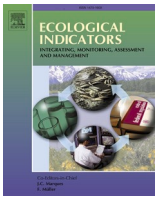
Ecological Indicators 130 (2021) 108055

Contents lists available at ScienceDirect

Ecological Indicators

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



Chlorophyll *a* as an indicator of microcystin: Short-term forecasting and risk assessment in Lake Erie

Song S. Qian [a,*], Craig A. Stow [b], Freya E. Rowland [c], Qianqian Liu [d], Mark D. Rowe [b], Eric J. Anderson [b], Richard P. Stumpf [e], Thomas H. Johengen [f]

[a] Department of Environmental Sciences, University of Toledo, Toledo, OH 43606, USA
[b] Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, Ann Arbor, MI 48018, USA
[c] Yale School the Environment, Yale University, New Haven, CT 06511, USA
[d] Department of Physics and Physical Oceanography, University of North Carolina Wilmington, Wilmington, NC 28403, USA
[e] National Centers for Coastal Ocean Science, National Oceanic and Atmospheric Administration, Silver Spring, MD 20910, USA
[f] Cooperative Institute for Great Lakes Research (CIGLR), University of Michigan, Ann Arbor, MI 48018, USA

ARTICLE INFO

Keyword:
Bayesian hierarchical model
harmful algal blooms
informative priors
predictive model
sequential updating

ABSTRACT

We developed a Bayesian hierarchical modeling framework to establish a short-term forecasting model of particulate cyanobacterial toxin concentrations in Western Lake Erie using chlorophyll *a* concentration as the predictor. The model evolves over time with additional data to reflect the changing dynamics of cyanobacterial toxin production. Specifically, parameters of the empirical relationship between the cyanobacterial toxin microcystin and chlorophyll *a* concentrations are allowed to vary annually and seasonally under a hierarchical framework. As such, the model updated using the most recent sampling data is suited to provide short-term forecasts. The reduced model predictive uncertainty makes the model a viable tool for risk assessment. Using data from the long-term Western Lake Erie harmful algal bloom monitoring program (2008–2018), we illustrate the model-building and model-updating process and the application of the model for short-term risk assessment. The modeling process demonstrates the use of the Bayesian hierarchical modeling framework for developing informative priors in Bayesian modeling.

1. Introduction

Large scale harmful and nuisance algal blooms (HABs) in lakes are recognized as a public health threat because of the risk of exposure to cyanobacterial toxins (various congeners of microcystin or *MC*) (Van Dolah et al., 2001). As lakes represent a significant resource providing services from recreation to water supply, forecasting high toxin levels is an important public service. Accurate prediction of toxin levels is, however, elusive because of the poorly understood processes of microcystin production and degradation. In addition to nutrient (phosphorus and/or nitrogen) enrichment and climate change and their impact on the composition of toxic and non-toxic genotypes for cyanobacteria species (Heisler et al., 2008; Davis et al., 2009; O'Neil et al., 2012; Paerl and Huisman, 2008; Suominen et al., 2017), studies have proposed triggers of toxin production ranging from the presence of phytoplanktivorous fish (Jang et al., 2004) to infectious virus (Steffen et al., 2017). Until we have a better understanding of the main causes of toxin production, an empirically-based risk assessment approach offers a reasonable starting point.

Many authors have explored the correlation between *MC* and other related variables such as chlorophyll *a* (hereafter, *Chla*) concentration. For example, several authors used the US EPA's National Lake assessemtn data to model *MC* at a continental scale (Yuan et al., 2014; Yuan and Pollard, 2017; Taranu et al., 2015) using nutrient concentration and a number of environmental variables. They used both classical and Bayesian statistical modeling approaches. Shan et al. (2019) divided *MC* concentrations into high, moderate, and low categories and used a Bayesian networks model to predict the likelihood of each category for three large lakes in China. While modeling *MC* across a large number of lakes based on lake's trophic status (nutrient concentrations) are often successful, such cross-lake models often do not reflect the patterns of change in *MC* within a lake, due to Simpson's paradox (Qian et al., 2019). Stumpf et al. (2016) discussed the use of *Chla* as a predictor of *MC* in Lake Erie to evaluate the spatial variability of *MC*. More recently Liu

* Corresponding author.
E-mail address: song.qian@utoledo.edu (S.S. Qian).

https://doi.org/10.1016/j.ecolind.2021.108055
Received 23 December 2020; Received in revised form 27 July 2021; Accepted 28 July 2021
Available online 3 August 2021

et al. (2020) used a hydrodynamic model and satellite-based *Chla* measurements to forecast near-term *Chla* concentrations in the Western Basin of Lake Erie. The forecasted *Chla* concentrations were then used to estimate *MC* based on simple linear regression models predicting *MC* from *Chla*, developed from recent *in situ* measurements. The model-predicted *MC* was linked to the measured *MC* data again to derive conditional probabilities of *MC* concentration exceeding several relevant environmental management thresholds (Liu et al., 2020). However, researchers are still exploring how to best predict *MC* concentration in Lake Erie.

We conducted an extensive exploratory analysis of the long-term monitoring data from Western Lake Erie that did not reveal any meaningful empirical relationship between *MC* concentration and various candidate predictors including nutrient concentrations. Only measures of phytoplankton biomass (e.g. *Chla* and particulate organic carbon) showed a consistent positive correlation with *MC* concentration. Although phytoplankton biomass is a necessary condition for *MC*, biomass measures, including *Chla*, do not capture the sufficient conditions for high *MC*. For example, the composition of toxic genotypes of *Micrositis spp* in bloom may be a result of factors such as temperature, nutrient (Davis et al., 2009), light limitation (Kardinaal et al., 2007), and/or carbon chemistry and pH (Van de Waal et al., 2011). These are the same factors that determine primary productivity. In other words, the connection between *MC* and *Chla* can be depicted as a latent structure where both *MC* and *Chla* are causally linked to some unknown common factors (or latent variables, Fig. 1). Accordingly, the apparent correlation is "direction-separated" (Pearl et al., 2016). The direction-separated correlation does not imply a direct causal link but can be useful for prediction if the underlying latent varibles are reasonably stable. However, because our lack of knowledge of factors influencing *MC* we expect that the *MC*–*Chla* relationship vary over time. A simple empirical model (e.g. a log–log linear model) fit to existing data may only be valid for a short time. As a result, a simple empirical *MC*–*Chla* relationship needs to be updated frequently to reflect the changing environment. We accommodated possible changes in the underlying structure using a Bayesian hierarchical framework over two temporal scales: allowing model parameters to differ yearly and well as weekly/biweekly within a year. Pooling data using a hierarchical structure (referred to as "partial-pooling" in the literature (Gelman and Hill, 2007)) reduces model parameter uncertainty by including many observations (Qian et al., 2015). Additionally, a hierarchical structure prevents misleading results that can arise when data are improperly aggregated, a phenomenon known as Simpson's paradox (Cha et al., 2016; Qian et al., 2019). While the hierarchical structure is a flexible framework to account for uncertainty due to external factors, a large part of uncertainty associated with *MC* data is related to the lab measurement method (Qian et al., 2015), which cannot be accounted for by external factors. Thus, predictive uncertainty of an empirical model of *MC* is most likely high. A successful predictive model should, therefore, aim at provide a tool for risk assessment, such as predicting the probabilities of *MC* concentration exceeding certain threshold values. Our Bayesian hierarchical model predicted *MC* concentration distribution can be readily used to derive such probabilities.

