ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Probabilistic prediction of algal blooms from basic water quality parameters by Bayesian scale-mixture of skew-normal model

To cite this article: Muyuan Liu *et al* 2023 *Environ. Res. Lett.* **18** 014034

IOP Publishing Environ. Res. Lett. 18 (2023) 014034 https://doi.org/10.1088/1748-9326/acaf11

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

Probabilistic prediction of algal blooms from basic water quality parameters by Bayesian scale-mixture of skew-normal model

Muyuan Liu[1], Jing Hu[1], Yuzhou Huang[1], Junyu He[1,2], Kokoette Effiong[1], Tao Tang[1], Shitao Huang[1], Yuvna Devi Perianen[1], Feier Wang[3], Ming Li[5] and Xi Xiao[1,4,6,*]

[1] Ocean College, Zhejiang University, #1 Zheda Road, Zhoushan, Zhejiang 316021, People's Republic of China
[2] Ocean Academy, Zhejiang University, #1 Zheda Road, Zhoushan, Zhejiang 316021, People's Republic of China
[3] College of Environmental & Resource Sciences, Zhejiang University, #866 Yuhangtang Road, Hangzhou, Zhejiang 310058, People's Republic of China
[4] Donghai Laboratory, Zhoushan, Zhejiang 316021, People's Republic of China
[5] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, People's Republic of China
[6] Key Laboratory of Watershed Non-Point Source Pollution Control and Water Eco-Security of Ministry of Water Resources, College of Environmental and Resources Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, People's Republic of China
* Author to whom any correspondence should be addressed.

E-mail: xi@zju.edu.cn

Keywords: algal blooms, water quality, Bayesian hierarchical model, scale-mixture of skew-normal (SMSN) distribution, risk assessments

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
28 September 2022

REVISED
20 December 2022

ACCEPTED FOR PUBLICATION
29 December 2022

PUBLISHED
12 January 2023

Abstract

The timeliness of monitoring is essential to algal bloom management. However, acquiring algal bio-indicators can be time-consuming and laborious, and bloom biomass data often contain a large proportion of extreme values limiting the predictive models. Therefore, to predict algal blooms from readily water quality parameters (i.e. dissolved oxygen, pH, etc), and to provide a novel solution to the modeling challenges raised by the extremely distributed biomass data, a Bayesian scale-mixture of skew-normal (SMSN) model was proposed. In this study, our SMSN model accurately predicted over-dispersed biomass variations with skewed distributions in both rivers and lakes (in-sample and out-of-sample prediction R^2 ranged from 0.533 to 0.706 and 0.412 to 0.742, respectively). Moreover, we successfully achieve a probabilistic assessment of algal blooms with the Bayesian framework (accuracy >0.77 and macro-F_1 score >0.72), which robustly decreased the classic point-prediction-based inaccuracy by up to 34%. This work presented a promising Bayesian SMSN modeling technique, allowing for real-time prediction of algal biomass variations and *in-situ* probabilistic assessment of algal bloom.

1. Introduction

The frequency and intensity of algal blooms have increased globally (Hallegraeff 1993, Ho *et al* 2019, Xiao *et al* 2019a). As one of the most serious environmental problems, algal blooms profoundly threaten local water safety and cause critical losses to aquatic biodiversity (Qin *et al* 2010, Olokotum *et al* 2020, Amorim and Moura 2021). Traditionally, algal biomass monitoring is of primary importance to aquatic health assessment and bloom management (Lee and Lee 1995, Ye *et al* 2014, Wu *et al* 2017). However, limited by the laborious and time-consuming process in field survey and lab analysis (Liu *et al* 2020), the timeliness of algal monitoring is often insufficient to meet demands, e.g. Lake Erhai and coastal waters in Hong Kong, China (Wang *et al* 2018, Guo *et al* 2020). In recent decades, with the expansion of *in-situ* water quality monitoring systems, mathematical approaches using basic physicochemical parameters become promising to improve water management outcomes by achieving real-time biomass predictions (Glibert *et al* 2010, Shmueli and Koppius 2010).

The ecological relationships between algal over-proliferation and basic water quality parameters, such as water temperature, pH, dissolved oxygen (DO), conductivity and clarity, have been widely observed (Weisse 2008, Mantzouki and Visser 2015, Visser *et al*

2016). For instance, the increase in water temperature can boost the algal blooms within an optimal range, and intensive algal photosynthesis usually alters the DO, pH and conductivity of water columns (Flynn *et al* 2015), and reduces water clarity due to the high biomass accumulations (Mantzouki and Visser 2015). Over past years, based upon such tight empirical links, a wide variety of methods have been successfully applied for predicting the algal variations and trends, including deep neural networks (Lee and Lee 2018, Lee *et al* 2022, Liu *et al* 2022), hybrid evolutionary algorithms (Recknagel *et al* 2014, Ye *et al* 2014), and support vector regressions (García-Nieto *et al* 2020). More recently, with the cheaper availability of computation, Bayesian regression has risen in great popularity (Qian *et al* 2019, Zhang *et al* 2019). Bayesian approaches usually have high power of prediction and allow the use of probabilistic paradigm to address the modeling uncertainties (He *et al* 2020). In practice, the employing of Bayesian regression model is often useful for bloom management efforts, especially for analyzing the exceeding risks of algal biomass according to different guidelines (Cha *et al* 2014, Mellios *et al* 2020).

However, algal data samples often have large variances and contain a big proportion of extreme values (Fletcher *et al* 2005), posing great challenges for empirical models (Gelman *et al* 2013, Cusack *et al* 2015, Haakonsson *et al* 2020). To overcome this issue, commonly the data pre-transformation can be a feasible way to scale the data range and eliminate the presence of extreme values, e.g. Box–Cox (Chung *et al* 2007). Recently, as an alternative, the scale-mixture of skew-normal (SMSN) modeling assumption (Branco and Dey 2001) also provides solutions to handle the irregular data characteristics (Benites *et al* 2019). With extra scale factor and shape parameters, the SMSN models can strongly accommodate occasional data and show generate robust modeling analysis in many other study fields (Montenegro and Branco 2016, Silva *et al* 2020, Mirfarah *et al* 2021). Nevertheless, to our knowledge, there have been no previous reports of utilizing this tool to predict algal variations.

In view of the above considerations, our main objectives were to explore the Bayesian SMSN regression to predict algal blooms, by (a) using only basic water quality parameters that are convenient to measure; (b) modeling biomass variations having extreme data distribution characteristics; (c) incorporating probabilistic framework to enhance the assessment accuracy of algal blooms. The Bayesian SMSN models were developed and validated using three ecological datasets with records spanning from 2012 to 2019, which were acquired from one large river system (Zhejiang, China) and two multi-lake systems (Wisconsin, USA) with cyanobacterial and chlorophyll-*a* (Chl-*a*) levels analyzed. The proposed approach achieved real-time prediction of algal biomass dynamics and *in-situ* assessment of algal blooms, supporting water environmental management.

