Use statistical machine learning to detect nutrient thresholds in Microcystis blooms and microcystin management

Use statistical machine learning to detect nutrient thresholds in *Microcystis* blooms and microcystin management

Kun Shan[a,b*], Xiaoxiao Wang[b,e], Hong Yang[d], Botian Zhou[a,b], Lirong Song[c,e], Mingsheng Shang[a,b]

[a] *Chongqing Key Laboratory of Big Data and Intelligent Computing, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China*

[b] *CAS Key Lab on Reservoir Environment, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China*

[c] *State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China*

[d] *Department of Geography and Environmental Science, University of Reading, Whiteknights, Reading, RG6 6AB, UK*

[e] *University of Chinese Academy of Sciences, Beijing 100049, China*

[*] Corresponding author.

E-mail addresses: shankun@cigit.ac.cn (K. Shan)

ABSTRACT

The frequency of toxin-producing cyanobacterial blooms has increased in recent decades due to nutrient enrichment and climate change. Because *Microcystis* blooms are related to different environmental conditions, identifying potential nutrient control targets can facilitate water quality managers to reduce the likelihood of microcystins (MCs) risk. However, complex biotic interactions and field data limitations have constrained our understanding of the nutrient-microcystin relationship. This study develops a Bayesian modelling framework with intracellular and extracellular MCs that characterize the relationships between different environmental and biological factors. This model was fit to the across-lake dataset including three bloom-plagued lakes in China and estimated the putative thresholds of total nitrogen (TN) and total phosphorus (TP). The lake-specific nutrient thresholds were estimated using Bayesian updating process. Our results suggested dual N and P reduction in controlling cyanotoxin risks. The total *Microcystis* biomass can be substantially suppressed by achieving the putative thresholds of TP (0.10 mg/L) in Lakes Taihu and Chaohu, but a stricter TP target (0.05 mg/L) in Dianchi Lake. To maintain MCs concentrations below 1.0 μg/L, the estimated TN threshold in three lakes was 1.8 mg/L, but the effect can be counteracted by the increase of temperature. Overall, the present approach provides an efficient way to integrate empirical knowledge into the data-driven model and is helpful for the management of water resources.

Keywords: Bayesian modelling; eutrophication; nutrient thresholds; cyanobacterial blooms; *Microcystis*; microcystin

1. Introduction

The harmful cyanobacterial blooms have been exacerbated across the world in the last decades with the growing threat from human activities and climate change (Harke et al., 2016； O'Neil et al., 2012). One of the cosmopolitan cyanobacterium genera is *Microcystis*, which has been reported to form blooms in more than 257 countries and territories, particularly in large lake ecosystems (Jankowiak et al., 2019). Blooms by *Microcystis* often cause serious environmental problems, such as degradation of water quality and illness or even death of other eukaryotic organisms, animals, and humans (MacKintosh et al., 1990; Singh et al., 2015), due to the production of hepatotoxic microcystins (MCs). As a result, identifying specific environmental conditions under which MCs in water columns can exceed the provisional World Health Organization (WHO) Guideline of 1.0 ug/L is of great importance for lake managers (Burch, 2008; WHO, 1998).

Previous studies have revealed that prevailing environmental conditions can result in high MCs events, including low nitrogen-phosphorus ratios (N:P) (Orihel et al., 2012), an imbalance in cellular carbon-nitrogen ratios (C:N) (Beversdorf et al., 2015), warm temperature (Bui et al., 2018), photosynthetically active radiation limitation (Wiedner et al., 2003), and iron limitation (Alexova et al., 2011). Compared to other environmental factors, N and P concentrations are more amenable to control. Furthermore, extensive cyanobacterial blooms and high MCs events are most prevalent in eutrophic and hypereutrophic lakes (Rigosi et al., 2015). Spatiotemporal patterns and ecophysiology of toxigenic *Microcystis* blooms are likely influenced by the distribution

of nutrients inside the lake (Otten et al., 2012). In this context, establishing nutrient thresholds or criteria for controlling an abrupt change or regime shift of toxic cyanobacterial blooms is a quantifiable and attractive approach (Xu et al., 2014; Zhang et al., 2006).

Setting nutrient control targets which are often being supported by scenario analysis using mechanistic models has been a challenge for lake managers (Recknagel et al., 2017). These process-based approaches can give insights into the biogeochemical cycling which are crucial to simulate how environmental conditions affect the composition and growth of phytoplankton (Reynolds and Irish, 1997). However, cyanotoxin production is highly variable in space and time and cannot be accurately predicted from cyanobacterial composition and abundance (Huisman et al., 2018). For instance, cyanobacterial blooms in natural water are often comprised of toxic and non-toxic strains, and changes in strain composition can, therefore, lead to major alterations in the toxin content (Kardinaal et al., 2007). It has also been reported that cyanotoxin production among taxa or even within strains of the same species is triggered by different environmental factors (Beaver et al., 2018; Davis et al., 2009). Complicating the prediction is that the majority of MCs remain intracellular in intact cells, and they are released into water columns when cells are lysed or damaged (Daly et al., 2007). Owing to the complex interaction between physical, chemical and biological factors, the capacity of mechanistic models to simulate the MCs dynamics remains poor.

In the last several years, data-intensive statistical models, have received increasing attention and several advanced methods have been developed to simulate MCs

concentrations. For example, the hierarchical zero-altered model (Taranu et al., 2017), the hierarchical Bayesian model (Yuan et al., 2017), and the Bayesian network (Yuan et al., 2019) have been fit to USEPA National Lake data and provided estimates of the potential relationships between different lake characteristics. Due to the large difference between areas, local water management teams want to explore lake-specific criteria. Therefore, a major challenge is to estimate the specific relationship for a lake from the limited samples. Bayesian inferential methods perform well when dealing with small sample sizes and limited data, leading to some ecological applications (Link and Barker, 2009). For instance, Kelly et al. (2019) estimated the environmental conditions associated with the probability of exceedance MCs levels in a eutrophic lake and the results help predict MCs risk. However, these approaches could not characterize the relationships between taxon-specific biomass and MCs production at the same time.

To remove the limitations, the present study presents a continuous variable Bayesian networks model, which develops from the basic model-developing strategy by Qian and Miltner (2015). The main research aim is to estimate the relationships between nutrient concentrations and potential MCs thresholds. Emphasis is given to the causal diagram, which combines cell-bound and dissolved MCs with different biotic and abiotic factors. To leverage knowledge from macro-scale data to enhance understanding of specific lakes, the Bayesian computation was applied to develop the across-lake model based on data from specific lakes. To showcase the modelling framework, this study applied data from three cyanobacterial bloom-plagued lakes in China and evaluated whether the nutrient control targets could be affected by warming

in the future.

2. *Methods*

2.1 Dataset description

Three typical cyanobacterial bloom-dominated lakes in China, including Lake Taihu (30°56'~31°33'N, 119°55'~120°54'E), Lake Chaohu (30°25'~31°43'N, 117°17'~117°52'E), and Lake Dianchi (24°29'~25°28'N, 102°29'~103°01'E), were selected for examination in this study. More detailed descriptions of sampling and laboratory methods are available in the study of Shan et al. (2019b). The following environmental variables were determined and included in this study: water temperature (WT, in °C), dissolved oxygen (DO, in mg/L), pH, electrical conductivity (EC, in S/m), Secchi disk (SD, in cm), wind speed (WS, in m/s), total P (TP in mg/L), dissolved inorganic P (DIP, in mg/L), total N (TN, in mg/L), and dissolved inorganic N (DIN = ammonium (NH_4^+) + nitrate (NO_3^-) + nitrite (NO_2^-), in mg/L). The following biological variables were measured and included: chlorophyll-*a* (Chl-*a*, in μg/L), cyanobacterial biomass (B_{cya}, in mg/L), total biomass of *Microcystis* (B_M, in mg/L), and the taxon-specific biomass of *Microcystis aeruginosa* ($B_{MA,}$ in mg/L). The MCs concentrations were measured across 17 sampling transects encompassing the entire lakes. Dissolved microcystins (dMCs, in μg/L) were measured by 96 wells filled for enzyme-linked immunosorbent assays. Cell-bound microcystins (cMCs, in mg/g dry weight) were extracted with 90% (v/v) aqueous methanol, and extracts have seeped through Sep Pak C18 cartridges. Finally, cell-bound MCs were eluted in solutions with 1 mL 50% (v/v)

chromatographic pure methanol (Thermo Fisher Scientific, Waltham, MA, USA) and stored at −20 °C for HPLC analysis. The details of cell counting and MCs measurements are available in the study of Hu et al. (2016) and Wu et al. (2014).

2.2 Model development

A Bayesian modelling framework was developed to link environmental factors, phytoplankton-related biomass, and MCs concentrations in the across-lake dataset. The steps of model development were summarized in **Fig. 1**.

2.2.1 LASSO regression

A regression model with the least absolute shrinkage and selection operator (LASSO) was used to build empirical regressions for developing the conceptual model of the Bayesian network (Tibshirani, 1996). Given a linear regression with predictor variable x_i and response variable y_i, the LASSO solves the l_1-penalized regression problem of finding $\beta = \{\beta_j\}$ to minimize the formula as follows:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x}_i^{'}\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (1)$$

where β_0 and $\boldsymbol{\beta}$ are the regression coefficients, and p corresponds to the number of covariates in the model. LASSO identifies parsimonious predictive models by gradually shrinking the absolute value of regression coefficients so that the sum of all coefficients is less than a prespecified threshold ($\sum_{j=1}^{p}|\beta_j| \le s$)(Yuan et al., 2014). In LASSO regression, shrinkage and variable selection are achieved simultaneously because the coefficients are linearly shrunk to exactly zero, thereby avoiding overestimation of

variables. Given collinear variables and limited observations for MCs, concepts of parameter shrinkage by penalized estimation can be used to fit interpretable models with reliable predictions (Dahlgren et al., 2010; Hooten and Hobbs, 2015).