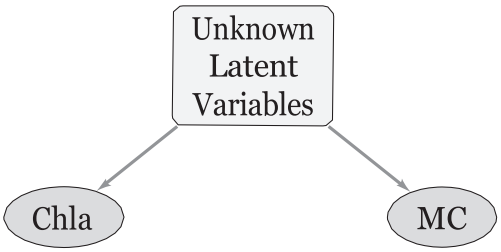


Fig. 1. Direction separated correlation induced by common factors.

2. Methods

2.1. Western Lake Erie sampling

Western Lake Erie was sampled approximately weekly at set sampling stations (Fig. 2) from late spring (April–May) until fall turnover (October) from 2008–2018 (regularly-sampled stations). Additional samples were collected either at the surface (0.75 m) or scums during large *Microcystis* blooms (bloom-chasing samples). Samples were collected with a modified clean Niskin bottle (Fahnenstiel et al., 2002), poured into acid-washed 4 L Nalgene containers, stored in coolers, and brought back to the Great Lakes Environmental Research Laboratory (GLERL) for immediate processing.

Pigment samples were filtered onto Whatman GF/F filters (0.7 μm nominal pore size) and frozen until extraction the next day. *Chla* was estimated with two replicate filters extracted with N, N-dimethylformamide (Speziale et al., 1984) and analyzed on a Turner Designs fluorometer calibrated with commercial *Chla* standards.

Two replicates for particulate microcystin (*MC*) were filtered onto Whatman GF/F 47-mm filter (0.7 μm nominal pore size, 2008–2015) or Isopore 47-mm membrane filters (3 μm pore size, 2016–2017) and frozen. Filters were placed in 70:30 methanol:water, sonicated and then extracted for 12 h at −20 °C (2008–2015). After removing the solvent extract, the extraction was repeated twice again for an hour (2008–2015) (Lawton et al., 1994). After 2015, toxins were extracted using commercially available Abraxis QuikLyse kits. We used the particulate *MC* because it is used by the US EPA in developing recreational ambient environmental criteria for Lake Erie (USEPA, 2019).

2.2. Exploratory data analysis

Our exploratory data analysis included a systematic search of potential predictors of *MC* using scatter plots for direct correlations and various conditional plots for interactive relations among multiple predictors (Cleveland, 1993). We also used classification and regression trees (CART) to search for potential predictive variables (Qian and Anderson, 1999), as well as the linear absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) to explore potential predictors. These exploratory models and plots suggested that concentrations of *MC* and *Chla* had a consistent log–log linear relationship (Fig. 3); this relationship was, however, weak at low *Chla* concentrations, likely due to the high measurement uncertainty of MC (Qian et al., 2015). This is partly due to the log-transformation of *MC* concentrations, which amplifies the uncertainty for low *MC* concentrations (by expanding the ranges for *MC* values below 1) while suppreses the uncertainty for high *MC* values (compressing the range for *MC* values above 1). Accordingly, we modeled the *MC*–*Chla* relationship using a piece-wise linear model (Qian and Richardson, 1997), also known as the hockey-stick model (Qian, 2016) and broken stick model (Chiu et al., 2006). This relationship is empirical. We expect that the hockey-stick pattern (especially the flat lower arm) would be weakened when the measurement uncertainty is reduced.

2.3. The hockey-stick model

The log–log linear relationship between *MC* and *Chla* resembled two joint line segments. When the *Chla* concentration was low, the correlation is weak, either because the high level of measurement uncertainty obscured the correlation or because no correlation is expected as relatively low *Chla* concentrations would occur in the spring and fall when the phytoplankton community is likely nontoxic. Consequently, the line segment at the low end of the *Chla* range was nearly flat (with a slope near 0). The line segment at the high end of the *Chla* range has a positive slope. The resulting model has three parameters: the height of the flat line segment (β_0, assuming $\beta_1 = 0$), the slope of the second line segment ($\beta_1 + \delta$), and the change point (ϕ, in log *Chla*) where the two line

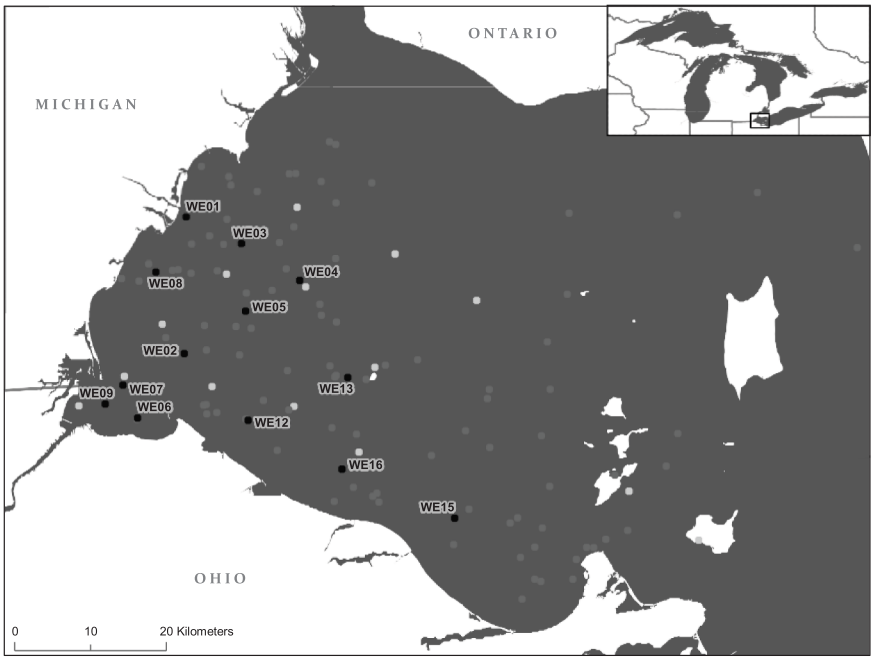


Fig. 2. GLERL HABs sampling sites in Western Lake Erie. Black dots are regularly sampled stations (more than 1 year of multiple site visits), the light gray are sites from 2008 (1 yr of multiple site visits), and dark gray are bloom chasing samples (grab samples only).

