Environmental Science and Ecotechnology 14 (2023) 100233

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Original Research

Algal community structure prediction by machine learning

Muyuan Liu [a], Yuzhou Huang [a], Jing Hu [a], Junyu He [a, b], Xi Xiao [a, c, d, e, *]

[a] Ocean College, Zhejiang University, #1 Zheda Road, Zhoushan, Zhejiang, 316021, China
[b] Ocean Academy, Zhejiang University, #1 Zheda Road, Zhoushan, Zhejiang, 316021, China
[c] Key Laboratory of Marine Ecological Monitoring and Restoration Technologies, Ministry of Natural Resources, Shanghai, 201206, China
[d] Donghai Laboratory, Zhoushan, Zhejiang, 316021, China
[e] Key Laboratory of Watershed Non-point Source Pollution Control and Water Eco-security of Ministry of Water Resources, College of Environmental and Resources Sciences, Zhejiang University, Hangzhou, Zhejiang, 310058, China

ARTICLE INFO

Article history:
Received 16 October 2022
Received in revised form
21 December 2022
Accepted 21 December 2022

Keywords:
Phytoplankton community
Random forests
Environmental driver
Meteorology
Hydrology
Model interpretability

ABSTRACT

The algal community structure is vital for aquatic management. However, the complicated environmental and biological processes make modeling challenging. To cope with this difficulty, we investigated using random forests (RF) to predict phytoplankton community shifting based on multi-source environmental factors (including physicochemical, hydrological, and meteorological variables). The RF models robustly predicted the algal communities composed by 13 major classes (Bray-Curtis dissimilarity = 9.2 ± 7.0%, validation NRMSE mostly <10%), with accurate simulations to the total biomass (validation R^2 > 0.74) in Norway's largest lake, Lake Mjosa. The importance analysis showed that the hydro-meteorological variables (Standardized MSE and Node Purity mostly >0.5) were the most influential factors in regulating the phytoplankton. Furthermore, an in-depth ecological interpretation uncovered the interactive stress-response effect on the algal community learned by the RF models. The interpretation results disclosed that the environmental drivers (i.e., temperature, lake inflow, and nutrients) can jointly pose strong influence on the algal community shifts. This study highlighted the power of machine learning in predicting complex algal community structures and provided insights into the model interpretability.

1. Introduction

Phytoplankton, the critical part of primary producers, form the foundation of the aquatic ecosystem and contribute to an irreplaceable source of biodiversity [1,2]. However, eutrophication and global warming have dramatically promoted the over-proliferation of phytoplankton in both frequency and magnitude [3–7], exacerbating harmful ecological impacts and leading to serious restrictions to the socioeconomic development [3,8–10].

Phytoplankton biomass changes are accompanied with shifts in community structure [11–13]. Often, a comprehensive analysis of the algal community is required for water management regarding its high susceptivity to environmental stressors [11,14,15]. However, ecological models are commonly algae-specific due to the diverse environment-algae processes [16–18]. Typically for the conventional process-driven approaches, their accuracy is highly species-specific owing to the uncertainty of transport and kinetics parameters [18–20]. Whereas the empirical statistical models usually lack adaptability to complicated ecological patterns [21,22]. Without novel techniques to develop robust ecological models, predicting shifts for the entire algal community will remain a formidable challenge [15,20,23].

Unfortunately, under a wide variety of direct and interactive environmental perturbations [9,24,25], the prediction of algal communities is becoming increasingly difficult. For instance, excessive nutrients are widely considered the fundamental factor boosting phytoplankton growth [4,26]. In addition, the incorporative effects of meteorology [27] and hydrology [28,29] can also control the algal variability and structures. Recent studies have shown that the hydrometeorological factors (notably, inlet-outlet runoff, rainfall, light, wind, and temperature) can vastly control epilimnion conditions (e.g., water mixing, warming and stratification, and resident time), thereby, jointly regulating the algal

* Corresponding author. Ocean College, Zhejiang University, #1 Zheda Road, Zhoushan, Zhejiang, 316021, China.
E-mail address: xi@zju.edu.cn (X. Xiao).

https://doi.org/10.1016/j.ese.2022.100233

production with nutrients [5,9,30–32]. This calls for consideration of multi-source environmental influences on phytoplankton dynamics, further complicating ecological modeling.

Given the recent developments in machine learning, state-of-the-art computational intelligent techniques are now suitable to overcome the overall difficulty, especially, the random forests (RF) [33]. The RF, among the wide range of machine learning approaches, such as neural networks [22,34], support vector regressions [35,36], and genetic programming [37], are well-known for the robust modeling of intricate ecosystems with ensemble strategies [38]. Moreover, current interpretable studies have provided deep insights into its interpretability weakness by disassembling the tree-based structure [39–41]. The emerging RF-based feature analysis is no longer limited to identifying the important environmental factors individually. Furthermore, it proposes a feasible framework to understand the interactive effects of factors on modeling [38,41,42]. This high accommodation of RF presents a promise to predict and describe the complicated response of algal communities to environmental variations. Yet, to our knowledge, experiences with machine-learning-driven exploration are still very limited for such a complex ecological problem.

Therefore, in this study, we explored the prediction of phytoplankton community structure using the machine-learning RF method, with the meteorological, hydrological, and physicochemical variables integrated as inputs of this model. The RF models were developed and validated using the long-term phytoplankton time series (n = 858, 1994–2021) in Norway's largest lake, Lake Mjosa. A total of 13 major algal classes and the total cell biovolume were considered. Furthermore, the interpretability of the RF models was investigated via analyzing the importance and interactions among environmental variables.

2. Material and methods

2.1. Study area

Lake Mjosa (Norwegian: Mjøsa) is the largest Norwegian lake, located in the north of Oslo city (southern Norway) (Fig. 1). It has a surface area of 369 km^2, a medium depth of 150 m, a maximum length of 117 km, the theoretical residence time of 4.89 years, and is surrounded by one of Norway's most important agricultural districts [43]. The lake catchment is mostly a mountainous area, receiving glaciers meltwater through main tributaries [44]. However, the lake also receives discharges from urban sewage treatment plants, as well as industrial and agricultural sources, which cause serious eutrophication and induced widespread surface cyanobacterial blooms (*Oscillatoria bornetii fo. Tenuis*) starting in the early 1970s [45]. This took a national campaign to regulate the lake by comprehensively reducing phosphorus loads during the 1980s, which substantially decreased lake phosphorus concentrations [46]. Since then, the total phytoplankton biomass has declined, and cyanobacteria no longer thrive. Nevertheless, diatom over-proliferation events still often occur and are a nuisance to local water resources by forming sticky layers [32].

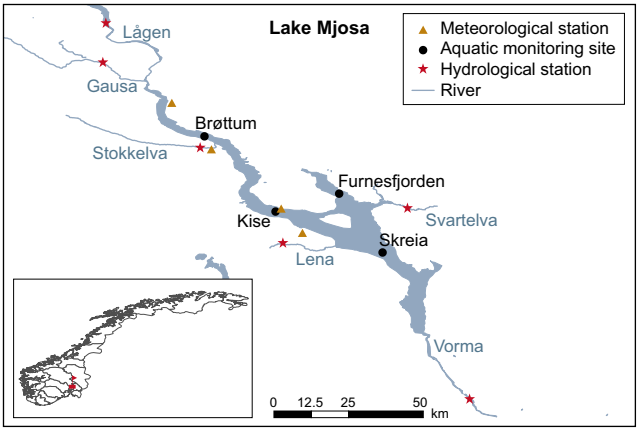

Fig. 1. Locations of aquatic, meteorological, and hydrological monitoring sites in the Lake Mjosa region.

2.2. Description and collection of modeling dataset

2.2.1. Phytoplankton community and water physicochemical data

The aquatic environment in Lake Mjosa is well monitored by the "AquaMonitor" project (https://aquamonitor.niva.no/mjosovervak) of the Norwegian Institute for Water Research (NIVA). To obtain long-term epilimnion records from the lake, four main sites were selected to continuously collect *in situ* water samples (0–10 m depth) at a regular frequency (bi-weekly or monthly during May–October; the ice-free season) (Fig. 1). Based on field collections, the phytoplankton samples were analyzed down to species level (only released at taxonomic class level, details in Table S1) under an inverted microscope and were quantified using cell biovolume (mm^3 m^{-3}) [43]. The phytoplankton composition was measured by the relative biovolume per taxonomic class. As shown in Table 1, four physicochemical variables related to the phytoplankton growth, including total nitrogen (TN), total phosphorus (TP), water transparency (Transparency), and water temperature (WT), which were also taken and analyzed by chemical or physical methods following a standardized methodology [43]. Using TN and TP concentrations, the mass ratio of nitrogen to phosphorus (N:P) was also calculated. In total, 858 data samples consisting of 13 taxonomical phytoplankton classes and five water physicochemical variables were collected, spanning 28 years of records from May 1994 to October 2021.