2. Material and methods

2.1. Monitoring data

The Hangjiahu Region Rivers are in the downstream reaches of Lake Taihu, located in Taihu Basin (figure 1). Taihu Basin is one of the most developed areas in China, surrounded by many large cities including Shanghai, Suzhou, Wuxi, and Hangzhou. Over the past decades, aquatic ecosystems in Taihu Basin have continuously suffered eutrophication and harmful cyanobacterial blooms (CyanoHABs) problems due to excessive nutrient inputs (Qin *et al* 2019). During 2018–2019, we sampled Hangjiahu Region rivers at a quarterly frequency (spring: 12–18 April 2018; summer: 13–19 June 2018; fall: 25–31 October 2018; and winter: 2–8 January 2019), in total, there were 31 sites and 124 collected samples (table 1). To acquire cyanobacterial abundance data, the riverine cyanobacteria samples were identified down to the species level (Hu 2006) using a microscope (BX53, Olympus Inc., Japan) in the laboratory and were quantified as cyanobacterial cell biomass. Physicochemical parameters including pH, turbidity, DO, conductivity, photosynthetically active radiation (PAR), water temperature, and water depth, were measured *in-situ* with portable multi-parameter analyzers (YSI EXO2, YSI Inc., U.S.A.), and water transparency was measured with Secchi disk (Secchi disk depth, SDD; Shanghai Changmu Environment Technology Ltd) (table 2).

The two multi-lake districts, i.e. Trout Lake Region and Madison Lakes Region, are located in northern and southern Wisconsin, respectively (figure 1). These lakes are monitored by the North Temperate Lakes-Long Term Ecological Research (NTL-LTER, https://lter.limnology.wisc.edu/) project, and are sampled every 2 weeks during the ice-free season (late March or early April through early September) and every 6 weeks during the ice-covered season. In this study, the Trout lakes dataset was collected in five lakes and two bog lakes from 2015 to 2018, and the Madison lakes dataset was collected in two lakes from 2013 to 2018 (table 1). The Chl-*a* concentrations were analyzed spectrophotometrically, and cyanobacterial samples were identified to species via microscope and were reported as cell biomass. Together, water physicochemical parameters including water temperature, SDD, pH, DO, and PAR were measured at each site with multi-parameter sondes (YSI EXO2, YSI Inc., U.S.A.) (table 2). All data for the two lake systems were obtained from the LTER website (https://lter.limnology.wisc.edu/about/overview).

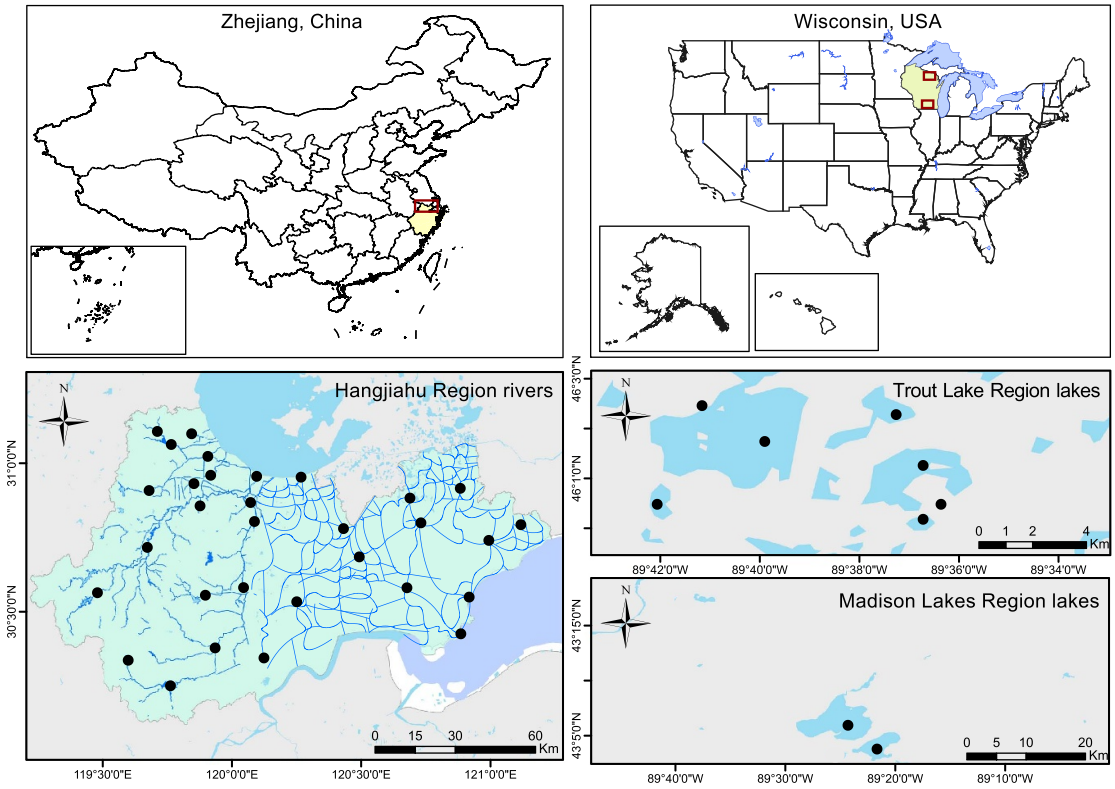

Figure 1. Locations of monitoring sites for the study areas of Hangjiahu, Trout and Madison.

Table 1. Overview of the monitoring datasets.

Study area (number of sites)	Frequency	Sampling period [a]	Samples [b]
Hangjiahu Region rivers (31)	Quarterly	04/2018–10/2018 (01/2019)	93 (31)
Trout Lake Region lakes (7)	Every 2 or 6 weeks	01/2015–11/2017 (01/2018–11/2018)	216 (70)
Madison Lakes Region lakes (2)	Every 2 or 6 weeks	02/2013–11/2016 (02/2017–11/2018)	103 (53)

[a,b] Calibration data and validation data (in parentheses).

2.2. Model development

2.2.1. Bayesian SMSN regression model

Continuous algal data, such as cell biomass, are commonly modeled as a normal or lognormal distribution. However, sampling data often have extreme values that can cause skewness, fat-tailedness, and even multimodality in the distribution, which violated Normal or lognormal assumptions. The use of the SMSN distribution family can well characterize these departures (Benites *et al* 2019). In this case, considering that response variable y (i.e. cyanobacterial biomass and Chl-*a* concentration) were all positive, we assumed it followed a log-SMSN distribution. Formally, the general class of SMSN distributions was given by the location μ, the scale σ^2, the positive scale factor s, and the shape (skewness) λ. Hereby, the log-SMSN model can be restrictively written as following hierarchical representations within the Bayesian framework (Marchenko and Genton 2010, Cabral *et al* 2012) (more details in text S1):

$$\log\,(y_i) \sim \mathrm{SN}\left(\mu_i, \frac{\sigma^2}{s}, \lambda\right) \tag{1}$$

where i denotes ith observation, the distribution of s determines the form of $\log\,(y_i)$. For example, when $s = 1$, the SMSN distribution degenerates to skew-normal (SN) distribution, and with both $s = 1$ and $\lambda = 0$, we retrieve the normal distribution. Here, we assigned s as following:

$$s \sim \mathrm{Gamma}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right) \tag{2}$$

where $\nu_1 > 0$ and $\nu_2 > 0$, and can be considered as two unknown 'degree of freedom' parameters that characterize the shape of the SMSN distribution. In fact, with the given s, the general SMSN case becomes a special generalized-skew-t case (Branco and Dey 2001) (or well-known as the skewed Pearson type VII case (Nadarajah and Gupta 2005, Shimizu and Iida 2006)), as we can have the usual skew-t case with $\nu_1 = \nu_2$.

Then we constructed linear model conditional on the location μ_i by including water quality parameters as predictors:

Table 2. Statistical information of monitored water quality parameters.