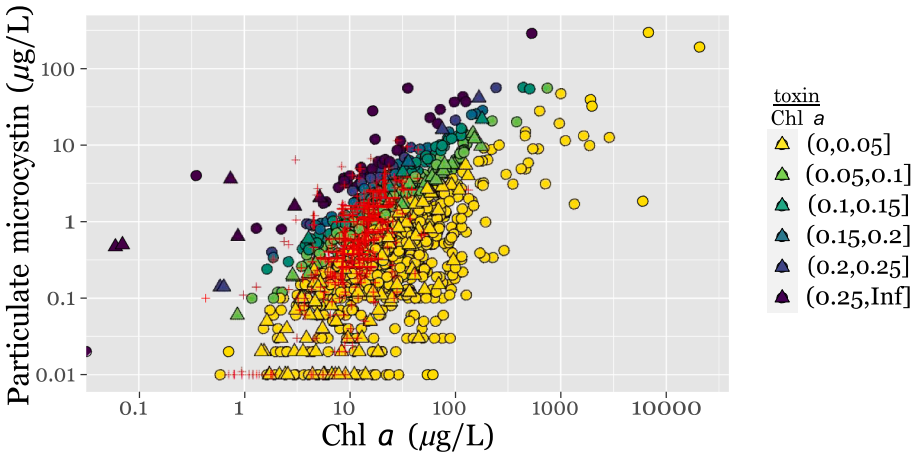


Fig. 3. Log *MC* concentrations are plotted against log *Chla* concentrations for all observations from 2008–2018. Circles depict regularly sampled stations; triangles depict bloom-chasing samples. Color of the data points indicates the *MC:Chla* ratio; note this ratio is independent of *Chla* concentration. The red crosses (+) are data from Saginaw Bay, showing a comparable pattern.

segments join. Combined with the observation/model error term ϵ, the model is:

$$\log(MC_{ij}) = \begin{cases} \beta_{0j} + \beta_{1j}(\log(Chla_{ij}) - \phi_j) & \text{if } \log(Chla_{ij}) \leqslant \phi_j \\ \beta_{0j} + (\beta_{1j} + \delta_j)(\log(Chla_{ij}) - \phi_j) & \text{otherwise} \end{cases} + \epsilon_{ij} \quad (1)$$

where the subscript ij represent the ith observation in year j. Moel parameters (β_0, β_1, ϕ, and δ are year-specific (denoted by the subscript j).

A concentration variable is most likely log-normal (Ott, 1995) (hence log*MC* is most likely normal), as a result, we assume that the error term ϵ has a normal distribution $N(0, \sigma^2)$. The log-transformation streches the range for concentration values below 1 and compresses the range for values above 1. As a result, a common residual variance (σ^2) assumption of the model can be justified.

The hockey stick model is a generalization of the simple log–log linear model because the change point (ϕ) is a parameter to be estimated (Qian, 2014). When the underlying pattern is log–log linear, the estimated change point will be near one of the two ends of the log *Chla* range. The log–log piecewise linear pattern of the *MC*–*Chla* relationship is also observed in Saginaw Bay (Fig. 3) and elsewhere (Kelly et al., 2019). A smoothed version of the hockey-stick model (2) was used for computational stability (see Online Supplementary Materials for details):

$$y = \beta_0 + \beta_1(x - \phi) + (\beta_1 + \delta)\lambda \log\left(1 + e^{\frac{x-\phi}{\lambda}}\right) + \epsilon \quad (2)$$

where $y = \log(MC), x = \log(Chla)$, and λ is a smoothness parameter.

There are 6% of *MC* concentration values below the method reporting limit of 0.01 μg/L (i.e., censored at 0.01). The censored data were addressed in our computer code, specifically, the likelihood function for a censored *MC* value is defined by the normal distribution

cumulative density function (cdf), while the likelihood function of an uncensored value is define by the normal distribution probability density function (pdf).

Because the *MC–Chla* relationship is not directly causal, we allowed it to vary over time, assuming that the latent structure includes variables operating at different scales. Thus, the base model form (Eq. (2)) stayed the same but model coefficients varied from year to year. Consequently, our prior idea (the direction-separated correlation) was properly reflected in the model, consistent with both frequentist (Stein, 1956; Efron and Morris, 1975) and Bayesian (Box and Tiao, 1973; Box, 1983; Gelman, 2005) theories (Cox, 1995). The natural modeling approach for this problem is the Bayesian hierarchical model. Specifically, we allowed model coefficients (β_0,β_1,δ, and ϕ) to vary among years without imposing other assumptions with respect to their relative magnitudes among years. For this work, we assume that $\beta_1 = 0$. As we have no additional information on the direction of change from one year to another, a common prior is imposed on these coefficients:

$$\begin{pmatrix} \beta_{0j} \\ \delta_j \\ \phi_j \end{pmatrix} \sim MVN\left[\begin{pmatrix} \mu_{\beta_0} \\ \mu_\delta \\ \mu_\phi \end{pmatrix}, \Sigma \right] \tag{3}$$

By fitting a BHM, our goal is to specify the prior distribution, a multivariate normal distribution parameterized by hyper-parameters $\mu_{\beta_0}, \mu_\delta, \mu_\phi$, and Σ, representing the long-term means of the respective coefficients and their (among year) variance–covariance matrix. The hyper-distribution can be estimated when data from sufficient number of years are available. Afterwards, the hockey-stick model can be evaluated sequentially: each time when new data are available, the hockey-stick model is fit as a (nonlinear) regression model using the most recently estimated hyper-distribution as the prior for model parameters. The posterior distributions of year-specific parameters are estimated, as well as the hyper-distribution. This sequential updating process is evaluated in this study as follows. We used data from 2008–2016 to fit the Bayesian hierarchical model (equations (1) and (3)) and the resulting posterior distribution of the hyper-distribution is used and updated sequentially to estimate the year-specific model parameters using data from 2017 and 2018. These sequentially updated year-specific models were used to make short-term predictions of *MC* probabilities of *MC* concentration exceeding selected thresholds under different *Chla* conditions.

