AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology

Environmental Microbiology | Full-Length Text

Potential health risk assessment of cyanobacteria across global lakes

Hang Hu,[1] Zhenyan Zhang,[1] Bingfeng Chen,[1] Qi Zhang,[2,3] Nuohan Xu,[2,3] Hans W. Paerl,[4] Tingzhang Wang,[5] Wenjie Hong,[5] Josep Penuelas,[6,7] Haifeng Qian[1]

AUTHOR AFFILIATIONS See affiliation list on p. 14.

ABSTRACT Cyanobacterial blooms pose environmental and health risks due to their production of toxic secondary metabolites. While current methods for assessing these risks have focused primarily on bloom frequency and intensity, the lack of comprehensive and comparable data on cyanotoxins makes it challenging to rigorously evaluate these health risks. In this study, we examined 750 metagenomic data sets collected from 103 lakes worldwide. Our analysis unveiled the diverse distributions of cyanobacterial communities and the genes responsible for cyanotoxin production across the globe. Our approach involved the integration of cyanobacterial biomass, the biosynthetic potential of cyanotoxin, and the potential effects of these toxins to establish potential cyanobacterial health risks. Our findings revealed that nearly half of the lakes assessed posed medium to high health risks associated with cyanobacteria. The regions of greatest concern were East Asia and South Asia, particularly in developing countries experiencing rapid industrialization and urbanization. Using machine learning techniques, we mapped potential cyanobacterial health risks in lakes worldwide. The model results revealed a positive correlation between potential cyanobacterial health risks and factors such as temperature, N_2O emissions, and the human influence index. These findings underscore the influence of these variables on the proliferation of cyanobacterial blooms and associated risks. By introducing a novel quantitative method for monitoring potential cyanobacterial health risks on a global scale, our study contributes to the assessment and management of one of the most pressing threats to both aquatic ecosystems and human health.

IMPORTANCE Our research introduces a novel and comprehensive approach to potential cyanobacterial health risk assessment, offering insights into risk from a toxicity perspective. The distinct geographical variations in cyanobacterial communities coupled with the intricate interplay of environmental factors underscore the complexity of managing cyanobacterial blooms at a global scale. Our systematic and targeted cyanobacterial surveillance enables a worldwide assessment of cyanobacteria-based potential health risks, providing an early warning system.

KEYWORDS lake, metagenome, cyanobacterial bloom, cyanotoxins, risk assessment, machine learning

Eutrophication, elevated CO_2 levels, and global warming are contributing to the widespread occurrence of potentially toxic cyanobacterial blooms (1–3). The frequency of global cyanobacterial blooms has been estimated to have increased by 44% from the 2000s to the 2010s (4, 5). The ongoing expansion of cyanobacteria blooms poses an escalating threat to aquatic ecosystems and human health.

Cyanobacteria are common phytoplankton constituents in aquatic ecosystems and play vital roles in elemental cycling and energy transformation (6, 7). Additionally, they

Editor John R. Spear, Colorado School of Mines, Golden, Colorado, USA

Address correspondence to Haifeng Qian, hfqian@zjut.edu.cn.

The authors declare no conflict of interest.

See the funding table on p. 14.

Received 3 October 2024
Accepted 9 October 2024
Published 4 November 2024

can significantly alter the community structure and diversity of plankton communities, particularly in water bodies undergoing eutrophication (6, 8). Predominant genera responsible for harmful cyanobacterial blooms include *Aphanizomenon*, *Cylindrospermopsis*, *Dolichospermum*, *Microcystis*, *Nodularia*, and *Planktothrix* (6), many of which thrive at temperatures >25°C (9, 10). Many of these genera can accumulate on the water surface via buoyancy, which is mediated by gas vesicles, thereby shading subsurface competing algal species and promoting cyanobacterial dominance (8, 11).

Cyanobacterial blooms can degrade water quality in freshwater ecosystems by producing various toxic secondary metabolites or cyanotoxins (12), including neurotoxins, hepatotoxins, cytotoxins, and dermatotoxins (13). These toxins have been detected globally, with liver-toxic microcystins frequently identified in 40%–75% of cyanobacterial blooms (14). Human exposure to cyanotoxins can occur through the consumption of cyanobacteria-based foods and contaminated drinking water (15, 16). Additional exposure routes include dermal contact or aerosol exposure during recreational activities in waters affected by toxic blooms (17). Owing to the health risks associated with cyanotoxins, the WHO has established threshold values for a few cyanotoxins (e.g., a limit of 1 μg/L for microcystin-LR [MC-LR]) in drinking water on the basis of toxicological data (18). However, thresholds have not been established for multiple cyanotoxins (8).

Over the past few decades, monitoring techniques for cyanobacterial blooms and their health risk assessment have evolved. Current assessment methods include reporting cyanobacterial bloom frequency and magnitude via remote sensing (19, 20). Remote sensing may detect and quantify cyanobacterial blooms on the basis of chlorophyll-a content, whereas low chlorophyll-a concentrations can make accurate measurements challenging (21). Another limitation of current assessment methods is their failure to differentiate cyanotoxin production and diverse toxic effects, which enhances human health risks (13).

Metagenomics enhances the comprehensiveness and accuracy of studies on the structure and dynamics of ecosystems (22), reducing the uncertainties of remote sensing. Several studies have been conducted using metagenomics to characterize the ecological risks of cyanobacterial blooms, including antibiotic resistance carried by cyanobacterial blooms (23, 24). Some studies have employed metagenomic techniques to identify cyanotoxin genes (25, 26); however, few studies have comprehensively assessed the health risks associated with cyanobacteria and their secreted cyanotoxins by integrating multiple relevant indicators.

In this study, we analyzed 750 metagenomes from samples from lakes worldwide to depict the distribution of cyanobacteria and cyanotoxin biosynthesis genes. We then devised a novel framework to calculate potential cyanobacterial health risks in lakes by integrating cyanobacterial biomass, cyanotoxin biosynthesis genes, and cyanotoxin effects. Using machine learning, we mapped potential cyanobacterial health risks in lakes across the globe and identified contributing risk factors. Our study introduces a new approach to assess potential cyanobacterial health risks and identifies priority areas for cyanobacterial health risk management.

MATERIALS AND METHODS

Collection of 750 metagenomes for samples

We conducted a literature search via the keywords “Lake,” “Waterbody,” and “Metagenome” on Web of Science, Google Scholar, and PubMed. Our study encompassed a total of 750 metagenomes for samples retrieved from 103 lakes across five continents and 19 countries. These data were sourced from the National Center for Biotechnology Information (NCBI) SRA database. There is no uniform definition for cyanobacterial blooms. Nearly 70% of the samples in our data were collected during summer (see Table S1). Summer is generally considered the peak season for cyanobacterial blooms. Notably, in certain lakes and under specific climatic conditions, blooms may not be restricted to the summer months. During data collection, we adhered to specific criteria: (i) inclusion of samples from lake water columns; (ii) provision of accurate coordinates and lake names; (iii) exclusion of samples from systems treated with chemical compounds; and (iv) avoidance of sampling times coinciding with significant severe climate events such as heavy precipitation and strong winds.

Annotation and abundance calculation of taxonomy and cyanotoxin biosynthesis genes

We preprocessed raw metagenomic data via FastQC (v0.11.5; https://github.com/s-andrews/FastQC) for data quality assessment, followed by trimming and quality filtering via Trimmomatic (v0.36) (27). Taxonomic annotation was accomplished through Kraken (v2.1.2) and the Nonredundant Protein Sequence Database (NR) at the phylum level (see Table S2). We focused on five cyanotoxins, namely, anatoxin, cylindrospermopsin, microcystin, nodularin, and saxitoxin. By conducting a literature search, we compiled a list of cyanotoxin biosynthesis genes (listed in Table S3) and acquired the corresponding protein sequences from the NCBI protein database. We established a comprehensive database encompassing genes related to the five classes of cyanotoxin biosynthesis. The BWA (v0.7.13) tool enabled the annotation of clean data, and unmapped reads were removed via SAMtools. The abundance of cyanotoxin biosynthesis genes was calculated as reads per kilobase per million mapped reads (RPKM). We characterized the global distribution of cyanotoxins in lakes by integrating the corresponding biosynthetic genes.

Defining potential cyanobacterial health risk

Acknowledging that not all cyanobacteria are toxin producers, we recognized that cyanobacterial biomass does not necessarily correspond to cyanobacterial toxicity. Consequently, we evaluated potential cyanobacterial health risks via three criteria: cyanobacterial biomass, cyanotoxin biosynthesis genes, and cyanotoxin toxicity. Cyanobacterial biomass was characterized by the relative abundance of cyanobacteria within bacterial communities at the phylum level. Toxicity indices were derived from the acute toxic effects of cyanotoxins, with higher median lethal dose (LD50) values indicating lower toxicity and lower health risk, and vice versa (28). We calculated the reciprocal of the LD50 values of the cyanotoxins. By taking the reciprocal, the resulting values are directly proportional to the level of health risk. These reciprocal values were then normalized to a 0–100 scale, with saxitoxin, the most toxic cyanotoxin, as the reference. This approach enables clear comparisons of relative toxicity levels. Notably, the toxicity index calculated on the basis of the reciprocal of acute toxicity refers to relative toxicity. We defined the toxicity index of saxitoxin as 100, which represents the highest acute toxicity of the five cyanotoxins (Table S6). The toxicity indices of the other cyanotoxins were standardized to equal proportions. We calculated the potential cyanobacterial health risk index for each sample. The cyanobacterial health risk index (RI) was then computed as follows:

$$\mathrm{RI} = \text{Cyanobacterial biomass} \times \sum_{i=1}^{n} \text{Abundance}_{\text{cyanotoxin biosynthesis gene}} \times \text{Toxicity index}$$

where n was the number of genes regulating cyanotoxin synthesis. $\text{Abundance}_{\text{cyanotoxin biosynthesis gene}}$ was the abundance of the cyanotoxin biosynthesis gene in each sample. For example, we summarized the abundance of 10 genes containing *mcyA-J* to characterize the biosynthetic potential of microcystins.