2.2.2. Meteorological and hydrological data

Based on the coordinates of the four aquatic monitoring sites, meteorological data were obtained from the closest weather stations of the Norwegian Meteorological Institute (MET) through the Norwegian Center for Climate Services (NCCS; https://seklima.met.no/observations). Together, four weather stations (Fig. 1) were selected to obtain meteorological conditions, including daily mean air temperature (AT), daily total precipitation (Precipitation), daily sunshine duration (Sun), and daily mean wind speed (Wind) (Table 1). Four sets of meteorological variables were created using the monitored data to better describe the meteorological status. For example, AT on the sample day (abbreviated as AT.0) and average AT for the days before the sampling day (including 3, 7, 14, 30 days; AT.3, AT.7, AT.14, AT.30). The same procedure was repeated to create the variable sets of sunshine duration and wind speed. On the other hand, precipitation variables were created as the precipitation on the sample day and the total precipitation for the days before the sampling day. Therefore, four variable sets, with five each, were created to measure meteorological conditions (20 in total, Table S2). Since monitored parameters available at each weather station were different, the meteorological variables assigned for each aquatic site were calculated from the average of observations at all selected weather stations (except precipitation, details in Table S1).

Hydrological data of the Lake Mjosa watershed were collected from the Norwegian Water Resources and Energy Directorate (NVE) delivered by the "Sildre" system (https://sildre.nve.no/map). To acquire the daily discharge conditions of Lake Mjosa, a total of five

Table 1
Overview of the modeling variables.

Monitoring dataset	Variable set	Abbr.	Unit	Range	Mean ± S.D.
Phytoplankton community structure	Biomass (total biovolume)	-	$mm^3\ m^{-3}$	7.2–2439.8	365.6 ± 334.3
	Composition (relative biovolume per class)	-	-	0–100%	-
Physicochemical factor	Total nitrogen	TN	$\mu g\ L^{-1}$	145.0–644.7	419.8 ± 88.1
	Total phosphorus	TP	$\mu g\ L^{-1}$	1.8–24.0	5.5 ± 1.8
	TN:TP ratio by mass	N:P	-	13.2–253.6	83.6 ± 34.2
	Water transparency	Transparency	m	0.9–18.0	8.0 ± 2.2
	Water temperature	WT	°C	3.2–20.8	12.0 ± 3.6
Hydrological factor	Inflow discharge (log)	Inflow	$m^3\ s^{-1}$	4.4–7.8	5.9 ± 0.6
	Outflow discharge (log)	Outflow	$m^3\ s^{-1}$	1.0–7.3	6.0 ± 0.7
	Discharge difference (log)	Q.Diff	$m^3\ s^{-1}$	−1.1–3.8	−0.1 ± 0.5
Meteorological factor	Total precipitation	Precipitation	mm	0.0–349.3	72.6 ± 38.0
	Air temperature	AT	°C	−4.9–23.8	11.7 ± 5.0
	Sunshine duration	Sun	hour	0.0–15.4	4.4 ± 3.4
	Wind speed	Wind	$m\ s^{-1}$	0.0–5.2	1.8 ± 0.8
Spatial factor	Latitude	-	°	60.6–61.0	-
	Longitude	-	°	10.6–11.1	-

main inlet rivers that can occupy over 90% of lake inflows [43,44] were selected to sum the daily inflow discharge (Inflow). While the lake daily outflow discharge (Outflow) was acquired from its outlet Vorma [43] (Fig. 1, hydrological station details in Table S1). The outlet of Lake Mjosa was regulated for hydro-electric production during the entire sampling period (water withdrawn in winter months) [44]. Therefore, to indicate the hydraulic fluctuations in the lake, the daily discharge exchange difference (Q.Diff) was also calculated using inflow subtracted by the outflow discharges (Table 1). Same as the above meteorological AT variables, three sets of hydrological variables were subsequently created (15 in total, Table S2). The flow discharge data were all on a logarithmic scale to reduce the variance of raw data for facilitating modeling.

2.3. Random forests

2.3.1. Model development

Random forests (RF) are a combination of tree predictors, with each built independently using a random set of bootstrap samples from the overall data [33]. To conduct prediction, the RF model aggregates the results of each tree and then votes on the final output, with majority votes for classification and average outcomes for regression. For applications, two important hyperparameters: the number of trees in the forest and the number of variables considered to split the nodes, should be optimized to minimize the generalization error [47].

To appropriately fit the RF model in the present study, our modeling procedure included two phases: calibration and validation. In the calibration phase, approximately 80% of the overall data were used as the constructing data to optimize the model. Although inherent out-of-bag validation provides the tolerance of RF against overfitting [38], the constructing data were used to conduct the 10-fold cross-validation to avoid over-fitting problems [48]. During cross-validation, constructing data were randomly split into ten subsets, with one subset for testing and the remaining nine subsets for training, looping ten times till every subset was tested. After completing the cross-validation, results were averaged, and the parameters that performed best on the testing set were recorded. In the validation phase, model prediction performances were finally evaluated. Here, RF models were re-trained using the entire constructing data based on recorded parameters to make in-sample (i.e., the 80% calibrating set) and out-of-sample (i.e., the remaining 20% unused validating set) predictions.

RF models were applied with the *sklearn* library in Python 3.8 software [49], using a grid-search scheme to tune the two hyperparameters during cross-validation (details in Table S3). The *correlation coefficient* (R^2), *root-mean-square-error* (RMSE), and *normalized-root-mean-square-error* (NRMSE) were used to measure the deviation of predictions from the observations for each algal class. Moreover, the Bray-Curtis dissimilarity index was applied to measure the deviation of predicted community composition from the observed (all calculations in Text S1).

2.3.2. Variable importance and interaction analysis

By-products of RF from the internal estimates that monitor the error, strength, and correlation are also useful for model interpretation [33]. To evaluate the relative importance of each variable, two measures provided by RF were both applied: the mean-square-error decrease in accuracy (MSE decrease) and the mean decrease in node purity (node purity decrease). Specifically, the MSE decrease of each variable was computed from the RF prediction error on out-of-bag data via permuting variables, and the node purity decrease was calculated based on changes in node impurity from splitting on the variables (measured by the residual sum of squares). Additionally, to identify the interactive variables, the conditional mean minimal depth of pairwise variables was calculated and used to measure the interaction strength [41]. Moreover, double-variable partial dependence analysis was applied to quantify the interactions, which enabled graphically checking the interactive effects of important variables on model responses (i.e., phytoplankton variations) [38,40,50]. Detailed information about these measures can be found in the Supporting Information (Text S2).

To conduct this RF-based analysis, the MSE decrease and node purity decrease were computed through the R package *randomForest* [51], the conditional mean minimal depth was calculated using the R package *randomForestExplainer* [52], and the partial dependence plots were applied via the R package *pdp* [53]. Since data-driven analysis can only be guaranteed within the range of training data [19,53] and predictions on new data were no longer required during this procedure, this RF feature analysis was performed based on the overall dataset to maximally reveal the information of the studied data and avoid uncertain extrapolations.

2.3.3. Feature processing and pre-filtration

To facilitate the training of RF models, the input variables (i.e., model features) were normalized by:

$$x_i' = \frac{x_i - \overline{x}}{x_{sd}} \quad (1)$$

where, x_i' is the normalized value of observed x_i; $\overline{x}$ and x_{sd} are the

mean and standard deviation of the observed variable, respectively. In addition, missing values of model features were replaced with the mean values of corresponding variables.

Nevertheless, the correlation analysis revealed the multicollinearity within the seven sets of meteorological and hydrological variables (Table S2 and Fig. S1). These redundant variables hampered the RF model from functionally detecting the important and interactive variables. Thereby, to filter the redundant information within hydro-meteorological variable sets, only the most important one in each set was reserved for the RF-based feature analysis (details are in Fig. S2). The pre-filtrated hydrometeorological variables were reported in Table S4.

3. Results

3.1. Variation of algal community in the past three decades

The sharp shifts within the algal composition, and the lack of strong linear correlation to environmental variables, highlighted the heterogeneous and complicated property of the community data and thus challenged its prediction (Fig. 2 and Fig. S1). In detail, based on the collected 858 water samples at four aquatic stations from 1994 to 2021, the phytoplankton in Lake Mjosa mainly belonged to 13 taxonomic classes (Fig. 2). The Bacillariophyceae class (commonly known as diatom) accounted for approximately 40% of the phytoplankton biomass and contributed over 50% of the occurrences (478 of 858 counts), followed by the Cryptophyceae, Chrysophyceae, and Cyanophyceae classes (Fig. 2a). The phytoplankton biomass was highly dynamic and mostly ranged between 100 and 1000 $mm^3\ m^{-3}$ (Fig. 2b). In addition, the majority of the algal classes were found in low correlations with the 42 environmental variables recorded in the same historic period (low Pearson's coefficients, majority of $|r| < 0.3$). Moreover, several classes, such as Cyanophyceae (commonly known as cyanobacteria) and Raphidiophyceae, showed weak linear relationships with the environmental variables ($|r| < 0.1$) due to the zero-inflated data caused by their limited detections (Fig. S1).