Study area	Parameter[a]	Unit	Calibration data		Validation data	
			Mean	SD	Mean	SD
Hangjiahu	Cyanobacterial biomass	$\mu g\,l^{-1}$	205.96	473.08	180.31	496.49
	Water temperature	°C	18.26	7.42	7.53	1.31
	pH	—	7.76	0.43	7.66	0.21
	Conductivity	$\mu s\,cm^{-1}$	360.93	214.84	234.69	102.91
	DO	$mg\,l^{-1}$	7.87	2.77	9.91	1.74
	Water depth	m	1.16	0.55	1.18	0.54
	Turbidity	NTU	55.08	64.86	40.32	38.93
	PAR	$\mu mol\,m^{-2}\,s^{-1}$	514.53	595.55	145.96	95.29
	SDD	m	0.43	0.30	0.56	0.37
Trout	Chl-*a* concentration	$\mu g\,l^{-1}$	19.02	51.62	24.65	57.47
	Water temperature	°C	9.61	5.46	8.40	5.64
	pH	—	6.55	1.12	6.56	1.12
	DO	$mg\,l^{-1}$	6.81	3.11	6.72	3.27
	PAR	$\mu mol\,m^{-2}\,s^{-1}$	916.26	533.43	811.71	532.98
	SDD	m	3.87	2.50	3.86	2.41
Madison	Cyanobacterial biomass	$\mu g\,l^{-1}$	2234.74	2626.89	2667.59	3723.68
	Water temperature	°C	13.02	4.67	13.11	4.47
	pH	—	8.17	0.19	8.40	0.18
	DO	$mg\,l^{-1}$	6.14	3.74	5.94	3.66
	SDD	m	2.47	1.58	2.94	1.80

[a] Parameters in bold denote the preserved model predictors via stepwise selection procedures.

$$\mu_i = \beta_0 + \beta_1 \cdot x_{1i} + \ldots + \beta_p \cdot x_{p_i} + \gamma_{j[i]} + \delta_{k[i]} \quad (3)$$

where $X = (x_1, \ldots, x_p)$ are predictor matrices; $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are vectors of regression parameters. Note that the monitoring data were collected across sites and dates, the spatial and temporal variations involved in data may largely influence the model estimates (Carstensen and Lindegarth 2016). Here, we also partially pooled the external site-specific and date-specific information to improve the posterior estimates. Thus, γ_j are site-specific random effects varying by site j; δ_k are date-specific random effects varying by date k. For the seasonal dataset of Hangjiahu rivers, $k = 1, 2, 3$; for the multi-weekly datasets of Trout and Madison lakes, monthly sales (i.e. $k = 1, \ldots, 12$) were well-fitted according to the previous study in this area (Xiao *et al* 2019b).

Additionally, our preliminary correlation analysis showed that some predictor variables were highly correlated (figure S1). Therefore, ridge regression was developed to address the potential multicollinearity problems in the linear model (Dormann *et al* 2013). Via ridge regression, additional regularization parameters $\tau = (\tau_0, \tau_1, \ldots, \tau_p)$ were taken to describe the prior precision of regression parameters β, thereby to restrict the overfitting of training data with the collinear variables (McElreath 2018). We assigned the priors for the ridge estimates of regression parameters as (Shi *et al* 2016, Assaf *et al* 2019):

$$\begin{aligned} \beta &\sim \text{Normal}\left(0, \tau^{-1}\right) \\ \tau &\sim \text{Gamma}\left(0.01, 0.01\right) \end{aligned} \quad (4)$$

For the random effects, the weakly informative priors were assigned as:

$$\begin{aligned} \gamma_j &\sim \text{Normal}\left(0, \sigma_\gamma^2\right) \\ \delta_k &\sim \text{Normal}\left(0, \sigma_\delta^2\right) \\ \sigma_\gamma^2, \sigma_\delta^2 &\sim \text{Normal}_+\left(0, 10\right) \end{aligned} \quad . \quad (5)$$

For two 'freedom' parameters characterizing the population-level data distribution, we considered the priors suggested by Rômulo Barbosa Cabral *et al* (2012):

$$\begin{aligned} \nu_1 &\sim \text{Exponential}\left(\varphi_1\right) \\ \nu_2 &\sim \text{Exponential}\left(\varphi_2\right) \\ \varphi_1, \varphi_2 &\sim \text{Unifrom}\left(0.1, 10\right) \end{aligned} \quad . \quad (6)$$

For the skewness and scale parameters, the weakly informative priors were also assigned as:

$$\begin{aligned} \sigma^2 &\sim \text{Normal}_+\left(0, 10\right) \\ \lambda &\sim \text{Normal}\left(0, 1\right) \end{aligned} \quad . \quad (7)$$

2.2.2. Computation procedures

All computations for Bayesian inference were programmed in the R environment (R Core Team 2020) using the RStan interface (Stan Development Team 2020) to Stan (Stan Development Team 2019). The Markov chain Monte Carlo (MCMC) algorithm was applied using the No-U-Turn sampler to sample for parameter posterior distributions. We ran four chains for 20 000 iterations, discarded the first 5000 (burning), and retained the second 15 000 (sampling) iterations per chain to obtain 60 000 MCMC samples

in total. We also pre-set the sampler controlling parameters (adapt_delta = 0.99, stepsize = 0.95, and max_treedepth = 25) and re-parameterized the Stan codes for efficient and stable computations in sampling procedures. The convergences of Markov chains were assured by R-hat statistics ($\hat{R}$ is maintained under 1.01).

The predictors were centered through standardization to achieve a reliable and stable posterior estimate. Model predictions were summarized as medians (point prediction) with credible intervals of the predictive posterior distributions (PPDs; probabilistic prediction). Predictions on new observations from new groups were obtained using the marginal of random effects (McElreath 2018). Moreover, for modeling simplification, a stepwise regression procedure was adopted to reduce predictor variables based on the five-fold cross-validation results using the calibration datasets (details in table S2).

2.2.3. Probabilistic assessments of algal blooms

To inform the algal blooms in three different waterbodies, two alert standards related to the health-based drinking water supplies were provided, as defined by the World Health Organization (2021). Two algal bloom thresholds were categorized according to either cyanobacterial biomass or Chl-*a* concentration in the water sample: alert level 1 (300 μg l^{-1} of cyanobacterial biomass or 1 μg l^{-1} of Chl-*a*); and alert level 2 (4000 μg l^{-1} of cyanobacterial biomass or 12 μg l^{-1} of Chl-*a*). Thus, we used the entire posterior distributions to calculate the probability of exceeding the two standards, denoted as the proportion of exceeded MCMC samples (more computation details in text S2).

2.3. Model evaluation

The model performance was assessed via both calibration data (in-sample) and validation data (out-of-sample). The regression model was evaluated with correlation-coefficient (R^2) and root-mean-square-error (RMSE), which were the measures of the deviation of predicted values from the observed values, and calculated as:

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

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

where n is the number of data points; $\hat{y}_i$ and y_i are the ith predicted and observed values; $\bar{y}$ is the mean of y_i.

The prediction performance on algal bloom stages was evaluated based on the confusion matrix in terms of accuracy and macro-F_1 score (F_1). The accuracy was normally used to accounts for the overall correct rates of classification, defined as:

$$\mathrm{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n}$$

where n is the number of data points.

The F_1 statistic considered both the true rate and false rate of classification when measuring the overall accuracy, and was calculated using the precision and recall as:

$$\mathrm{Precision}\ (P)_b = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\mathrm{Recall}\ (R)_b = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$F_1\mathrm{score}_b = 2 \times \frac{P \times R}{P + R}$$

$$\mathrm{Macro} - F_1\ \mathrm{score}\ (F_1) = \sum_{b=1}^{B} F_1\ \mathrm{score}_b$$

where B is the number of classes from the confusion matrix.