Experimental design

To explore the relationship between cyanotoxin biosynthesis genes and cyanotoxin concentrations, we conducted laboratory simulations of varying cyanobacterial bloom levels. For this purpose, we selected Meiliang Bay, situated in the northern part of Lake Taihu, China's third largest lake; this bay experiences eutrophication and toxic cyanobacterial blooms (29). On 28 June 2023, we collected water samples from two sites (31°31′16″N, 120°13′51″E and 31°31′57″N, 120°11′12″E) at a depth of 0.5 m. Subsequently, we added cultured *Microcystis aeruginosa* (FACHB-905) to untreated Lake Taihu water at various concentrations of the culture to simulate different levels of cyanobacterial blooms, creating concentrations of 2, 3, 4, and 5 mL for *M. aeruginosa* culture and 3, 2, 1, and 0 mL of sterile water in 200 mL of lake water, in triplicate, for each treatment. To ensure a concentration gradient of *M. aeruginosa* among the treatment groups, *M. aeruginosa* cultures were added from the same culture bottle for each treatment. We shook the cultures before dispensing to ensure homogeneity. After addition, we measured the cell density at 680 nm via a spectrophotometer to determine the biomass gradient. The *M. aeruginosa* strain was procured from the Institute of Hydrobiology, the Chinese Academy of Sciences (Wuhan, China). The cultures were maintained under controlled conditions of 25 ± 0.5°C with a fluorescent light intensity of 46 μmol $m^{-2}s^{-1}$ and a 12-hour light/12-hour dark cycle. The cultures were manually agitated three times daily to ensure uniform cyanobacterial distribution during the 5-day incubation period.

DNA was extracted from the mixture via 0.2-μm polycarbonate filter membranes and a SPINeasy DNA kit for soil (MP Biomedicals LLC, Ohio, USA). DNA concentration and quality were assessed via spectrophotometric analysis (30), with the extracted DNA stored at −20°C for subsequent analysis. High-throughput quantitative PCR of microcystin biosynthesis genes was conducted via the StepOnePlus Real-time PCR system. We utilized a total of 10 primer sets targeting 10 microcystin biosynthesis genes (*mcyA-J*) and the 16S rRNA gene as a housekeeping gene for the relative quantification of 10 genes. The reaction mixture was heated for 10 minutes at 95°C, followed by 40 cycles of 1 minute at 60°C and 20 seconds at 72°C. The relative gene copy number was calculated according to the methodology proposed by Zhu et al. (31).

Microcystin-LR analysis was performed via an Agilent 1290 Infinity II high-performance liquid chromatograph coupled with a 6540 quadrupole time-of-flight mass spectrometry system. A C18 column (1.8 μm, 50 mm × 2.1 mm) was employed at a temperature of 30°C, and microcystin-LR concentrations were measured via tandem mass spectrometry. Electrospray ionization was used; the scanning mode was positive ion mode; the drying gas temperature was 350°C at a flow rate of 10 L/min; the capillary voltage was 4,000 V; and the monitoring mode was multiple reaction monitoring.

Machine learning algorithms for predicting and controlling potential cyanobacterial health risk

We used a geographic information system to derive information on climate change and anthropogenic activities across all lakes (see Table S4). The variable inflation factor (VIF) of the independent variables was calculated via the R package “car,” resulting in the selection of 23 independent variables with VIF values below 10 (32). This step minimized the impact of multicollinearity from independent variables on the predictive model. The machine learning model was developed alongside the risk index via different methods, including four linear regression types and four nonlinear regression types. Linear regression encompassed models with and without stepwise selection, as well as models such as least angle regression (33) and elastic net (34). Nonlinear regression approaches include random forest (RF) (35), the boosted tree model (36), the model bagged tree (37), and the cubist model (38). Model performance and fit were evaluated through 10-fold cross-validation. The original data set was divided into 10 equal subsets, with nine serving as training sets and one serving as the test set, generating 10 results to assess algorithm accuracy (39, 40). The RF model was selected as the final prediction model because of its superior accuracy.

To quantify the contribution of environmental drivers to potential cyanobacterial health risk, we employed the R package “rfPermute” to calculate the increase in the mean squared error. By randomly assigning values to predictor variables, we identified variables of greater importance as those that, when replaced with random values, led to a greater increase in model prediction error. Moreover, we conducted a partial dependence analysis via the R package “pdp” to elucidate the impact of each independent factor on the predicted risk within the range of possible values indicated by the RF model. The partial dependence plot allowed visualization of the average partial relationship between the predicted response and one or more features (41).

Global map of potential cyanobacterial health risks in lakes

We gathered coordinates and shoreline polygons of approximately 1.4 million lakes worldwide from HydroLAKES V1.0 (https://www.hydrosheds.org/products/hydrolakes). Smaller lakes, which are more influenced by multiple environmental drivers and tend to exhibit less stable cyanobacterial blooms, were excluded if their size was less than 1 km^2. Additionally, lakes situated at high latitudes were omitted due to insufficient high-latitude lake samples were available for model development. Following these criteria, we identified 73,030 lakes globally (see Table S5). The environmental factors for each lake were extracted on the basis of coordinates, and a global prediction of potential cyanobacterial health risk for lakes was generated via an RF algorithm. The k-means method (42) was employed to categorize the risk values of the 73,030 lakes into 10 ranks, ranging from the highest risk (rank 10) to the lowest risk (rank 1). After the risk for each lake was determined, the data were visualized via ArcGIS (v10.8) to represent the global distribution of cyanobacterial risk in the lakes.

Statistical analysis

The analyses were primarily conducted via R version 4.1.1 (R Foundation for Statistical Computing) with relevant software packages. Alpha diversity metrics, including the Shannon index and richness, were calculated via the “vegan” and “picante” packages at the genus level for cyanobacterial communities, with additional assessment of the diversity of cyanotoxin biosynthesis genes. Significant differences ($P < 0.05$) were identified via Kruskal−Wallis tests implemented in IBM SPSS Statistics (v20.0.0). Nonmetric multidimensional scaling (NMDS) based on Bray−Curtis distances was generated via the R “ggplot2” package. Regression and stacking analyses were performed via the same package. Bar and line graphs were generated via GraphPad Prism 8 and Origin 2021.

RESULTS

Global distribution of cyanobacteria in lakes

On the basis of the metagenomic annotation results, cyanobacteria were the third-largest taxonomic group of bacteria in the lake habitat, accounting for 9.50% of bacteria (Fig. S1). Cyanobacterial abundance in the US-Canada Great Lakes, East Asia, and South Asia reached the highest level (Fig. 1a). Although the cyanobacterial community composition in global lakes at the order level was mainly composed of *Synechococcales*, *Oscillatoriales*, *Gloeobacterales*, *Nostocales*, and *Chroococcales*, they displayed geographically specific patterns. For example, *Chroococcales*, *Nostocales*, and *Synechococcales* were in highest abundance in Southeastern Asia, North America, and South America, respectively (Fig. 1b). *Microcystaceae* was predominant (>70%) in the Asian region. Among the US-Canada Great Lakes, we found that only Lake Erie had a higher abundance of *Microcystaceae*, while the rest were dominated by *Synechococcaceae* (Fig. S2).

We calculated the alpha diversity of cyanobacteria in each sample and found that it varied on different continents. The highest alpha diversity of cyanobacteria was detected in the European region (Fig. S3a and b). The similarity of cyanobacterial composition in the 750 samples was evaluated via NMDS, which revealed that the cyanobacterial community structure differed between continents ($P < 0.001$, $R^2 = 0.07$) and between countries ($P < 0.001$, $R^2 = 0.15$) (Fig. S3c and d). The geographic characteristics of beta diversity were consistent with the geographic variation in cyanobacterial composition and alpha diversity (Fig. S2 and S3).

