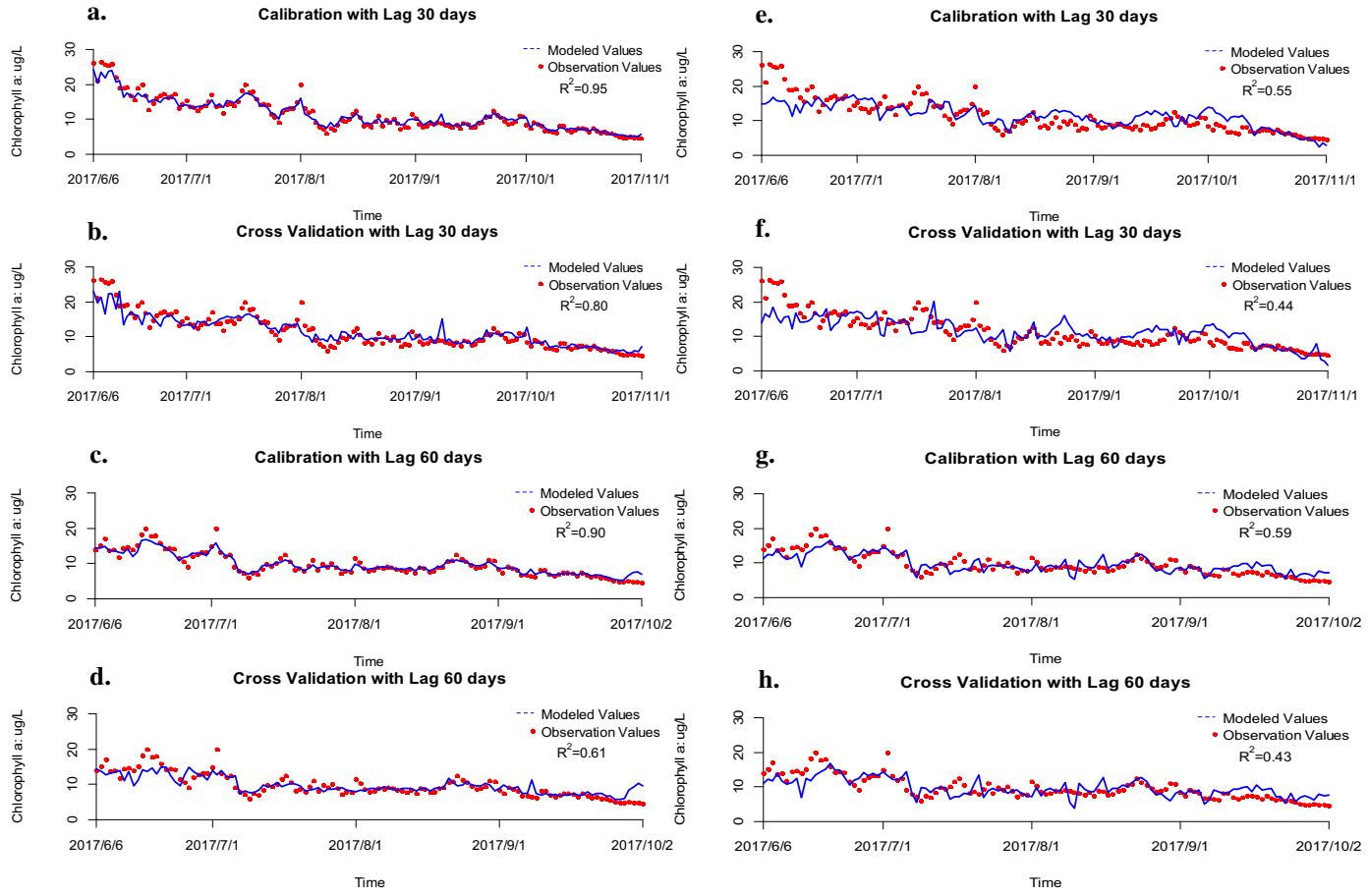


Fig. 5. Performance of the calibration and cross-validation of the random forest model and generalized additive model for forecasting lags of 30 and 60 days: (a) calibration of the random forest model for a lag of 30 days, (b) cross-validation of the random forest model for a lag of 30 days, (c) calibration of the random forest model for a lag of 60 days, (d) cross-validation of the random forest model for a lag of 60 days, (e) calibration of the generalized additive model for a lag of 30 days, (f) cross-validation of the generalized additive model for a lag of 30 days, (g) calibration of the generalized additive model for a lag of 60 days, and (h) cross-validation of the generalized additive model for a lag of 60 days.

importance. With different time lags, both models showed that temperature was the major predictor (Fig. 7). However, the relative importance of the other predictors for different time lags differed between the two models but