All physical-chemical (WT, EC, SD, DO, pH, WS, TN, DIN, TP, DIP, TN:TP, and DIN:TP) and biological (phytoplankton-related) variables were log-transformed before further analyze. Explanatory variables in LASSO regression were all standardized with zero as mean value and one as standard deviation. Given that λ controls the amount of shrinkage induced, different values of λ produced various models. The explanatory variables contributing to the model decrease with the increase in the value of λ. This study used a 10-fold cross-validation procedure to calculate the standard error of models along the gradient of λ and selected the value of λ based on the "one-standard-error" rule (Breiman et al., 1984). The "glmnet" package in the R library was used to implement LASSO regression (Friedman et al., 2010).

2.2.2 Bayesian network

Based on the results of LASSO, three following regressions can be linked together as a directed acyclic diagram (DAG) of the Bayesian network (BN) model (**Fig. 2a**). To deal with continuous variables, the initial DAG model was revised to connect data and unknow parameters (**Fig. 2b**). The first model was a regression for predicting *Microcystis* biomass, where WT, TP, DIN, SD, and pH were used as predictors. The second model was for predicting cell-bound MCs concentrations using the biomass of *Microcystis* and environmental variables including WT, DIN, and DIP. The third model

was for predicting dissolved MCs using cell-bound MCs concentrations and environmental variables including pH, WS, and TN. All these regression models were briefly summarized.

(1) *The Microcystis biomass model*

The *Microcystis* biomass model can be written as follow:

$$\log(B_M) = \beta_0^c + \beta_1^c \log(SD) + \beta_2^c \log(TP) + \beta_3^c \log(WT) + \beta_4^c \log(DIN) + \beta_5^c \log(pH) + \varepsilon^c. \quad (2)$$

This model can be replaced with the probability distribution of $\log(B_M)$, and that is

$$\log(B_M) \sim \mathrm{N}(\mu_M, \sigma_M^2) \quad (3)$$

$$\mu_M = \beta_0^c + \beta_1^c \log(SD) + \beta_2^c \log(TP) + \beta_3^c \log(WT) + \beta_4^c \quad (4)$$

(2) *The Cell-bound MCs model*

The cell-bound MCs model can be expressed as follow:

$$\log(cMCs) = \beta_0^d + \beta_1^d \log(WT) + \beta_2^d \log(DIN) + \beta_3^d \log(DIP) + \beta_4^d \mu_M + \varepsilon^d \quad (5)$$

This model can be changed with the probability distribution of log(*cMCs*), and that is

$$\log(cMCs) \sim \mathrm{N}(\mu_{cMCs}, \sigma_c^2) \quad (6)$$

$$\mu_c = \beta_0^d + \beta_1^d \log(WT) + \beta_2^d \log(DIN) + \beta_3^d \log(DIP) + \beta_4^d \mu_M \quad (7)$$

(3) *The Dissolved MCs model*

The dissolved MCs model can be listed as follow:

$$\log(dMCs) = \beta_0^e + \beta_1^e \log(pH) + \beta_2^e \log(WS) + \beta_3^e \log(TN) + \beta_4^e \mu_c + \varepsilon^e \quad (8)$$

This model can be revised with the probability distribution of log($dMCs$), and that is

$$\log(dMCs) \sim \mathrm{N}\,(\mu_d, \sigma_d^2) \tag{9}$$

$$\mu_d = \beta_0^e + \beta_1^e \log(pH) + \beta_2^e \log(WS) + \beta_3^e \log(TN) + \beta_4^e \mu_c \tag{10}$$

2.2.3 Gibbs sampler

Once these empirical models are established, they can be linked to form the joint probabilistic distribution of all parameters. The purpose of our Bayesian approach is to replace the conditional probability tables in traditional BN with a set of conditional probability distributions (Qian and Miltner, 2015). Estimating all unknown parameters result in the following likelihood function:

$$\begin{aligned}
&\mathrm{L}(\log(B_M), \log(cMCs), \log(dMCs) \mid \theta) \\
&= \frac{1}{(2\pi\sigma_M^2)^{\frac{1}{2}}} e^{-\frac{(\mu_M - \beta_0^c - \beta_1^c \log(SD) - \beta_2^c \log(TP) - \beta_3^c \log(WT) - \beta_4^c \log(DIN) - \beta_5^c \log(pH))}{2\pi\sigma_M^2}} \\
&\times \frac{1}{(2\pi\sigma_c^2)^{\frac{1}{2}}} e^{-\frac{(\mu_c - \beta_0^d - \beta_1^d log(WT) - \beta_2^d log(DIN) - \beta_3^d log(DIP) - \beta_4^d \mu_M)}{2\pi\sigma_c^2}} \\
&\times \frac{1}{(2\pi\sigma_d^2)^{\frac{1}{2}}} e^{-\frac{(\mu_d - \beta_0^e - \beta_1^e log(pH) - \beta_2^e log(WS) - \beta_3^e log(TN) - \beta_4^e \mu_c)}{2\pi\sigma_C^2}}
\end{aligned} \tag{11}$$

where θ represents a set of regression parameters. All model coefficients were defined in Equations 2, 5, and 8. Model coefficients were estimated simultaneously by the Gibbs sampler which was implemented using the Bayesian inference software JAGS (Plummer, 2003; Qian, 2016).

2.2.4 Monte Carlo simulations

Random variates as the model inputs (e.g., SD, TP, WT, DIN, DIP, TN, pH, and WS) were considered to follow log-transformation normal distributions. Based on the Pearson correlation coefficients, two nutrient groups (TP and DIP, TN and DIN) were assumed to follow the bivariate normal distributions. After the joint distribution of all coefficients was estimated by the Gibbs sampler, the statistical inference could subsequently be made through Monte Carlo simulations (Whitehead and Young, 1979). According to the management targets of *Microcystis* biomass and MCs concentrations (**Table 1)**, the conditional distributions of TN and TP that were associated with acceptable low risks of toxic cyanobacterial blooms and be derived.

2.2.5 Bayesian updating

Using the Bayesian updating method, the across-lake model was updated using data from specific lakes. In this present work, the estimated distributions of the coefficient from the across-lake model were applied as the prior distributions of coefficients from a lake-specific model. Qian and Reckhow (2007) suggested that improvement could be achieved by the Bayesian updating process if a priori parameter distribution across similarly sampling sites is known. Those updated models would be lake-specific and provide an insight into water quality management for local government.

2.3 Statistical analysis

To address the high spatiotemporal variation of dissolved and cell-bound MCs

concentrations, this study identified four components of the variation by intercept-only model: (1) inter-lake variation or variation in concentrations between different lakes; (2) inter-site within-lake variation or variation in concentrations collected at different sites in the same lake; (3) intra-year variation or variation in concentrations between different sampling months; and (4) residual error which includes variation due to measurement error and others.

The concentration of MCs in response to different forms of nutrient was assessed using generalized additive models (GAM). To examine the potential interactions between nitrogen and phosphorus, GAM models were applied using the combination of TN and TP, or DIN and DIP as two continuous explanatory variables:

$$\log(MC_{ij}+1) = \alpha_j + S_j(\text{Nitrogen}_i,\ \text{Phosphorus}_i) + \varepsilon_{ij},\ \ \varepsilon_{ij} \sim N(0,\sigma^2) \qquad (12)$$

where *i* and *j* are indices for the observations (monthly MCs and nutrient concentrations) and the studied lakes, respectively. The smoothing function (S_j) is the covariates between nitrogen and phosphorus. For two forms of MCs, the contour plots were used to visualize the function S_j. The "mgcv" package was used to implement GAM by optimizing the amount of cubic spline smoothing (Wood, 2001).

3. Results

3.1 Spatiotemporal variation of MC and development of the BN conceptual model

This study identified four components by the intercept-only model to compare the spatiotemporal variation in MCs distribution (**Table 2**). Considering the cell-bound MCs, the standard deviation from sampling month variation was 0.249, accounting for

the largest proportion of variance (75.6%). The standard deviations from inter- and intra-lake variations were 0.09 and 0.057, which accounted for 9.8% and 3.6% of the total variation in cell-bound MCs, respectively. Change in cell-bound MCs would be expected to exhibit regularly temporal trends due to that variation from inter-lake was likely stronger than that from intra-lake. By contrast, the standard deviation from sampling month variation was 0.049, which took up 38% of the total variation in dissolved MCs. The standard deviations of inter- and intra-lake variations accounted for the remaining 14.6% and 9.2% of the dissolved MCs variation, respectively. The residual variation implied uncertainty in predicting dissolved MCs (38.2%), which was considerably larger than that in predicting cell-bound MCs (11%).

LASSO regression was applied in exploring the relationship among a subset of abiotic and biotic variables and MCs concentrations to develop the conceptual linkage (**Fig. S1**). First, B_M achieved higher predictive accuracy than other biological factors, including Chl-*a*, B_{cya}, and B_{MA}. Utilizing cross-validation, the best model for predicting B_M, balancing parsimony, and predictive accuracy was the group of environmental variables including pH, TP, WT, DIN, and SD (**Table 3**). Second, relationships between cell-bound MCs and environmental factors were tested under the condition of combining different biotic factors (**Table 4)**. When the model selected the variables including B_M, WT, DIN, and DIP, it achieved an accurate prediction of cell-bound MCs concentrations (MSPE = 0.037). Third, the best predictive model for dissolved MCs was achieved (MSPE = 0.109) with selected variables including cell-bound MCs, TN, pH, and WS. Based on the aforementioned results, a four-layer structure BN model was

constructed to incorporate different biotic and abiotic variables for predicting the risk of MCs.