3. Results

3.1. Distribution of algal biomass data

For each of the three algal datasets, the distributions of response variables (i.e. Chl-*a* concentration or cyanobacterial biomass) were drawn by the histogram method (figure 2). The detailed description was also listed in table S1, with the statistics of Shapiro–Wilk normality statistic (W), coefficient of Skewness (SK), and coefficient of Pearson's kurtosis (K). In general, over-skewed and over-dispersed characteristics were found in all three algal datasets, and a large proportion of outliers can be also observed due to the inclusion of numerous extreme values (figure 3). For instance, Chl-*a* concentration in Trout lakes (figure 3(a)) presented the most extreme skewed and leptokurtic features (highest SK value of 6.49, highest K value of 49.9), showing a violation of normality (lowest W value of 0.34). Moreover, although the excessive extreme values were well reduced in the logarithmic scale, the logarithmic distribution still presented asymmetry and multimodality (figure 3(b)). This again justified the use of SMSN assumption for modeling the irregular algal samples.

3.2. Predicting algal biomass variations

Three sets of optimal modeling predictors were identified via stepwise regression procedures (table S2). In general, variables including pH, conductivity, water depth, water temperature, DO, and SDD showed stronger relevance to the algal variations, and pH and conductivity were the common inclusion of model predictors for three cases. In addition, the variance components of site-specific variation estimated by the model were nearly two orders of magnitude larger than the temporal variation, showing larger modeling uncertainties in spatial factors as compared to temporal factors (table S3).

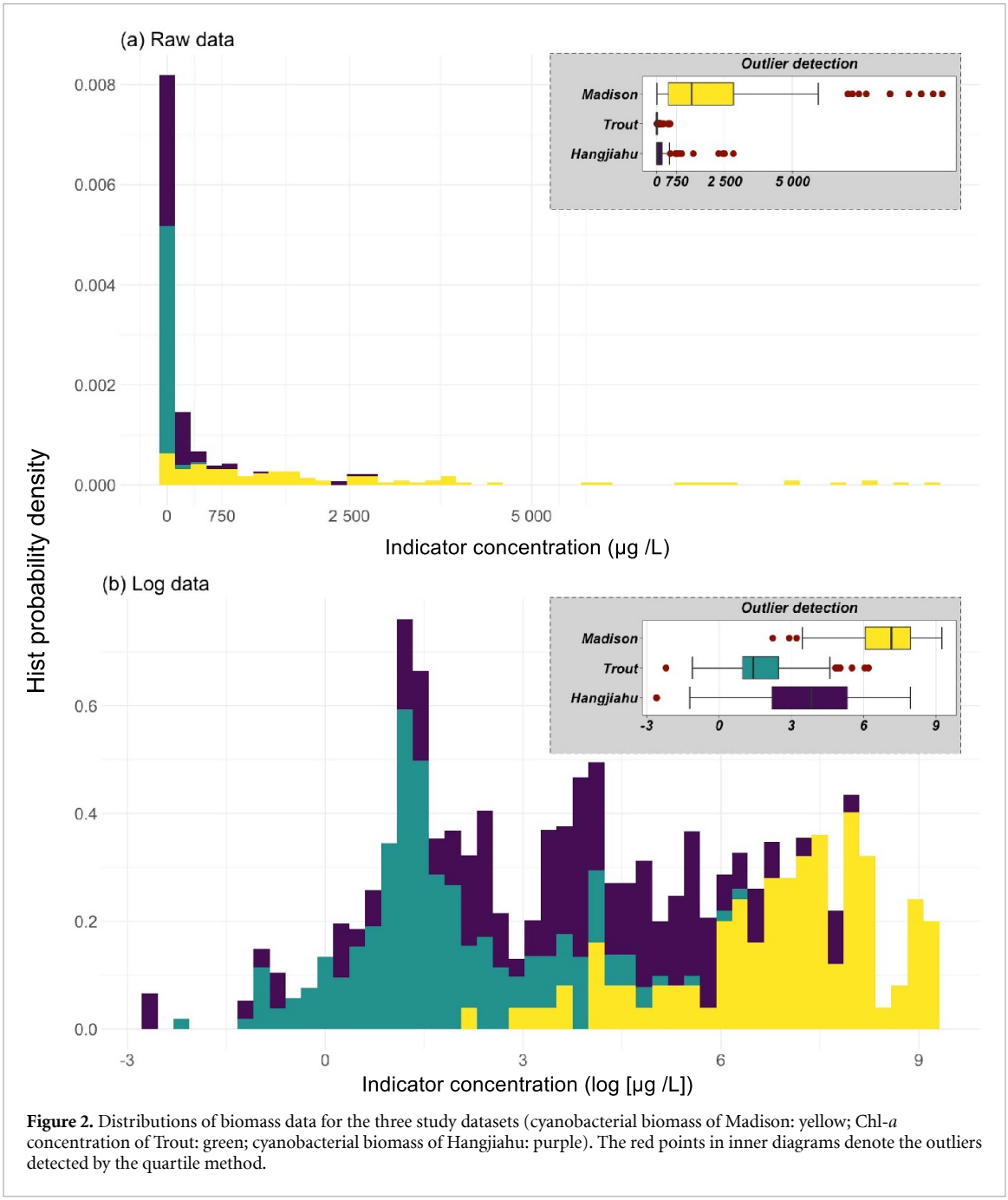

Figure 2. Distributions of biomass data for the three study datasets (cyanobacterial biomass of Madison: yellow; Chl-*a* concentration of Trout: green; cyanobacterial biomass of Hangjiahu: purple). The red points in inner diagrams denote the outliers detected by the quartile method.

The predictive performance of SMSN models was shown in figure 3. Overall, calibration R^2 values ranged from 0.533 to 0.706, indicating that the SMSN model presented successful goodness-of-fit for the algal biomass data violating the normality assumption. When comparing the out-of-sample predictions, the SMSN models still achieved acceptable accuracy, with validation R^2 values ranging from 0.412 to 0.742. The three case studies indicated that the developed Bayesian SMSN models presented a reliable tool to predict algal variability.

3.3. Probabilistic prediction of algal blooms

The model prediction performance of algal bloom was shown in table 3. Results of all three probabilistic models presented accurate assessment with low false classification rates, as revealed by the calculated accuracy rate and F_1 score (accuracy >0.758 and $F_1 > 0.725$). Taking Trout lakes as an example, in the early alert level 1 (figure 4(a)), 15 out of 22 non-exceedances and 192 out of 194 exceedances were correctly predicted, with 95.8% accurate rate at the probability threshold of 0.61 (1 μg l^{-1} of Chl-*a*). When further assessing the alert level 2 (12 μg l^{-1} of Chl-*a*) (figure 4(c)), the model predicted 130 out of 139 non-exceedances (nine false exceedances) and 45 out of 53 exceedances (eight false non-exceedances) at a probability threshold of 0.56. In general, 45 out of 53 exceedances of level 2 (84.9%), 130 out of 141 exceedances of level 1 (no level 2 exceeded; 92.2%), and 15 out of 22 non-exceedances of level 1 (no alert required; 68.2%) were correctly predicted