remained consistent for all tested time lags among the models (Fig. 7). For example, in the two models, the importance of temperature showed a decreasing trend, whereas that of wind showed an increasing trend. Furthermore, we found that the mean decrease in the MSE of TP was comparable to those of TN and NH_4^+-N.

The relative importance of the five drivers at a monthly scale and daily scale identified temperature as the most important predictor in driving the daily phytoplankton fluctuations and TP as the most important predictor in driving the monthly phytoplankton fluctuations (Figs. 6 and 7).

We also generated a weekly dataset from the daily dataset (Appendix S3). The R^2 of the CV were all unsatisfactory (when lag > 1 week, all R^2 < 40%, Appendix S4) for both the RF and GAM. In addition,

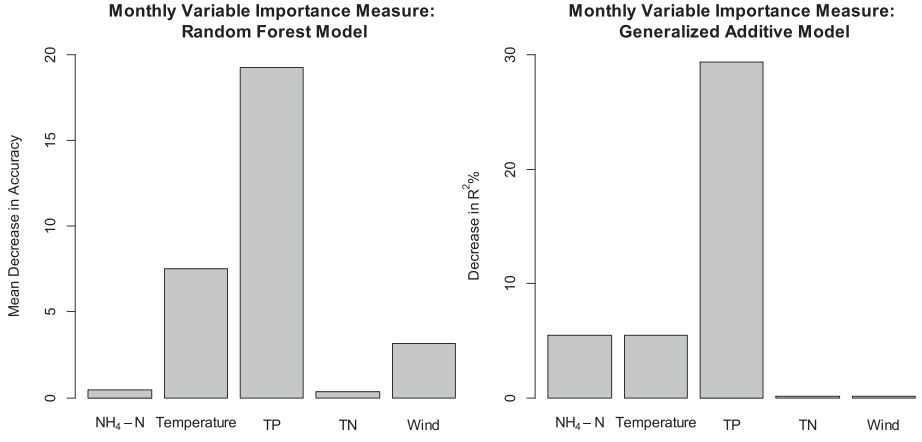


Fig. 6. Predicted relative importance of the five selected variables at a monthly resolution by the random forest and generalized additive models.