3.2. Prediction of the algal community structure

The RF approach showed good performance in predicting phytoplankton community structure from the environmental variables (Figs. 3 and 4). No overfitting or underfitting was observed during the 10-fold cross-validation training processes (Fig. 3a and Table S5). The RF model accurately captured the dynamic patterns of the phytoplankton community, as indicated by the low cross-validation NRMSE (mostly <5%). In addition, using the pre-filtrated environmental variables (Table S4), the RF model also provided comparable accuracy with NRMSE of less than 10% (Fig. 3a). When making predictions for the final evaluations, satisfactory performance can be found both in in-sample calibrating data (high R^2, majority above 0.90) and out-of-sample validating data (R^2 ranged from 0.75 to 0.41) for total biomass and the individual composition of four dominant phytoplankton classes (Fig. 3b–f, Fig. S3, and Table S6). For other algal classes with low occurrences, although the excessive zeros largely biased the R^2 statistics, the extremely low RMSE and NRMSE (mostly below 0.05 and 10%) can still support the high prediction accuracy (Table S6).

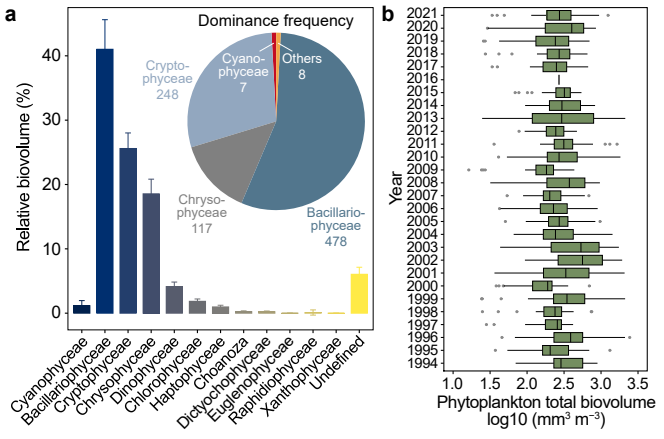

Fig. 2. Phytoplankton community in Lake Mjosa. **a**, The pie chart shows the statistics of dominance occurrences for 13 detected taxonomic phytoplankton classes. The bar chart with error bars represents the mean with standard error. **b**, The annual dynamics of total phytoplankton biomass.

Low dissimilarity errors were also found in predicting the phytoplankton composition (Fig. 4). The overall calibrating errors with outliers excepted were less than 17%, and the median error was only 6%. The validating prediction errors were slightly higher but still below 25%, with a median error of 17%. Moreover, it was worth noting that the low errors were often accompanied by the dominance occurrences of four dominant classes in the phytoplankton composition (Fig. 4b and c).

Therefore, the present RF method overcame the modeling challenges posed by the weak correlation between the phytoplankton community and environmental variables and can be used as a reliable approach to predict phytoplankton structure and identify the driving factors.

3.3. Identification and interpretation of the environmental drivers

On top of the robust prediction, the RF approach was also suggested because of its high interpretability. As an illustration, Fig. 5 presented the results of variable importance and interaction analysis by the RF model for the total phytoplankton biomass. For the prediction of phytoplankton biomass, it was evident that temperature was the major driver, followed by the hydrological and nutrient conditions, while the spatial factor showed limited influence (Fig. 5a). Moreover, strong interactions between the water physicochemical and meteorological variables were discovered, such as temperature–nutrient and temperature–wind (Fig. 5b–f). This observation was especially prevalent when the water temperature exceeded 15 °C. Analogically, the variable analyses for the other four dominant phytoplankton classes unrevealed a similar phenomenon (Figs. S4–S7).

The importance of the environmental variables on phytoplankton structure was also summarized in Fig. 6a. Notably, the hydrological and meteorological factors far outperformed the nutrients in shaping the individual composition, especially in the four dominant classes. As for the low occurrent phytoplankton classes, the meteorological variables such as sunshine duration and air temperature can also be the primary driver. In addition, the water temperature was shown to be the most important driver in regulating the total biomass. In general, water temperature and inflow discharge can be the two top drivers affecting phytoplankton community structure shifts.

The combined effects of the major drivers (i.e., WT and Inflow) and nutrients (TN and TP) were also shown in Fig. 6b–k. Interestingly, due to the leading competitive edge of the Bacillariophyceae class, the proportions of the remaining phytoplankton classes were complementary to its variation (Fig. 6b–k). The warm water temperature and low inflow discharge favored the increase in phytoplankton biomass and Bacillariophyceae composition. However, no significant nutrient limitations were observed for the phytoplankton communities. For instance, the localized phosphorus limitations were only observed for Cyanophyceae and

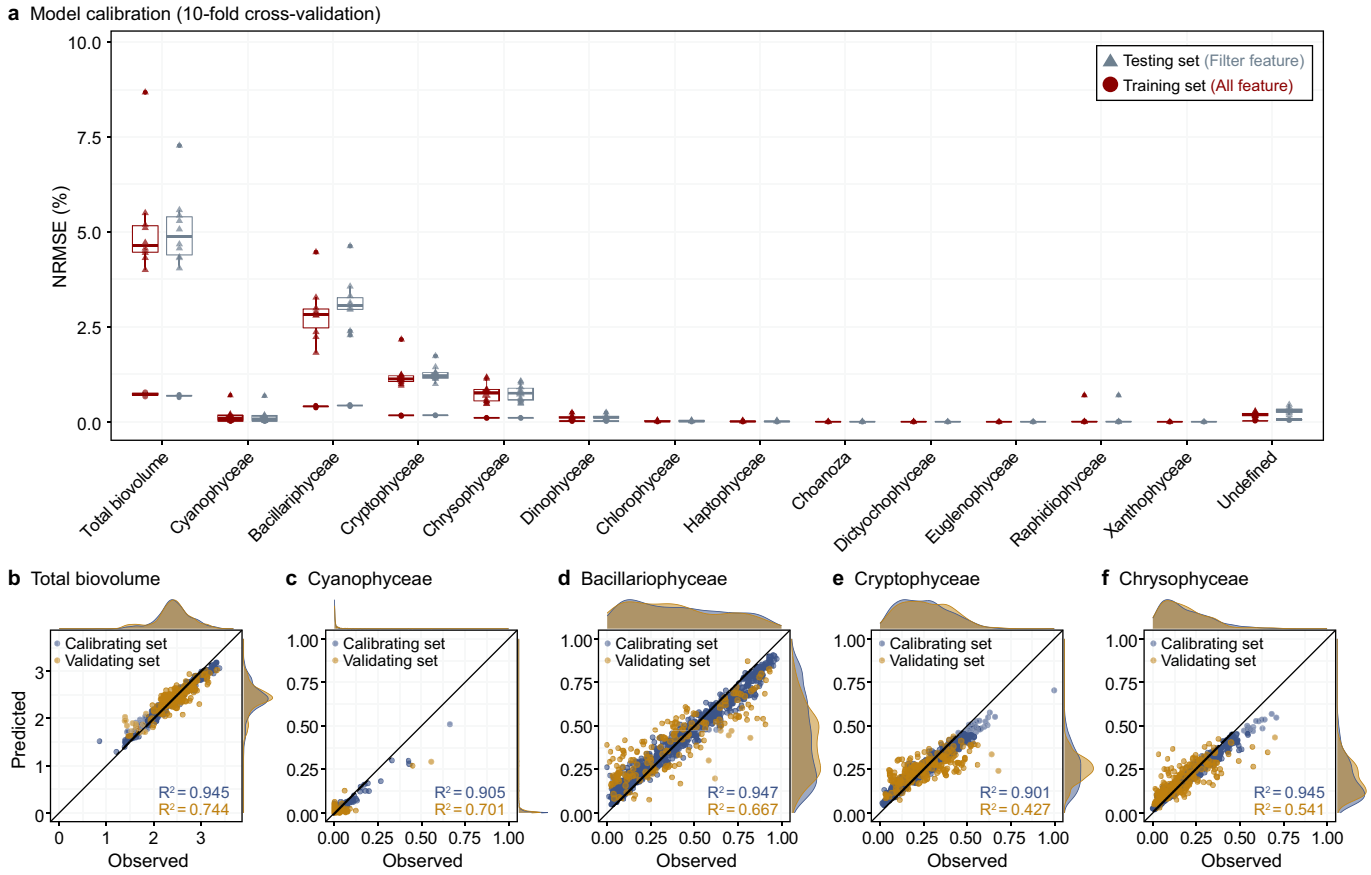

Fig. 3. Calibration and validation performance of RF models. **a**, The 10-fold cross-validation results for predicting total phytoplankton biomass (log_{10} biovolume) and individual class composition (relative biovolume per class). The dots and triangles represent the training and testing sets, respectively. **b**–**f**, Final validation for total algal biovolume (**b**) and four dominant classes: Cyanophyceae (**c**), Bacillariophyceae (**d**), Crytophyceae (**e**), and Chrysophyceae (**f**), displayed by the scatterplots with data distribution properties. The validation was based on all features.