3.2 Effects of environmental and biological factors on MC concentrations

The coefficients of the fitted joint models were presented in **Tables S1–S3.** When the variables are log-transformed, the slop represents a change in the response variable under per unit change in the predictor. For instance, the estimated β_3^c was 0.56 (**Table S1**) which represented an approximately 0.56% increase in B_M for a 1% increase in TP. When TP had a 1% increase, B_M increased by 0.59%, 0.55%, and 0.54% in Lakes Taihu, Chaohu, and Dianchi, respectively. When TN had a 1% increase, dissolved MC concentrations increased by 0.24%, 0.16%, and 0.15% in Lakes Taihu, Chaohu, and Dianchi, respectively, which were reflected by the value of β_3^e in **Table S3**.

The effect of nutrients on MCs was considered to change via network structure (**Table 5**). Given the increase in P concentrations from the 25th to the 75th percentile, B_M was predicted to increase by 91.8%, whereas cell-bound MCs and dissolved MCs concentrations decreased by 7.5% and 4.4%, respectively. When N concentrations increased from the 25th to the 75th percentile, the cell-bound MCs concentrations decreased from 0.312 to 0.276, whereas dissolved MCs increased from 0.676 to 0.713. The concentration of MCs in response to N and P was assessed using the GAM approach (**Fig. 3**). The model results suggested that MCs concentrations depended heavily on the interaction between TP and TN. More specifically, cell toxin quota was sensitive to the conditions of low DIN concentrations and high TN:TP ratios, whereas

the probability of dissolved MCs was higher at the increase of TN and SRP than that of TP.

3.3 Evaluation of nutrient control targets

Monte Carlo simulation was repeated until 10,000 TP and TN values were accepted. Given the condition of B_M < 0.6 mg/L and MCs < 0.4 μg/L, a histogram of the discrete distribution indicated that the means of TN and TP in conditional distribution were lower than those in marginal distribution, although both had similar variance (**Fig. 4**). If this calculated conditional distribution can be considered as the "reference" distribution, the U.S. EPA's recommendation can be set as the nutrient criterion at the 75^{th} percentile (U.S. EPA, 2000). The criterion was 0.16 mg/L for TP (marginal: 0.24 mg/L) and 3.12 mg/L for TN (marginal: 3.78 mg/L), respectively. Alternatively, the U.S. EPA also recommends that the 25^{th} percentile in all sampling data can be accepted as the nutrient criterion when "reference" distributions are unavailable. The TN criterion in our data had a 25^{th} percentile of 1.8 mg/L, and the TP criterion had a 25^{th} percentile of 0.1 mg/L. This threshold of TP is close to the empirical value in eutrophication management. However, the significant difference in the estimated values from a 75^{th} percentile and a 25^{th} percentile may largely be attributed to the large spatiotemporal variations in B_M and MCs in the studied lakes.

Furthermore, the goal of setting a single nutrient criterion for different bloom-dominated lakes is likely impractical. After the process of Bayesian updating, those updated models could be used to evaluate lake-specific nutrient thresholds by achieving

the management targets (e.g., B_M < 0.6 mg/L or MCs < 1.0 μg/L in **Fig. 5)**. Despite a considerable interaction that exists, the models responded to changes of P more rapidly than changes of N, and they predicted the high probabilities of achieving water quality objectives at low nutrients concentrations. However, *Microcystis* biomass or MCs concentrations in the three lakes differed in their response to nutrients. The probability of meeting the B_M objectives of 0.6 mg/L at the target TP concentration (0.10 mg/L) in Lake Dianchi was approximately 0.3, while the probabilities in Lakes Taihu and Chaohu were nearly close to 1.0 (**Fig. 5a**). At a stricter TP target (0.05 mg/L), the probability in Lake Dianchi increased to 0.6. On the other hand, the probability of meeting the MCs objectives at the target TN concentration in Lake Dianchi was higher than those in the other two lakes. When the estimated TN target was set to 1.8 mg/L, the probabilities of meeting the provisional guidelines of WHO (MCs < 1.0 μg/L) in three studied lakes were above 0.8 (**Fig. 5b**).

Nutrients in water are not the only key factors influencing the proliferation of *Microcystis* and the production of MCs. Thus, other factors, such as water temperature, were considered to achieve the desired water quality goals. Updated model coefficients were used to estimate the probability of B_M < 1.5 mg/L and MCs < 1.0 μg/L as a function of both TN or TP and WT, with all other variables taken their respective observed means (**Fig. 6**). Simulated scenarios of nutrient enrichment and temperature warming suggested that toxic cyanobacterial blooms may be more sensitive to synergistic effects rather than individual effects alone. For instance, the effects of interactions between TP and WT were evident when WT exceeded 20 °C. On the

contrary, the effect of TN will be counteracted by the fluctuations of water temperature when separating the joint effects of different forms of N and P.

4. Discussion

4.1 The potential factors influencing the spatiotemporal variations in MCs

The key challenge for the risk management of MCs is the large spatiotemporal variation, lack of sufficient field measurement, and complicated relationships between different forms of nutrients. From the results of the intercept-only model, most variations in observed MCs were quite dependent upon the temporal sampling scale. Thus, seasonal variation in environmental conditions should be considered when setting the targets of nutrient control (Tong et al., 2019). Variations from inter-lake were likely stronger than those from different sites inside the lake, which indicated the importance of the regional effect. However, the variations from inter-lake in predicting dissolved MCs were likely stronger than those in predicting intracellular MCs. This was in line with the previous study in the Midwestern US that found a positive association between MCs and lake latitude (Graham et al., 2004).

Insights gained from linkage among multiple variables could be sharpened by considering the simultaneous effects of biotic and abiotic conditions. The LASSO regression may provide more accurate predictions of MCs concentrations under the conditions of collinearity (Yuan et al., 2014). When all environmental variables were considered together, the total biomass of *Microcystis* was the best biological variable for predicting toxin quota; thereby implying that all detected MC-producing genotypes

were likely to belong to the cyanobacterium *Microcystis* (Ye et al., 2009). In addition，the biomass of toxic *Microcystis aeruginosa* did not achieve the same prediction accuracy as the total *Microcystis* biomass. It is reasonable to infer that other morphospecies such as *Microcystis viridis* might also produce considerable amounts of MCs (Shan et al., 2019a; Wu et al., 2017).

Because of multiple sampling sites within lakes, causal relationships between environmental drivers and MCs showed stronger evidence than analyses of a single lake or a snapshot sampling. The biomass of *Microcystis* was found to be positively correlated with TP and negatively with DIN, buttressing previous findings in San Francisco Bay by Lehman et al. (2013). On the other hand, the trends from multivariate analysis also reinforced that the tradeoff between the costs and benefits of MCs production as N-rich secondary metabolites reduced disproportionately under N-limitation (Horst et al., 2014; Monchamp et al., 2014). In agreement with results from the analysis of the US continental-scale data, TN contributed a higher proportion of the variation in MCs in water columns than TP (Beaver et al., 2014; Yuan et al., 2017). There is, however, strong evidence that the relationships between MCs and nutrients were more complex rather than a hypothesized linear response due to the variations in strain within species (Shan et al., 2019a).

4.2 Rationality and limitations of the proposed framework

It is usually difficult to imitate the dynamics of MCs *in situ* by mathematical equations, because of their complicated fate in the aquatic environment (Wörmer et al.,

2011). In this study, a Bayesian modelling framework that accommodates rigorous uncertainty analysis was proposed to quantify the risk of MCs. The iterative nature of the Bayesian theorem can incorporate existing knowledge and update the joint distribution as new information, thereby developing a site-specific model using a local dataset (Arhonditsis et al., 2008; Cha et al., 2014). Our proposed model was developed based on the across-lake dataset and therefore achieve the necessary statistical purpose by increasing the sample size (Malve and Qian, 2006). Multi-lake data are incorporated into empirical models to broaden the sample size and stabilize the inference, while the resulting model may not be very relevant to anyone lake.

Biotic and abiotic variables are constantly numeric, and discretizing continuous variables into a finite set of states is a key step in implementing the Bayesian network. In previous studies, the discretization of a continuous variable has relied on expert experience, recognized thresholds, and frequency distribution of response nodes (Lucena-Moya et al., 2015). Because the conclusions may rely on the choice of discretization method, Nojavan et al. (2017) suggested that the discretization of continuous variables should be avoided if possible. This study took advantage of a series of conditional probability distributions to replace the conditional probability tables, to avoid discretizing continuous variables (Qian and Miltner, 2015).

For heuristic purposes, this study established a Bayesian network that represents the hypothesized causal connections among environmental factors, biological biomasses, and MCs concentrations. However, this approach has a limitation that empirical regression allows variables to be modelled with linear relationships. The

Bayesian inferential method provides an entire predictive distribution for the response variables over which inference can be made instead of using a point estimate, such as the mean value (Stow et al, 2006). In addition, a stereotype of the method without considering lake-specific environmental gradients could lead to some problems (Taranu et al., 2012), because the structure of the DAG model is a dominant source of uncertainty. Hence, the repetition of the building process in other lakes may be preferable to the indiscriminate use of it.

4.3 Implications for future research and water quality management

Water pollution in China poses a huge threat to the environment and human health (Yang et al., 2013). The Chinese government has invested a large amount of money on it. For example, ~100 billion RMB (~US $14 billion) has been invested in Lake Taihu ecosystem restoration. However, nutrient concentrations and cyanobacterial blooms have not been mitigated as quickly as expected (Qin et al., 2019). Long-term nutrient trends confirmed that TP concentrations were relatively stable or has possibly increased over the last decade, despite the decline of TN concentrations (Xu et al., 2017). Considering the long hydraulic residence time in three studied lakes, a legacy of the internal loading, especially P, is a formidable problem for the rapid recovery of water quality (Shan et al., 2014). In general, TP concentration was the principal force driving cyanobacteria's contribution to total algal biomass (Wagner and Adrian, 2009). Our results reinforced the viewpoint that P is the main element regulating *Microcystis* biomass, whereas N may influence the overall toxicity of blooms. We recommend the

importance of dual N and P reduction in the future management of toxic cyanobacterial blooms.