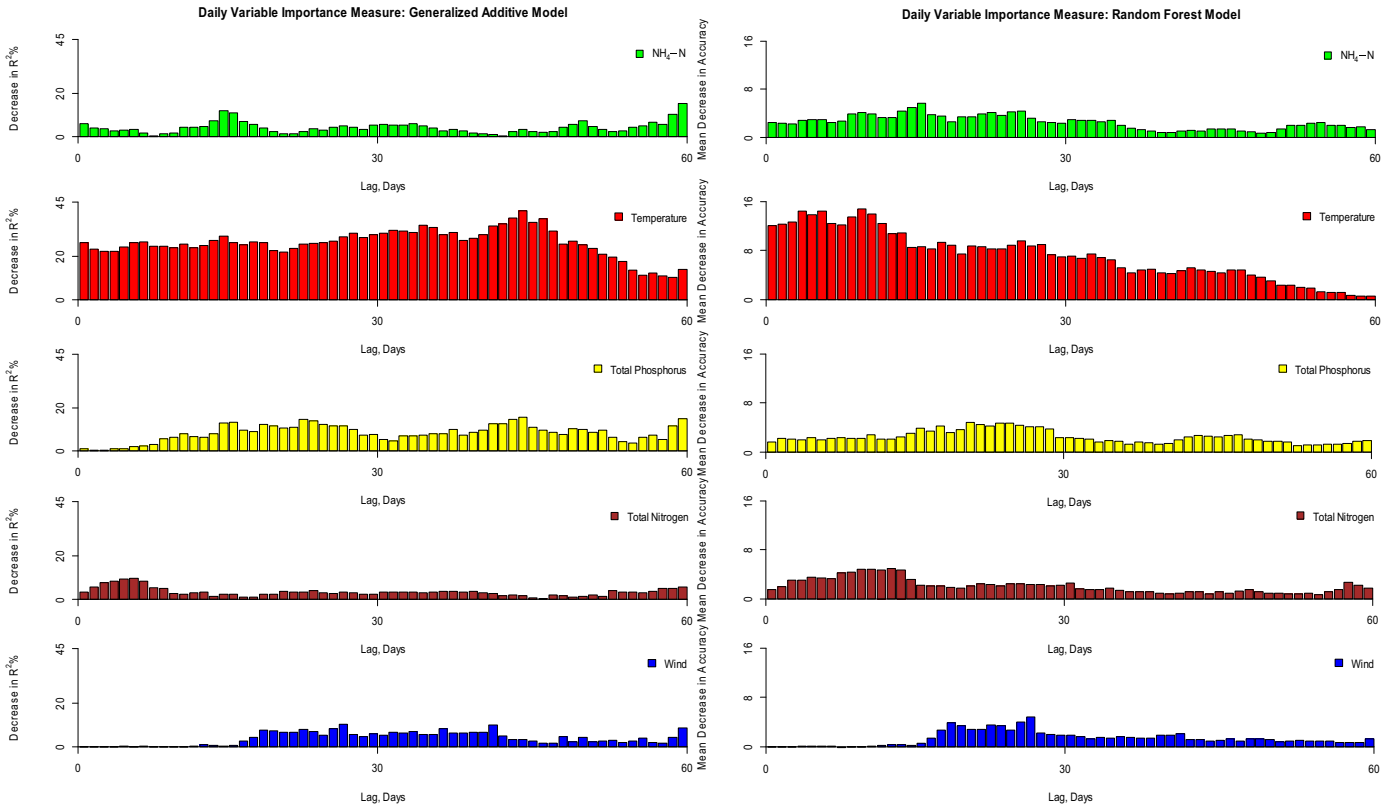


Fig. 7. Predicted relative importance of the five selected variables at a daily resolution by the random forest and generalized additive models. The variable importance results measured for a forecasting lag time ranging from 0 to 60 days are also shown.

temperature was identified the best predictor of weekly phytoplankton fluctuations and exceeded the impact of nutrients (Appendix S5).