Bacillariophyceae classes when the TN concentration reached certain levels (>500 and 200–400 μg L^{-1}, respectively; Fig. 6g–k). Despite that, nutrient influences on the phytoplankton composition were minimal relative to the major divers. Noticeably, the water temperature was highly positively correlated with previous air temperature conditions (correlation r = 0.5–0.7, Fig. S1A). This observation suggested that the hydro-meteorological factors in Lake Mjosa could strongly regulate its surface water characteristics and pose a joint impact on phytoplankton production. Moreover, the impacts of antecedent hydro-meteorological conditions (particularly the previous 30-day period status) were stronger than the current (Fig. S2 and Table S4).

4. Discussion

4.1. Machine learning for predicting phytoplankton community structure

As our RF models suggested, this machine learning (ML) approach accurately predicted the algal variability under intensive environmental perturbations (validation $R^2 > 0.740$), even for those communities that were weakly correlated with any explanatory variables (e.g., the cyanobacteria class (correlation $|r| < 0.05$, Fig. S1; prediction $R^2 > 0.701$, Fig. 3c and Table S6). Given the complexity of ecosystems [17], developing an accurate algal predictive model from high-dimensional and interactive environmental variables remains difficult [19]. Traditional statistical models often fail to capture such sophisticated ecological patterns due to the upfront assumption and linear form basis [21,22]. Comparatively, with the self-learning and self-organizing properties, ML has gradually been proven to be useful for modeling complex algal dynamics [19,40,54,55]. The RF technique is such a representative ML algorithm that fully characterizes the available variable interactions and reduces the biases by sharing and integrating the ability of internal decision trees [33]. Current studies have widely pointed out that this characteristic makes RF display superior capability in predicting erratic algal fluctuations under massive environmental influences [15,19,47]. This coincides with our findings and again suggests that ML-based techniques could be reliable tools for applying predictive models of algae.

Remarkably, the RF models presented powerful accommodation for a total of 13 different algal classes (mostly NRMSE <10%), with robust prediction to the shifts in phytoplankton structure among four monitoring sites (Bray-Curtis dissimilarity = 9.2 ± 7.0%). Historically, predicting algal composition shifts was a big challenge due to the heterogeneous environment-algae processes [20,56,57], especially for those process models with site-specific and species-specific parameters [18,20]. This highlights another major feature of the ML method, that is, it requires few prior ecological knowledge and is purely modeled upon data information, which is a good supplementary to the process-driven models [19]. As the prediction rule is exclusively extracted from data, such flexibility provides unexpected generality for the contemporary ML-based aquatic models, which usually involve a large-scale ecological dataset with

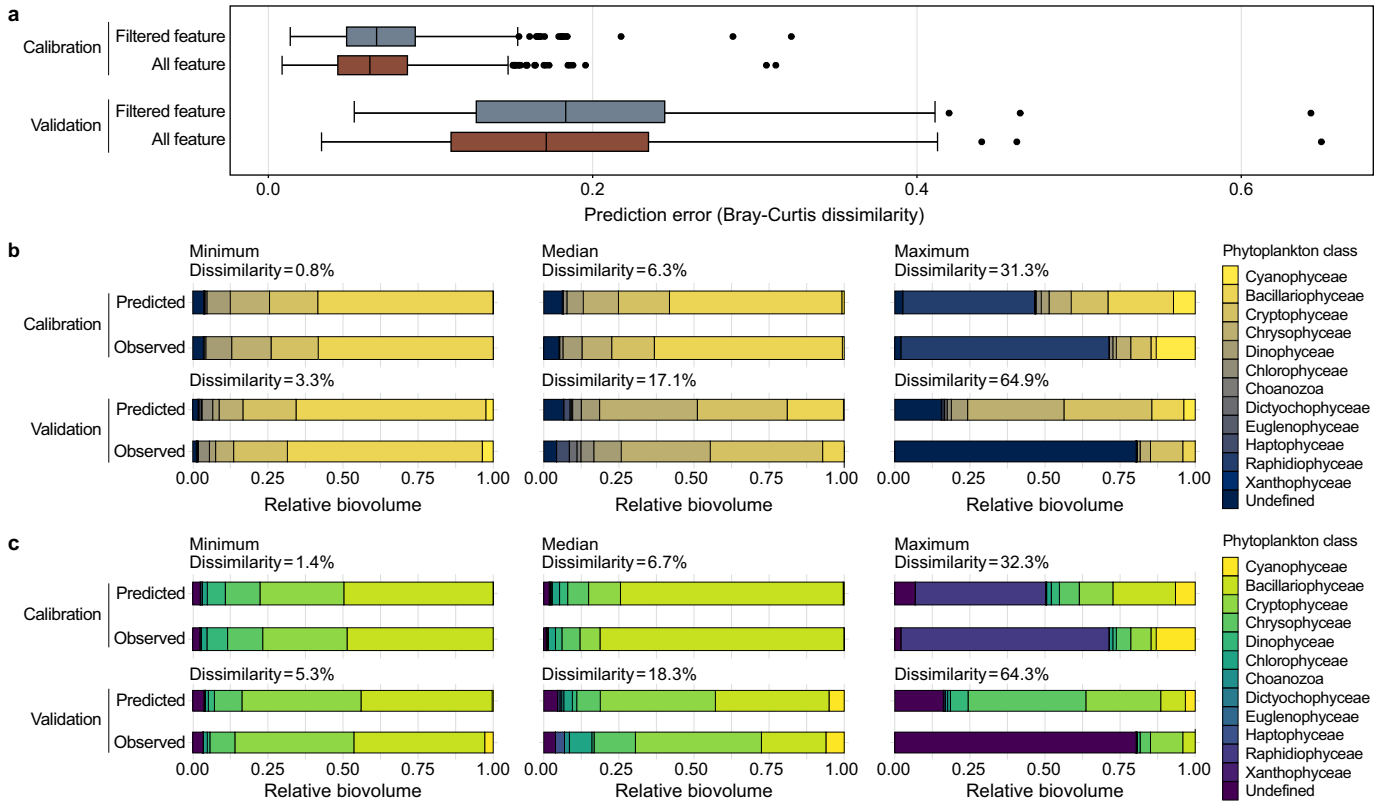

Fig. 4. Performance of RF models for predicting phytoplankton composition. **a**, Performance statistics for in-sample calibration and out-of-sample validation. **b**–**c**, The predicted and observed algal composition corresponding to the minimal, median, and maximum prediction error (including outliers), based on the models using all feature (**b**) and filtered feature (**c**), respectively.

substantial algal communities [20,57]. Notably, recent studies have shown that integrating prior knowledge could enhance algal prediction via facilizing variable selections in data-driven models [58,59]. In this sense, data information and domain knowledge can complement each other for future ML-based phytoplankton composition modeling. Therefore, this study has shown a promising direction for digging the ecological values of ML (e.g., modeling the algal community shifting) due to its high accuracy and general practicality.

4.2. Deep exploration for the ecological interpretation of machine learning

More importantly, this work presented a reproducible workflow for the in-depth ecological interpretation of ML approaches. As an emerging tool, the explanation of ML-learned relationships can be crucial since it validates modeling reliability and informs future studies [60]. But this has always been a difficulty due to the inflated parameters and complex structure of ML models [61]. Fortunately, with the advancement in ML interpretability, the importance measure for individual input variables has become less of a challenge. For instance, Gebler et al. [62] used sensitivity analysis in the neural network to evaluate the impact of river ecological variables. Panidhapu et al. [63] identified the important environmental variables for predicting microbes based on the strength of influence in the Bayesian network. Specifically, simple monitoring for internal errors (i.e., MSE decrease and node purity decrease) prevents the RF from suffering such shortcomings (Fig. 6a). In fact, this feature has already made RF a widely used method for variable selection [38,40,42].

This work is also one of the earliest ML-driven studies in uncovering the joint variable effects driving the phytoplankton community shifts. We showed the calculation of conditional depth largely benefited from the detection of interactive variables in RF models (Fig. 5 and Figs. S4–S7). The partial dependence further facilitated the quantification of important interactions in affecting algal structure shifts, such as the temperature–discharge and TN-TP (Fig. 6). In practical terms, such variable interaction analysis is vital to support water management, e.g., delineating nutrient guidelines [64], and discovering nonlinear environmental effects on algal blooms [65]. However, in previous research, the interpretability exploration of most ML studies often stops at revealing importance and ignores variable interactions [41,66,67]. Nowadays, given the high-speed development in RF computational procedures, increasing ML-driven environment subfields have begun to focus on variable interaction analysis [41,50,68]. Despite that, to our knowledge, there are still very limited interpretation explorations for ML-based algal prediction tasks [15,40]. Hence, we strongly advocate for using the RF model due to its robustness and stable analysis procedures [38]. Moreover, the interaction analysis can be constructed for any ML method via partial dependence plots [69]. With this range of abilities, ML could offer a promising alternative to traditional statistical methods for analyzing high-stakes ecological management in the future [38].