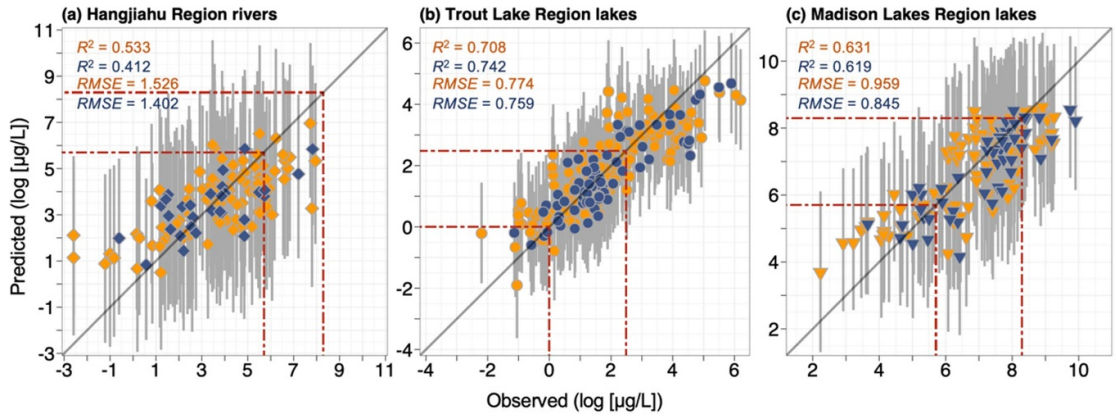

Figure 3. Observed and predicted cyanobacterial biomass (a), (c) and Chl-*a* concentration (b) based on Bayesian SMSN models. Data used for model calibration and model validation were in yellow and blue, respectively. The red dotted lines indicated two alert levels. The gray bars indicated the PPDs of the regression estimates with 95% credible intervals.

by the model, achieving an overall accuracy rate of 88.1% and an overall F_1 score as high as 0.841. In addition, the probabilistic method achieved high classifying accuracy in the validation data as good as in the calibration process (figures 4(b) and (d)). Similarly, the other two models for Hangjiahu rivers and Madison lakes also suggested satisfactory performances (figures S2, S3 and table 3).

As a view of comparison, the classification performance of directly using point predictions was also listed in table 3. The results showed very comparable accuracy, however, when the linear models did not yield strong regression performance, the point-prediction-based classification presented high misjudgments for the actual exceedances of alert standards (e.g. Hangjiahu rivers in figure 4(a) and table 3). Moreover, the point-prediction-based classification resulted in more false positives and false negatives in all case studies, as indicated by its lower F_1 scores (table 3).

4. Discussion

In this study, we presented a promising Bayesian SMSN approach to predict algal biomass from the aquatic environment fluctuations. Differing from many of previous works, our model was conducted only based on the standard water physicochemical parameters (e.g. DO, pH and conductivity). Comparatively, this is an advantage over predictive models that require time-consuming predictors (e.g. nutrients), since these parameters can be rapidly measured *in-situ* with portable sensors and have been included in the basic monitoring of most waterbodies. Using the historical observations, our method can therefore facilitate future algal monitoring via achieving real-time and reliable biomass variation estimates. We also found the relatively strong effects of water temperature, conductivity, and DO on the algal variation (table S3), which was in line with many relevant studies (Cha *et al* 2014, Xiao *et al* 2019a, Haakonsson *et al* 2020, Liu *et al* 2023). This highlighted that the selection of such indicators that were closely related to the algae growth can be critical for future analogue modeling. To date, increasing studies have been aware of this importance. In Australia, Recknagel *et al* (2014) successfully predicted algal dynamics in three sub-tropical reservoirs with conductivity, turbidity, DO and water temperature; in China, an early warning system for phytoplankton blooms was developed based on *in-situ* automated online sondes in Xiangxi Bay (Ye *et al* 2014); and in South America, coastal cyanobacterial blooms were accurately predicted ($R^2 = 0.82$) from water temperature and conductivity conditions (Haakonsson *et al* 2020). Moreover, with the cheaper availability of sonde technologies in the future, using a simplified parameter approach could further cut down the costs of algal monitoring systems, which would greatly benefit aquatic environment management.

Interestingly, our model worked well not only for lakes, but also for the riverine system. Compared to the relatively static lakes, the hydrological condition plays a more important role in algal distributions in river systems (Qu *et al* 2018, Wu *et al* 2018). The stream flow can cause unexpected changes to the relationships between algal biomass and environmental factors from site to site (Smith *et al* 1999, Jaiswal and Pandey 2019). This high spatial variation may lead to a problem of Simpson's paradox in a cross-sectional ecological modeling (Qian *et al* 2015), and often make the linear estimates of riverine algal variations less applicable (Cha *et al* 2016). In our preliminary analysis for Hangjiahu rivers (figure S1(a)), the low correlations between cyanobacterial biomass and water quality parameters posed great challenges to conducting linear models. Nevertheless, as presented in this study and many other previous works (Malve and Qian 2006, Gronewold and Borsuk 2010, Cha *et al* 2016, Qian *et al* 2019, Seis *et al* 2020), partial

Table 3. Model performance for probabilistic prediction of algal blooms.

		Model calibration[a]				Model validation[a]			
Study area	Alert level	Accuracy		F_1		Accuracy		F_1	
Hangjiahu	Level 1	0.903	0.827 (−8.4%)	0.758	0.552 (−27.2%)	0.903	0.870 (−3.7%)	0.949	0.631 (−33.5%)
	Level 2	1.000	1.000 (−0.0%)	1.000	1.000 (−0.0%)	1.000	1.000 (−0.0%)	1.000	1.000 (−0.0%)
	Overall	0.903	0.827 (−8.4%)	0.839	0.552 (−34.2%)	0.903	0.871 (−3.7%)	0.975	0.631 (−35.3%)
Trout	Level 1	0.958	0.935 (−2.4%)	0.873	0.808 (−7.4%)	0.957	0.957 (−0.0%)	0.852	0.852 (−0.0%)
	Level 2	0.912	0.912 (−0.0%)	0.889	0.883 (−0.1%)	0.889	0.928 (4.3%)	0.856	0.865 (0.1%)
	Overall	0.881	0.856 (−2.8%)	0.841	0.787 (−6.4%)	0.857	0.886 (3.3%)	0.804	0.832 (0.3%)
Madison	Level 1	0.893	0.883 (−1.1%)	0.857	0.832 (−2.9%)	0.868	0.868 (−0.0%)	0.826	0.805 (−2.5%)
	Level 2	0.861	0.864 (0.3%)	0.750	0.669 (−10.8%)	0.861	0.906 (5.2%)	0.765	0.771 (0.7%)
	Overall	0.796	0.747 (−6.1%)	0.736	0.654 (−11.1%)	0.774	0.773 (−0.1%)	0.725	0.713 (−1.7%)

[a] The values with parentheses represent the classification directly using point-predictions and their relative changes to the probabilistic classifications.