4. Discussion

The difficulty associated with the use of daily data is access to continuous in situ data provided by a buoy. Although such buoys and continuous recordings are becoming more sophisticated and accurate and cover an increasing number of parameters, they are still far from having the ability to include all relevant parameters. In addition, the models investigated in this study performed well, even potential determinants of phytoplankton development, such as solar radiation, orthophosphate and grazing were not included in this study. This finding indicates that the five selected drivers can satisfactorily explain phytoplankton fluctuations and other potential factors might not be the main drivers. In general, daily data performed better than monthly data, and this finding was obtained for both the RF and GAM. The R^2 of the CV had peaks at 14, 32, 45 and 57 lag days (Fig. 4a), suggesting that time-lag models have the potential to provide reliable predictions, even over a 2-month period. To some extent, this finding could indicate that daily weather is not completely random, i.e., that there are periods of high and low temperatures (Appendix S6-e). These results might also suggest that phytoplankton growth in a given period is better reflected by the internal stores of P than the daily nutrient inputs. Another possibility is simply that nutrients are found at excess concentrations during most of the growing season (Appendix S6) due to strong external inputs or turbulent mixing from the sediments, whereas the air temperature is more variable. Moreover, light and photoperiods combined with water temperature might explain the time lag between phytoplankton and water temperature fluctuation. For example, temperature is suitable for phytoplankton growth, whereas photoperiods cannot support the phytoplankton growth. When both temperature and light are favorable for phytoplankton growth, the phytoplankton will grow rapidly, which causes a time lag between the phytoplankton and temperature fluctuations. Furthermore, the temperature profiles in lakes depend on solar radiation, air temperature and wind mixing, and although the phytoplankton growth rates may exceed 1 d^{-1}, there will always be a time lag between temperature changes and biomass responses. Changes in phytoplankton growth also depend on temperature, and although increases in the low or intermediate temperatures promote cell division and biomass development (provided there are sufficient nutrients), increases in temperature above a maximum threshold might reduce biomass if the respiratory losses exceed the C-fixation rates. Furthermore, temperature is always considered a proxy for a longer photoperiod, solar radiation, high atmospheric pressure, and low hydrodynamic for phytoplankton growth.

The RF model is a black box model that cannot separate positive or negative effects and simply provides the net impact of temperature. Furthermore, although water temperatures reflect air temperatures, these dampen the temporal variations (Piccolroaz et al., 2013), as observed in this study (Appendix S6-e).

At both monthly and daily scales, the RF had a higher R^2 than the GAM, and this finding was obtained for both the CV and the calibration. This finding suggests that the RF showed increased predictive power than the GAM in a freshwater environment. At a monthly scale, the performance of the CV decreased with increases in the time lag and performed poorly with a time lag of 3 month. This result contrasts with those obtained in coastal research (Blauw et al., 2018), which could be due to differences between marine and freshwater sites, the inclusion of the effects of predictor time lag in our study and site-specific properties.

Monthly monitoring programs prevail in many countries and have been implemented as routine elements of environmental monitoring. These programs are helpful to reveal long-term trends of eutrophication and phytoplankton fluctuations. With monthly datasets, we can study the impact of global changes on phytoplankton fluctuations because

the timescale enables the removal of environmental noise. However, according to the findings of this study, monthly monitoring programs are insufficient for predicting an algal bloom in advance. Therefore, the fluctuation of drivers at a short scale should not be dismissed as environmental noise, and a daily or shorter than daily monitoring program is necessary for forecasting and managing algal blooms.

The comparison of the relative importance of five drivers at a monthly scale with that at a daily scale revealed that temperature was the most important predictor of phytoplankton fluctuations at a daily scale and that TP was the key predictor of phytoplankton fluctuations at a monthly scale. These findings are in agreement with previous studies (Rigosi et al., 2014). Although nutrients are undoubtedly the ultimate drivers of algal biomass and thus determine seasonal biomass (e.g., spring blooms) and inter-lake differences, the temperature has two major roles in determining biomass. On the one hand, at a given nutrient level, increases in temperature promote the metabolic rate and cell division of phytoplankton and boost productivity, but these effects are only observed when the temperatures are below an optimal level for phytoplankton growth. Air temperatures might show substantial fluctuations at a daily timescale, but the water temperature fluctuations are dampened. Thus, a few days of high temperatures will result in the accumulation of heat in the upper water layers and might thereby promote phytoplankton growth and biomass build-up over extended periods. On the other hand, the temperature might be influenced by light or photoperiods. Light impacts the thermal balance and thus indirectly influences the temperature, and the resulting temperature determines the phytoplankton fluctuation. Elevated temperatures could also affect long-term phytoplankton development by promoting a stronger thermocline and thus reducing the vertical mixing of nutrients (Jacob, 2002). Notably, the peak Chl-a on 1st July was followed by a marked increase in TP and a decline in NH_4^+-N (Fig. 2b and Appendix S6b-c). This might be explained by the input of nonpoint source N and P into the YQR in summer (Chun et al., 2017) and the absorption of P onto particles, which would result in a lack of bioavailable P (Zhang et al., 2017a). Moreover, the second peak of Chl-a on 1st September was followed by a marked increase in TN (Fig. 2b Appendix S6a), which might be explained by the fact that water transfer was the largest TN source in autumn in the YQR (Chun et al., 2017): TN was introduced through upstream water transfer into the YQR in autumn, and these water-transferred TN stimulated the growth of phytoplankton.