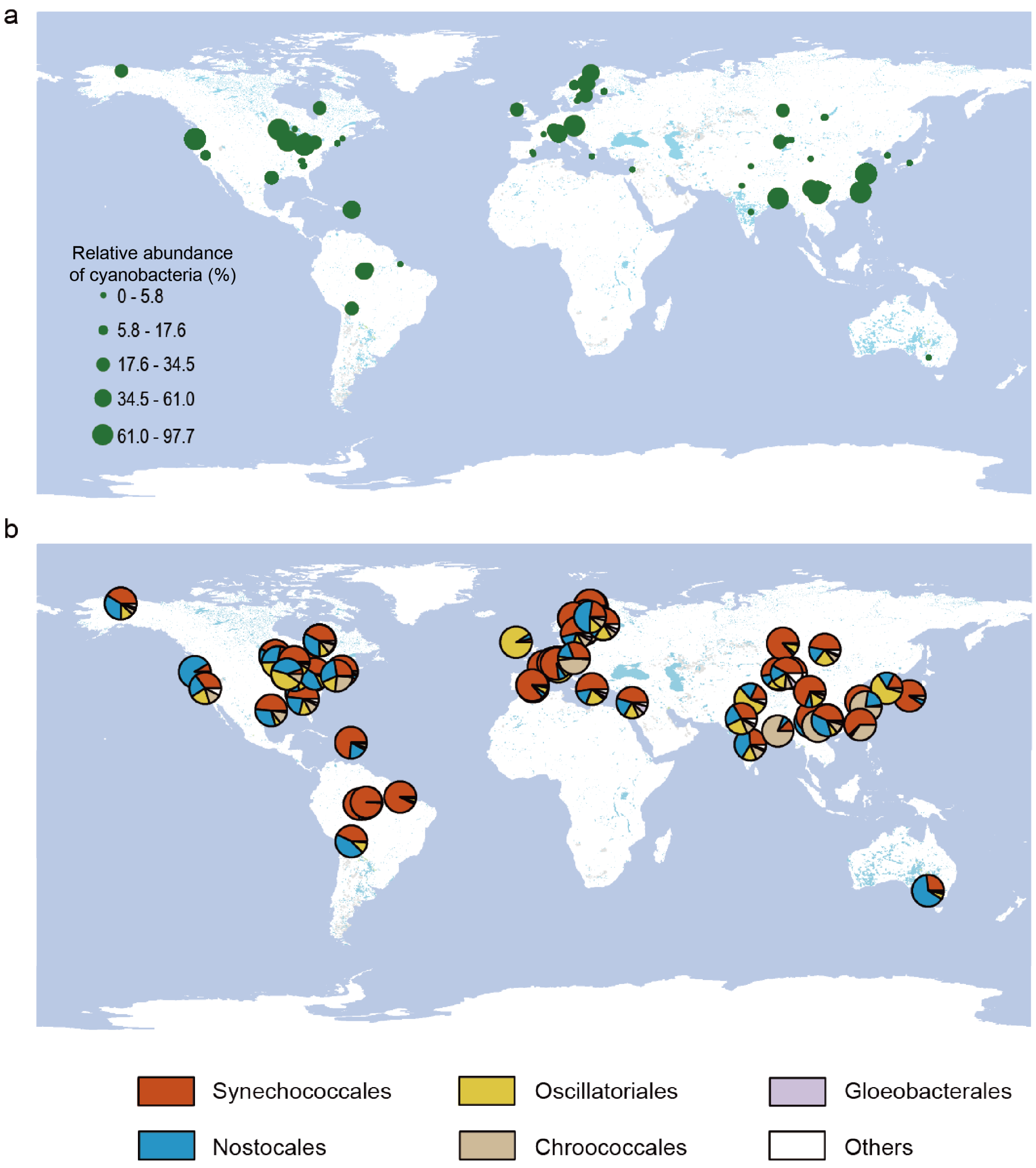

FIG 1 Cyanobacterial distribution patterns in lakes worldwide. (a) Map of the relative abundance of cyanobacteria in all samples. (b) Global map of prominent cyanobacterial orders and their relative abundance in lakes. Base maps are from the default world hydrography map provided by ArcGIS v10.8 software.

Global distribution of cyanotoxins in lakes

The distribution of the abundance of cyanotoxin biosynthesis genes among various lakes was similar to the relative abundance of cyanobacteria. The US-Canada Great Lakes region and Southeast Asia were potentially the areas with the highest cyanotoxin production (Fig. 2a). The global composition of cyanotoxin biosynthesis genes in lakes

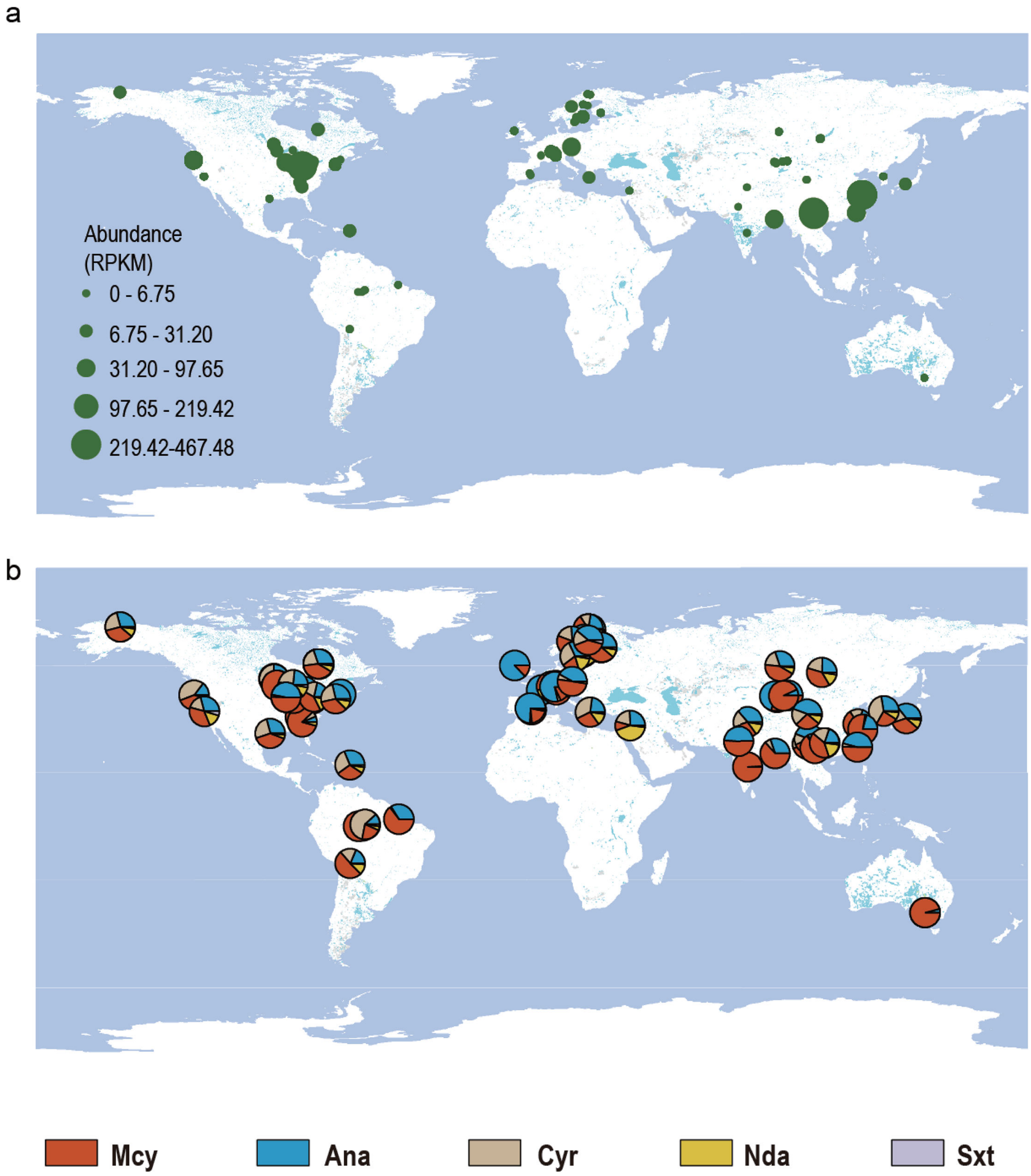

FIG 2 Global distribution patterns of cyanotoxin biosynthesis genes in lakes. (a) Map of the abundance of cyanotoxin biosynthesis genes (RPKMs). (b) Global map of cyanotoxin biosynthesis genes in lakes. The *mcy*, *ana*, *cyr*, *nda*, and *sxt* genes encode the biosynthesis of Microcystin (Mcy), Anatoxin (Ana), Cylindrospermopsin (Cyr), Nodularin (Nda), and Saxitoxin (Sxt), respectively. Alpha diversities of the cyanotoxin biosynthesis genes in all the samples. Base maps are from the default world hydrography map provided by ArcGIS v10.8 software.

displayed significant geographic variation, whereas *mcy* and *ana*, which are responsible for the biosynthesis of microcystin and anatoxin, respectively, were the dominant cyanotoxin biosynthesis genes in lakes worldwide. The *cyr* gene (encoding a gene for cylindrospermopsin biosynthesis) was more abundant in North America than in other regions (Fig. 2b; Fig. S4).

The alpha diversity of cyanotoxin biosynthetic genes differed in lakes from different continents. For example, cyanotoxin biosynthetic genes from European samples presented greater alpha diversity than those from other continents did (Fig. S5a and b). The similarity of cyanotoxin synthesis genes in the 750 samples was evaluated via NMDS, which revealed that the composition of cyanotoxin synthesis in the lakes also differed between continents ($P < 0.001$, $R^2 = 0.14$) and between countries ($P < 0.001$, $R^2 = 0.31$) (Fig. S5c and d).

Potential cyanobacterial health risk assessment framework

The toxicity of different cyanotoxins on the basis of the intraperitoneal acute LD50 varied considerably, with a minimum toxicity index of 0.37 for cylindrospermopsin and a maximum toxicity index of 100 for saxitoxin (Fig. S6). To develop a comprehensive framework for potential cyanobacterial health risk assessment, we integrated the relative abundance of cyanobacteria, the abundance of cyanotoxin synthesis genes, and the toxicity index of each cyanotoxin (Table S6; details can be found in the Methods). Our experimental results confirmed that the abundance of cyanotoxin biosynthesis genes was significantly positively correlated with cyanotoxin concentration; thus, it is feasible to characterize potential cyanotoxin concentrations in terms of cyanotoxin biosynthesis gene abundance (Fig. 3b). Importantly, while our cultures expressed microcystin and presented this correlation, it may not hold true for all cyanotoxins across different natural environments. This limitation should be considered when interpreting the outputs of our algorithm, as the dynamics of cyanotoxin production can vary significantly in diverse ecological contexts. The risk map revealed East Asia and South Asia as the areas with the highest risk (Fig. 3a). The relative abundance of *Synechococcaceae* was greater in low-risk samples than in high-risk samples, whereas *Microcystaceae* was significantly enriched in high-risk samples (Fig. 3c), and the composition of cyanotoxin biosynthesis genes was similar across risk ranks (Fig. S7). To characterize the risk rank visually, we discretized the data via the k-means method and then classified the samples into 10 ranks on the basis of risk (rank 10 for the highest risk and rank 1 for the lowest risk). The risk rank here represented relative risk. In total, 618 (82.4%) of the 750 samples were ranked first, and the other 17.6% had a higher risk (ranked 2 to 10).