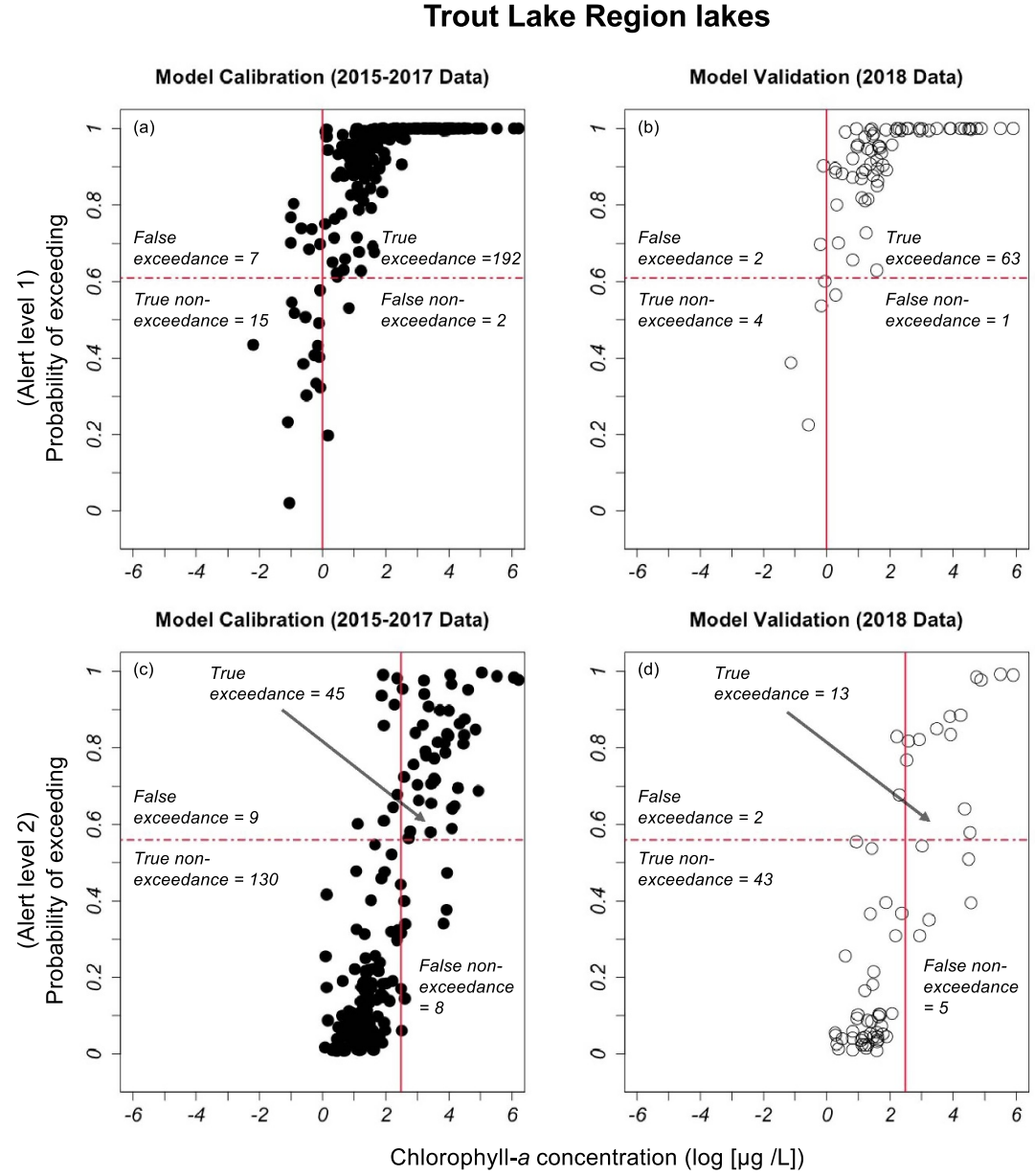

Figure 4. Observed log Chl-*a* concentration versus the probability of exceeding the alert levels in Trout. The red vertical line represented the two alert standards in terms of log Chl-*a* concentration. The red horizontal dot line represented the optimal probability threshold determined by the model.

pooling of such site-specific heterogeneity as external information can succeed in addressing this issue. The spatial hierarchical structure appears to be useful to improve the overall fitting ability of a model based on large-scale monitoring data. In addition, although spatial heterogeneity was substantially stronger than temporal heterogeneity for affecting algal biomass in our three cases, considering that the distribution of algal community usually has high variation in both spatial and temporal scales (Kolber and Falkowski 1995), temporal hierarchy is still important and need to care within a time-varying algal prediction model. In recent years, given the increasing monitoring of spatiotemporal scales for ecology science (Xiao *et al* 2014), there are broader applications of the proposed approach.

Under favorable conditions such as climate warming (Xiao *et al* 2019a, 2019b), stable hydraulic status (Park *et al* 2015, Cha *et al* 2017), and high nutritional level (Heisler *et al* 2008, Beaulieu *et al* 2013), algal biomass can increase dramatically from low values to blooming levels in a short time. As such, extreme and occasional values frequently show up in the algal samples. This can prevent standard Normal linear models from correct predictions as their parameter inferences are typically sensitive to the occasional values (Gelman *et al* 2013). Similar to this study, increasingly improved models have been used to overcome the challenges caused by the extremely distributed biomass data. For example, the hurdle model (Cha *et al* 2014), zero-inflated mixture model (Cusack *et al* 2015), and compound Poisson-gamma model (Haakonsson *et al* 2020) were successively developed to address the 'excessive zeros' problem in cyanobacterial bloom predictions; and the over-dispersed Poisson model was developed for fitting the abundance data with large variance (Cha *et al* 2017). For our cases, the SMSN models accurately predicted the skewed algal variations with intensive extreme points (table 3 and figure 3). Further, since the family of SMSN distribution conceptually allows for the possibility of outliers in the data distribution, namely that the SMSN regression model is robust to avoid parameter inference bias (Silva *et al* 2020). In practice, a robust and high fault-tolerance approach with powerful predictability could better support the decision-making works such as the bloom management.

We showed that the incorporation of probabilistic framework benefits the assessment of algal bloom stages, which successfully addressed the high false rate problem when employing regression point-prediction (table 3). For ecological studies, regression estimates are often useful for classification purposes. However, this utilization often tends to show appreciable false rates even it has good overall accuracy (Motamarri and Boccelli 2012), since regression-based outcomes are typically point-wise and inevitably involve a large amount of uncertainties (Zhao and Kockelman 2002, Carstensen and Lindegarth 2016, Hutorowicz and Pasztaleniec 2021). The uncertainty may come from the spatiotemporal variations, inaccuracy and mistakes in measurements when collecting the source data (He and Kolovos 2018), or resulted from the statistical models (Carstensen and Lindegarth 2016). Nevertheless, the effect of uncertainty on data-driven ecological research receives less attention (Carstensen and Lindegarth 2016), though it has been informed that the uncertainty will propagate through the input to the output of models (Zhao and Kockelman 2002) and may bias a modeling analysis if without prior acknowledged (He *et al* 2020). Fortunately, in a Bayesian model, the overall uncertainties can be propagated forward to the entire PPDs via inference (McElreath 2018). The PPDs approximate the probability of true values within a creditable interval, offering a natural uncertainty assessment framework to the parameter and outcome estimates (Qian *et al* 2004, He and Christakos 2018). In our case, the PPDs of algal biomass estimates were applied for the probabilistic prediction of algal bloom stages. Using this method, all of the categorizations presented high accuracy even when the regression models performed poorly (figure 4 and table 3). Additionally, this development gave us direct information about the probability of water samples in exceeding certain alert standards, which can further be used as the scientific basis for the lake or river managers to build bloom-warning advisories.

5. Conclusion

This work presented a promising and efficient Bayesian probabilistic SMSN modeling technique, allowing for the real-time prediction of algal variations and *in-situ* assessments of algal bloom stages, which:

(a) Required only basic physiochemical water quality parameters.
(b) Had good prediction performance on biomass data having over-dispersed characteristics and containing a big proportion of extreme values.
(c) Achieved robust prediction accuracy of algal blooms through combining probabilistic framework.