Furthermore, prudent sustainable management of MCs will require the consideration of the background of limnologic conditions and effect of increasing water temperature, due to that the efforts of nutrient reductions in controlling toxic cyanobacterial blooms may be counteracted by the effect of increasing temperature (Lürling et al., 2017; Richardson et al., 2018). This partly conforms to the field observations in Grand Lake St. Marys, western Ohio, U.S. (Walls et al., 2018) and field survey in 137 European lakes (Mantzouki et al., 2018). Monte Carlo simulation indicated that the highly hazardous risk of *Microcystis* and microcystins were controlled by achieving the TN and TP thresholds at below 0.8 mg/L and 0.05 mg/L, which were previously estimated by a nutrient dilution bioassay in Lake Taihu (Xu et al., 2014). Nevertheless, managing all lakes to a single TN and TP concentration is infeasible. Due to differences between lake ecoregions in China, effective management strategies require a good understanding of the influence of nutrients in different regions (Liang et al., 2019).

In contrast to the effect of TN on MCs, TP thresholds under a range of possible windows exhibited significant differences between Lake Dianchi and the other two lakes. The simulation results indicated that it was important to implement stricter control objectives of TP in Lake Dianchi. In comparison, the low concentrations of dissolved MCs in Lake Dianchi might be attributed to the photodegradation under high UV radiation; however, toxigenic genera could form the MC-protein complexes that

prevent proteolytic degradation within the cell (Melssner et al., 2013; Su et al., 2019). Our results suggested controlling toxic cyanobacterial blooms in lakes within low latitudes should strictly control nutrients and focus on the cell quota instead of extracellular toxin in water columns alone.

5. Conclusions

In this study, a Bayesian modelling framework incorporating biotic and abiotic factors was proposed to predict the risk of MCs. Using data from three bloom-dominated lakes in China, our approach can aid in understanding the causal link between key factors and MCs concentrations, by which researchers and decision-makers can partly infer and predict future MCs scenarios. The results demonstrate the estimated TP thresholds are crucial for reducing the biomass of *Microcystis*. More importantly, the estimated TN thresholds for controlling cyanotoxin can be counteracted by the effect of increasing temperature.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No.51609229; 41701247; 51979262), Chongqing Science and Technology Commission (No. cstc2017jcyjAX0241; cstc2018jscx-msyb1133) and National Key Scientific and Technological Project of China (2014ZX07104-006). The field data in three studied lakes of China was financed by the National Basic Research Program of China (2008CB418006).

References

Arhonditsis, G.B., Papantou, D., Zhang, W., Perhar, G., Massos, E., Shi, M., 2008. Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. Journal of Marine Systems 73(1-2), 8-30.

Alexova, R., Fujii, M., Birch, D., Cheng, J., Waite, T.D., Ferrari, B.C., Neilan, B.A., 2011. Iron uptake and toxin synthesis in the bloom-forming *Microcystis aeruginosa* under iron limitation. Environmental Microbiology 13(4), 1064-1077.

Beaver, J.R., Manis, E.E., Loftin, K.A., Graham, J.L., Pollard, A.I., Mitchell, R.M., 2014. Land use patterns, ecoregion, and microcystin relationships in US lakes and reservoirs: a preliminary evaluation. Harmful Algae 36, 57-62.

Beaver, J.R., Tausz, C.E., Scotese, K.C., Pollard, A.I., Mitchell, R.M., 2018. Environmental factors influencing the quantitative distribution of microcystin and common potentially toxigenic cyanobacteria in US lakes and reservoirs. Harmful Algae 78, 118-128.

Beversdorf, L.J., Miller, T.R., McMahon, K.D., 2015. Long-term monitoring reveals carbon–nitrogen metabolism key to microcystin production in eutrophic lakes. Frontiers in Microbiology 6, 456.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees. CRC press.

Bui, T., Dao, T.S., Vo, T.G., Lürling, M., 2018. Warming Affects Growth Rates and Microcystin Production in Tropical Bloom-Forming *Microcystis* Strains. Toxins 10(3), 123.

Burch, M.D., 2008. Effective doses, guidelines & regulations. In Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs (pp. 831-853). Springer, New York, NY.

Cha, Y., Stow, C.A., 2014. A Bayesian network incorporating observation error to predict phosphorus and chlorophyll a in Saginaw Bay. Environmental Modelling & Software 57, 90-100.

Dahlgren, J.P., 2010. Alternative regression methods are not considered in Murtaugh (2009) or by ecologists in general. Ecology Letters 13(5), E7-E9.

Daly, R.I., Ho, L., Brookes, J.D., 2007. Effect of chlorination on *Microcystis aeruginosa* cell integrity and subsequent microcystin release and degradation. Environmental Science & Technology 41(12), 4447-4453.

Davis, T.W., Berry, D.L., Boyer, G.L., Gobler, C.J., 2009. The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of *Microcystis* during cyanobacteria blooms. Harmful Algae 8(5), 715-725.

Falconer, I.R., Burch, M.D., Steffensen, D.A., Choice, M., Coverdale, O.R., 1994. Toxicity of the blue-green alga (cyanobacterium) *Microcystis aeruginosa* in drinking water to growing pigs, as an animal model for human injury and risk assessment. Environmental Toxicology and Water Quality 9(2), 131-139.

Friedman, J., Hastie, T., Tibshirani, R., 2009. glmnet: Lasso and elastic-net regularized

generalized linear models. R package version, 1(4).

Graham, J.L., Jones, J.R., Jones, S.B., Downing, J.A., Clevenger, T.E., 2004. Environmental factors influencing microcystin distribution and concentration in the Midwestern United States. Water Research 38(20), 4395-4404.

Harke, M.J., Steffen, M.M., Gobler, C.J., Otten, T.G., Wilhelm, S.W., Wood, S.A., Paerl, H.W., 2016. A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, *Microcystis* spp. Harmful Algae 54, 4-20.

Hooten, M.B., Hobbs, N.T., 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs 85(1), 3-28.

Horst, G.P., Sarnelle, O., White, J.D., Hamilton, S.K., Kaul, R.B., Bressie, J.D., 2014. Nitrogen availability increases the toxin quota of a harmful cyanobacterium, *Microcystis aeruginosa*. Water Research 54, 188-198.

Hu, L., Shan, K., Lin, L., Shen, W., Huang, L., Gan, N., Song, L., 2016. Multi-year assessment of toxic genotypes and microcystin concentration in northern lake Taihu, China. Toxins 8 (1), 23.

Huisman, J., Codd, G.A., Paerl, H.W., Ibelings, B.W., Verspagen, J.M., Visser, P.M., 2018. Cyanobacterial blooms. Nature Reviews Microbiology 16(8), 471-483.

Izydorczyk, K., Carpentier, C., Mrówczyński, J., Wagenvoort, A., Jurczak, T., Tarczyńska, M., 2009. Establishment of an Alert Level Framework for cyanobacteria in drinking water resources by using the Algae Online Analyser for monitoring cyanobacterial chlorophyll a. Water Research 43, 989-996.

Jankowiak, J., Hattenrath‐Lehmann, T., Kramer, B.J., Ladds, M., Gobler, C.J., 2019.

Deciphering the effects of nitrogen, phosphorus, and temperature on cyanobacterial bloom intensification, diversity, and toxicity in western Lake Erie. Limnology and Oceanography 64, 1347-1370.

Kardinaal, W.E.A., Janse, I., Kamst-van Agterveld, M., Meima, M., Snoek, J., Mur, L. R., Huisman, J., Zwart, G., Visser, P.M., 2007. *Microcystis* genotype succession in relation to microcystin concentrations in freshwater lakes. Aquatic Microbial Ecology 48 (1), 1-12.

Kelly, N.E., Javed, A., Shimoda, Y., Zastepa, A., Watson, S., Mugalingam, S., Arhonditsis, G.B., 2019. A Bayesian risk assessment framework for microcystin violations of drinking water and recreational standards in the Bay of Quinte, Lake Ontario, Canada. Water Research 162, 288-301.

Liang, Z., Qian, S.S., Wu, S., Chen, H., Liu, Y., Yu, Y., Yi, X., 2019. Using Bayesian change point model to enhance understanding of the shifting nutrients-phytoplankton relationship. Ecological Modelling 393, 120-126.

Lehman, P.W., Marr, K., Boyer, G. L., Acuna, S., Teh, S.J., 2013. Long-term trends and causal factors associated with *Microcystis* abundance and toxicity in San Francisco Estuary and implications for climate change impacts. Hydrobiologia 718(1), 141-158.

Link, W.A., Barker, R.J., 2009. Bayesian inference: with ecological applications. Academic Press.

Lucena-Moya, P., Brawata, R., Kath, J., Harrison, E., ElSawah, S., Dyer, F., 2015. Discretization of continuous predictor variables in Bayesian networks: an

ecological threshold approach. Environmental Modelling & Software 66, 36-45.

Lürling, M., Van Oosterhout, F., Faassen, E., 2017. Eutrophication and warming boost cyanobacterial biomass and microcystins. Toxins 9(2), 64.

MacKintosh, C., Beattie, K.A., Klumpp, S., Cohen, P., Codd, G.A., 1990. Cyanobacterial microcystin-LR is a potent and specific inhibitor of protein phosphatases 1 and 2A from both mammals and higher plants. FEBS Lett 264, 187–192.