We tried to incorporate temperature explicitly (online supporting materials) because it is a potential factor affecting cyanobacterial growth and toxin production (Pearl et al., 2016). However, our model was unable to separate the temperature effect from the annual variation reflected in the annual-based hierarchical structure. This is likely because low temperatures often occur in spring and late fall when *Chla* is also low. As a result, variation in *MC* due to temperature change was partially accounted for by changes in *Chla*. More importantly, we do not have information on the changes in cyanobacteria genotypes which may be more important than the effect of temperature. As a result, we summarize only the results without the temperature effect explicitly modeled (Eq. (2)).

2.4. Short-term forecasting and long-term updating

Within a given year, we fit the hockey stick model on a weekly/biweekly basis, using the most recent monitoring data and a prior distribution for the model parameters cumulatively updated based on the previous years' data. The prior distribution parameters were incrementally updated to reflect the short-term (or seasonal) variation. These incrementally updated models were used to forecast the risk of high *MC* events for the subsequent one or two weeks before the next sampling event (prediction period). The predicted *MC* concentrations were used to derive risks (probabilities) of *MC* concentration exceeding relevant thresholds as a function of *Chla*. At the end of each year, the data from the current year are included in the hierarchical model to update the hyper-distribution. Specifically, to evaluate this process, we divide each year to periods, each with at least 8 observations, representing one to two weeks of time. We used the model developed using data from 2010 to 2016 to predict the *MC* concentrations observed in the first period of 2017. After the prediction, the model is updated using the data from the first period. The updated model was used to predict *MC* concentrations in the next period. At each step, we evaluate the prediction accuracy. (Computational details are in the online supplementary materials.).

This iterative process of model-fitting, forecasting, and model-updating allowed us to approximate the changes in the factors affecting both *MC* and *Chla* without knowing exactly what factors are involved. Furthermore, the use of the hyper-distribution as the prior distribution enabled a meaningful short-term forecast model based on a limited sample size. The annual updating step ensures that we gradually improve the specification of the priors of the hockey-stick model parameters.

2.5. Model checking using the predictive quantile

We followed the general principle of the Bayesian predictive simulation approach for model assessment (Gelman et al., 1996; Gelman et al., 2014) and developed a probabilistic measure for evaluating our model's short-term predictive accuracy. Specifically, we derived a predictive log *MC* concentration distribution for each observation in a prediction period using the posterior model updated immediately prior to the prediction period (without using the *MC* concentrations we are predicting). A good predictive distribution should capture the intended future observation. One way to evaluate the predictive distribution is to compare each observation to the respective predictive distribution and calculate the probability of observing an observation larger than or equal to the observed data point (the upper tail area) as a measure of the predictive accuracy (see Figs. 9.3 and 9.4 in Qian, 2016). A good model should lead to predictive distributions that can include a target within the 90% credible interval 90% of the time. In other words, a good predictive model should have upper tail areas between 0.05 and 0.95 for 90% of the data.

3. Results

3.1. The MC–Chla relationship

Fitting the existing monitoring data to the hierarchical hockey-stick model (Fig. 3) shows the year-to-year variation of the *MC–Chla* relationship. The hockey-stick model relationship is not always obvious when only data from a single year are plotted (Fig. 4), because of the range of *MC* varies from year to year. In years without very low concentrations of *Chla*, the flat line segment may not materialize. The fitted year-specific models largely reflect the data (reducing the fitted models to a linear one when appropriate). The flat line segment serves as a lower bound of estimated *MC* reflecting the high level of uncertainty associated with the ELISA method (Qian et al., 2015). Although precisely predicting *MC* levels is difficult due in part to the measurement uncertainty, the Bayesian representation of the *MC–Chla* relationship appropriately quantifies the likely distribution of *MC* concentration at a given *Chla* concentration. As a result, the model is suitable as a risk assessment tool for predicting the probability of *MC* concentration exceeding various thresholds under a given *Chla* level.

3.2. Weekly updating and prediction

Using the estimated hyper-parameters based on data from 2010–2016 to specify the prior model, we updated the models for 2017 and 2018 sequentially using incremental weekly/biweekly data. Each time when the model was updated with data collected up to the sampling date we used the resulting posterior model for risk assessment