Predictors of potential cyanobacterial health risks

We ultimately obtained 23 factors with a VIF less than 10 to construct a machine learning model together with the risk index (Fig. S8). To construct the best model for predicting cyanobacterial health risk, regression modeling was conducted, including four different linear and four different nonlinear regression types (Fig. S9). Tenfold cross-validation revealed that the RF model explained the highest percentage ($R^2 = 0.74$), indicating a good prediction of cyanobacterial health risk (Fig. 4b; Fig. S9). We quantified the contributions of environmental factors to potential cyanobacterial health risks (Fig. 4a). The results indicated that temperature was the most important predictor in the model, followed by the human influence index and N_2O emissions (Table S7). Extreme climatic conditions and anthropogenic activities can greatly increase the potential health risk of cyanobacteria (Fig. 4c; Fig. S10). For example, the wind speed and mean temperature of the wettest quarter were strongly positively correlated with the risk index at high levels. Similar trends were observed for agricultural yield and N_2O emissions (Fig. 4c). In addition, we found that climatic factors exhibited more complex patterns. The extent of the impact of the climate parameters varied considerably over a range of values (Fig. S10).

A global map of potential cyanobacterial health risks in lakes

The predicted results were discretized into the aforementioned 10 ranks. On the basis of the predictions and discretization results, we mapped the potential cyanobacterial health risk in 73,030 lakes across the globe (Fig. 5a). East Asia, South Asia, and southern

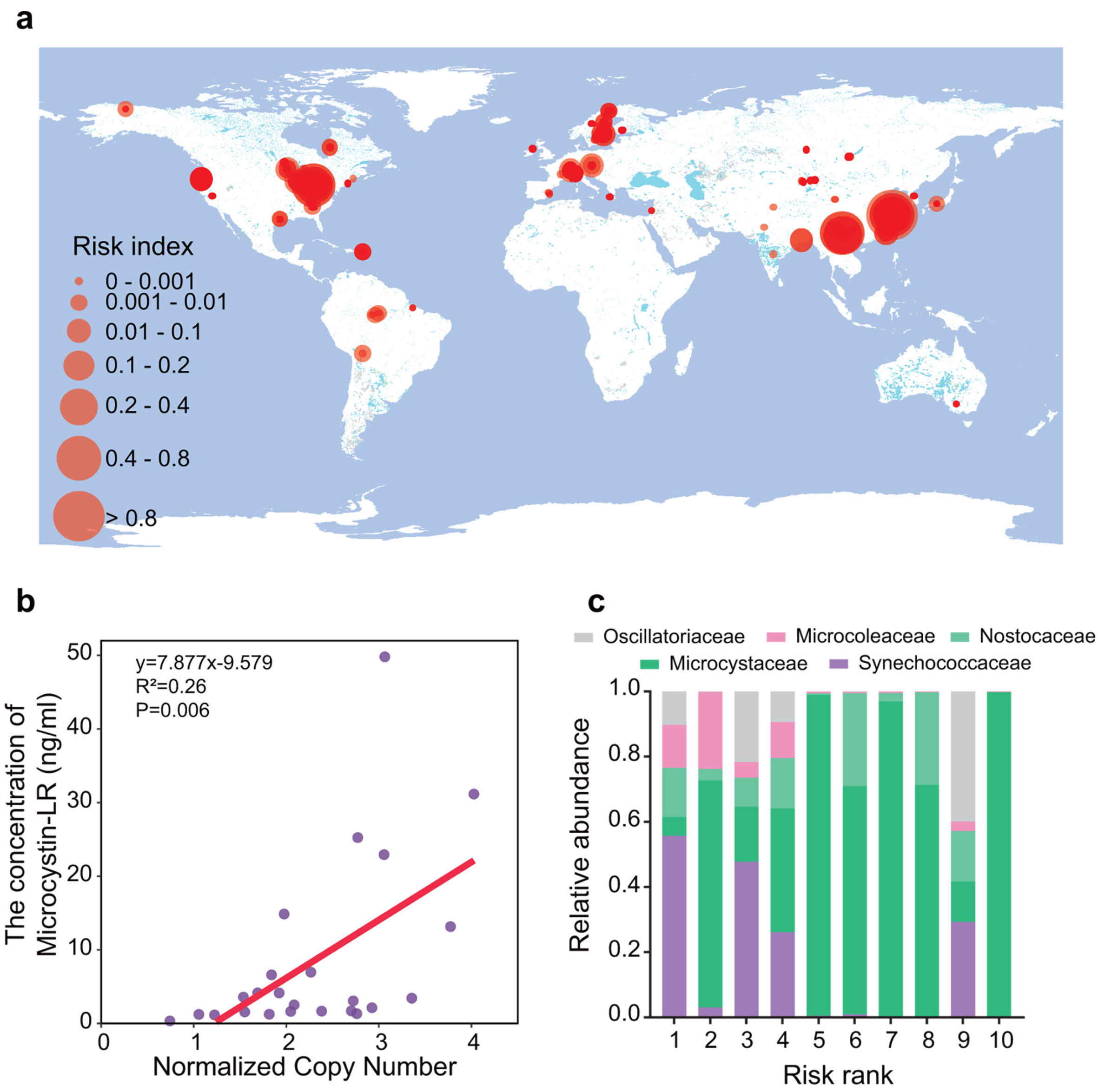

FIG 3 Characterization of potential cyanobacterial health risks for all samples. (a) Potential cyanobacterial health risk map for 750 samples. $\mathrm{RI} = \text{Cyanobacterial biomass} \times \sum_{i=1}^{n} \text{Abundance}_{\text{cyanotoxin biosynthesis gene}} \times \text{Toxicity index}$. The base map is from the default world hydrography map provided by ArcGIS v10.8 software. (b) Linear regression of the normalized copy number (sum of the microcystin biosynthesis genes *mcyA-J*) and the concentration of microcystin-LR. (c) The five cyanobacterial families with the highest average abundance of all samples. The risk index was discretized via the k-means method and was divided into 10 ranks according to the size of the risk index.

Africa were the areas with the highest potential cyanobacterial health risk ranks. North America and Western Europe, the most economically developed regions, presented the lowest potential cyanobacterial health risk. The potential cyanobacterial health risk along the southeast coast of Australia was lower than that in the rest of Australia. We further divided the prediction results into low- (ranks 1, 2, and 3), medium- (ranks 4, 5, and 6), and high-risk (ranks 7, 8, 9, and 10) lakes, accounting for 51.29%, 37.88%, and 10.82% of the global lakes, respectively (Fig. 5b). In terms of risk distribution by continent, Europe presented the lowest potential cyanobacterial health risk, with more than 60% of the lakes at low risk. Africa was the highest risk region, with nearly half of the lakes being at high risk (Fig. 5c).

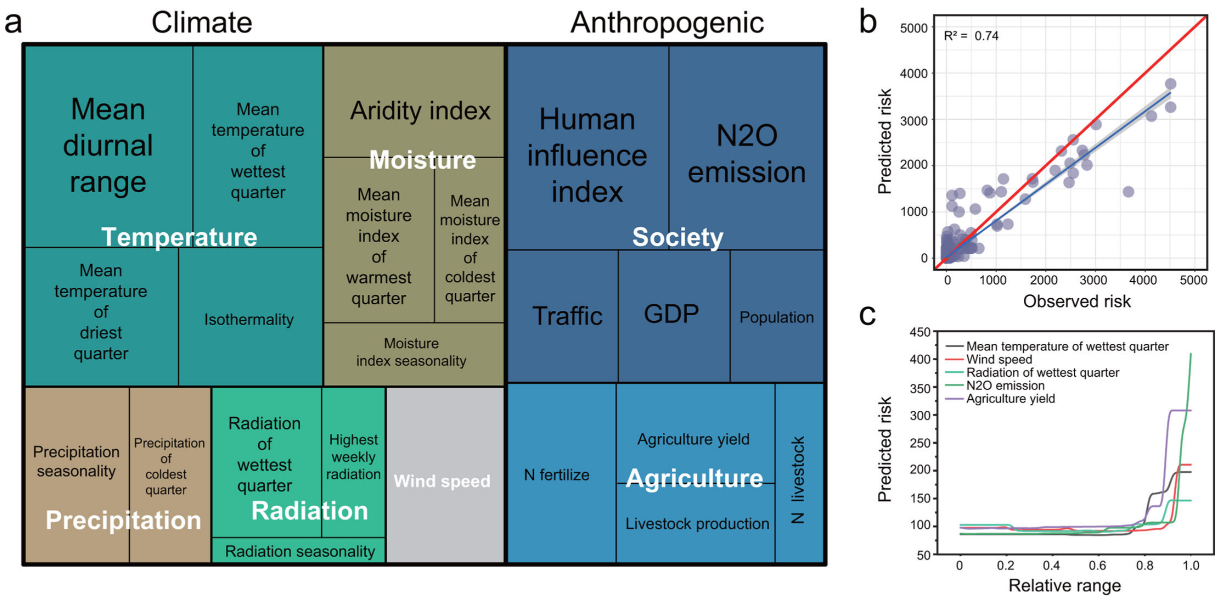

FIG 4 Predictors of potential cyanobacterial health risks. (a) Importance of different factors for potential cyanobacterial health risk. The larger the grid square is, the more important the factor. (b) Performance of random forest models in predicting potential cyanobacterial health risks. The slope of the red line is 1. The closer the fitted straight line is to y = x, the better the fit of the model. The dots represent the samples used by the random forest models. (c) Partial dependence diagram illustrating the effects of various factors on the potential health risk of cyanobacteria.