4.3. Important driving factors affecting shifts in algal community structure

The present results suggested that antecedent hydro-meteorological factors (mostly 30-day period) were the most

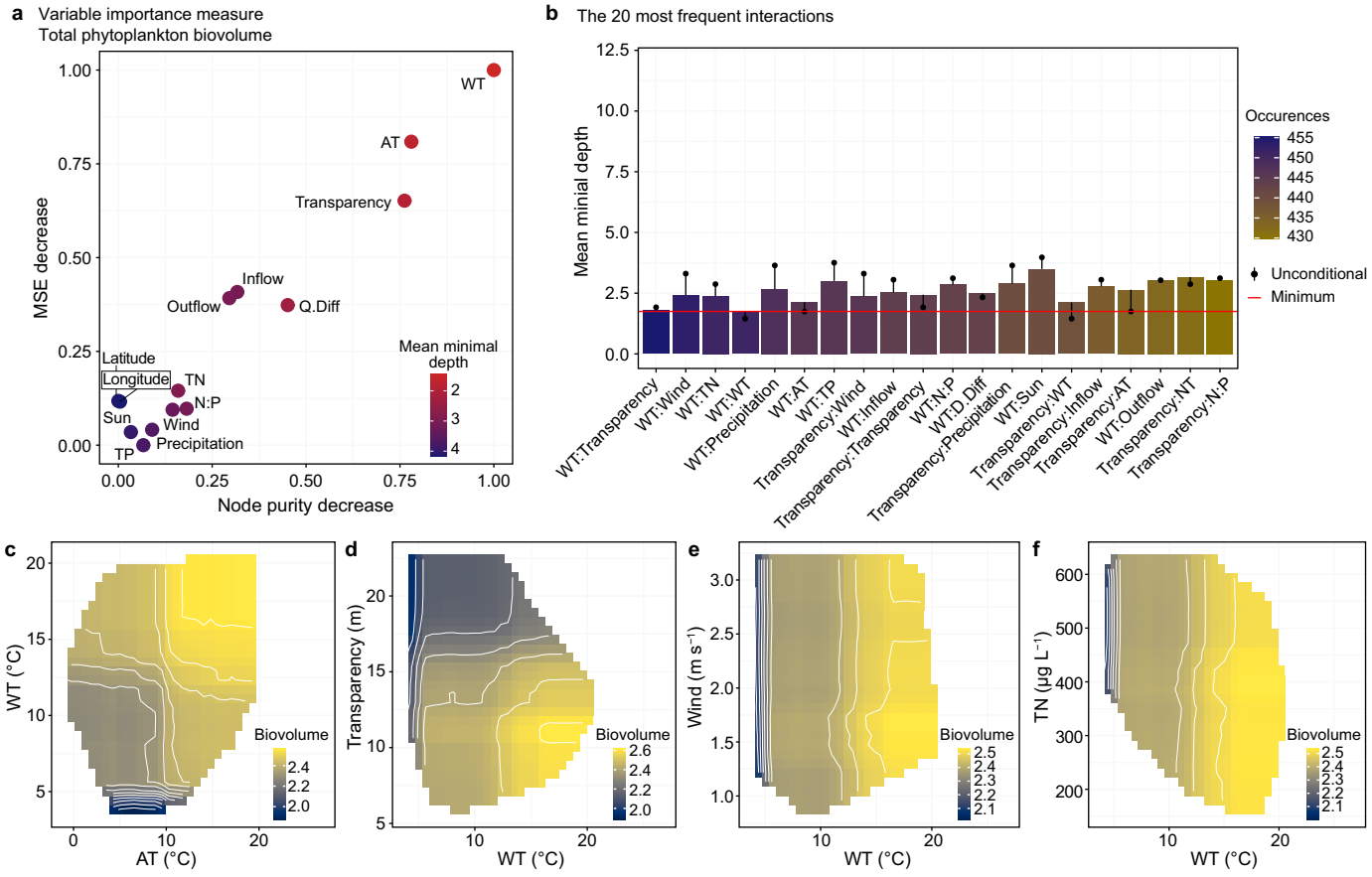

Fig. 5. Variable importance and interaction analysis by RF model for total phytoplankton biomass (log_{10} biovolume). **a**, Variable importance analysis combining the MSE decrease and node purity decrease. The standardized MSE decrease and standardized node purity decrease were employed to measure the variable importance. **b**, The discovered top 20 interactive variables via interaction analysis using the conditional mean minimal depth. The occurrences refer to the interaction occurrences between two features. The variable analysis was based on the filtered features. **c**–**f**, Bivariate partial dependence analysis on phytoplankton biomass for the two most important variables (**c**) and the three strongest pairs of interacted variables (**d**–**f**).

important factors in affecting algal biomass and composition shifts since they mainly controlled the behaviors of diatoms (as mentioned in section 3.3). The high surface water temperature and mild inflows promoted both the algal proliferation and diatoms dominance and can jointly cause approximately ten times wider biomass changes than other variables (Fig. 6). Noticeably, the relatively high biomass in Lake Mjosa (>300 $mm^3\ m^{-3}$) appeared only when water temperature raised and exceeded 15 °C (Fig. 5c), which typically can only be reached in warm summer months and when stable vertical thermal stratification occurs [43,46]. The analogical results have been shown in many previous studies. For example, Cha et al. [70] revealed a stronger cyanobacterial sensitivity to the low outflow and high water temperature (6-day before) than the nutrients in a South Korean reservoir. In addition, algal growth was found to mostly prefer a stable antecedent water fluctuation (10-day period) in the largest tributary of the Yangtze River, China [40]. Interestingly, the steady and stratified epilimnion environment could favor the floating of cyanobacteria, where they have better access to light and shade non-buoyant classes [9,71], especially the diatoms with heavier and larger cells [70]. However, the huge absence of cyanobacterial colonies due to the extremely low phosphorus content was largely eliminated in Lake Mjosa. In contrast, the widespread diatoms, along with cryptophytes and chrysophytes, took full advantage of favorable conditions and thus dominated phytoplankton. Similar composition shifts have been reported in many lakes. For example, Lake Oswego in the USA [72], and 35 lakes in North America and Europe [11]. This indicates that the hydro-meteorological characteristics are also important monitoring components throughout the water management period, as they could largely alter phytoplankton community structure.

Contrary to our expectations, phytoplankton growth was not sensitive to the single variation of nutrients (biomass changed only from 250 to 280 $mm^3\ m^{-3}$). No significant phosphorus limitations existed in the TN-TP dependence analysis, despite the whole lake TP condition (1.8–24.0 $\mu g\ L^{-1}$) was below the thresholds proposed by most empirical studies (e.g., 200 $\mu g\ L^{-1}$ in Lake Taihu, China [73], 100 $\mu g\ L^{-1}$ in 137 Iowa lakes, USA [74], 30 $\mu g\ L^{-1}$ in 221 lakes from 14 countries [75]). In contrast, the high levels of phosphorus and nitrogen interact with other hydro-meteorological factors, including temperature and discharges, more readily to promote algal growth (Fig. 5 and Figs. S4–S7). Recently, many studies have claimed this collaboration may be performed through the short-term nutrient enrichments induced by the intense hydro-meteorological activities [27,76–78]. Some specific algae (e.g., *Microcystis*) could also rapidly utilize incremental nutrients under such favorable condition and cause folds of increase in the overall biomass [73,79]. Nevertheless, the internal nutrient replenishment and recycling in deep lakes are still complicated and require further investigation [76,80]. Despite that, management of nutrient inputs to the watershed is the key to mitigating lake eutrophication and can be a radical approach to control the proliferation of phytoplankton community [4,9,25].