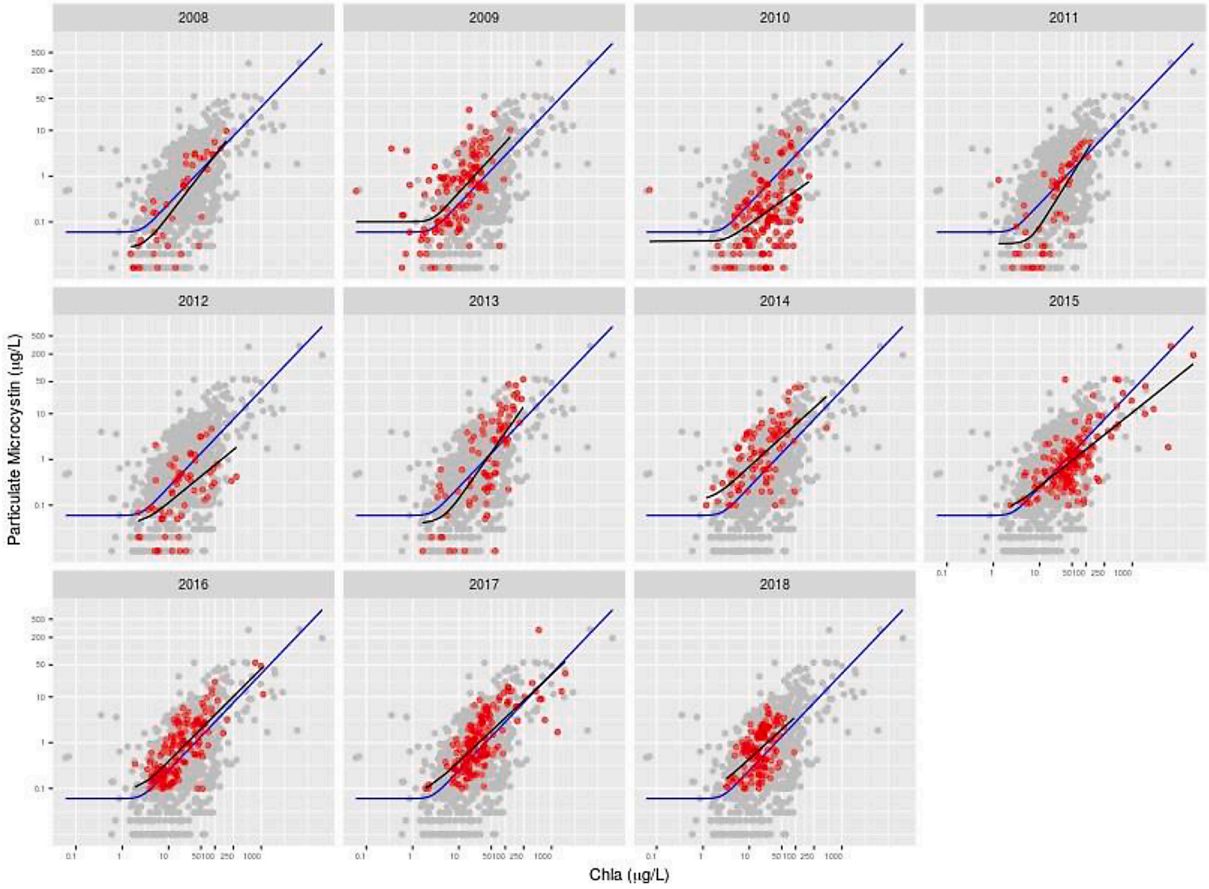


Fig. 4. The hierarchically fit annual hockey stick models (black lines) are compared to the overall average model and the annual data points (red dots) and combined data (gray dots).

(Figs. 5 and 6). Large cyanobacteria blooms were observed in 2017. With wide ranges of *MC* and *Chla*, the posterior *MC–Chla* model captured the general pattern after two updates, after which the sequentially updated models accurately predicted *MC* of the next sampling event until October, when the posterior model began to over-predict. In 2018, Lake Erie experienced a relatively calm year in terms of algal blooms. As observed *MC* were largely below 5 µg/L for the entire year, the noise (uncertainty levels in *MC*) to signal (the *MC–Chla* relationship) ratio is high, accurate prediction is unlikely. However, as a risk assessment tool, the model is adequate for consistently predicting low risk (probability) of MC exceeding 5 µg/L.

3.3. Model evaluation

Data from 2017 and 2018 were used for model evaluation, using the posterior distribution of the hyper-parameters of the hierarchical model fitted based on data up to 2016 as priors. In each step of the sequential updating, we use the posterior model updated for the time step to predict the next sampling period (the prediction period). In each prediction period, we have at least 8 observations. We calculate the upper tail areas for these data points and present them using a simplified boxplot for each prediction period. All upper tail areas are well within the range of 0.05 and 0.95 (Fig. 7).

4. Discussion

Statistical models are based on assumptions, and the validity of these assumptions determines the utility of the models. Using a simple regression of *MC* against *Chla*, we normally either fit the model using data from multiple years or data from a single sampling event. When fitting the model using data from multiple years, the resulting model represents the long-term average of the *MC–Chla* relationship. When using data from a single sampling event, the fitted model likely embeds substantial temporal variability. Neither approach is likely to produce a reliable short-term risk assessment. However, without an accurate scientific understanding of the factors that trigger the production of microcystin, we are unlikely to develop a process-based model for the task. The Bayesian hierarchical modeling framework we adopted is a natural compromise where we can include annual variation in the model when building the initial model using historical data. Using the same hierarchical modeling approach, we also included seasonal variation for short-term forecasting. The short-term forecasting model is based on the most recent sampling data with the long-term average *MC–Chla* relationship serving as an anchor to avoid the common pitfalls of fitting a regression model with a small sample size (e.g., fitting noise).

We used *MC* concentration exceeding a specific threshold as an adverse event because the occurrence of such an event may trigger certain management practices (e.g., initiating costly carbon filtration in a local drinking water plant). Our model can be readily integrated into a risk assessment process for developing strategies for mitigating anticipated occurrence of such adverse event. As a risk assessment tool, our model is well suited for predicting the probability of occurrence (risk) of such an event. We predict the risk as a function of *Chla* because of the availability of *Chla* concentrations reported from real-time sensors in important locations in Western Basin of Lake Erie, such as the drinking water intake of the Toledo drinking water plant.

Our model is limited to predicting the *MC* concentrations at a spatial scale of the Western Lake Erie, as an empirical model is always limited by data. Because the uneven spatial distribution of HABs, location-specific forecasting is often more important (e.g., for locations with