DISCUSSION

We identified and quantified cyanobacterial communities and cyanotoxin biosynthesis genes in 750 metagenomes from samples from lakes worldwide. Our findings revealed distinct geographical patterns in both cyanobacterial communities and cyanotoxin biosynthesis genes. The dominance of particular cyanobacterial species varies significantly across different regions, presenting a substantial challenge for the management of cyanobacterial-infested water bodies (8, 43). Notably, the prevalence of cyanobacteria in the US-Canada Great Lakes region and Southeast Asia aligns with previous remote sensing-based monitoring results of cyanobacterial blooms (5). Furthermore, cyanotoxin biosynthesis genes are widespread in these regions. Some studies highlight the use of remote sensing for monitoring cyanobacterial blooms (8, 44–47). However, genome-based assessments of lake environmental toxicity serve as a crucial complement to these methods. Utilization of metagenomic data enhances the precision of cyanobacterial community characterization, circumventing the uncertainties associated with remote sensing techniques in identifying phytoplankton (44), including the differentiation of dominant cyanobacteria across different regions (45, 46). Our approach enhances cyanobacterial toxicity evaluation beyond satellite monitoring, offering more detailed insights that can improve early warning systems for lake cyanobacterial pollution.

The ongoing global surveillance of cyanobacterial blooms (18) relies primarily on assessing bloom frequency and area, which does not convey the full extent of potential cyanobacterial health risk (48–50). Additionally, the abundance of cyanobacteria may not necessarily correlate with their toxicity, given that not all cyanobacterial taxa produce toxins (28). Our study introduces a comprehensive framework that evaluates cyanobacteria by linking common cyanotoxin toxic effects with cyanotoxin biosynthesis genes, offering a fresh perspective on potential cyanobacterial health risk assessment. Intriguingly, despite the high cyanobacterial abundance and cyanotoxin gene presence in North America, this region did not present elevated potential cyanobacterial health risks. This is attributable to the high abundance of *cyr* genes in North America, which corresponds to the lowest toxicity index (cylindrospermopsin, synthesized by *cyr* genes), contributing to the overall low toxicity of North American lakes. Importantly, however,

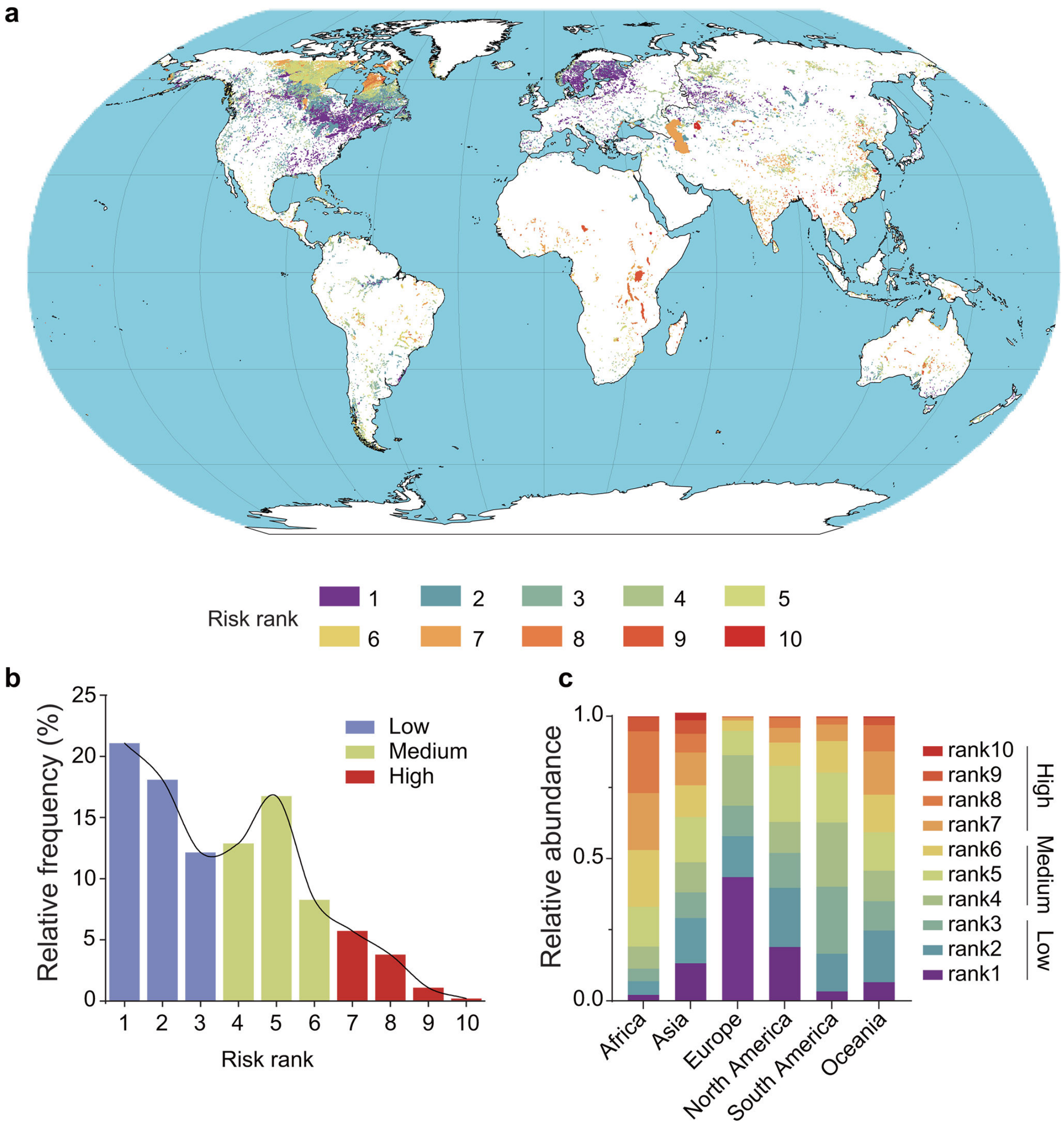

FIG 5 A global map of potential cyanobacterial health risks in lakes. (a) Potential cyanobacterial health risk maps for 73,030 lakes worldwide via machine learning model predictions; the coordinates and shoreline polygons of the lakes were obtained from HydroLAKES V1.0. The base map is from the default world hydrography map provided by ArcGIS v10.8 software. (b) Global distribution of potential cyanobacterial health risk in lakes. We divided the 10 risk levels into three categories: low, medium, and high. (c) Composition of potential cyanobacterial health risk ranks in lakes by continent.

our assessment system relies on the qualification of five major classes of common cyanotoxins. There are substantial differences in toxicity between structural variants within the same class of cyanotoxins, so we characterized the toxic effects of the five major classes of cyanotoxins as the mean values. More accurate and comprehensive information on cyanotoxins would further strengthen the capabilities of our assessment system. Furthermore, our assessment represents only potential risks, as the cyanotoxin concentration is influenced by gene abundance and expression. Our experiments confirmed a positive correlation between MC-LR concentrations and the relative gene copy number of *mcy*. A previous study also revealed a positive correlation between cyanotoxin concentrations measured via enzyme-linked immunosorbent assay and the copy number of related cyanotoxin biosynthesis genes (51). However, the existence of this positive correlation in the biosynthesis of all cyanotoxins remains to be demonstrated, and the increase in such studies in the future will greatly optimize our assessment framework. Moreover, we solely considered potential health risks linked to cyanobacterial toxicity, overlooking ecological risks tied to community structure changes, such as hypoxia potential and disruption of food webs.

In this study, we classified lakes into larger regions (e.g., Asia and North America) to increase the sample size and statistical power. However, this broad classification may obscure ecological differences across varying elevations and latitudes. Future research should analyze the influences of environmental factors on lake ecosystems and cyanotoxin distributions at a finer geographical scale. After completing the potential cyanobacterial health risk assessment, we delved into the potential determinants of cyanobacterial health risk. Although the factors driving potential cyanobacterial health risk are multifaceted, our machine learning model identified temperature as the most pivotal factor. Warm climates and human activities have emerged as major contributors to cyanobacterial toxicity, which aligns with findings by Paerl and Paul (2). The human influence index, the second most influential factor in our model, offers a quantitative measure of human activity impact encompassing population pressure, land use, and infrastructure distribution (52). Potential cyanobacterial health risks exhibited heightened levels in regions marked by pronounced anthropogenic activities, such as the Indian Peninsula and Eastern China. However, compared with anthropogenic activities, climatic factors exhibited complex and more dominant patterns in the predictive model. The contributions of these factors to potential cyanobacterial health risk underscore the necessity of robust human interventions to counteract climatic effects. As anthropogenic activities such as fossil fuel combustion significantly influence climate change, they present a more manageable target than the manipulation of climate conditions does (53). Previous research on China's Lake Taihu revealed that a 46.3% reduction in phosphorus could alleviate the risk of extensive cyanobacterial blooms (48). Furthermore, our study underscores that extreme climatic conditions (elevated temperatures) significantly increase potential cyanobacterial health risk, echoing findings that climate extremes fuel cyanobacterial blooms (54–57). In summary, the management of cyanobacterial blooms is increasingly complex in the face of more frequent extreme weather events (58). However, we did not consider nutrients such as P or N when exploring potential determinants of potential cyanobacterial health risk, even though they are known to be decisive factors in cyanobacterial blooms. This is because few studies have uploaded complete water quality parameters and different methods of determining nutrients; for example, nitrogen is characterized as nitrate, nitrite, ammonium nitrogen, and total nitrogen. We were unable to obtain water quality data that were standardized and adequate for analysis. However, nutrient concentrations and loads are broadly linked to the intensity of human activity.