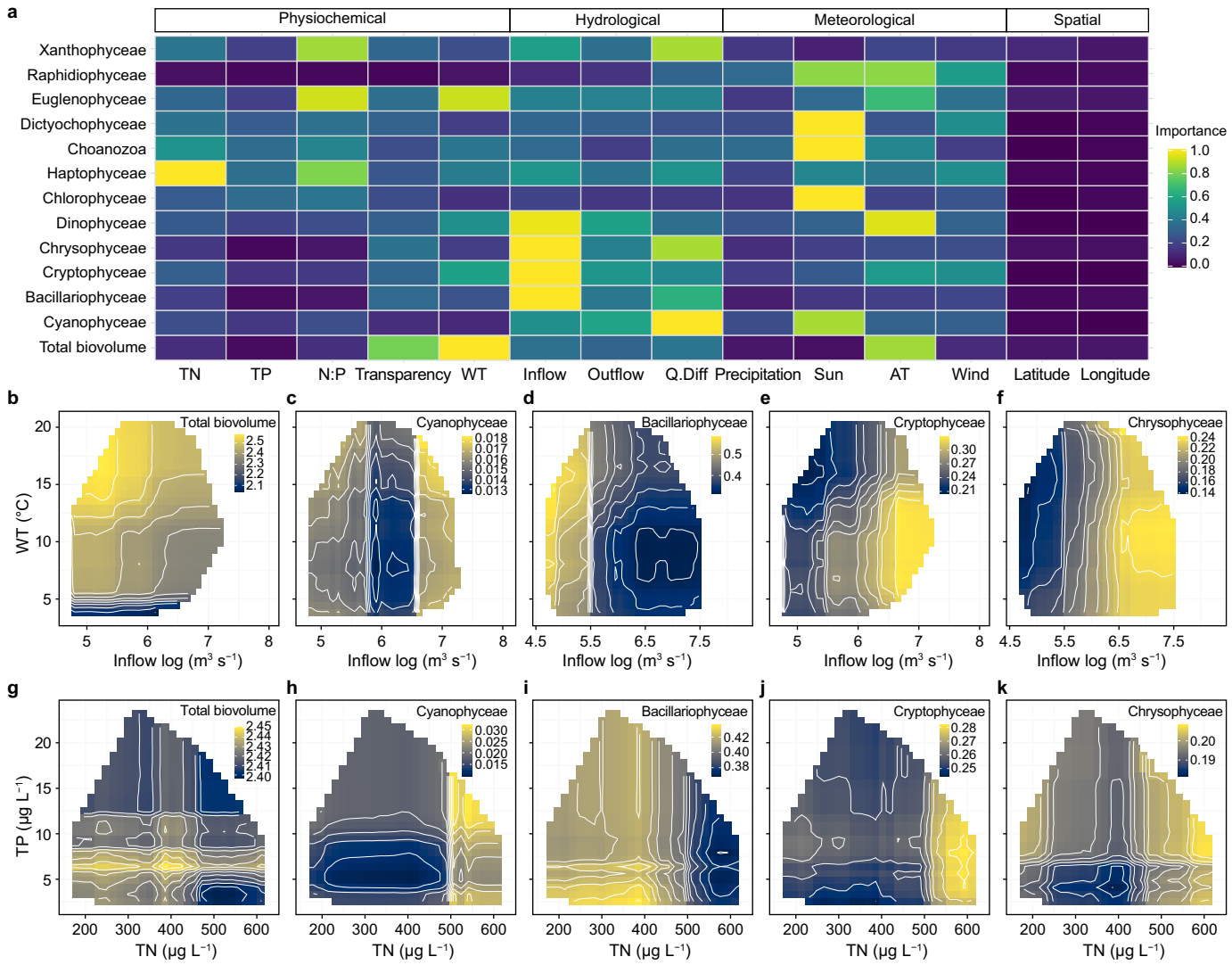

Fig. 6. Importance measure of each variable type for total phytoplankton biomass ($\log_{10}$ biovolume) and individual class composition (relative biovolume per class). **a**, Variable importance plotted by the average index of standardized MSE decrease and standardized node purity decrease. **b**–**k**, Bivariate partial dependence plots of total biomass and four dominant classes, based on the water temperature and inflow discharge (**b**–**f**) and nutrients (**g**–**k**).

5. Conclusion

This study investigated the use of machine-learning random forests (RF) method to predict and interpret phytoplankton structure shifts from water physicochemical, meteorological, and hydrological variables. The RF models presented robust and accurate predictions on both phytoplankton biomass and community composition. The deep RF-based explanation showed that the hydro-meteorological factors outperformed nutrients in affecting phytoplankton community structure. Especially, the temperature, lake inflow, and nutrients jointly posed strong regulations to the algal community shifts in this nutrient-poor system.

As such, this work provided a reproducible machine learning workflow to handle the challenging phytoplankton structure predictions and reveal the underlying complicated ecological relationships. In the future, this workflow could be expanded to the species level from the current class level and include anthropogenic stress as environmental variables, which were limited by the dataset as shown in this study.

CRediT authorship contribution statement

Muyuan Liu: Conceptualization, Software, Visualization, Writing. **Yuzhou Huang:** Visualization. **Jing Hu:** Visualization. **Junyu He:** Software. **Xi Xiao:** Conceptualization, Supervision.

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.

Acknowledgement

This study was financially supported by the National Natural Science Foundation of China (21876148), the Zhejiang Provincial Natural Science Foundation/Funds for Distinguished Young Scientists (LR22D06003), the Key Laboratory of Marine Ecological Monitoring and Restoration Technologies of the Ministry of Natural

Resources of China (MEMRT202102), Science Foundation of Donghai Laboratory (DH-2022KF01021) and Fundamental Research Funds for the Central Universities (226-2022-00119) and Funding for ZJU Tang Scholar to X. X. The authors acknowledge the data sharing from the Norwegian Institute for Water Research (NIVA). We appreciate those who participated in the NIVA "AquaMonitor" project (https://aquamonitor.niva.no/mjosovervak) by collecting and measuring such high-quality records for their dedicated works. We are also grateful for the meteorological and hydrological data sharing from the Norwegian Meteorological Institute (MET), and the Norwegian Water Resources and Energy Directorate (NVE). We thank Dr. Kokoette Effiong and ZJU PhD student Yuvna Devi Perianen for their assistance in the written English.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2022.100233.