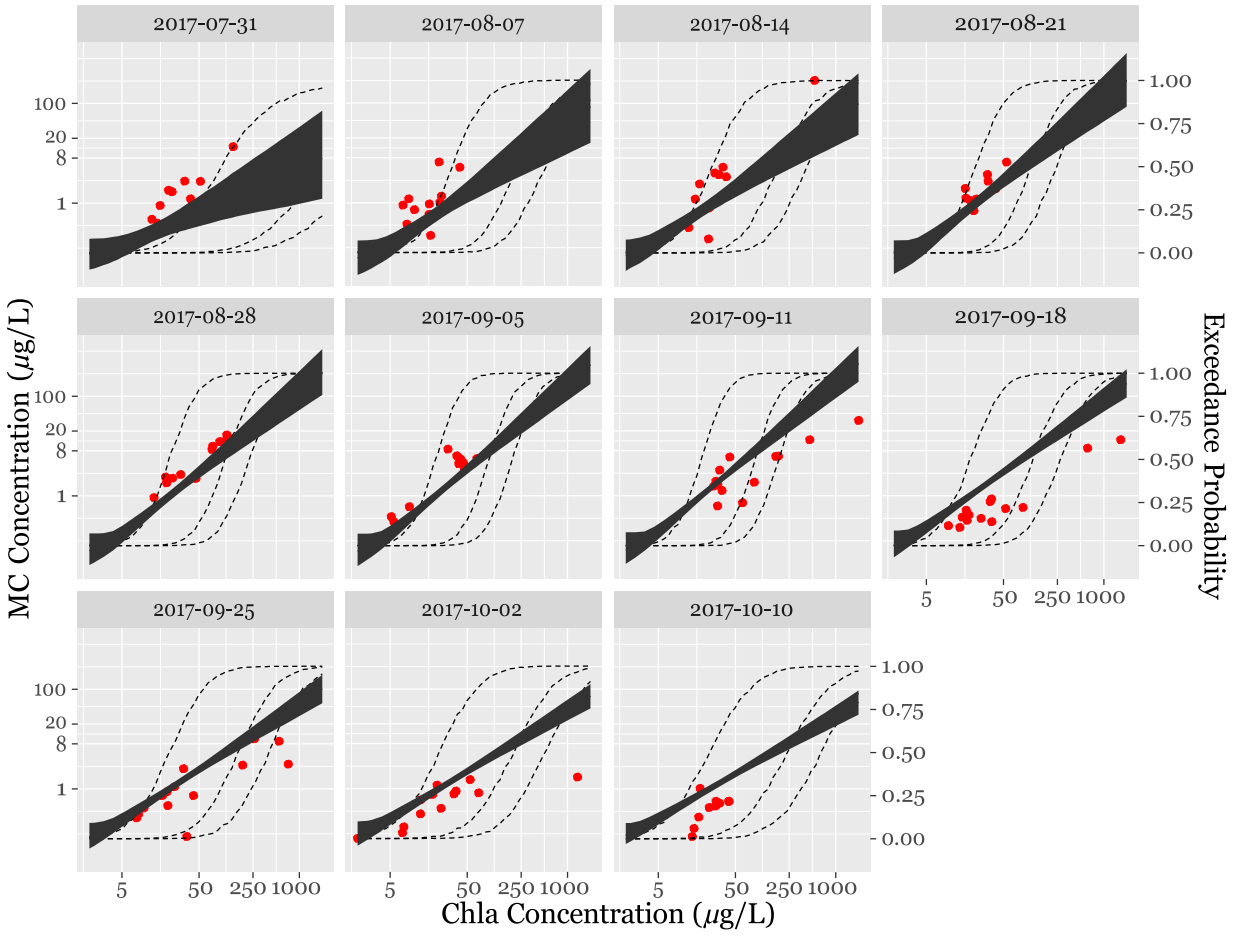


Fig. 5. Near-term forecast of the probability of *MC* exceeding selected thresholds for the year 2017. Red dots in each panel were the observations made during the sampling period beginning at the date shown on the top of the panel. The black lines are the fitted model along with the 95% predictive interval (gray shaded polygons). The dashed, broken, and dotted lines are the predicted probabilities of MC exceeding 1, 8, and 20 μg/L, respectively.

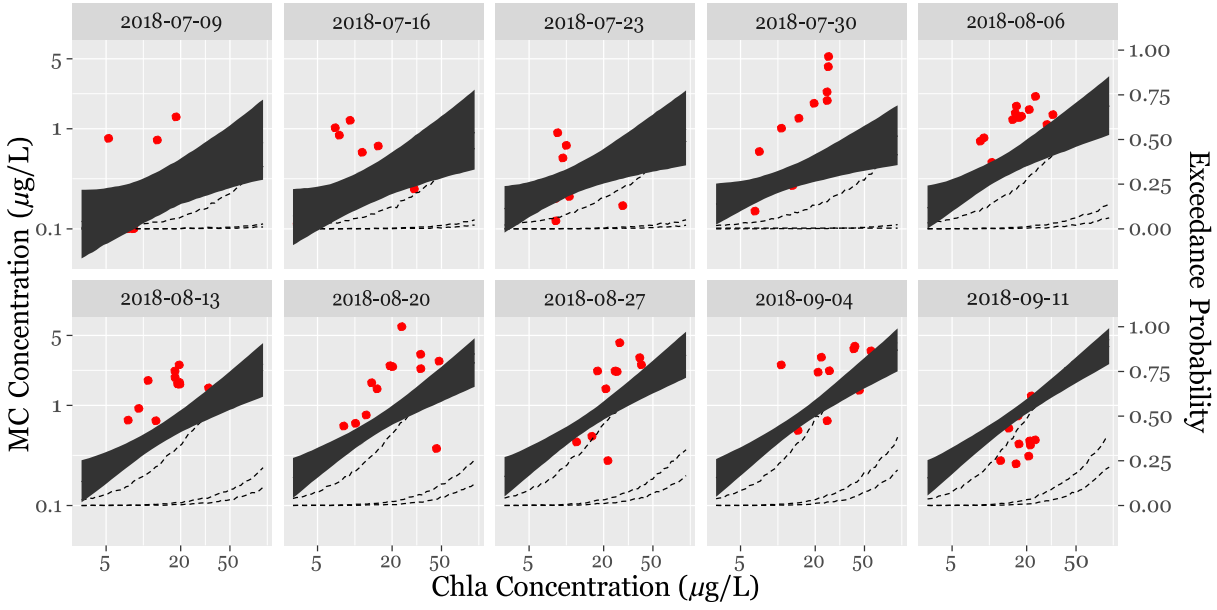


Fig. 6. Near-term forecast of the probability of *MC* exceeding selected thresholds (1, 5, and 8 μg/L) for the year 2018. See Fig. 5.

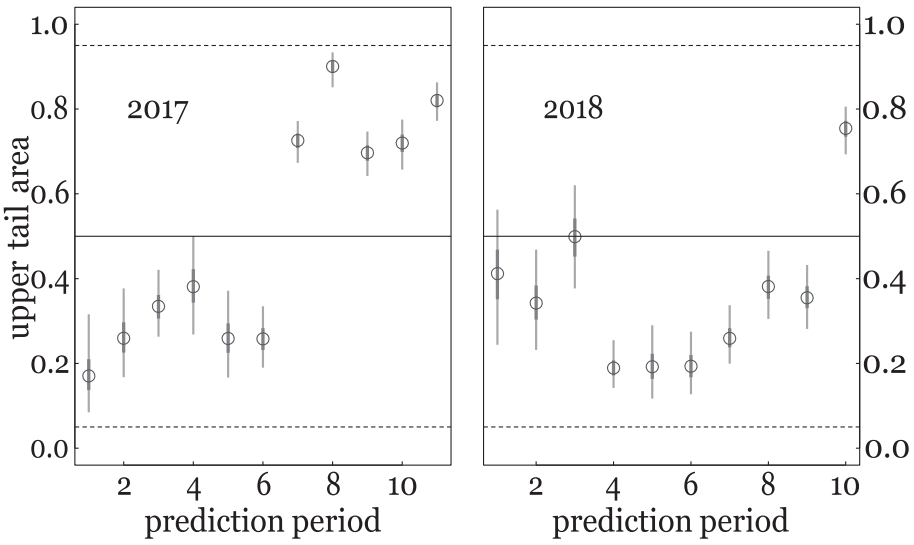


Fig. 7. Predictive upper tail areas, estimated for data points in each prediction period (red dots in Figs. 5 and 6), are presented using simplified box-plots. The open circles represent the median, thick dark lines are the interquartile range, and the thin gray lines are the 90% credible intervals.

drinking water intakes and areas of recreational values). In principle, we can extend our modeling framework to include specific locations if regular monitoring data for these locations are available. Specifically, a location-specific model can take the same functional form as the model we presented. Instead of using the hyper-parameter distribution as the prior (summarizing annual variation), we can use the year-specific model parameter distribution as the basis for developing prior distribution for the location-specific model at first. As we accumulate more location-specific data, each location can develop a location-specific sequential updating process.