Our global risk map underscores the critical situation of potential cyanobacterial health risk in lakes. High-risk areas often cluster in underdeveloped regions such as southern Africa and Southeastern Asia, where rapid industrialization and urbanization threaten the ecological integrity of lakes (59). Negative repercussions stemming from potential cyanobacterial health risks can lead to substantial economic costs, particularly those associated with recreation and tourism (60). Moreover, underdeveloped regions with limited healthcare face challenges in managing cyanotoxin-related illnesses such as liver damage and cancer. While large-scale cyanobacterial epidemics are rare, localized outbreaks from contaminated water can cause gastrointestinal, hepatic, and skin issues, posing serious public health risks where water treatment is lacking (61). Therefore, our study stresses the need for prioritizing cyanobacterial bloom management efforts in underdeveloped regions and the reinforcement of measures in other areas. Our global risk maps based on predictions have reference significance for the monitoring and management of cyanobacteria. However, we acknowledge that our prediction model is limited by sample numbers, particularly the scarcity of samples from Africa. Although the lack of suitable African samples led to uncertainties, our predictions were similar to those of existing reports. A review of cyanobacterial blooms in Africa spanning a decade revealed that the most cosmopolitan genus is *Microcystis*, which is similar to the characteristics of the high-risk areas we identified (62). In Mozambique, where only 50% of people have access to safe drinking water, MCs (hepatotoxins) are not monitored; consequently, the population may be exposed to MCs. The monitoring results for MCs in some areas of Africa were very high, approximately seven times above the maximum limit recommended by the WHO (63). In addition, our sampling of large lakes was limited by temporal and spatial discontinuities. The uneven distribution of sampling points might cause deviations in the assessment results. However, the impact of deviation from such sampling points is low in terms of the prediction results. For example, Lake Taihu has suffered from recurring cyanobacterial blooms for more than two decades, particularly since the 1990s, due to increasing nutrient pollution and eutrophication. The limitations of the Lake Taihu samples did not affect its ability to predict a high potential cyanobacterial health risk lake. To summarize, we call for a global collaborative system to gather high-quality metagenomic data from lakes subjected to cyanobacterial bloom monitoring.

On the basis of the global map of potential cyanobacterial health risks, we find medium–high risk in western China and eastern Hudson Bay in Canada. There is evidence that cyanobacteria can be quite toxic in lakes and fjords at high latitudes, which would certainly include Hudson Bay (64–67). High-latitude water bodies show surprisingly high toxicity, probably because of long summer day-length conditions allowing plenty of light for cyanotoxin synthesis. According to our risk assessment framework, the biosynthetic potential of cyanotoxins and their high toxicity indices (such as microcystin and nodularin) resulted in a relatively high potential cyanobacterial health risk in Hudson Bay (ranked in the top 30% of a total of 750 samples). Our model, after adjusting its parameters, achieved an R^2 of 74%. However, similar to other studies, uncertainties remain due to limitations in environmental data, potential model assumptions, and inherent variability (68, 69). Addressing these uncertainties is essential for improving the robustness of future predictions. The precision of the environmental factors was one of the main reasons for the prediction uncertainty. We obtained environmental factors for each sample from publicly available databases, but the resolution of the data from different databases was variable, ranging from 30 arc seconds to 10 minutes. Therefore, the accuracy of the environmental factors via GIS may deviate, even if they were extracted from similar latitude and longitude coordinates.

Conclusion

Our research introduces a novel approach to potential cyanobacterial health risk assessment, offering insights into the risk from a toxicity perspective. The distinct geographical variations in cyanobacterial communities coupled with the intricate interplay of environmental factors underline the complexity of managing cyanobacterial blooms at a global scale. Our systematic and targeted cyanobacterial surveillance enables a worldwide assessment of cyanobacteria-based potential health risks, providing an early warning system. In conclusion, the combined impact of extreme climatic events and anthropogenic activities amplifies the health risk posed by cyanobacteria. Urgent action is needed to address cyanobacterial bloom mitigation in underdeveloped regions, whereas robust strategies are crucial to combat blooms in a warmer, more climatically extreme world. We recommend a global collaborative system to gather more high-quality and continuous metagenomic data from lakes monitored for cyanobacterial blooms. Moreover, harmonized standards for the measurement of environmental parameters in the context of global cooperation are needed.

ACKNOWLEDGMENTS

H.F.Q., Q.Z., H.W.P., and J.P. acquired funding for this project. This work was financially supported by the National Natural Science Foundation of China (21976161, 21777144, 42307158, and 22376187), the Zhejiang Provincial Natural Science Foundation of China (Grant No. LZ23B070001), the China Postdoctoral Science Foundation (2023M743100), the Zhejiang Provincial Postdoctoral Science Foundation (ZJ2023058), and the Key Scientific and Technological Program of Hangzhou (2023SZD0058). H.P. was supported by the US National Science Foundation (1831096, 1803697, 2108917) and the National Institutes of Health (1P01ES028939-01). J.P.'s research was supported by the Spanish Government grant TED2021-132627B-I00 funded by MCIN, AEI/10.13039/501100011033, and the European Union NextGenerationEU/PRTr; the Fundación Ramón Areces grant CIVP20A6621; and the Catalan Government grant SGR2021-1333.

H.H. designed the study with guidance from H.F.Q. H.H. wrote the first draft of the manuscript, and H.F.Q., H.W.P., and J.P. contributed substantially to the revisions. W.J.H. contributed to all the functional annotations of the metagenome-assembled genomes. Z.Y.Z., Q.Z., and T.Z.W. performed all the metagenomic analyses. N.H.X. and B.F.C. were responsible for machine learning model construction and related data analysis. H.H. performed the visualization of all the data and the artistic design of all the figures. H.F.Q., H.W.P., and J.P. acquired funding for this project.

AUTHOR AFFILIATIONS

[1]College of Environment, Zhejiang University of Technology, Hangzhou, Zhejiang, China
[2]The Institute for Advanced Studies, Shaoxing University, Shaoxing, China
[3]College of Chemistry & Chemical Engineering, Shaoxing University, Shaoxing, China
[4]Institute of Marine Sciences, University of North Carolina at Chapel Hill, Morehead City, North Carolina, USA
[5]Key Laboratory of Microbial Technology and Bioinformatics of Zhejiang Province, Hangzhou, China
[6]CSIC, Global Ecology Unit CREAF-CSIC-UAB, Barcelona, Catalonia, Spain
[7]CREAF, Campus Universitat Autònoma de Barcelona, Barcelona, Catalonia, Spain

AUTHOR ORCIDs

Bingfeng Chen http://orcid.org/0009-0002-4090-7224
Haifeng Qian http://orcid.org/0000-0003-0807-9991

FUNDING

Funder	Grant(s)	Author(s)
MOST | National Natural Science Foundation of China (NSFC)	21976161, 21777144, 22376187	Haifeng Qian
Fundación Ramón Areces (FRA)	CIVP20A6621	Josep Penuelas
the Catalan Government	SGR2021-1333	Josep Penuelas
MOST | National Natural Science Foundation of China (NSFC)	42307158	Qi Zhang
MOST | NSFC | NSFC-Zhejiang Joint Fund | 浙江省科学技术厅 | Natural Science Foundation of Zhejiang Province (ZJNSF)	LZ23B070001	Haifeng Qian
China Postdoctoral Science Foundation (China Postdoctoral Foundation Project)	2023M743100	Qi Zhang
Zhejiang Provincial Postdoctoral Science Foundation	ZJ2023058	Qi Zhang
the Key Scientific and Technological Program of Hangzhou	2023SZD0058	Qi Zhang
National Science Foundation (NSF)	1831096, 1803697, 2108917	Hans W. Paerl
Foundation for the National Institutes of Health (FNIH)	1P01ES028939-01	Hans W. Paerl
the Spanish Government	TED2021-132627B-I00	Josep Penuelas

DATA AVAILABILITY

We collected 96 factors on climate change and anthropogenic activity. There were 37 climate change factors, 19 from WorldClim (https://www.worldclim.org/data/worldclim21.html), 16 from CliMond (https://www.climond.org/BioclimRegistry.aspx#BioclimFAQ), and 2 from CGIAR-CSI (https://cgiarcsi.community/2019/01/24/global-aridity-index-and-potential-evapotranspiration-climate-database-v2/). Livestock production: https://dataverse.harvard.edu/dataverse/glw_4 (Food and Agriculture Organization of the United Nations). Proportion of feed nitrogen: http://www.fao.org/geonetwork/srv/en/ (FAO GeoNetwork). Agricultural yield: https://cgiarcsi.community/2019/01/04/global-spatially-disaggregated-crop-production-statistics-data-for-2010/ (CGIAR-CSI). Fertilizer use: https://beta.sedac.ciesin.columbia.edu/search/data?contains=Nitrogen+Fertilizer+Application (EarthData). Travel time to cities and ports: https://cgiarcsi.community/2019/01/28/travel-time-to-cities-and-ports-2015/ (CGIAR-CSI). Population density of the world: https://beta.sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-adjusted-to-2015-unwpp-country-totals (UN-Adjusted Population Density). Human influence index: https://beta.sedac.ciesin.columbia.edu/data/set/wildareas-v2-human-influence-index-geographic (EarthData). Per capital GDP: https://datadryad.org/stash/dataset/doi:10.5061/dryad.dk1j0 (Dryad). Global emissions of polluting gases: https://edgar.jrc.ec.europa.eu/dataset_ghg70 (Emissions Database for Global Atmospheric Research). Wind speed: An artificial intelligence reconstruction of global gridded surface winds (70). The codes associated with this study are publicly available at https://github.com/Huhangupup/Cyanobacterial-risk.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental material (AEM01936-24-s0001.docx). Figures S1 to S10.
Table S1 (AEM01936-24-s0002.xlsx). Description of all metagenomic samples.
Table S2 (AEM01936-24-s0003.csv). Cyanobacterial abundance of all samples.
Table S3 (AEM01936-24-s0004.xlsx). Abundance of cyanotoxin biosynthesis genes in all samples.
Table S4 (AEM01936-24-s0005.xlsx). Metadata for the environmental factors.
Table S5 (AEM01936-24-s0006.csv). Description of 73,030 lakes.
Table S6 (AEM01936-24-s0007.xlsx). Potential cyanobacterial health risk of all samples.
Table S7 (AEM01936-24-s0008.xlsx). Increase in mean squared error of 23 factors.

