EPA Public Access

Author manuscript

Harmful Algae. Author manuscript; available in PMC 2020 April 10.

About author manuscripts

Submit a manuscript

Published in final edited form as:

Harmful Algae. 2019 April ; 84: 75–83. doi:10.1016/j.hal.2019.02.009.

Combining national and state data improves predictions of microcystin concentration

Lester L. Yuan*, Amina I. Pollard

Office of Water, U.S. Environmental Protection Agency, Washington DC 20460, USA.

Abstract

Data collected from lakes at national (regional) scales and state (local) scales can provide different insights regarding relationships between environmental factors and biological responses, and combining these two types of data can potentially yield more precise and accurate understanding of ecological phenomena. National data can include many measures, cover large spatial areas, and span broad environmental gradients. Because of these characteristics, analyses of these data can yield accurate estimates of relationships among different lake characteristics. However, the number of samples in a national data set that is available for estimating a relationship specific to waterbodies within a smaller region, like a single state, is limited. Conversely, state monitoring data provide intensive sampling of lakes within a smaller area, but these data span a narrower range of conditions and may only include a subset of relevant measurements. Here, Bayesian network model is described that represents the causal linkages between observations of chlorophyll *a* concentration, cyanobacterial biovolume, and microcystin concentration. This network model was fit to national data and provided a context for modeling observations of chlorophyll *a* and microcystin collected from lakes in Iowa. Using the knowledge inherent in the national network model improved the accuracy of predictions of microcystin concentrations in Iowa compared to a model based only on Iowa data.

Keywords

cyanotoxin; microcystin; chlorophyll; cyanobacteria; Bayesian network

1. Introduction

Large amounts of environmental data are becoming increasingly available for investigating ecological patterns at regional and local spatial scales. These data can be used to explore relationships between environmental factors and biological responses (see, for example, www.data.gov), and estimates of these relationships can be used to inform environmental management decisions (Yuan and Pollard, 2017). Combining different data sets increases the number of samples available for analysis and can expand the spatial and temporal extent of the data, improving the ability to estimate broadly applicable relationships. Indeed, analysis of several different data sets can be the only way to examine certain large scale phenomena,

*Corresponding author: yuan.lester@epa.gov.

Declarations of interest: None

as few organizations have the resources to coordinate, collect, and manage extremely large amounts of data (Soranno and Schimel, 2014).

Unique challenges are associated with analyzing combined data. First, the initial process of combining data is nontrivial. Even when disparate data are housed in a single repository, the quality of individual data sets must be confirmed, and different data formats must be standardized (Sprague et al., 2017). Then, measurements of similar parameters must be available in each of the data sets, a restriction that often limits analysis to commonly collected parameters. For example, in lakes, total phosphorus, total nitrogen, and chlorophyll *a* (chl *a*) are most frequently measured by research programs and local to national agencies, and therefore, analyses of combined datasets have often focused on relationships among these three parameters (Phillips et al., 2008; Soranno et al., 2010; Wagner et al., 2011). Differences in the spatial scale and location of individual data sets should also be considered when analyzing combined data because biases associated with spatial location can exist. More specifically, environmental covariates can vary systematically in space, and relationships estimated from a data set collected from one location may reflect the influence of an environmental covariate that is absent from another location. Analyses that treat combined data as a single pooled dataset implicitly assume that the effects of spatially-varying covariates are weak relative to the main effects being estimated. In contrast, multilevel statistical models can explicitly account for these differences, and application of these types of models have often yielded additional insights (Wagner et al., 2011).

Analysis of water quality monitoring data collected by national and state organizations in the U.S. typifies the challenges and potential benefits of combining data sets. Water quality management in the U.S. is directed at local and state scales. As a result, resource managers often want to explore relationships between environmental factors and biological responses at these smaller, local spatial scales. Depending on the question, opportunities might be available to leverage knowledge from broader regional scales (e.g., national scale) to enhance local understanding. National (e.g., continental-scale) data span broad environmental gradients (US EPA, 2010), and in the case of the National Lakes Assessment, provide a comprehensive set of measurements collected from large numbers of sites with identical protocols (Pollard et al., 2018; US EPA, 2011). The availability of consistent measures from lakes spanning broad gradients facilitates accurate, national estimates of relationships of interest. However, the number of samples within that national data set that are available to estimate relationships within any single state is small. As a result, it can be difficult to rely solely on a national data set to estimate relationships specific to a state. In contrast, monitoring conducted by state agencies can yield more intensive temporal sampling over more sites, and hence, relationships estimated from these data more readily inform local management decisions within that state. However, data collected at the state level can be limited in terms of the parameters that are measured, and the range of environmental conditions that is sampled is limited by conditions occurring within the state boundaries. Here, a Bayesian network model is described that provides a framework for combining national-scale and state-scale data to inform environmental decision making within an individual state.

The model focuses specifically on the problem of predicting algal toxin concentrations in lakes. Given reports of increasing algal blooms and cyanotoxin occurrences (Paerl and Huisman, 2009, 2008; Taranu et al., 2015), environmental scientists have been pressed to estimate relationships and develop strategies to understand management options (Yuan et al., 2014; Yuan and Pollard, 2017). These relationships and strategies should reflect both general information that can be transferred among different systems and specific information that pertains to a particular location of interest. In the conterminous United States, the National Lakes Assessment provides co-located chl *a*, cyanobacterial biovolume, and microcystin concentration data from a broad range of lakes and conditions (US EPA, 2010). In contrast, local, state monitoring data provide the advantage of increased spatial resolution at a local scale, albeit over a limited set of conditions. Recognizing these limitations, we posit that joining these data sets with a Bayesian network can highlight how a regional data set can provide broad context for local data, resulting in more precise and accurate estimates of predictive relationships for microcystin within a particular state.