Although spring TP loading is the main predictor for annual peak algal bloom size (Stumpf et al., 2012; Obenour et al., 2014), spring TP loading from the main source (Maumee River) was not strongly related to lake *MC* and *Chla* concentrations (Rowland et al., 2020). The lack of loading-concentration association appears to suggest that the nutrient loading determines the spatial extent of elevated nutrient concentrations, while algal growth is largely determined by nutrient concentration, which further indicates the need for a separate modeling effort for evaluating toxin production.

Our Bayesian hierarchical modeling approach leverages the theoretical understanding of the shrinkage estimator from both classical and Bayesian studies. By comparing the empirical Bayes literature and studies on Stein's paradox, we suggest that the Bayesian hierarchical modeling framework provides a descriptive means for deriving informative prior distribution as the distribution of the parameter of interest among exchangeable units. In our study, the parameters of interest are the coefficients of the hockey-stick model and the exchangeable units are multiple years.

5. Conclusions

- A predictive model of cyanotoxin microcystin is an important tool for communities affected by severe eutrophication worldwide. Statistical modeling is often used because of the poorly understood mechanisms of microcystin production. Applications of such models are necessarily at an individual-lake level. Consequently, an effective empirical model must account for the inevitable temporal variation in the model due to changing causes of microcystin production.
- We used a Bayesian hierarchical modeling approach to develop a simple predictive model of microcystin using the commonly measured chlorophyll-*a* concentration as the sole predictor. The hierarchical structure at two temporal scales (over multiple years and over weekly/biweekly sampling periods within a given year) allows model coefficients to be updated over time to make short-term forecasts of microcystin concentrations in Western Lake Erie.
- The basic model form of a hockey-stick *MC–Chla* is not unique to Lake Erie.
- We fit the predictive model using long-term monitoring data from NOAA-GLERL from 2008 to 2018. We tested the model's predictive performance using the same model fit with data from 2008 to 2016 and making short-term predictions for 2017 and 2018. Our model can compliment the on-going Western Lake Erie HABS forecasting provided by NOAA by adding a weekly/biweekly toxin advisory.

CRediT authorship contribution statement

Song S. Qian: Conceptualization, Formal analysis, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. **Craig A. Stow:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Freya E. Rowland:** Data curation, Formal analysis, Software, Validation, Visualization. **Qianqian Liu:** Writing - review & editing. **Mark D. Rowe:** Writing - review & editing. **Eric J. Anderson:** Writing - review & editing. **Richard P. Stumpf:** Writing - review & editing. **Thomas H. Johengen:** Writing - review & editing, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank L. Mason for providing sampling location information. Funding for this research was provided partially by NOAA, NOAA-GLERL contribution number: 1986

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ecolind.2021.108055.

References

Box, G.E.P., 1983. An appology for ecumenism in statistics. In: Sci. Inference Data Anal. Robust. Academic Press, pp. 51–84.

Box, G.E.P., Tiao, G.C., 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, MA.

Cha, Y.K., Park, S.S., Lee, H.W., Stow, C.A., 2016. A Bayesian hierarchical approach to model seasonal algal variability along an upstream to downstream river gradient. Water Resour. Res. 52, 348–357.

Chiu, G., Lockhart, R., Routledge, R., 2006. Bent-cable regression theory and applications. J. Am. Stat. Assoc. 101 (474), 542–553.

Cleveland, W.S., 1993. Visualizing Data. Hobart Press, Summit, NJ.

Cox, D.R., 1995. The relation between theory and application in statistics [with discussions]. Test 4 (2), 207–261.

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. https://doi.org/10.1016/j.hal.2009.02.004.

Efron, B., Morris, C., 1975. Data analysis using Stein's estimator and its generalizations. J. Am. Stat. Assoc. 70 (350), 311–319.

Fahnenstiel, G.L., Beckmann, C., Lohrenz, S.E., Millie, D.F., Schofield, O.M.E., McCormick, M.J., 2002. Standard niskin and van dorn bottles inhibit phytoplankton photosynthesis in lake michigan. Internationale Vereinigung für Theoretische und Angewandte Limnologie: Verhandlungen 28 (1), 376–380.

Gelman, A., 2005. Analysis of variance – why it is more important than ever (with discussions). Ann. Stat. 33 (1), 1–53.

Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York.

Gelman, A., Meng, X.L., Stern, H.S., 1996. Posterior predictive assessment of model fitness via realized discrepencies (with discussions). Statist. Sin. 6, 733–807.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, David B., Vehtari, Aki, Rubin, D.B., 2014. Bayesian Data Analysis, 3rd edition,. CRC Press, Boca Raton, Florida.

Heisler, J., Glibert, P.M., Burkholder, J.M., Anderson, D.M., Cochlan, W., Dennison, W. C., Dortch, Q., Gobler, C.J., Heil, C.A., Humphries, E., Lewitus, A., Magnien, R., Marshall, H.G., Sellner, K., Stockwell, D.A., Stoecker, D.K., Suddleson, M., 2008. Eutrophication and harmful algal blooms: A scientific consensus. Harmful Algae 8, 3–13.

Jang, M.H., Ha, K., Lucas, M.C., Joo, G.J., Takamura, N., 2004. Changes in microcystin production by Microcystis aeruginosa exposed to phytoplanktivorous and omnivorous fish. Aquat. Toxicol. 68 (2004), 51–59.

Kardinaal, W.E.A., Tonk, L., Janse, I., Hol, S., Slot, P., Huisman, J., and Visser, P.M. 2007. Competition for light between toxic and nontoxic strains of the harmful cyanobacterium microcystis. Appl. Environ. Microbiol., 73(9): 2939–2946. ISSN 0099–2240. doi:10.1128/AEM.02892-06. url:https://aem.asm.org/content/73/9/2939.