REFERENCES

1. Visser PM, Verspagen JMH, Sandrini G, Stal LJ, Matthijs HCP, Davis TW, Paerl HW, Huisman J. 2016. How rising CO_2 and global warming may stimulate harmful cyanobacterial blooms. Harmful Algae 54:145–159. https://doi.org/10.1016/j.hal.2015.12.006
2. Paerl HW, Paul VJ. 2012. Climate change: links to global expansion of harmful cyanobacteria. Water Res 46:1349–1363. https://doi.org/10.1016/j.watres.2011.08.002
3. Erratt KJ, Creed IF, Lobb DA, Smol JP, Trick CG. 2023. Climate change amplifies the risk of potentially toxigenic cyanobacteria. Glob Chang Biol 29:5240–5249. https://doi.org/10.1111/gcb.16838
4. Ho JC, Michalak AM, Pahlevan N. 2019. Widespread global increase in intense lake phytoplankton blooms since the 1980s. Nat New Biol 574:667–670. https://doi.org/10.1038/s41586-019-1648-7
5. Hou X, Feng L, Dai Y, Hu C, Gibson L, Tang J, Lee Z, Wang Y, Cai X, Liu J, Zheng Y, Zheng C. 2022. Global mapping reveals increase in lacustrine algal blooms over the past decade. Nat Geosci 15:130–134. https://doi.org/10.1038/s41561-021-00887-x
6. Huisman J, Codd GA, Paerl HW, Ibelings BW, Verspagen JMH, Visser PM. 2018. Cyanobacterial blooms. Nat Rev Microbiol 16:471–483. https://doi.org/10.1038/s41579-018-0040-1
7. Reinl KL, Harris TD, Elfferich I, Coker A, Zhan Q, De Senerpont Domis LN, Morales-Williams AM, Bhattacharya R, Grossart H-P, North RL, Sweetman JN. 2022. The role of organic nutrients in structuring freshwater phytoplankton communities in a rapidly changing world. Water Res 219:118573. https://doi.org/10.1016/j.watres.2022.118573
8. Paerl HW, Otten TG. 2013. Harmful cyanobacterial blooms: causes, consequences, and controls. Microb Ecol 65:995–1010. https://doi.org/10.1007/s00248-012-0159-y
9. Paerl HW, Huisman J. 2009. Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. Environ Microbiol Rep 1:27–37. https://doi.org/10.1111/j.1758-2229.2008.00004.x
10. Richardson J, Feuchtmayr H, Miller C, Hunter PD, Maberly SC, Carvalho L. 2019. Response of cyanobacteria and phytoplankton abundance to warming, extreme rainfall events and nutrient enrichment. Glob Chang Biol 25:3365–3380. https://doi.org/10.1111/gcb.14701

11. Gu P, Zhang G, Luo X, Xu L, Zhang W, Li Q, Sun Y, Zheng Z. 2021. Effects of different fluid fields on the formation of cyanobacterial blooms. Chemosphere 283:131219. https://doi.org/10.1016/j.chemosphere.2021.131219
12. Zhang Y, Whalen JK, Cai C, Shan K, Zhou H. 2023. Harmful cyanobacteria-diatom/dinoflagellate blooms and their cyanotoxins in freshwaters: a nonnegligible chronic health and ecological hazard. Water Res 233:119807. https://doi.org/10.1016/j.watres.2023.119807
13. Janssen E-L. 2019. Cyanobacterial peptides beyond microcystins - a review on co-occurrence, toxicity, and challenges for risk assessment. Water Res 151:488–499. https://doi.org/10.1016/j.watres.2018.12.048
14. Corbel S, Mougin C, Bouaïcha N. 2014. Cyanobacterial toxins: modes of actions, fate in aquatic and soil ecosystems, phytotoxicity and bioaccumulation in agricultural crops. Chemosphere 96:1–15. https://doi.org/10.1016/j.chemosphere.2013.07.056
15. Lee J, Lee S, Jiang X. 2017. Cyanobacterial toxins in freshwater and food: important sources of exposure to humans. Annu Rev Food Sci Technol 8:281–304. https://doi.org/10.1146/annurev-food-030216-030116
16. Wood R. 2016. Acute animal and human poisonings from cyanotoxin exposure - a review of the literature. Environ Int 91:276–282. https://doi.org/10.1016/j.envint.2016.02.026
17. Plaas HE, Paerl HW. 2021. Toxic cyanobacteria: a growing threat to water and air quality. Environ Sci Technol 55:44–64. https://doi.org/10.1021/acs.est.0c06653
18. Ibelings BW, Backer LC, Kardinaal WEA, Chorus I. 2014. Current approaches to cyanotoxin risk assessment and risk management around the globe. Harmful Algae 40:63–74. https://doi.org/10.1016/j.hal.2014.10.002
19. Guan Q, Feng L, Hou X, Schurgers G, Zheng Y, Tang J. 2020. Eutrophication changes in fifty large lakes on the Yangtze Plain of China derived from MERIS and OLCI observations. Remote Sens Environ 246:111890. https://doi.org/10.1016/j.rse.2020.111890
20. Li Y, Tao J, Zhang Y, Shi K, Chang J, Pan M, Song L, Jeppesen E, Zhou Q. 2023. Urbanization shifts long‐term phenology and severity of phytoplankton blooms in an urban lake through different pathways. Glob Chang Biol 29:4983–4999. https://doi.org/10.1111/gcb.16828
21. Shi K, Zhang Y, Qin B, Zhou B. 2019. Remote sensing of cyanobacterial blooms in inland waters: present knowledge and future challenges. Sci Bull Sci Found Philipp 64:1540–1556. https://doi.org/10.1016/j.scib.2019.07.002
22. Faust K, Raes J. 2012. Microbial interactions: from networks to models. Nat Rev Microbiol 10:538–550. https://doi.org/10.1038/nrmicro2832
23. Volk A, Lee J. 2023. Cyanobacterial blooms: a player in the freshwater environmental resistome with public health relevance? Environ Res 216:114612. https://doi.org/10.1016/j.envres.2022.114612
24. Chen H, Jing L, Yao Z, Meng F, Teng Y. 2019. Prevalence, source and risk of antibiotic resistance genes in the sediments of Lake Tai (China) deciphered by metagenomic assembly: a comparison with other global lakes. Environ Int 127:267–275. https://doi.org/10.1016/j.envint.2019.03.048
25. Linz DM, Sienkiewicz N, Struewing I, Stelzer EA, Graham JL, Lu J. 2023. Metagenomic mapping of cyanobacteria and potential cyanotoxin producing taxa in large rivers of the United States. Sci Rep 13:2806. https://doi.org/10.1038/s41598-023-29037-6
26. Romanis CS, Pearson LA, Neilan BA. 2021. Cyanobacterial blooms in wastewater treatment facilities: significance and emerging monitoring strategies. J Microbiol Methods 180:106123. https://doi.org/10.1016/j.mimet.2020.106123
27. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
28. Merel S, Walker D, Chicana R, Snyder S, Baurès E, Thomas O. 2013. State of knowledge and concerns on cyanobacterial blooms and cyanotoxins. Environ Int 59:303–327. https://doi.org/10.1016/j.envint.2013.06.013
29. Qin B, Zhu G, Gao G, Zhang Y, Li W, Paerl HW, Carmichael WW. 2010. A drinking water crisis in Lake Taihu, China: linkage to climatic variability and lake management. Environ Manage 45:105–112. https://doi.org/10.1007/s00267-009-9393-6
30. Zhang Q, Lei C, Jin M, Qin G, Yu Y, Qiu D, Wang Y, Zhang Z, Zhang Z, Lu T, Peijnenburg WJGM, Gillings M, Yao Z, Qian H. 2024. Glyphosate disorders soil *Enchytraeid* gut microbiota and increases its antibiotic resistance risk. J Agric Food Chem 72:2089–2099. https://doi.org/10.1021/acs.jafc.3c05436
31. Zhu D, An X-L, Chen Q-L, Yang X-R, Christie P, Ke X, Wu L-H, Zhu Y-G. 2018. Antibiotics disturb the microbiome and increase the incidence of resistance genes in the gut of a common soil collembolan. Environ Sci Technol 52:3081–3090. https://doi.org/10.1021/acs.est.7b04292
32. Zhang Q, Xu N, Lei C, Chen B, Wang T, Ma Y, Lu T, Penuelas J, Gillings M, Zhu Y-G, Fu Z, Qian H. 2023. Metagenomic insight into the global dissemination of the antibiotic resistome. Adv Sci (Weinh) 10:2303925. https://doi.org/10.1002/advs.202303925
33. Hirose Y, Komaki F. 2010. An extension of least angle regression based on the information geometry of dually flat spaces. J Comput Graph Stat 19:1007–1023. https://doi.org/10.1198/jcgs.2010.09064
34. Zou H, Hastie T. 2005. Regularization and variable selection via the Elastic Net. J R Stat Soc Ser B 67:301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
35. Breiman L. 2001. Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
36. Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. Ann Statist 29:1189–1232. https://doi.org/10.1214/aos/1013203451
37. Breiman L. 1996. Bagging predictors. Mach Learn 24:123–140. https://doi.org/10.1007/BF00058655
38. Quinlan JR. 1992. Learning with continuous classes. 5th Australian Joint Conference on Artificial Intelligence, Vol. 92, p 343–348
39. Ban Z, Yuan P, Yu F, Peng T, Zhou Q, Hu X. 2020. Machine learning predicts the functional composition of the protein corona and the cellular recognition of nanoparticles. Proc Natl Acad Sci U S A 117:10492–10499. https://doi.org/10.1073/pnas.1919755117
40. Zhang Z, Zhang Q, Chen B, Yu Y, Wang T, Xu N, Fan X, Penuelas J, Fu Z, Deng Y, Zhu Y-G, Qian H. 2024. Global biogeography of microbes driving ocean ecological status under climate change. Nat Commun 15:4657. https://doi.org/10.1038/s41467-024-49124-0
41. Goldstein A, Kapelner A, Bleich J, Pitkin E. 2015. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24:44–65. https://doi.org/10.1080/10618600.2014.907095
42. Sinaga KP, Yang M-S. 2020. Unsupervised k-means clustering algorithm. IEEE Access 8:80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796
43. Carey CC, Ibelings BW, Hoffmann EP, Hamilton DP, Brookes JD. 2012. Eco-physiological adaptations that favour freshwater cyanobacteria in a changing climate. Water Res 46:1394–1407. https://doi.org/10.1016/j.watres.2011.12.016
44. Dörnhöfer K, Oppelt N. 2016. Remote sensing for lake research and monitoring – recent advances. Ecol Indic 64:105–122. https://doi.org/10.1016/j.ecolind.2015.12.009
45. Bertone E, Burford MA, Hamilton DP. 2018. Fluorescence probes for real-time remote cyanobacteria monitoring: a review of challenges and opportunities. Water Res 141:152–162. https://doi.org/10.1016/j.watres.2018.05.001
46. Stumpf RP, Davis TW, Wynne TT, Graham JL, Loftin KA, Johengen TH, Gossiaux D, Palladino D, Burtner A. 2016. Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. Harmful Algae 54:160–173. https://doi.org/10.1016/j.hal.2016.01.005
47. Bullerjahn GS, McKay RM, Davis TW, Baker DB, Boyer GL, D'Anglada LV, Doucette GJ, Ho JC, Irwin EG, Kling CL, Kudela RM, Kurmayer R, Michalak AM, Ortiz JD, Otten TG, Paerl HW, Qin B, Sohngen BL, Stumpf RP, Visser PM, Wilhelm SW. 2016. Global solutions to regional problems: collecting global expertise to address the problem of harmful cyanobacterial blooms. A Lake Erie case study. Harmful Algae 54:223–238. https://doi.org/10.1016/j.hal.2016.01.003
48. Deng J, Shan K, Shi K, Qian SS, Zhang Y, Qin B, Zhu G. 2023. Nutrient reduction mitigated the expansion of cyanobacterial blooms caused by climate change in Lake Taihu according to Bayesian network models. Water Res 236:119946. https://doi.org/10.1016/j.watres.2023.119946
49. Wang S, Zhang X, Wang C, Chen N. 2023. Multivariable integrated risk assessment for cyanobacterial blooms in eutrophic lakes and its spatiotemporal characteristics. Water Res 228:119367. https://doi.org/10.1016/j.watres.2022.119367