References

[1] P. Falkowski, Ocean Science: the power of plankton, Nature 483 (2012) S17–S20, https://doi.org/10.1038/483S17a.
[2] C.B. Field, M.J. Behrenfeld, J.T. Randerson, P. Falkowski, Primary production of the biosphere: integrating terrestrial and oceanic components, Science (80-281 (1998) 237–240, https://doi.org/10.1126/science.281.5374.237.
[3] G.M. Hallegraeff, A review of harmful algal blooms and their apparent global increase, Phycologia 32 (1993) 79–99, https://doi.org/10.2216/i0031-8884-32-2-79.1.
[4] J. Heisler, P.M. Glibert, J.M. Burkholder, D.M. Anderson, W. Cochlan, W.C. Dennison, Q. Dortch, C.J. Gobler, C.A. Heil, E. Humphries, A. Lewitus, R. Magnien, H.G. Marshall, K. Sellner, D.A. Stockwell, D.K. Stoecker, M. Suddleson, Eutrophication and harmful algal blooms: a scientific consensus, Harmful Algae 8 (2008) 3–13, https://doi.org/10.1016/J.HAL.2008.08.006.
[5] H.W. Paerl, J. Huisman, Climate change: a catalyst for global expansion of harmful cyanobacterial blooms, Environ. Microbiol. Rep. 1 (2009) 27–37, https://doi.org/10.1111/j.1758-2229.2008.00004.x.
[6] X. Xiao, S. Agustí, Y. Pan, Y. Yu, K. Li, J. Wu, C.M. Duarte, Warming amplifies the frequency of harmful algal blooms with eutrophication in Chinese coastal waters, Environ. Sci. Technol. 53 (2019a) 13031–13041, https://doi.org/10.1021/acs.est.9b03726.
[7] X. Xiao, J. He, Y. Yu, B. Cazelles, M. Li, Q. Jiang, C. Xu, Teleconnection between phytoplankton dynamics in north temperate lakes and global climatic oscillation by time-frequency analysis, Water Res. 154 (2019b) 267–276, https://doi.org/10.1016/j.watres.2019.01.056.
[8] H. Huang, X. Xiao, J. Shi, Y. Chen, Structure-activity analysis of harmful algae inhibition by congeneric compounds: case studies of fatty acids and thiazolidinediones, Environ. Sci. Pollut. Res. 21 (2014) 7154–7164, https://doi.org/10.1007/s11356-014-2626-0.
[9] J. Huisman, G.A. Codd, H.W. Paerl, B.W. Ibelings, J.M.H. Verspagen, P.M. Visser, Cyanobacterial blooms, Nat. Rev. Microbiol. 16 (2018) 471–483, https://doi.org/10.1038/s41579-018-0040-1.
[10] X. Xiao, C. Li, H. Huang, Y.P. Lee, Inhibition effect of natural flavonoids on red tide alga Phaeocystis globosa and its quantitative structure-activity relationship, Environ. Sci. Pollut. Res. 26 (2019c) 23763–23776, https://doi.org/10.1007/s11356-019-05482-7.
[11] E. Jeppesen, M. Søndergaard, J.P. Jensen, K.E. Havens, O. Anneville, L. Carvalho, M.F. Coveney, R. Deneke, M.T. Dokulil, B. Foy, D. Gerdeaux, S.E. Hampton, S. Hilt, K. Kangur, J. Köhler, E.H.H.R. Lammens, T.L. Lauridsen, M. Manca, M.R. Miracle, B. Moss, P. Nõges, G. Persson, G. Phillips, R. Portielje, S. Romo, C.L. Schelske, D. Straile, I. Tatrai, E. Willén, M. Winder, Lake responses to reduced nutrient loading - an analysis of contemporary long-term data from 35 case studies, Freshw. Biol. 50 (2005) 1747–1771, https://doi.org/10.1111/j.1365-2427.2005.01415.x.
[12] E. Mette, M. Vanni, J. Newell, M.J. González, Phytoplankton communities and stoichiometry are interactively affected by light, nutrients, and fish, Limnol. Oceanogr. 56 (2011) 1959–1975, https://doi.org/10.4319/lo.2011.56.6.1959.
[13] A. Zingone, H. Oksfeldt Enevoldsen, The diversity of harmful algal blooms: a challenge for science and management, Ocean Coast Manag. 43 (2000) 725–748, https://doi.org/10.1016/S0964-5691(00)00056-9.
[14] K. Rao, X. Zhang, M. Wang, J. Liu, W. Guo, G. Huang, J. Xu, The relative importance of environmental factors in predicting phytoplankton shifting and cyanobacteria abundance in regulated shallow lakes, Environ. Pollut. 286 (2021), 117555, https://doi.org/10.1016/j.envpol.2021.117555.
[15] X. Zhao, S. Drakare, R.K. Johnson, Use of taxon-specific models of phytoplankton assemblage composition and biomass for detecting impact, Ecol. Indicat. 97 (2019) 447–456, https://doi.org/10.1016/j.ecolind.2018.10.026.
[16] C. Gameiro, P. Cartaxana, V. Brotas, Environmental drivers of phytoplankton distribution and composition in Tagus Estuary, Portugal, Estuar. Coast Shelf Sci. 75 (2007) 21–34, https://doi.org/10.1016/j.ecss.2007.05.014.
[17] J. Huisman, F.J. Weissing, Biodiversity of plankton by species oscillations and chaos, Nature 402 (1999) 407–410, https://doi.org/10.1038/46540.
[18] L.V. Lucas, J.K. Thompson, L.R. Brown, Why are diverse relationships observed between phytoplankton biomass and transport time? Limnol. Oceanogr. 54 (2009) 381–390, https://doi.org/10.4319/lo.2009.54.1.0381.
[19] B.Z. Rousso, E. Bertone, R. Stewart, D.P. Hamilton, A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes, Water Res. 182 (2020), 115959, https://doi.org/10.1016/j.watres.2020.115959.
[20] J. Shen, Q. Qin, Y. Wang, M. Sisson, A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading, Ecol. Model. 398 (2019) 44–54, https://doi.org/10.1016/j.ecolmodel.2019.02.005.
[21] S. Lek, M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, S. Aulagnier, Application of neural networks to modelling nonlinear relationships in ecology, Ecol. Model. 90 (1996) 39–52, https://doi.org/10.1016/0304-3800(95)00142-5.
[22] M. Liu, J. He, Y. Huang, T. Tang, J. Hu, X. Xiao, Algal bloom forecasting with time-frequency analysis: a hybrid deep learning approach, Water Res. 219 (2022a), 118591, https://doi.org/10.1016/j.watres.2022.118591.
[23] S. Michel-Mata, X. Wang, Y. Liu, M.T. Angulo, Predicting microbiome compositions from species assemblages through deep learning, iMeta (2022) 1–13, https://doi.org/10.1002/imt2.3.
[24] Y. Qu, N. Wu, B. Guse, N. Fohrer, Riverine phytoplankton shifting along a lentic-lotic continuum under hydrological, physiochemical conditions and species dispersal, Sci. Total Environ. 619–620 (2018) 1628–1636, https://doi.org/10.1016/J.SCITOTENV.2017.10.139.
[25] V.H. Smith, G.D. Tilman, J.C. Nekola, Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems, Environ. Pollut. 100 (1999) 179–196, https://doi.org/10.1016/S0269-7491(99)00091-3.
[26] C. Xu, Z. Ge, C. Li, F. Wan, X. Xiao, Inhibition of harmful algae Phaeocystis globosa and Prorocentrum donghaiense by extracts of coastal invasive plant Spartina alterniflora, Sci. Total Environ. 696 (2019), https://doi.org/10.1016/j.scitotenv.2019.133930.
[27] Z. Yang, M. Zhang, X. Shi, F. Kong, R. Ma, Y. Yu, Nutrient reduction magnifies the impact of extreme weather on cyanobacterial bloom formation in large shallow Lake Taihu (China), Water Res. 103 (2016) 302–310, https://doi.org/10.1016/j.watres.2016.07.047.
[28] J. Huisman, J. Sharples, J.M. Stroom, P.M. Visser, W.E.A. Kardinaal, J.M.H. Verspagen, B. Sommeijer, Changes in turbulent mixing shift competition for light between phytoplankton species, Ecology 85 (2004) 2960–2970, https://doi.org/10.1890/03-0763.
[29] A. Sharpley, H.P. Jarvie, A. Buda, L. May, B. Spears, P. Kleinman, Phosphorus legacy: overcoming the effects of past management practices to mitigate future water quality impairment, J. Environ. Qual. 42 (2013) 1308–1326, https://doi.org/10.2134/jeq2013.03.0098.
[30] O. Anneville, S. Gammeter, D. Straile, Phosphorus decrease and climate variability: mediators of synchrony in phytoplankton changes among European peri-alpine lakes, Freshw. Biol. 50 (2005) 1731–1746, https://doi.org/10.1111/j.1365-2427.2005.01429.x.
[31] J.A. Ferris, J.T. Lehman, Interannual variation in diatom bloom dynamics: roles of hydrology, nutrient limitation, sinking, and whole lake manipulation, Water Res. 41 (2007) 2551–2562, https://doi.org/10.1016/j.watres.2007.03.027.
[32] A. Hobaek, J.E. Løvik, T. Rohrlack, S.J. Moe, M. Grung, H. Bennion, G. Clarke, G.T. Piliposyan, Eutrophication, recovery and temperature in Lake Mjøsa: detecting trends with monitoring data and sediment records, Freshw. Biol. 57 (2012) 1998–2014, https://doi.org/10.1111/j.1365-2427.2012.02832.x.
[33] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[34] X. Xiao, J. He, H. Huang, T.R. Miller, G. Christakos, E.S. Reichwaldt, A. Ghadouani, S. Lin, X. Xu, J. Shi, A novel single-parameter approach for forecasting algal blooms, Water Res. 108 (2017) 222–231, https://doi.org/10.1016/J.WATRES.2016.10.076.
[35] P.J. García-Nieto, E. García-Gonzalo, F. Sánchez Lasheras, J.R. Alonso Fernández, C. Díaz Muñiz, A hybrid DE optimized wavelet kernel SVR-based technique for algal atypical proliferation forecast in La Barca reservoir: a case study, J. Comput. Appl. Math. 366 (2020), 112417, https://doi.org/10.1016/j.cam.2019.112417.
[36] J. He, Y. Chen, J. Wu, D.A. Stow, G. Christakos, Space-time chlorophyll-a retrieval in optically complex waters that accounts for remote sensing and modeling uncertainties and improves remote estimation accuracy, Water Res. 171 (2020), 115403, https://doi.org/10.1016/j.watres.2019.115403.
[37] S. Ding, C. Su, J. Yu, An optimizing BP neural network algorithm based on genetic algorithm, Artif. Intell. Rev. 36 (2011) 153–162, https://doi.org/10.1007/s10462-011-9208-z.
[38] D.R. Cutler, T.C. Edwards, K.H. Beard, A. Cutler, K.T. Hess, J. Gibson, J.J. Lawler, Random forests for classification in ecology, Ecology 88 (2007) 2783–2792, https://doi.org/10.1890/07-0539.1.
[39] X. Hu, J.H. Belle, X. Meng, A. Wildani, L.A. Waller, M.J. Strickland, Y. Liu, Estimating PM2.5 concentrations in the conterminous United States using the random forest approach, Environ. Sci. Technol. 51 (2017) 6936–6944, https://doi.org/10.1021/acs.est.7b01210.