In the future, the modeling could be enhanced via involving more diverse predictor variables such as hydrometeorological and anthropogenic factors, which were limited by the dataset as shown in the current study.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The authors acknowledge the North Temperate Lakes-Long Term Ecological Research (NTL-LTER, https://lter.limnology.wisc.edu/) project for sharing the monitoring data used in this study. We appreciate those who participated in LTER program by collecting and measuring the high-quality records for their dedicated work.

We thank master and bachelor students Chao Li, Lihao Shi, Fanchang Ding, Fangyi Wei, Xutao Ye, Chen Wang, and Qijun Li for their assistance in field sampling and on-site analysis. This study was financially supported by the National Natural Science Foundation of China (21876148), the Zhejiang Provincial Natural Science Foundation/Funds for Distinguished Young Scientists (LR22D06003), the Key Laboratory of Marine Ecological Monitoring and Restoration Technologies of the Ministry of Natural Resources of China (MEMRT202102), Science Foundation of Donghai Laboratory (DH-2022KF01021), the Key Research and Development Program of Guangxi Province (AB22080099) and Funding for ZJU Tang Scholar to X. X.

References

Amorim C A and Do Nascimento Moura A 2021 Ecological impacts of freshwater algal blooms on water quality, plankton biodiversity, structure, and ecosystem functioning *Sci. Total Environ.* **758** 143605

Assaf A G, Tsionas M and Tasiopoulos A 2019 Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression *Tour. Manage.* **71** 1–8

Beaulieu M, Pick F and Gregory-Eaves I 2013 Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes data set *Limnol. Oceanogr.* **58** 1736–46

Benites L, Maehara R, Lachos V H and Bolfarine H 2019 Linear regression models using finite mixtures of skew heavy-tailed distributions *Chil. J. Stat.* **10** 21–41

Branco M D and Dey D K 2001 A general class of multivariate skew-elliptical distributions *J. Multivariate Anal.* **79** 99–113

Cabral C R B, Hugo Lachos V and Regina Madruga M 2012 Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the random-effects population *J. Stat. Plan. Inference* **142** 181–200

Carstensen J and Lindegarth M 2016 Confidence in ecological indicators: a framework for quantifying uncertainty components from monitoring data *Ecol. Indic.* **67** 306–17

Cha Y, Cho K H, Lee H, Kang T and Kim J H 2017 The relative importance of water temperature and residence time in predicting cyanobacteria abundance in regulated rivers *Water Res.* **124** 11–19

Cha Y, Park S S, Kim K, Byeon M and Stow C A 2014 Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model *Water Resour. Res.* **50** 2518–32

Cha Y, Park S S, Lee W and Stow C A 2016 A Bayesian hierarchical approach to model seasonal algal variability along an upstream to downstream river gradient *Water Resour. Res.* **52** 348–57

Chung S H, Pearn W L and Yang Y S 2007 A comparison of two methods for transforming non-normal manufacturing data *Int. J. Adv. Manuf. Technol.* **31** 957–68

Cusack C, Mouriño H, Moita M T and Silke J 2015 Modelling Pseudo-nitzschia events off southwest Ireland *J. Sea Res.* **105** 30–41

Dormann C F *et al* 2013 Collinearity: a review of methods to deal with it and a simulation study evaluating their performance *Ecography* **36** 27–46

Fletcher D, MacKenzie D and Villouta E 2005 Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression *Environ. Ecol. Stat.* **12** 45–54

Flynn K J, Clark D R, Mitra A, Fabian H, Hansen P J, Glibert P M, Wheeler G L, Stoecke D K, Blackford J C and Brownlee C 2015 Ocean acidification with (de)eutrophication will alter future phytoplankton growth and succession *Proc. R. Soc.* B **282** 2–7

García-Nieto P J, García-Gonzalo E, Sánchez Lasheras F, Alonso Fernández J R and Díaz Muñiz C 2020 A hybrid DE optimized wavelet kernel SVR-based technique for algal atypical proliferation forecast in La Barca reservoir: a case study *J. Comput. Appl. Math.* **366** 112417

Gelman A, Carlin J B, Stern H S, Dunson D B, Vehtari A and Rubin D B 2013 *Bayesian Data Analysis* 3rd edn (Boca Raton, FL: CRC Press)

Glibert P M, Allen J I, Bouwman A F, Brown C W, Flynn K J, Lewitus A J and Madden C J 2010 Modeling of HABs and eutrophication: status, advances, challenges *J. Mar. Syst.* **83** 262–75

Gronewold A D and Borsuk M E 2010 Improving water quality assessments through a hierarchical Bayesian analysis of variability *Environ. Sci. Technol.* **44** 7858–64

Guo J, Dong Y and Lee J H W 2020 A real time data driven algal bloom risk forecast system for mariculture management *Mar. Pollut. Bull.* **161** 111731

Haakonsson S, Rodríguez M A, Carballo C, Del Carmen Pérez M, Arocena R and Bonilla S 2020 Predicting cyanobacterial biovolume from water temperature and conductivity using a Bayesian compound Poisson-gamma model *Water Res.* **176** 115710

Hallegraeff G M 1993 A review of harmful algal blooms and their apparent global increase *Phycologia* **32** 79–99

He J, Chen Y, Wu J, Stow D A and Christakos G 2020 Space-time chlorophyll-a retrieval in optically complex waters that accounts for remote sensing and modeling uncertainties and improves remote estimation accuracy *Water Res.* **171** 115403

He J and Christakos G 2018 Space-time $PM_{2.5}$ mapping in the severe haze region of Jing-Jin-Ji (China) using a synthetic approach *Environ. Pollut.* **240** 319–29

He J and Kolovos A 2018 Bayesian maximum entropy approach and its applications: a review *Stoch. Environ. Res. Risk Assess.* **32** 859–77

Heisler J *et al* 2008 Eutrophication and harmful algal blooms: a scientific consensus *Harmful Algae* **8** 3–13

Ho J C, Michalak A M and Pahlevan N 2019 Widespread global increase in intense lake phytoplankton blooms since the 1980s *Nature* **574** 667–70

Hu H 2006 *The Freshwater Algae of China: Systematics, Taxonomy and Ecology* 1st edn, vol 16 (Beijing: Science Press) p 1023

Hutorowicz A and Pasztaleniec A 2021 Uncertainty in phytoplankton-based lake ecological status classification: implications of sampling frequency and metric simplification *Ecol. Indic.* **127** 107754

Jaiswal D and Pandey J 2019 An ecological response index for simultaneous prediction of eutrophication and metal pollution in large rivers *Water Res.* **161** 423–38

Kolber Z and Falkowski P G 1995 Variations in chlorophyll fluorescence yields in phytoplankton in the World Oceans *Aust. J. Plant Physiol.* **22** 341–55

Lee D, Kim M, Lee B, Chae S, Kwon S and Kang S 2022 Integrated explainable deep learning prediction of harmful algal blooms *Technol. Forecast. Soc. Change* **185** 122046

Lee H S and Lee J H W 1995 Continuous monitoring of short term dissolved oxygen and algal dynamics *Water Res.* **29** 2789–96

Lee S and Lee D 2018 Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models *Int. J. Environ. Res. Public Health* **15** 1322

Liu J Y, Zeng L H, Ren Z H, Du T M and Liu X 2020 Rapid *in situ* measurements of algal cell concentrations using an artificial neural network and single-excitation fluorescence spectrometry *Algal Res.* **45** 101739

Liu M, He J, Huang Y, Tang T, Hu J and Xiao X 2022 Algal bloom forecasting with time-frequency analysis: a hybrid deep learning approach *Water Res.* **219** 118591