Although the full mechanistic underpinnings for the predictive strength of daily observations remains to be settled, the predictions from the two independent models established in this study are fairly robust and suggest a strong importance of temperature in the predictions of phytoplankton (with nutrients remaining the ultimate long-term determinants of actual phytoplankton levels). Thus, predictions should pay more attention to temperature when studying short-term fluctuations in phytoplankton levels, whereas nutrients are more important in regulating phytoplankton fluctuations at longer timescales (e.g., monthly).

We also analyzed a weekly dataset (Appendix S3), but the weekly results need to be interpreted with caution because the data size at the weekly scale was less than that at a daily scale (daily scale: $n = 245$, weekly scale: $n = 35$). As shown in Appendix S4, the weekly dataset resulted in poor predictability, regardless of the model used, which was similar to the findings obtained at the monthly scale. These results indicate that short-term fluctuations facilitate the prediction of phytoplankton fluctuations that cannot be captured by the weekly fluctuations of drivers. In both the RF and GAM, temperature is a more important predictor in driving weekly phytoplankton fluctuations than nutrients (Appendix S5), and this result is consistent with that obtained at the daily scale.

Another outcome of this study is that in the YQR, TP was found to be a stronger predictor than TN and NH_4^+-N (Fig. 6) at a monthly scale. However, at a daily scale, the importance of TP was generally comparable to that of N, which indicates that the temporal scale might impact the relative roles of P and N. Although the roles of N and P limitation in autotroph growth have been continuously debated (Conley et al., 2009; Elser et al., 2009; Gardner et al., 2017; Schindler, 2012; Schindler et al., 2016; Schindler et al., 2017), previous studies have omitted the impact of temporal scales on the relative roles of P and N. The findings from this study could also reflect different uptake and cellular turnover rates for N and P, but because we only have TP data, the mechanistic causality of this temporal effect remains speculative.

Although our findings are based on a specific reservoir and the applicability of these results for other freshwaters should be judged with some caution, we believe that our findings have general relevance for large, shallow and nutrient-rich water bodies.

5. Conclusion

Using the RF and GAM methods, this study assessed the predictability of phytoplankton fluctuations and compared the relative importance of five drivers at monthly and daily scales. The results revealed that both the RF and GAM demonstrated better phytoplankton predictability at a daily scale than at a monthly scale. Moreover, at both the monthly and daily scales, the RF exhibited a higher R^2 for both the CV and calibration than the GAM. The relative importance assessment results indicated that water temperature was a more important predictor than nutrients of daily phytoplankton fluctuations, and this finding was obtained with both the RF and GAM. However, at a monthly scale, nutrients (TP) were identified as a more important predictor than water temperature in regulating phytoplankton fluctuations. In addition, at a monthly scale, P was found to be a more important predictor than N, but at a daily scale, the importance of P and that of N were comparable in the freshwater reservoir. These results provide a reference for the monitoring of phytoplankton fluctuations and for understanding the mechanisms underlying phytoplankton fluctuations at different timescales.

Acknowledgements

This study was supported by the National Water Pollution Control and Treatment Science and Technology Major Project (2017ZX07301-001, 2017ZX07301-002) and National Key Research and Development Program of China (2018YFC1406400). We also thank Dag Olav Hessen, the editor and the anonymous referees, whose comments and suggestions greatly improved the quality of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2018.08.146.

References

Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J., Elston, D.A., 2010. Regression analysis of spatial data. Ecol. Lett. 13, 246–264. https://doi.org/10.1111/j.1461-0248.2009.01422.x.