Mantzouki, E., Lürling, M., Fastner, J., de Senerpont Domis, L., Wilk-Woźniak, E., Koreivienė, J., ... & Walusiak, E., 2018. Temperature effects explain continental scale distribution of cyanobacterial toxins. Toxins 10(4), 156.

Malve, O., Qian, S.S., 2006. Estimating nutrients and chlorophyll a relationships in Finnish lakes. Environmental Science & Technology 40(24), 7848-7853.

Meissner, S., Fastner, J., Dittmann, E., 2013. Microcystin production revisited: conjugate formation makes a major contribution. Environmental Microbiology, 15(6), 1810-1820.

Monchamp, M.E., Pick, F.R., Beisner, B.E., Maranger, R., 2014. Nitrogen forms influence microcystin concentration and composition via changes in cyanobacterial community structure. PloS One 9(1): e85573.

O'neil, J.M., Davis, T.W., Burford, M.A., Gobler, C.J., 2012. The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. Harmful Algae 14, 313-334.

Orihel, D.M., Bird, D.F., Brylinsky, M., Chen, H., Donald, D.B., Huang, D.Y., Giani,

A., Kinniburgh, D., Kling, H., Kotak, B.G., Leavitt, P.R., Nielsen, C.C., Reedyk, S., Rooney, R.C., Watson, S.B., Zurawell, R.W., Vinebrooke, R.D., 2012. High microcystin concentrations occur only at low nitrogen-to-phosphorus ratios in nutrient-rich Canadian lakes. Canadian Journal of Fisheries and Aquatic Sciences 69(9), 1457-1462.

Otten, T.G., Xu, H., Qin, B., Zhu, G., Paerl, H.W., 2012. Spatiotemporal patterns and ecophysiology of toxigenic *Microcystis* blooms in Lake Taihu, China: implications for water quality management. Environmental Science & Technology 46 (6), 3480-3488.

Plummer, M., 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing (Vol. 124, No. 125.10).

Qian, S.S., Reckhow, K.H., 2007. Combining model results and monitoring data for water quality assessment. Environmental Science & Technology 41(14), 5008-5013.

Qian, S.S., Miltner, R.J., 2015. A continuous variable Bayesian networks model for water quality modeling: a case study of setting nitrogen criterion for small rivers and streams in Ohio, USA. Environmental Modelling & Software 69, 14-22.

Qian, S.S., 2016. Environmental and ecological statistics with R. Chapman and Hall/CRC.

Qin, B., Paerl, H.W., Brookes, J.D., Liu, J., Jeppesen, E., Zhu, G., ... & Deng, J., 2019. Why Lake Taihu continues to be plagued with cyanobacterial blooms through

10 years (2007–2017) efforts. Science Bulletin 64(6), 354-356.

Recknagel, F., Orr, P.T., Bartkow, M., Swanepoel, A., Cao, H., 2017. Early warning of limit-exceeding concentrations of cyanobacteria and cyanotoxins in drinking water reservoirs by inferential modelling. Harmful algae 69, 18-27.

Reynolds, C.S., Irish, A.E., 1997. Modelling phytoplankton dynamics in lakes and reservoirs: the problem of in-situ growth rates. Hydrobiologia 349(1-3), 5-17.

Richardson, J., Miller, C., Maberly, S.C., Taylor, P., Globevnik, L., Hunter, P., ... & Søndergaard, M., 2018. Effects of multiple stressors on cyanobacteria abundance vary with lake type. Global change biology 24(11), 5044-5055.

Rigosi, A., Hanson, P., Hamilton, D.P., Hipsey, M., Rusak, J. A., Bois, J., Sparber, K., Chorus, I., Watkinson, A.J., Qin, B., Kim, B., Brookes, J.D., 2015. Determining the probability of cyanobacterial blooms: the application of Bayesian networks in multiple lake systems. Ecological Applications 25, 186-199.

Singh, S., Rai, P.K., Chau, R., Ravi, A.K., Neilan, B.A., Asthana, R.K., 2015. Temporal variations in microcystin-producing cells and microcystin concentrations in two fresh water ponds. Water Research 69, 131-142.

Shan, K., Li, L., Wang, X., Wu, Y., Hu, L., Yu, G., Song, L., 2014. Modelling ecosystem structure and trophic interactions in a typical cyanobacterial bloom-dominated shallow Lake Dianchi, China. Ecological Modelling 291, 82-95.

Shan, K., Shang, M., Zhou, B., Li, L., Wang, X., Yang, H., Song, L., 2019a. Application of Bayesian network including *Microcystis* morphospecies for microcystin risk assessment in three cyanobacterial bloom-plagued lakes, China. Harmful

Algae 83, 14-24.

Shan, K., Song, L., Chen, W., Li, L., Liu, L., Wu, Y., Jia, Y., Zhou, Q., Peng, L., 2019b. Analysis of environmental drivers influencing interspecific variations and associations among bloom-forming cyanobacteria in large, shallow eutrophic lakes. Harmful algae 84, 84-94.

Stow, C.A., Reckhow, K.H., Qian, S.S., 2006. A Bayesian approach to retransformation bias in transformed regression. Ecology 87(6), 1472-1477.

Su, M., Andersen, T., Burch, M., Jia, Z., An, W., Yu, J., Yang, M., 2019. Succession and interaction of surface and subsurface cyanobacterial blooms in oligotrophic/mesotrophic reservoirs: A case study in Miyun Reservoir. Science of the Total Environment 649, 1553-1562.

Taranu, Z.E., Zurawell, R.W., Pick, F., Gregory-Eaves, I., 2012. Predicting cyanobacterial dynamics in the face of global change: the importance of scale and environmental context. Global Change Biology 18, 3477-3490.

Taranu, Z.E., Gregory-Eaves, I., Steele, R.J., Beaulieu, M., Legendre, P., 2017. Predicting microcystin concentrations in lakes and reservoirs at a continental scale: A new framework for modelling an important health risk factor. Global Ecology and Biogeography 26(6), 625-637.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

Tong, Y., Xu, X., Zhang, S., Shi, L., Zhang, X., Wang, M., Qi, M., Chen, C., Wen, Y., Zhao, Y., Zhang, W., Lu, X., 2019. Establishment of season-specific nutrient

thresholds and analyses of the effects of nutrient management in eutrophic lakes through statistical machine learning. Journal of Hydrology 578, 124079.

U.S. EPA, 2000. Nutrient Criteria Technical Guidance Manual. Technical Report EPA-822-B00-022. United States Environmental Protection Agency, Office of Water, Washington, D.C.

Wagner, C., Adrian, R., 2009. Cyanobacteria dominance: quantifying the effects of climate change. Limnology and Oceanography 54(6part2), 2460-2468.

Walls, J.T., Wyatt, K.H., Doll, J.C., Rubenstein, E.M., Rober, A.R., 2018. Hot and toxic: Temperature regulates microcystin release from cyanobacteria. Science of the Total Environment 610, 786-795.

Whitehead, P., Young, P., 1979. Water quality in river systems: Monte-Carlo analysis. Water Resources Research 15(2), 451-459.

Wiedner, C., Visser, P.M., Fastner, J., Metcalf, J.S., Codd, G.A., Mur, L.R., 2003. Effects of light on the microcystin content of *Microcystis* strain PCC 7806. Applied and Environmental Microbiology 69 (3), 1475-1481.

Widder, S., Allen, R.J., Pfeiffer, T., Curtis, T.P., Wiuf, C., Sloan, W.T., ... Kettle, H., 2016. Challenges in microbial ecology: building predictive understanding of community function and dynamics. The ISME journal 10, 2557–2568.

Wood, S.N., 2001. mgcv: GAMs and generalized ridge regression for R. R news 1(2), 20-25.

World Health Organization (WHO), 1998. World Health Organization (WHO) Guidelines for Drinking Water Quality, vol. 1, second ed., Recommendations

/World Health Organization, World Health Organization, Geneva, Switzerland (1998) 36 pp.

Wörmer, L., Cirés, S., Quesada, A., 2011. Importance of natural sedimentation in the fate of microcystins. Chemosphere 82(8), 1141-1146.

Wu, Y., Li, L., Gan, N., Zheng, L., Ma, H., Shan, K., Liu, J., Xiao, B., Song, L., 2014. Seasonal dynamics of water bloom-forming *Microcystis* morphospecies and the associated extracellular microcystin concentrations in large, shallow, eutrophic Dianchi Lake. Journal of Environmental Sciences 26, 1921-1929.

Wu, Z., Liu, Y., Liang, Z., Wu, S., Guo, H., 2017. Internal cycling, not external loading, decides the nutrient limitation in eutrophic lake: A dynamic model with temporal Bayesian hierarchical inference. Water Research 116, 231-240.

Xu, H., Paerl, H.W., Qin, B., Zhu, G., Hall, N.S., Wu, Y., 2014. Determining critical nutrient thresholds needed to control harmful cyanobacterial blooms in eutrophic Lake Taihu, China. Environmental Science & Technology 49(2), 1051-1059.

Xu, H., Paerl, H.W., Zhu, G., Qin, B., Hall, N.S., Zhu, M., 2017. Long-term nutrient trends and harmful cyanobacterial bloom potential in hypertrophic Lake Taihu, China. Hydrobiologia 787, 229-242.

Yang, H., Flower, R.J., Thompson, J.R., 2013. Sustaining China's water resources. Science 339(6116), 141-141.

Ye, W., Liu, X., Tan, J., Li, D., Yang, H., 2009. Diversity and dynamics of microcystin-Producing cyanobacteria in China's third largest lake, Lake Taihu. Harmful

Algae 8(5), 637-644.

Yuan, L.L., Pollard, A.I., Pather, S., Oliver, J.L., D'Anglada, L., 2014. Managing microcystin: identifying national-scale thresholds for total nitrogen and chlorophyll a. Freshwater Biology 59(9), 1970-1981.