Liu M, Huang Y, Hu J, He J and Xiao X 2023 Algal community structure prediction by machine learning *Environ. Sci. Technol.* **14** 100233

Malve O and Qian S S 2006 Estimating nutrients and chlorophyll a relationships in Finnish lakes *Environ. Sci. Technol.* **40** 7848–53

Mantzouki E and Visser P M 2015 Understanding the key ecological traits of cyanobacteria as a basis for their management and control in changing lakes *Aquat. Ecol.* **50** 333–50

Marchenko Y V and Genton M G 2010 Multivariate log-skew-elliptical distributions with applications to precipitation data *Environmetrics* **21** 318–40

McElreath R 2018 *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Boca Raton, FL: CRC Press)

Mellios N K, Moe S J and Laspidou C 2020 Using Bayesian hierarchical modelling to capture cyanobacteria dynamics in Northern European lakes *Water Res.* **186** 116356

Mirfarah E, Naderi M and Chen D 2021 Mixture of linear experts model for censored data: a novel approach with scale-mixture of normal distributions *Comput. Stat. Data Anal.* **158** 107182

Montenegro C and Branco M 2016 Bayesian state-space approach to biomass dynamic models with skewed and heavy-tailed error distributions *Fish. Res.* **181** 48–62

Motamarri S and Boccelli D L 2012 Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms *Water Res.* **46** 4508–20

Nadarajah S and Gupta A K 2005 A skewed truncated Pearson type VII *J. Japan Stat. Soc.* **35** 61–71

Olokotum M, Mitroi V, Troussellier M, Semyalo R, Bernard C, Montuelle B, Okello W, Quiblier C and Humbert J F 2020 A review of the socioecological causes and consequences of cyanobacterial blooms in Lake Victoria *Harmful Algae* **96** 101829

Park Y, Pachepsky Y A, Cho K H, Jeon D J and Kim J H 2015 Stressor–response modeling using the 2D water quality model and regression trees to predict chlorophyll-a in a reservoir system *J. Hydrol.* **529** 805–15

Qian S S, Donnelly M, Schmelling D C, Messner M, Linden K G and Cotton C 2004 Ultraviolet light inactivation of protozoa in drinking water: a Bayesian meta-analysis *Water Res.* **38** 317–26

Qian S S, Stow C A and Cha Y 2015 Implications of Stein's paradox for environmental standard compliance assessment *Environ. Sci. Technol.* **49** 5913–20

Qian S S, Stow C A, Nojavan A F, Stachelek J, Cha Y, Alameddine I and Soranno P 2019 The implications of Simpson's paradox for cross-scale inference among lakes *Water Res.* **163** 114855

Qin B, Paerl H W, Brookes J D, Liu J, Jeppesen E and Zhu G 2019 Why Lake Taihu continues to be plagued with cyanobacterial blooms through 10 years (2007–2017) efforts *Sci. Bull.* **64** 354–6

Qin B, Zhu G, Gao G, Zhang Y, Li W, Paerl H W and Carmichael W W 2010 A drinking water crisis in Lake Taihu, China: linkage to climatic variability and lake management *Environ. Manage.* **45** 105–12

Qu Y, Wu N, Guse B and Fohrer N 2018 Riverine phytoplankton shifting along a lentic-lotic continuum under hydrological, physiochemical conditions and species dispersal *Sci. Total Environ.* **619–620** 1628–36

R Core Team 2020 R: a lanavailable at: guage and environment for statistical computing *R Foundation for Statistical Computing* (Vienna, Austria) (available at: www.r-project.org/)

Recknagel F, Orr P T and Cao H 2014 Inductive reasoning and forecasting of population dynamics of cylindrospermopsis raciborskii in three sub-tropical reservoirs by evolutionary computation *Harmful Algae* **31** 26–34

Seis W, Rouault P and Medema G 2020 Addressing and reducing parameter uncertainty in quantitative microbial risk assessment by incorporating external information via Bayesian hierarchical modeling *Water Res.* **185** 116202

Shi Q, Abdel-Aty M and Lee J 2016 A Bayesian ridge regression analysis of congestion's impact on urban expressway safety *Accid. Anal. Prev.* **88** 124–37

Shimizu K and Iida K 2006 Pearson type VII distribution on spheres **0926**

Shmueli G and Koppius O R 2010 Predictive analytics in information systems research *MIS Q. Manage. Inf. Syst.* **35** 553–72

Silva N B, Prates M O, Gonçalves F B and Prates M O 2020 Bayesian linear regression models with flexible error distributions distributions *J. Stat. Comput. Simul.* **90** 2571–91

Smith V H, Tilman G D and Nekola J C 1999 Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems *Environ. Pollut.* **100** 179–96

Stan Development Team 2019 Stan modeling language users guide and reference manual, version 2.25 (available at: https://mc-stan.org)

Stan Development Team 2020 RStan: the R interface to Stan pp 1–22

Visser P M, Verspagen J M H, Sandrini G, Stal L J, Matthijs H C P, Davis T W, Paerl H W and Huisman J 2016 How rising CO 2 and global warming may stimulate harmful cyanobacterial blooms *Harmful Algae* **54** 145–59

Wang H, Zhu R, Zhang J, Ni L, Shen H and Xie P 2018 A novel and convenient method for early warning of algal cell density by chlorophyll fluorescence parameters and its application in a highland lake *Front. Plant Sci.* **9** 1–13

Weisse T 2008 Limnoecology: the ecology of lakes and streams *J. Plankton Res.* **30** 489–90

World Health Organization 2021 *Toxic Cyanobacteria in Water* ed I Chorus and M Welker (Boca Raton, FL: CRC Press) (available at: www.taylorfrancis.com/books/9781000262025)

Wu N, Dong X, Liu Y, Wang C, Baattrup-Pedersen A and Riis T 2017 Using river microalgae as indicators for freshwater biomonitoring: review of published research and future directions *Ecol. Indic.* **81** 124–31

Wu N, Qu Y, Guse B, Makarevičiūtė K, To S, Riis T and Fohrer N 2018 Hydrological and environmental variables outperform spatial factors in structuring species, trait composition, and beta diversity of pelagic algae *Ecol. Evol.* **8** 2947–61

Xiao X, Agustí S, Pan Y, Yu Y, Li K, Wu J and Duarte C M 2019a Warming amplifies the frequency of harmful algal blooms with eutrophication in Chinese coastal waters *Environ. Sci. Technol.* **53** 13031–41

Xiao X, He J, Yu Y, Cazelles B, Li M, Jiang Q and Xu C 2019b Teleconnection between phytoplankton dynamics in north temperate lakes and global climatic oscillation by time-frequency analysis *Water Res.* **154** 267–76

Xiao X, Sogge H, Lagesen K, Tooming-Klunderud A, Jakobsen K S and Rohrlack T 2014 Use of high throughput sequencing and light microscopy show contrasting results in a study of phytoplankton occurrence in a freshwater environment *PLoS One* **9** e106510

Ye L, Cai Q, Zhang M and Tan L 2014 Real-time observation, early warning and forecasting phytoplankton blooms by integrating *in situ* automated online sondes and hybrid evolutionary algorithms *Ecol. Inform.* **22** 44–51

Zhang X, Dong Q, Costa V and Wang X 2019 A hierarchical Bayesian model for decomposing the impacts of human activities and climate change on water resources in China *Sci. Total Environ.* **665** 836–47

Zhao Y and Kockelman K M 2002 The propagation of uncertainty through travel demand models: an exploratory analysis *Ann. Reg. Sci.* **36** 145–63