Bininda-Emonds, O.R.P., Purvis, A., 2012. Comment on "Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification". Science 337. https://doi.org/10.1126/science.1220012.

Blauw, A.N., Benincà, E., Laane, R.W.P.M., Greenwood, N., Huisman, J., 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.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1023/a:1018054314350.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/a:1010933404324.

Carpenter, S.R., Cole, J.J., Pace, M.L., Wilkinson, G.M., 2016. Response of plankton to nutrients, planktivory and terrestrial organic matter: a model analysis of whole-lake experiments. Ecol. Lett. 19, 230–239. https://doi.org/10.1111/ele.12558.

Cha, Y., Park, S.S., Kim, K., Byeon, M., Stow, C.A., 2014. Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model. Water Resour. Res. 50, 2518–2532. https://doi.org/10.1002/2013WR014372.

Cha, Y., Cho, K.H., Lee, H., Kang, T., Kim, J.H., 2017. The relative importance of water temperature and residence time in predicting cyanobacteria abundance in regulated rivers. Water Res. 124, 11–19. https://doi.org/10.1016/j.watres.2017.07.040.

Chun, C., Dongmei, S., Ping, F., Miao, Z., Ning, G., 2017. Impacts of nonpoint source pollution on water quality in the Yuqiao reservoir. Environ. Eng. Sci. 34, 418–432. https://doi.org/10.1089/ees.2016.0124.

Cloern, J.E., Jassby, A.D., 2010. Patterns and scales of phytoplankton variability in estuarine–coastal ecosystems. Estuar. Coasts 33, 230–241. https://doi.org/10.1007/s12237-009-9195-3.

Conley, D.J., Paerl, H.W., Howarth, R.W., Boesch, D.F., Seitzinger, S.P., Havens, K.E., et al., 2009. ECOLOGY controlling eutrophication: nitrogen and phosphorus. Science 323, 1014–1015. https://doi.org/10.1126/science.1167755.

Elser, J.J., Andersen, T., Baron, J.S., Bergstrom, A.K., Jansson, M., Kyle, M., et al., 2009. Shifts in Lake N:P stoichiometry and nutrient limitation driven by atmospheric nitrogen deposition. Science 326, 835–837. https://doi.org/10.1126/science.1176199.

Gardner, W.S., Newell, S.E., McCarthy, M.J., Hoffman, D.K., Lu, K.J., Lavrentyev, P.J., et al., 2017. Community biological ammonium demand: a conceptual model for cyanobacteria blooms in eutrophic lakes. Environ. Sci. Technol. 51, 7785–7793. https://doi.org/10.1021/acs.est.6b06296.

Hessen, D.O., Faafeng, B.A., Brettum, P., Andersen, T., 2006. Nutrient enrichment and planktonic biomass ratios in lakes. Ecosystems 9, 516–527. https://doi.org/10.1007/s10021-005-0114-6.

Hollister, J.W., Milstead, W.B., Kreakie, B.J., 2016. Modeling lake trophic state: a random forest approach. Ecosphere 7. https://doi.org/10.1002/ecs2.1321.

Hu, X.F., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., et al., 2017. Estimating PM2.5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944. https://doi.org/10.1021/acs.est.7b01210.

Jacob, K., 2002. Limnology: Inland Water Ecosystems: Pearson Education (US).

Jakobsen, H.H., Markager, S., 2016. Carbon-to-chlorophyll ratio for phytoplankton in temperate coastal waters: seasonal patterns and relationship to nutrients. Limnol. Oceanogr. 61, 1853–1868. https://doi.org/10.1002/lno.10338.

Kehoe, M., O'Brien, K., Grinham, A., Rissik, D., Ahern, K.S., Maxwell, P., 2012. Random forest algorithm yields accurate quantitative prediction models of benthic light at intertidal sites affected by toxic Lyngbya majuscula blooms. Harmful Algae 19, 46–52. https://doi.org/10.1016/j.hal.2012.05.006.