Yuan, L.L., Pollard, A.I., 2017. Using National-scale data to develop nutrient – microcystin relationships that guide management decisions. Environmental Science & Technology 51(12), 6972-6980.

Yuan, L.L., Pollard, A.I., 2019. Combining national and state data improves predictions of microcystin concentration. Harmful algae 84, 75-83.

Zhang, M., Xie, P., Xu, J., Liu, B., Yang, H., 2006. Spatiotemporal variations of internal P loading and the related mechanisms in the large shallow Lake Chaohu. Sci. China Ser. D Earth Sci. 49, 72–81.

Fig. 1. The logic of the modelling framework used to predict *Microcystis* and microcystin risks based on across-lake and lake-specific dataset.

Fig. 2. (a) Conceptual model that links environmental and biological factors to MC concentration. (b) When observation data are available, the initial nodes can be divided into predictor variables (observed accurately) and response variables (observed with errors).

Fig. 3. The contour lines provided a visualization of the GAM with relationships between nutrients and MCs concentrations. Filled circles: observed values of different forms of nutrients; Contours: predicted mean cell-bound MCs (Fig. 3a and b) and dissolved MCs (Fig. 3c and d) by the GAM associated with each combination of N and P.

Fig. 4. Marginal distribution (hollow histograms) of (a) Log TP and (b) Log TN are compared with their conditional distribution - the distribution corresponding to B_M < 0.6 mg/L and MCs < 0.4 μg/L (grey histograms). The 75^{th} percentiles of the distribution are shown as a black solid line for marginal distribution and grey dashed line for conditional distribution. The 25th percentiles of the distribution of all data are shown as a black dashed line.

Fig. 5. Probability of achieving the low risks of *Microcystis* biomass and microcystin

concentration in different lakes over a range of TP and TN concentrations using updated lake-specific models. The solid line represents the results of the across-lake model. The orange solid and dash lines represent the potential TN or TP thresholds.

Fig. 6. Predicted probabilities of *Microcystis* biomass < 1.5 mg/L and MC concentration < 1.0 μg/L are shown as a function of water temperature and (a) TP or (b) TN.

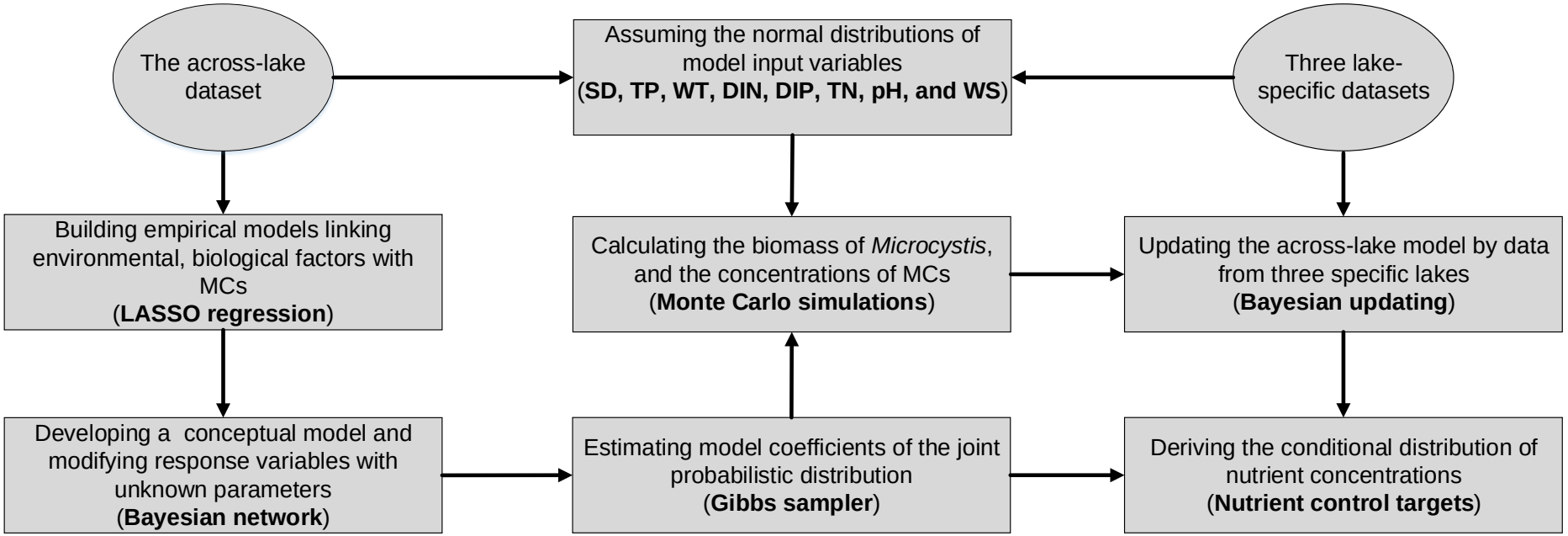


Fig. 1. The logic of the modelling framework used to predict *Microcystis* and microcystin risks based on across-lake and lake-specific dataset.

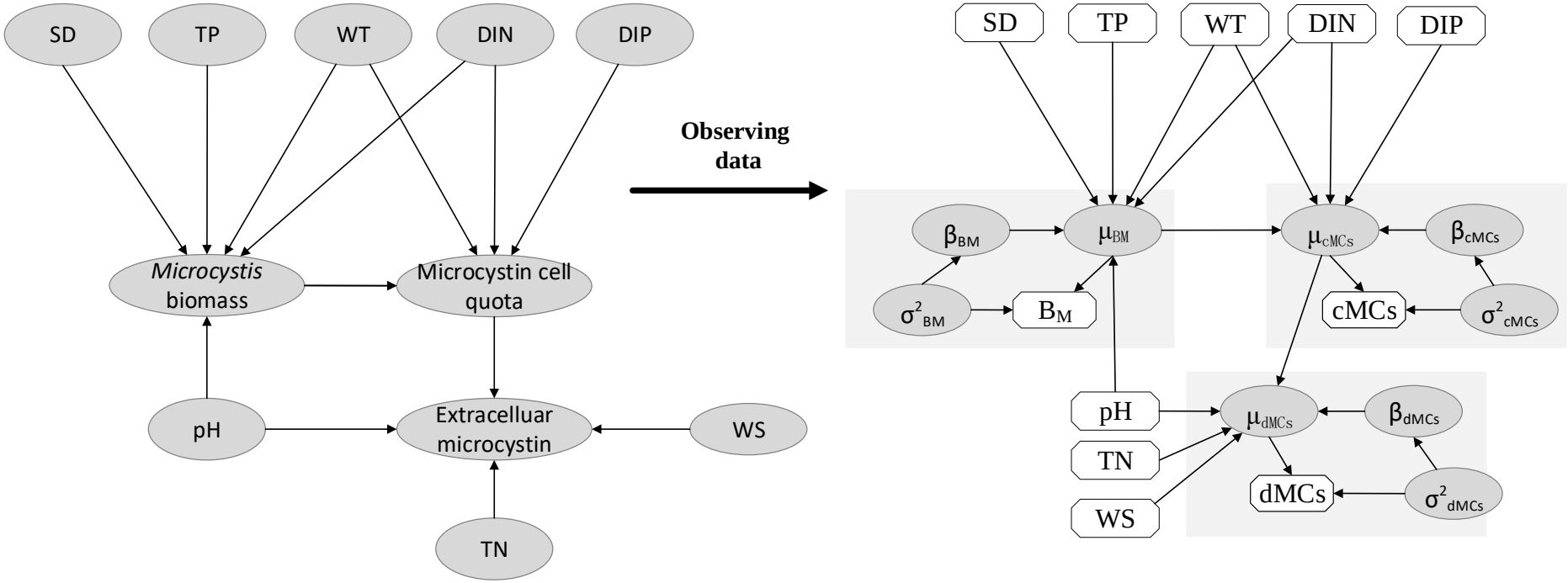


Fig. 2. (a) Conceptual model that links environmental and biological factors to microcystin concentration. (b) When observation data are available, the initial nodes can be divided into predictor variables (observed accurately) and response variables (observed with errors).