[40] R. Xia, G. Wang, Y. Zhang, P. Yang, Z. Yang, S. Ding, X. Jia, C. Yang, C. Liu, S. Ma, J. Lin, X. Wang, X. Hou, K. Zhang, X. Gao, P. Duan, C. Qian, River algal blooms are well predicted by antecedent environmental conditions, Water Res. 185 (2020), 116221, https://doi.org/10.1016/j.watres.2020.116221.
[41] F. Yu, C. Wei, P. Deng, T. Peng, X. Hu, Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles, Sci. Adv. 7 (2021a) 1–15, https://doi.org/10.1126/sciadv.abf4130.
[42] Z. Ban, Q. Zhou, A. Sun, L. Mu, X. Hu, Screening priority factors determining and predicting the reproductive toxicity of various nanoparticles, Environ. Sci. Technol. 52 (2018) 9666–9676, https://doi.org/10.1021/acs.est.8b02757.
[43] J.-E. Thrane, A. Økelsrud, B. Skjelbred, S.B. Ranneklev, J.P. Håll, M.R. Kile, Tiltaksorientert overvåking i vannområde Mjøsa. Årsrapport for 2020, NIVA Rep. (2021) 167, 7622–2021.
[44] G. Milbrink, Oligochaetes and water pollution in two deep Norwegian lakes, Hydrobiologia 278 (1994) 213–222, https://doi.org/10.1007/BF00142329.
[45] H. Holtan, The Lake Mjøsa story, Arc. Hydrobiol. Beih 13 (1979) 242–258.
[46] J.E. Løvik, G. Kjellberg, Long-term changes of the crustacean zooplankton community in Lake Mjøsa, the largest lake in Norway, J. Limnol. 62 (2003) 143–150, https://doi.org/10.4081/jlimnol.2003.143.
[47] J. Peters, B. De Baets, N.E.C. Verhoest, R. Samson, S. Degroeve, P. De Becker, W. Huybrechts, Random forests as a tool for ecohydrological distribution modelling, Ecol. Model. 207 (2007) 304–318, https://doi.org/10.1016/J.ECOLMODEL.2007.05.011.
[48] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Int. Jt. Conf. Artif. Intell. (1995).
[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, Machine learning in Python, J. Mach. Learn. Res. 12 (2019) 128–154, https://doi.org/10.4018/978-1-5225-9902-9.ch008.
[50] T. Allen, K.A. Murray, C. Zambrana-Torrelio, S.S. Morse, C. Rondinini, M. Di Marco, N. Breit, K.J. Olival, P. Daszak, Global hotspots and correlates of emerging zoonotic diseases, Nat. Commun. 8 (2017) 1124, https://doi.org/10.1038/s41467-017-00923-8.
[51] A. Liaw, M. Wiener, Classification and regression by randomForest, R. News 2 (2002) 18–22, https://doi.org/10.1023/A:1010933404324.
[52] A. Paluszynska, P. Biecek, Y. Jiang, Explaining and visualizing Random Forests in terms of variable importance, R. News (2020), https://doi.org/10.1198/jasa.2009.tm08622.
[53] B.M. Greenwell, pdp: an R package for constructing partial dependence plots, R J 9 (2017) 421–436, https://doi.org/10.32614/rj-2017-016.
[54] T.D. Harris, J.L. Graham, Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset, Lake Reservoir Manag. 33 (2017) 32–48, https://doi.org/10.1080/10402381.2016.1263694.
[55] P. Yu, R. Gao, D. Zhang, Z. Liu, Predicting coastal algal blooms with environmental factors by machine learning methods, Ecol. Indicat. 123 (2021b), 107334, https://doi.org/10.1016/j.ecolind.2020.107334.
[56] H.S. Lee, J.H.W. Lee, Continuous monitoring of short term dissolved oxygen and algal dynamics, Water Res. 29 (1995) 2789–2796, https://doi.org/10.1016/0043-1354(95)00126-6.
[57] Y. Park, K.H. Cho, J. Park, S.M. Cha, J.H. Kim, Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea, Sci. Total Environ. 502 (2015) 31–41, https://doi.org/10.1016/j.scitotenv.2014.09.005.
[58] Q. Chen, A.E. Mynett, Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake, Ecol. Model. 162 (2003) 55–67, https://doi.org/10.1016/S0304-3800(02)00389-7.
[59] R. Fornarelli, S. Galelli, A. Castelletti, J.P. Antenucci, C.L. Marti, An empirical modeling approach to predict and understand phytoplankton dynamics in a reservoir affected by interbasin water transfers, Water Resour. Res. 49 (2013) 3626–3641, https://doi.org/10.1002/wrcr.20268.
[60] S. Zhong, K. Zhang, M. Bagheri, J.G. Burken, A. Gu, B. Li, X. Ma, B.L. Marrone, Z.J. Ren, J. Schrier, W. Shi, H. Tan, T. Wang, X. Wang, B.M. Wong, X. Xiao, X. Yu, J.-J. Zhu, H. Zhang, Machine learning: new ideas and tools in environmental science and engineering, Environ. Sci. Technol. acs.est (2021), 1c01339, https://doi.org/10.1021/acs.est.1c01339.
[61] X. Liu, D. Lu, A. Zhang, Q. Liu, G. Jiang, Data-driven machine learning in environmental pollution: gains and problems, Environ. Sci. Technol. 56 (2022b) 2124–2133, https://doi.org/10.1021/acs.est.1c06157.
[62] D. Gebler, G. Wiegleb, K. Szoszkiewicz, Integrating river hydromorphology and water quality into ecological status modelling by artificial neural networks, Water Res. 139 (2018) 395–405, https://doi.org/10.1016/j.watres.2018.04.016.
[63] A. Panidhapu, Z. Li, A. Aliashrafi, N.M. Peleato, Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks, Water Res. 170 (2020), 115349, https://doi.org/10.1016/j.watres.2019.115349.
[64] O. Malve, S.S. Qian, Estimating nutrients and chlorophyll a relationships in Finnish lakes, Environ. Sci. Technol. 40 (2006) 7848–7853, https://doi.org/10.1021/es061359b.
[65] S. Haakonsson, M.A. Rodríguez, C. Carballo, M. Pérez, C. del, R. Arocena, S. Bonilla, Predicting cyanobacterial biovolume from water temperature and conductivity using a Bayesian compound Poisson-Gamma model, Water Res. 176 (2020), 115710, https://doi.org/10.1016/j.watres.2020.115710.
[66] J. Bobbin, F. Recknagel, Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms, Ecol. Model. 146 (2001) 253–262, https://doi.org/10.1016/S0304-3800(01)00311-8.
[67] T. Cordier, P. Esling, F. Lejzerowicz, J. Visco, A. Ouadahi, C. Martins, T. Cedhagen, J. Pawlowski, Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning, Environ. Sci. Technol. 51 (2017) 9118–9126, https://doi.org/10.1021/acs.est.7b01518.
[68] Z. Zhang, Q. Zhang, T. Wang, N. Xu, T. Lu, W. Hong, J. Penuelas, M. Gillings, M. Wang, W. Gao, H. Qian, Assessment of global health risk of antibiotic resistance genes, Nat. Commun. 13 (2022) 1553, https://doi.org/10.1038/s41467-022-29283-8.
[69] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2009.
[70] Y. Cha, S.S. Park, K. Kim, M. Byeon, C.A. Stow, Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model, Water Resour. Res. (2014) 2518–2532, https://doi.org/10.1002/2013WR014372.
[71] H.W. Paerl, V.J. Paul, Climate change: links to global expansion of harmful cyanobacteria, Water Res. 46 (2012) 1349–1363, https://doi.org/10.1016/j.watres.2011.08.002.
[72] Y. Grund, Y. Pan, M. Rosenkranz, E. Foster, Long-term phosphorus reduction and phytoplankton responses in an urban lake (USA), Water Biol. Syst. 1 (2022), 100010, https://doi.org/10.1016/j.watbs.2022.100010.
[73] H. Xu, H.W. Paerl, B. Qin, G. Zhu, G. Gao, Nitrogen and phosphorus inputs control phytoplankton growth in eutrophic Lake Taihu, China, Limnol. Oceanogr. 55 (2010) 420–432, https://doi.org/10.4319/lo.2010.55.1.0420.
[74] C.T. Filstrup, J.A. Downing, Relationship of chlorophyll to phosphorus and nitrogen in nutrient-rich lakes, Inl. Waters 7 (2017) 385–400, https://doi.org/10.1080/20442041.2017.1375176.
[75] J.A. Downing, E. McCauley, The nitrogen : phosphorus relationship in lakes, Limnol. Oceanogr. 37 (1992) 936–945, https://doi.org/10.4319/lo.1992.37.5.0936.
[76] H. Cyr, S.K. McCabe, G.K. Nürnberg, Phosphorus sorption experiments and the potential for internal phosphorus loading in littoral areas of a stratified lake, Water Res. 43 (2009) 1654–1666, https://doi.org/10.1016/j.watres.2008.12.050.
[77] I.-I. Lin, Typhoon-induced phytoplankton blooms and primary productivity increase in the western North Pacific subtropical ocean, J. Geophys. Res. Ocean. 117 (2012), https://doi.org/10.1029/2011JC007626 n/a-n/a.
[78] B.J. Robson, D.P. Hamilton, Summer flow event induces a cyanobacterial bloom in a seasonal Western Australian estuary, Mar. Freshw. Res. 54 (2003) 139–151, https://doi.org/10.1071/MF02090.
[79] D. Yue, Y. Peng, X. Qian, L. Xiao, Spatial and seasonal patterns of size-fractionated phytoplankton growth in Lake Taihu, J. Plankton Res. 36 (2014) 709–721, https://doi.org/10.1093/plankt/fbt131.
[80] G.K. Nürnberg, B.D. LaZerte, Modeling the effect of development on internal phosphorus load in nutrient-poor lakes, Water Resour. Res. 40 (2004) 1–9, https://doi.org/10.1029/2003WR002410.