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, 10984–10992. https://doi.org/10.1021/acs.est.5b00979.

Ma, J., Qin, B., Paerl, H.W., Brookes, J.D., Hall, N.S., Shi, K., et al., 2016. The persistence of cyanobacterial (*Microcystis* spp.) blooms throughout winter in Lake Taihu, China. Limnol. Oceanogr. 61, 711–722. https://doi.org/10.1002/lno.10246.

McQuatters-Gollop, A., Vermaat, J.E., 2011. Covariance among North Sea ecosystem state indicators during the past 50 years — contrasts between coastal and open waters. J. Sea Res. 65, 284–292. https://doi.org/10.1016/j.seares.2010.12.004.

N. Wood, S., 2006. Generalized Additive Models: An Introduction With R. vol. 66.

O'Reilly, C.M., Sharma, S., Gray, D.K., Hampton, S.E., Read, J.S., Rowley, R.J., et al., 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophys. Res. Lett. 42.

Ottersen, G., Planque, B., Belgrano, A., Post, E., Reid, P.C., Stenseth, N.C., 2001. Ecological effects of the North Atlantic oscillation. Oecologia 128, 1–14. https://doi.org/10.1007/s004420100655.

Paerl, H.W., Huisman, J., 2008. Blooms like it hot. Science 320, 57–58. https://doi.org/10.1126/science.1155398.

Page, T., Smith, P.J., Beven, K.J., Jones, I.D., Elliott, J.A., Maberly, S.C., et al., 2018. Adaptive forecasting of phytoplankton communities. Water Res. 134, 74–85. https://doi.org/10.1016/j.watres.2018.01.046.

Peter, J.S.F., 2005. Plankton patchiness, turbulent transport and spatial spectra. Mar. Ecol. Prog. Ser. 294, 295–309 (doi).

Piccolroaz, S., Toffolon, M., Majone, B., 2013. A simple lumped model to convert air temperature into surface water temperature in lakes. Hydrol. Earth Syst. Sci. 17, 3323–3338. https://doi.org/10.5194/hess-17-3323-2013.

Richardson, A.J., Schoeman, D.S., 2004. Climate impact on plankton ecosystems in the Northeast Atlantic. Science 305, 1609–1612 (doi).

Rigosi, A., Carey, C.C., Ibelings, B.W., Brookes, J.D., 2014. The interaction between climate warming and eutrophication to promote cyanobacteria is dependent on trophic state and varies among taxa. Limnol. Oceanogr. 59, 99–114. https://doi.org/10.4319/lo.2014.59.1.0099.

Rigosi, A., Hanson, P., Hamilton, D.P., Hipsey, M., Rusak, J.A., Bois, J., et al., 2015. Determining the probability of cyanobacterial blooms: the application of Bayesian networks in multiple lake systems. Ecol. Appl. 25, 186–199. https://doi.org/10.1890/13-1677.1.

Scavia, D., DePinto, J.V., Bertani, I., 2016. A multi-model approach to evaluating target phosphorus loads for Lake Erie. J. Great Lakes Res. 42, 1139–1150. https://doi.org/10.1016/j.jglr.2016.09.007.

Schindler, D.W., 1977. Evolution of phosphorus limitation in lakes. Science 195, 260–262. https://doi.org/10.1126/science.195.4275.260.

Schindler, D.W., 2012. The dilemma of controlling cultural eutrophication of lakes. Proc. R. Soc. B Biol. Sci. 279, 4322–4333. https://doi.org/10.1098/rspb.2012.1032.

Schindler, D.W., 2014. Unravelling the complexity of pollution by the oil sands industry. Proc. Natl. Acad. Sci. U. S. A. 111, 3209–3210. https://doi.org/10.1073/pnas.1400511111.

Schindler, D.W., Carpenter, S.R., Chapra, S.C., Hecky, R.E., Orihel, D.M., 2016. Reducing phosphorus to curb lake eutrophication is a success. Environ. Sci. Technol. 50, 8923–8929. https://doi.org/10.1021/acs.est.6b02204.