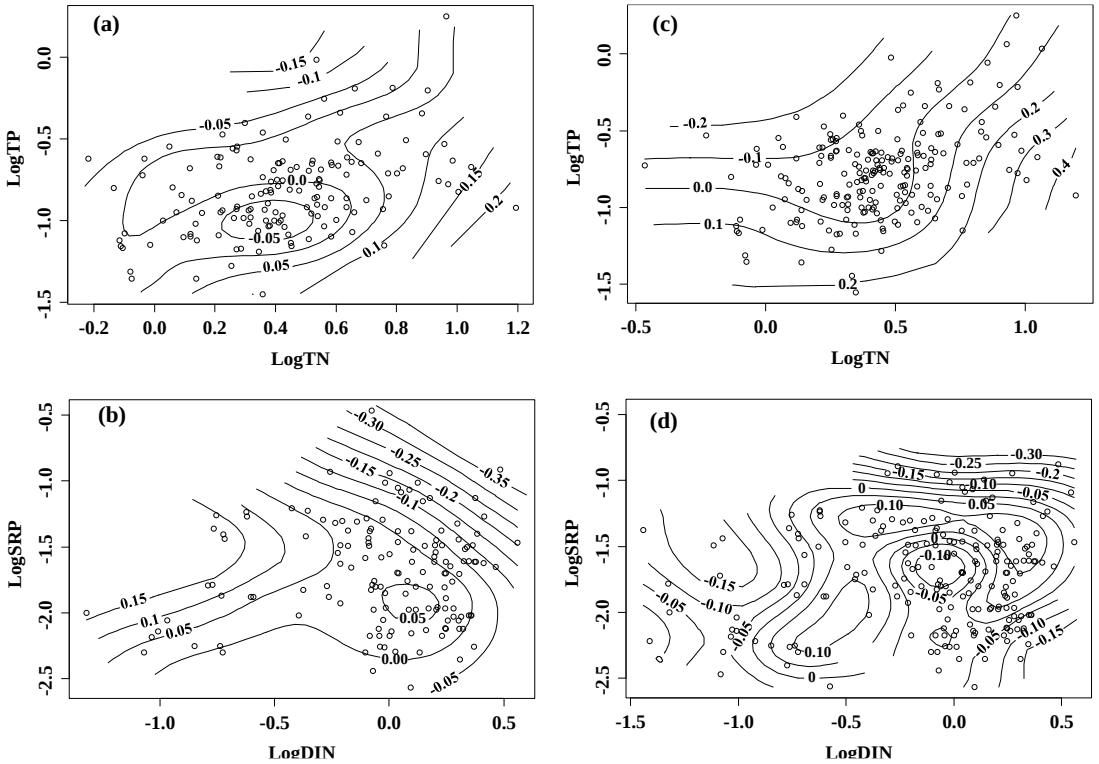


Fig. 3. Relationships between nutrients and MCs concentrations. Filled circles: observed values of TN, TP, DIN, and DIP; Contours: predicted mean cell-bound MCs (Fig. 4a and b) and dissolved MCs (Fig. 4c and d) associated with each combination of N and P. The contour plot shows two variables smoothly fitted by general addictive models.

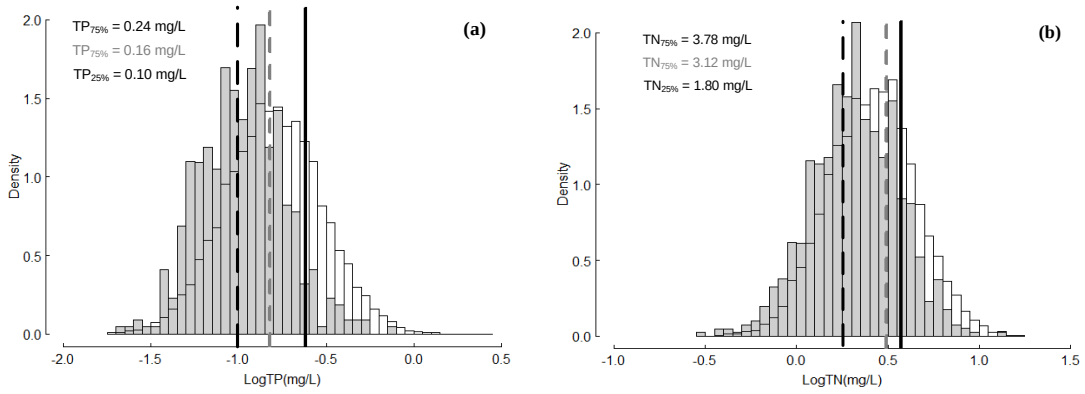


Fig. 4. Marginal distribution (hollow histograms) of (a) LogTP and (b) LogTN are compared with their conditional distribution - the distribution corresponding to $B_M < 0.6$ mg/L and MCs < 0.4 μg/L (grey histograms). The 75th percentiles of the distribution are shown as black solid line for marginal distribution and grey dashed line for conditional distribution. The 25th percentiles of the distribution of all data are shown as black dashed line.

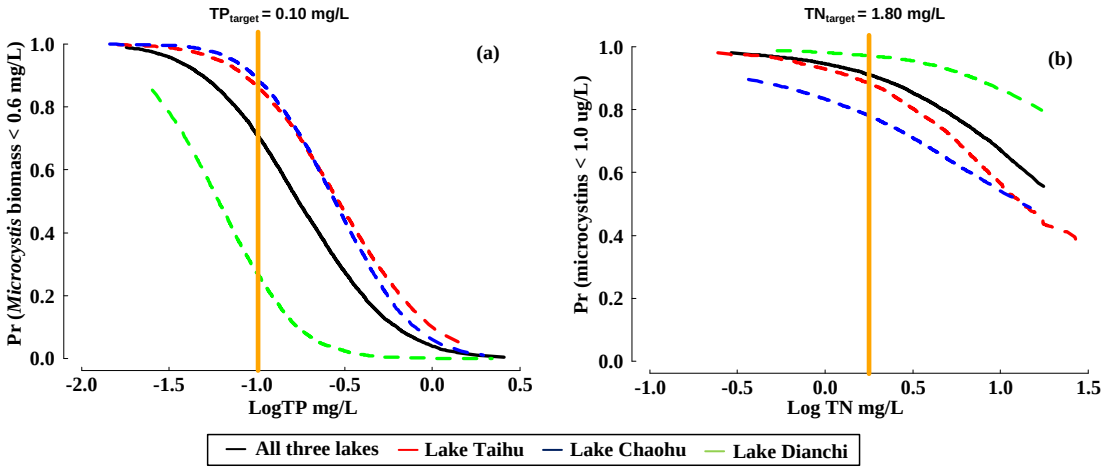
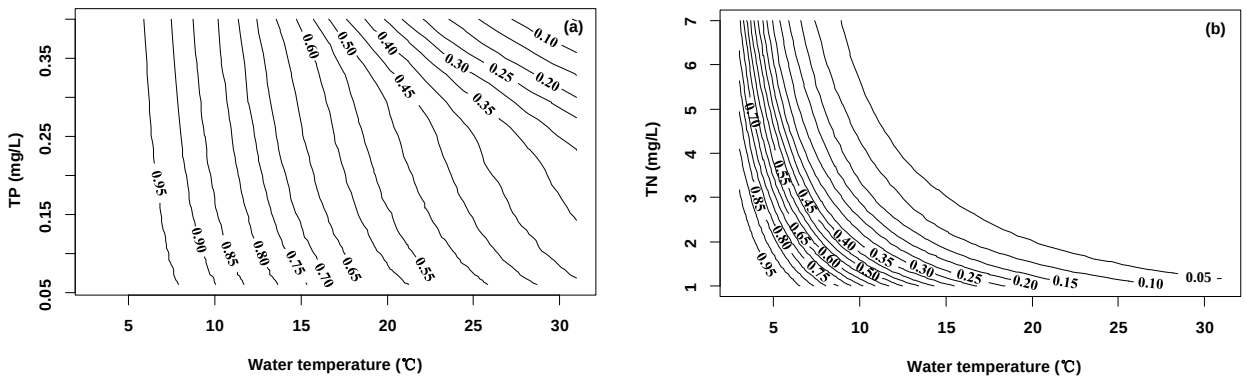


Fig.5. Probability of achieving the low risks of *Microcystis* biomass and microcystin concentration in different lakes over a range of TP and TN concentrations using updated lake-specific models. Solid line represents the results from the across-lake model. The orange bar represents the critical nutrient thresholds for TP (0.10 mg/L) and TN (1.80 mg/L).

23



24

25 **Fig. 6.** Predicted probabilities of *Microcystis* biomass < 1.5 mg/L and MC concentration < 1.0

26 μg/L are shown as a function of water temperature (WT) and (a) TP or (b) TN.

27

28

Table 1 Thresholds for *Microcystis* biomass and microcystin in the Alert Levels Framework

Node	Definition	Units	Thresholds	Level	References
MCs	Microcystin concentrations	μg/L	MCs<0.4		
			0.4≤MCs<1.0	Alert Level 1	(Falconer et al., 1994)
			MCs≥1.0	Alert Level 2	
B_M	Biomass of total *Microcystis* (*Note:*1000 cell/mL = 0.3 mg/L)	mg/L	B_M<0.6		
			0.6≤B_M<1.5	Alert Level 1	(Izydorczyk et al., 2009)
			B_M≥1.5	Alert Level 2	

Table 2 Observed variation in microcystin concentrations were partitioned into four different components by "intercept-only" models.

Response variables	Different component of variation	Standard Deviation	Variation	Proportion in total variation
Cell-bound microcystin	Lake	0.090	0.008	9.8%
	Site/Lake	0.057	0.003	3.6%
	Month	0.249	0.062	75.6%
	Residual	0.096	0.009	11.0%
Dissolved microcystin	Lake	0.137	0.0187	14.6%
	Site/Lake	0.108	0.0117	9.2%
	Month	0.220	0.0486	38.0%
	Residual	0.221	0.0488	38.2%

Table 3 Results of applying the least absolute shrinkage and selection operator techniques to explore the relationship between a subset of environmental variables and different levels of biological factors or microcystin concentrations. All water quality variables were log (x) transformed, and standardized to a mean value of zero and a standard deviation of 1. Entries for the predictors are regression coefficients for the variables included in the model. The value of the cross-validation mean squared prediction error (MSPE) and its standard deviation reported for each of models. R^2 is the coefficient of determination for a model containing all of the candidate predictors.

Model	Response variables	Explanatory variables									Cross-val. MSPE	R^2
		WT	SD	TN	TP	DIN	DIP	pH	DO	WS		
1	Chl *a*	0.084	-0.019		0.267	-0.026		0.383	0.014		0.589±0.032	0.420
2	B_{cya}	0.095			0.084	-0.057		0.354		-0.053	0.593±0.024	0.342
3	B_M	0.115	-0.006		0.156	-0.040		0.297			0.233±0.014	0.518
4	B_{MA}	0.011			0.033			0.024			0.089±0.007	0.094
5	cMCs	0.152				-0.018	-0.009	0.014			0.045±0.005	0.492
6	dMCs	0.073	0.022	0.037				-0.025		-0.030	0.110±0.012	0.159

Table 4 Results of combining different biological factors with environmental factors, predicting concentrations of cell-bound microcystin in three large-shallow lakes. In each row, the MSPE and R^2 were calculated according to aforementioned methods, and the coefficients are linearly shrunk to exactly zero were excluded from models. The models represent the different biological biomass as explanatory variables in the order, including Chl-*a*, B_{cya}, B_M and B_{MA}, respectively.