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 Res. 162, 288–301.

Lawton, L.A., Edwards, C., Codd, G.A., 1994. Extraction and high-performance liquid chromatographic method for the determination of microcystins in raw and treated waters. Analyst 119 (7), 1525–1530.

Liu, Q., Rowe, M.D., Anderson, E.J., Stow, C.A., Stumpf, R.P., Johengen, T.H., 2020. Probabilistic forecast of microcystin toxin using satellite remote sensing, in situ observations and numerical modeling. Environ. Modell. Softw. 128, 104705.

Obenour, D.R., Gronewold, A.D., Stow, C.A., Scavia, D., 2014. Using a Bayesian hierarchical model to improve Lake Erie cyanobacteria bloom forecasts. Water Resour. Res. 50, 7847–7860.

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.

Ott, W.R., 1995. Environmental Statistics and Data Analysis. Lewis Publishers, Boca Raton.

Paerl, H.W., Huisman, J., 2008. Blooms like it hot. Science 320, 57–58.

Pearl, J., Glymour, M., Jewell, N.P., 2016. Causal Inference in Statistics. Wiley, Chichester, UK.

Qian, S.S., 2014. Ecological threshold and environmental management: A note on statistical methods for detecting thresholds. Ecol. Ind. 38, 192–197.

Qian, S.S. 2016. Environmental and Ecological Statistics with R. Chapman and Hall/CRC Press, 2nd edition.

Qian, S.S., Anderson, C.W., 1999. Exploring factors controlling variability of pesticide concentrations in the Willamette River Basin using tree-based models. Environ. Sci. Technol. 33, 3332–3340.

Qian, S.S., Richardson, C.J., 1997. Estimating the long-term phosphorus accretion rate in the Everglades: A Bayesian approach with risk assessment. Water Resour. Res. 33 (7), 1681–1688.

Qian, S.S., Chaffin, J.D., DuFour, M.R., Sherman, J.J., Golnick, P.C., Collier, C.D., Nummer, S.A., Margida, M.G., 2015. Quantifying and reducing uncertainty in estimated microcystin concentrations from the ELISA method. Environ. Sci. Technol. 49 (24), 14221–14229.

Qian, S.S., Stow, C.A., Cha, Y.K., 2015. Implications of Stein's Paradox for environmental standard compliance assessment. Environ. Sci. Technol. 49 (10), 5913–5920.

Qian, S.S., Stow, C.A., Farnaz, A., Nojavan, 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. https://doi.org/10.1016/j.watres.2019.114855.

Rowland, F.E., Stow, C.A., Johengen, T.H., Burtner, A.M., Palladino, D., Gossiaux, D.C., Davis, T.W., Johnson, L.T., Ruberg, S., 2020. Recent patterns in Lake Erie phosphorus and chlorophyll a concentrations in response to changing loads. Environ. Sci. Technol. 54 (2), 835–841.

Shan, K., Shang, M., Zhou, B., Li, L., Wang, X., Yang, H., Song, L., 2019. Application of Bayesian network including microcystis morphospecies for microcystin risk assessment in three cyanobacterial bloom-plagued lakes. Harmful Algae 83, 14–24.

Speziale, B.J., Schreiner, S.P., Giammatteo, P.A., Schindler, J.E., 1984. Comparison of n, n-dimethylformamide, dimethyl sulfoxide, and acetone for extraction of phytoplankton chlorophyll. Can. J. Fish. Aquat. Sci. 41 (10), 1519–1522.

Steffen, M.M., Davis, T.W., McKay, R.M.L., Bullerjahn, G.S., Krausfeldt, L.E., Stough, J. M., et al., 2017. Ecophysiological examination of the Lake Erie Microcystis bloom in 2014: linkages between biology and the water supply shutdown of Toledo, OH. Environ. Sci. Technol. 51 (12), 6745–6755.

Stein, C. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 197–206. University of California Press.

Stumpf, R.P., Wynne, T.T., Baker, D.B., Fahnenstiel, G.L., 2012. Interannual variability of cyanobacterial blooms in Lake Erie. PLoS ONE 7 (8), e42444.

Stumpf, R.P., Davis, T.W., Wynne, T.T., Graham, J.L., Loftin, K.A., Johengen, T.H., Gossiaux, D., Palladino, D., Burtner, A., 2016. Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. Harmful Algae 54, 160–173.

Suominen, S., Brauer, V.S., Rantala-Ylinen, A., Sivonen, K., Hiltunen, T., 2017. Competition between a toxic and a non-toxic microcystis strain under constant and pulsed nitrogen and phosphorus supply. Aquat. Ecol. 51 (1), 117–130. https://doi.org/10.1007/s10452-016-9603-2.

Taranu, Z.E., Gregory-Eaves, I., Steele, R.J., Beaulieu, M., Legendre, P., 2015. Predicting microcystin concentrations in lakes and reservoirs at a continental scale: A new framework for modelling an important health risk factor. Glob. Ecol. Biogeogr. 26 (425), 625–637.

Tibshirani, Robert, 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. (B) 58 (1), 267–288.

USEPA, 2019. Recommended human health recreational ambient water quality criteria or swimming advisories for microcystins and cylindrospermopsin. Technical report, Office of Water, EPA 822-R-19-001. US Environmental Protection Agency, Washington, D.C.

Van de Waal, D.B., Verspagen, J.M.H., Finke, J.F., Vournazou, V., Immers, A.K., Kardinaal, W.E.A., Tonk, L., Becker, S., Van Donk, E., Visser, P.M., Huisman, J., 2011. Reversal in competitive dominance of a toxic versus non-toxic cyanobacterium in response to rising co_2. The ISME J. 5, 1438–1450.

Van Dolah, F.M., Roelke, D., Greene, R.M., 2001. Health and ecological impacts of harmful algal blooms: Risk assessment needs. Human Ecol. Risk Assess. 7 (5), 1329–1345.

Yuan, L.L., Pollard, A.I., 2017. Using national-scale data to develop nutrient-microcystin relationships that guide management decisions. Environ. Sci. Technol. 433 (51), 6972–6980.

Yuan, L.L., Pollard, A.I., Pather, S., Oliver, J.L., D'Anglada, L.D., 2014. Managing microcyctin: identifying national-scale thresholds for total nitrogen and chlorophyll a. Freshw. Biol. 59, 1970–1981.