Schindler, D.W., Carpenter, S.R., Chapra, S.C., Hecky, R.E., Orihel, D.M., 2017. Response to the letter, nitrogen is not a "house of cards". Environ. Sci. Technol. 51, 1943-1943. https://doi.org/10.1021/acs.est.6b06106.

Segura, A.M., Piccini, C., Nogueira, L., Alcantara, I., Calliari, D., Kruk, C., 2017. Increased sampled volume improves *Microcystis aeruginosa* complex (MAC) colonies detection and prediction using Random Forests. Ecol. Indic. 79, 347–354. https://doi.org/10.1016/j.ecolind.2017.04.047.

Sharples, J., Ross, O.N., Scott, B.E., Greenstreet, S.P.R., Fraser, H., 2006. Inter-annual variability in the timing of stratification and the spring bloom in the North-western North Sea. Cont. Shelf Res. 26, 733–751. https://doi.org/10.1016/j.csr.2006.01.011.

Sommer, U., Adrian, R., Domis, L.D.S., Elser, J.J., Gaedke, U., Ibelings, B., et al., 2012. Beyond the Plankton Ecology Group (PEG) model: mechanisms driving plankton succession. In: Futuyma, D.J. (Ed.), Annual Review of Ecology, Evolution, and Systematics. vol. 43.43, pp. 429–448.

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol. Methods 14, 323–348. https://doi.org/10.1037/a0016973.

Sun, J.-M., Lu, L., Liu, K.-K., Yang, J., Wu, H.-X., Liu, Q.-Y., 2018. Forecast of severe fever with thrombocytopenia syndrome incidence with meteorological factors. Sci. Total Environ. 626, 1188–1192. https://doi.org/10.1016/j.scitotenv.2018.01.196.

Trolle, D., Nielsen, A., Rolighed, J., Thodsen, H., Andersen, H.E., Karlsson, I.B., et al., 2015. Projecting the future ecological state of lakes in Denmark in a 6 degree warming scenario. Clim. Res. 64, 55–72 (doi).

Winder, M., Cloern, J.E., 2010. The annual cycles of phytoplankton biomass. Philos. Trans. R. Soc. B Biol. Sci. 365, 3215–3226. https://doi.org/10.1098/rstb.2010.0125.

Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Am. Stat. Assoc. 99, 673–686. https://doi.org/10.1198/016214504000000980.

Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol. Model. 157, 157–177.

Xiao, X., He, J., Huang, H., Miller, T.R., Christakos, G., Reichwaldt, E.S., et al., 2017. A novel single-parameter approach for forecasting algal blooms. Water Res. 108, 222–231. https://doi.org/10.1016/j.watres.2016.10.076.

Zhang, C., Lai, S., Gao, X., Xu, L., 2015a. Potential impacts of climate change on water quality in a shallow reservoir in China. Environ. Sci. Pollut. Res. Int. 22, 14971–14982. https://doi.org/10.1007/s11356-015-4706-1.

Zhang, Y., Huang, J.J., Chen, L., Qi, L., 2015b. Eutrophication forecasting and management by artificial neural network: a case study at Yuqiao Reservoir in North China. J. Hydroinf. 17, 679–695. https://doi.org/10.2166/hydro.2015.115.

Zhang, C., Zhang, W., Huang, Y., Gao, X., 2017a. Analysing the correlations of long-term seasonal water quality parameters, suspended solids and total dissolved solids in a shallow reservoir with meteorological factors. Environ. Sci. Pollut. Res. 24, 6746–6756. https://doi.org/10.1007/s11356-017-8402-1.

Zhang, H., Wu, P.B., Yin, A.J., Yang, X.H., Zhang, M., Gao, C., 2017b. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: a comparison of multiple linear regressions and the random forest model. Sci. Total Environ. 592, 704–713. https://doi.org/10.1016/j.scitotenv.2017.02.146.