Model	Biological variables	Environmental variables									Cross-val. MSPE	R^2
		WT	SD	TN	TP	DIN	DIP	pH	DO	WS		
1		0.152				-0.018	-0.009	0.014			0.044±0.004	0.492
2	0.012	0.162		0.008		-0.028	-0.021	0.014			0.042±0.004	0.532
3	0.063	0.148				-0.018	-0.025				0.037±0.004	0.576
4	0.030	0.137				-0.003	-0.004	0.015			0.044±0.005	0.501

54 **Table 5** The mean of *Microcystis* biomass, cell-bound and dissolved microcystin in
55 response to different quantile of nitrogen and phosphorus in studied lakes.

Nutrients	Quantile	Biomass of *Microcystis* (mg/L)	Cell-bound microcystin (mg/g DW)	Dissolved microcystin (μg/L)
Phophorus	5%	3.472	0.328	0.745
	25%	5.446	0.306	0.716
	50%	7.474	0.292	0.698
	75%	10.446	0.283	0.685
	95%	17.694	0.271	0.669
Nitrogen	5%	13.945	0.362	0.662
	25%	9.737	0.312	0.676
	50%	8.018	0.288	0.689
	75%	7.167	0.276	0.713
	95%	6.249	0.261	0.793

56

57

Supplementary information

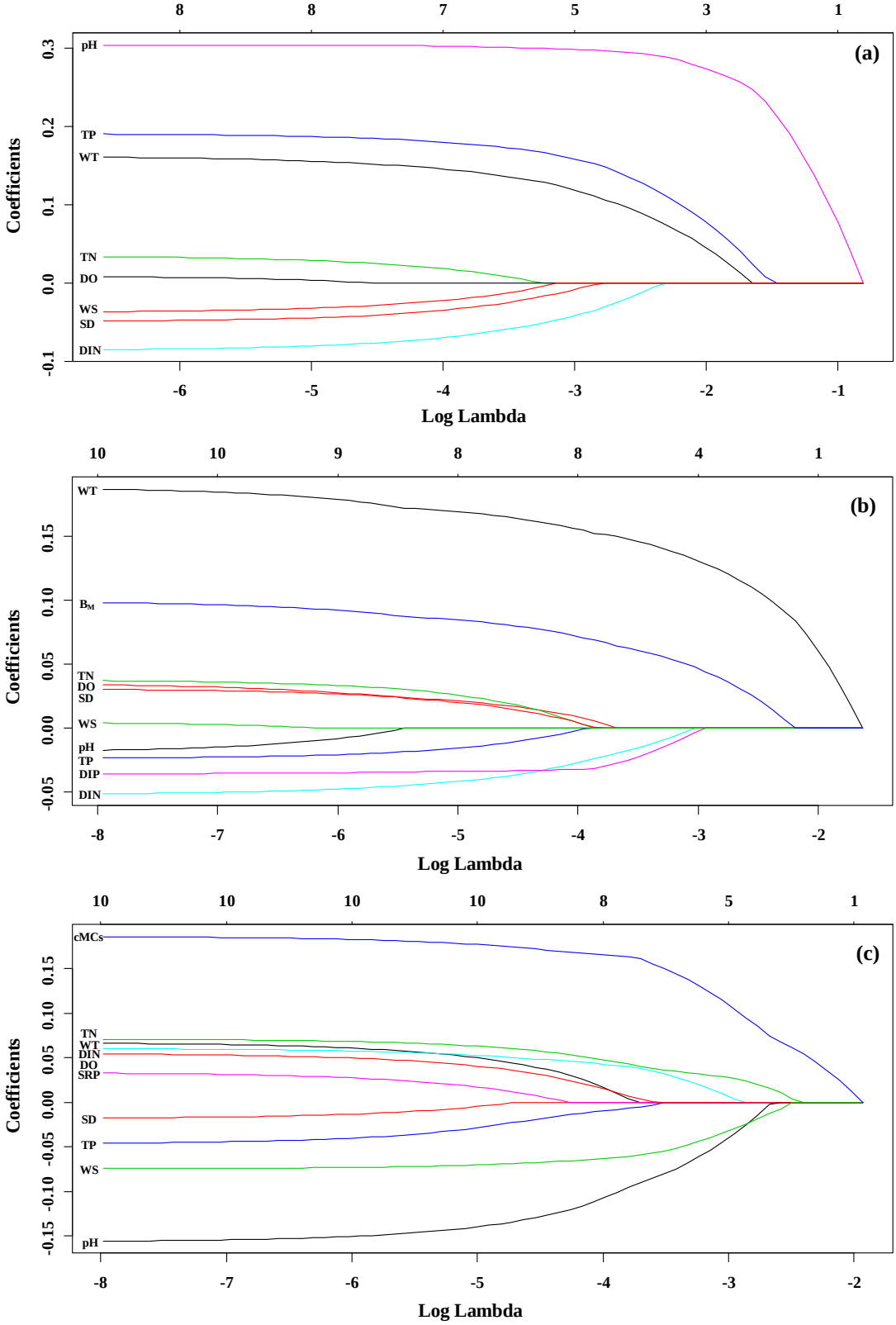


Fig. S1. Profiles of regression coefficient values for different explanatory variables and different values of log λ during LASSO regression for predicting the value of (a) total *Microcystis* biomass, B_M, (b) cell-bound MCs, cMCs, and (c) dissolved MCs, dMCs.

Table S1 The estimated biomass of *Microcystis* (B_M) in Equation (2).

Coef	cBN estimates				Bayesian updating — Taihu				Bayesian updating — Chaohu				Bayesian updating — Dianchi			
	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
β_0^c	-6.65	0.32	-7.28	-6.02	-6.25	0.38	-7	-5.5	-6.33	0.39	-7.09	-5.57	-7.12	0.39	-7.92	-6.36
β_1^c	-0.14	0.04	-0.22	-0.05	-0.1	0.05	-0.19	0	-0.18	0.05	-0.28	-0.07	-0.14	0.05	-0.24	-0.04
β_2^c	0.83	0.04	0.74	0.9	0.77	0.05	0.67	0.87	0.86	0.05	0.76	0.95	0.83	0.05	0.73	0.92
β_3^c	0.56	0.04	0.49	0.64	0.59	0.05	0.5	0.69	0.55	0.05	0.46	0.65	0.54	0.06	0.43	0.64
β_4^c	-0.27	0.03	-0.32	-0.22	-0.26	0.03	-0.33	-0.2	-0.28	0.03	-0.34	-0.21	-0.25	0.03	-0.32	-0.19
β_5^c	8.11	0.35	7.43	8.79	7.52	0.42	6.68	8.34	7.86	0.42	7.01	8.66	8.71	0.43	7.86	9.56
σ_{BM}	0.48	0.01	0.46	0.49	0.49	0.01	0.48	0.51	0.47	0.01	0.45	0.49	0.46	0.01	0.44	0.48

Table S2 The estimated concentration of cell-bound microcystins (cMCs) in Equation (5).

Coef	cBN estimates				Bayesian updating — Taihu				Bayesian updating — Chaohu				Bayesian updating — Dianchi			
	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
β_0^d	-1.46	0.06	-1.59	-1.34	-1.41	0.07	-1.56	-1.27	-1.42	0.07	-1.57	-1.28	-1.59	0.08	-1.75	-1.44
β_1^d	0.52	0.04	0.44	0.6	0.5	0.04	0.41	0.59	0.51	0.04	0.43	0.60	0.55	0.05	0.45	0.65
β_2^d	-0.08	0.03	-0.14	-0.03	-0.08	0.04	-0.15	-0.01	-0.1	0.03	-0.16	-0.04	-0.09	0.04	-0.16	-0.02
β_3^d	-0.12	0.03	-0.18	-0.07	-0.12	0.03	-0.19	-0.05	-0.09	0.03	-0.15	-0.03	-0.17	0.03	-0.23	-0.1
β_4^d	0.11	0.02	0.07	0.15	0.1	0.02	0.06	0.15	0.12	0.02	0.08	0.16	0.11	0.02	0.07	0.16
σ_{cMCs}	0.18	0.01	0.17	0.2	0.19	0.01	0.17	0.2	0.17	0.01	0.15	0.18	0.18	0.01	0.16	0.2

Table S3 The estimated concentration of dissolved microcystins (dMCs) in Equation (8).

Coef	cBN estimates				Bayesian updating — Taihu				Bayesian updating — Chaohu				Bayesian updating — Dianchi			
	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
β_0^e	3.15	0.66	1.85	4.46	2.71	0.85	0.99	4.33	4.03	0.99	2.08	5.94	3.3	0.76	1.87	4.83
β_1^e	-3.28	0.68	-4.61	-1.9	-2.77	0.88	-4.43	-0.99	-4.13	1.02	-6.14	-2.12	-3.43	0.77	-4.97	-1.95
β_2^e	-0.16	0.04	-0.23	-0.09	-0.12	0.04	-0.2	-0.04	-0.22	0.05	-0.32	-0.13	-0.17	0.05	-0.26	-0.09
β_3^e	0.19	0.06	0.07	0.32	0.24	0.07	0.11	0.39	0.16	0.08	-0.01	0.32	0.15	0.08	-0.01	0.31
β_4^e	0.56	0.09	0.38	0.74	0.62	0.1	0.42	0.8	0.58	0.1	0.38	0.78	0.59	0.11	0.38	0.8
σ_{dMCs}	0.32	0.01	0.3	0.34	0.3	0.01	0.28	0.33	0.32	0.02	0.29	0.35	0.31	0.02	0.28	0.34