50. Wang S, Zhang X, Wang C, Chen N. 2023. Temporal continuous monitoring of cyanobacterial blooms in Lake Taihu at an hourly scale using machine learning. Sci Total Environ 857:159480. https://doi.org/10.1016/j.scitotenv.2022.159480

51. Al-Tebrineh J, Merrick C, Ryan D, Humpage A, Bowling L, Neilan BA. 2012. Community composition, toxigenicity, and environmental conditions during a cyanobacterial bloom occurring along 1,100 kilometers of the Murray River. Appl Environ Microbiol 78:263–272. https://doi.org/10.1128/AEM.05587-11

52. Chi Y, Shi H, Zheng W, Sun J, Fu Z. 2018. Spatiotemporal characteristics and ecological effects of the human interference index of the Yellow River Delta in the last 30 years. Ecol Indic 89:880–892. https://doi.org/10.1016/j.ecolind.2017.12.025

53. Yindong T, Xiwen X, Miao Q, Jingjing S, Yiyan Z, Wei Z, Mengzhu W, Xuejun W, Yang Z. 2021. Lake warming intensifies the seasonal pattern of internal nutrient cycling in the eutrophic lake and potential impacts on algal blooms. Water Res 188:116570. https://doi.org/10.1016/j.watres.2020.116570

54. Jöhnk KD, Huisman J, Sharples J, Sommeijer B, Visser PM, Stroom JM. 2008. Summer heatwaves promote blooms of harmful cyanobacteria. Glob Chang Biol 14:495–512. https://doi.org/10.1111/j.1365-2486.2007.01510.x

55. Paerl HW, Pinckney JL, Steppe TF. 2000. Cyanobacterial-bacterial mat consortia: examining the functional unit of microbial survival and growth in extreme environments. Environ Microbiol 2:11–26. https://doi.org/10.1046/j.1462-2920.2000.00071.x

56. Paerl HW, Gardner WS, Havens KE, Joyner AR, McCarthy MJ, Newell SE, Qin B, Scott JT. 2016. Mitigating cyanobacterial harmful algal blooms in aquatic ecosystems impacted by climate change and anthropogenic nutrients. Harmful Algae 54:213–222. https://doi.org/10.1016/j.hal.2015.09.009

57. Yang J, Lv H, Yang J, Liu L, Yu X, Chen H. 2016. Decline in water level boosts cyanobacteria dominance in subtropical reservoirs. Sci Total Environ 557–558:445–452. https://doi.org/10.1016/j.scitotenv.2016.03.094

58. Hamilton DP, Salmaso N, Paerl HW. 2016. Mitigating harmful cyanobacterial blooms: strategies for control of nitrogen and phosphorus loads. Aquat Ecol 50:351–366. https://doi.org/10.1007/s10452-016-9594-z

59. Ferronato N, Torretta V. 2019. Waste mismanagement in developing countries: a review of global issues. Int J Environ Res Public Health 16:1060. https://doi.org/10.3390/ijerph16061060

60. Camargo JA, Alonso A. 2006. Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: a global assessment. Environ Int 32:831–849. https://doi.org/10.1016/j.envint.2006.05.002

61. Buratti FM, Manganelli M, Vichi S, Stefanelli M, Scardala S, Testai E, Funari E. 2017. Cyanotoxins: producing organisms, occurrence, toxicity, mechanism of action and human health toxicological risk evaluation. Arch Toxicol 91:1049–1130. https://doi.org/10.1007/s00204-016-1913-6

62. Ndlela LL, Oberholster PJ, Van Wyk JH, Cheng PH. 2016. An overview of cyanobacterial bloom occurrences and research in Africa over the last decade. Harmful Algae 60:11–26. https://doi.org/10.1016/j.hal.2016.10.001

63. Tamele IJ, Vasconcelos V. 2020. Microcystin Incidence in the drinking water of mozambique: challenges for public health protection. Toxins (Basel) 12:368. https://doi.org/10.3390/toxins12060368

64. España Amórtegui JC, Pekar H, Retrato MDC, Persson M, Karlson B, Bergquist J, Zuberovic-Muratovic A. 2023. LC-MS/MS analysis of cyanotoxins in bivalve mollusks-method development, validation and first evidence of occurrence of nodularin in mussels (*Mytilus edulis*) and oysters (*Magallana gigas*) from the West Coast of Sweden. Toxins (Basel) 15:329. https://doi.org/10.3390/toxins15050329

65. Halstvedt CB, Rohrlack T, Andersen T, Skulberg O, Edvardsen B. 2007. Seasonal dynamics and depth distribution of *Planktothrix* spp. in Lake Steinsfjorden (Norway) related to environmental factors. J Plankton Res 29:471–482. https://doi.org/10.1093/plankt/fbm036

66. Karlson B, Andersen P, Arneborg L, Cembella A, Eikrem W, John U, West JJ, Klemm K, Kobos J, Lehtinen S, Lundholm N, Mazur-Marzec H, Naustvoll L, Poelman M, Provoost P, De Rijcke M, Suikkanen S. 2021. Harmful algal blooms and their effects in coastal seas of Northern Europe. Harmful Algae 102:101989. https://doi.org/10.1016/j.hal.2021.101989

67. Mantzouki E, Campbell J, van Loon E, Visser P, Konstantinou I, Antoniou M, Giuliani G, Machado-Vieira D, Gurjão de Oliveira A, Maronić DŠ, et al. 2018. A European multi lake survey dataset of environmental variables, phytoplankton pigments and cyanotoxins. Sci Data 5:180226. https://doi.org/10.1038/sdata.2018.226

68. Guerra CA, Berdugo M, Eldridge DJ, Eisenhauer N, Singh BK, Cui H, Abades S, Alfaro FD, Bamigboye AR, Bastida F, et al. 2022. Global hotspots for soil nature conservation. Nat New Biol 610:693–698. https://doi.org/10.1038/s41586-022-05292-x

69. Araújo MB, Anderson RP, Márcia Barbosa A, Beale CM, Dormann CF, Early R, Garcia RA, Guisan A, Maiorano L, Naimi B, O'Hara RB, Zimmermann NE, Rahbek C. 2019. Standards for distribution models in biodiversity assessments. Sci Adv 5:eaat4858. https://doi.org/10.1126/sciadv.aat4858

70. Zhou L, Liu H, Jiang X, Ziegler AD, Azorin-Molina C, Liu J, Zeng Z. 2022. An artificial intelligence reconstruction of global gridded surface winds. Sci Bull Sci Found Philipp 67:2060–2063. https://doi.org/10.1016/j.scib.2022.09.022