2. Materials and methods

2.1. Data

2.1.1. National data.—Data from lakes and reservoirs in the United States were collected by the U.S. Environmental Protection Agency's National Lakes Assessment (NLA) in the summer (May-September) of 2012 (<https://www.epa.gov/national-aquatic-resource-surveys/nla>). The NLA consisted of a stratified randomized sample of lakes from the conterminous United States. Lakes greater than 1 ha were selected within the conterminous United States using a stratified random sampling design (US EPA, 2012a). The final data set was supplemented by a small number of hand-picked lakes and reservoirs that were identified a priori as being less disturbed by human activities (US EPA, 2010). Approximately 10% lakes of the lakes were randomly selected and resampled on a different day after the initial visit. The timing of the second visit varied among lakes, but on average, the second sample was collected approximately 46 days after the first.

During each visit to a selected lake, an extensive suite of abiotic and biological variables was measured. Only brief details on sampling protocols are provided here regarding the parameters used in this analysis; more extensive descriptions of sampling methodologies are available (US EPA, 2011). At each lake, a sampling location was established in open water at the deepest point of each lake (up to a maximum depth of 50 m) or in the mid-point of reservoirs. An additional sampling location for collection of microcystin, algae, and chl *a* data was established in the littoral zone approximately 10 m out from a randomly selected point on the shoreline.

At the open water site, a water sample was collected using a vertical, depth-integrated methodology that collected water from the photic zone of the lake (to a maximum depth of 2 m). Multiple sample draws were combined in a rinsed, 4 liter (L) cubitainer. When full, the cubitainer was gently inverted to mix the water, and an aliquot was taken as the water chemistry sample. This subsample was placed on ice and shipped overnight to the Willamette Research Station in Corvallis, Oregon. A second aliquot was taken for

characterizing the phytoplankton community and was preserved with a small amount of Lugol's solution.

At the littoral zone site, two grab water samples were collected 0.3 m below the surface at a depth of at least 1 m using a 2 L brown bottle. The first sample was split into one subsample for quantifying algal toxin concentration and a second subsample sample was preserved with a small amount of Lugol's solution and used to characterize the phytoplankton community. The second grab sample collected with the 2 L bottle was used to quantify chl *a* concentration.

To measure chl *a* concentration, 250 mL of lake water from the lake center and from the littoral zone site were pumped through a glass fiber filter in the field, and quantified in the lab. The lower reporting limit was 0.5 µg/L for chl *a*.

Cyanobacterial biovolume was quantified from the field samples in the laboratory. Samples collected from both open water and littoral zone locations were examined by taxonomists, who identified at least 400 natural algal units under 1000× magnification to the lowest practical taxonomic resolution (usually species). Cyanobacteria were aggregated and abundance was calculated as cells per mL. In each sample, the dimensions of the taxa that accounted for the largest proportions of the observed assemblage were measured and used to estimate biovolume. Biovolumes of the most abundant taxa were based on the average of measurements from at least 10 individuals, while biovolumes for less abundant taxa were based on somewhat fewer measurements (US EPA, 2012b). Approximately 5% of the phytoplankton samples were randomly selected and re-identified and measured by a second taxonomy lab. These re-identified samples provided a basis for estimating lab measurement error.

Microcystin sample processing began with three sequential freeze/ thaw cycles to lyse cyanobacteria (Loftin et al., 2008). Processed samples were filtered using 0.45 micron polyvinylidene difluoride membrane syringe filters and stored frozen until analysis. The concentration of microcystin in the filtered water sample was measured with a polyclonal Enzyme-Linked Immuno-Sorbent Assay (ELISA) using an Abraxis kit for Microcystin-ADDA, which refers to an amino acid that is unique to microcystin and other similar compounds. The binding mechanism of the Microcystin-ADDA assay is specific to the microcystins, nodularins, and their congeners; therefore, results from this assay may include contributions from any compound within the ADDA functional group (Fischer et al., 2001). The minimum reporting level for the assay was 0.1 µg/L as microcystin-LR.

2.1.2. State data.—Chl *a* measurements in IA were collected as part of an ambient lake monitoring program conducted by Iowa Department of Natural Resources (IDNR). Water samples were collected above the thermocline when present, to a maximum depth of 2 m at the historic deepest point of each lake with an integrated water column sampler. Lake water samples were collected in the summer (May - September). An aliquot of the water sample was analyzed for chl *a* in the lab by non-acidified fluorometry after filtering water samples through GF/C filters. In a separate IDNR monitoring program, MC concentrations are sampled regularly at swimming beaches in Iowa during the summer. This sampling effort

includes many state park beaches and locally managed beaches across the state. MC concentration was quantified in composite water samples collected at nine different locations on three transects spanning the swimming beach. On each transect samples were collected at depths of 0.15, 0.5 and 1.0 m. MC and chl *a* samples were matched by lake and sampling date for use in the analysis. To maximize the available data, MC and chl *a* measurements collected within one day of each other were included as matched samples.

2.2. Statistical analysis

The structure of a statistical model that accommodates data collected at different spatial scales must be defined such that the available data appropriately inform model estimates. Consider the case of a large national data set of ~1000 samples and a state data set of ~50 samples. If these two datasets were pooled, the national data would dominate the state data simply because of its larger sample size, and the state data would exert a weak influence on the model. However, in any single state, only ~20 samples from the national data may be available, and we would expect the state data to dominate estimates. Defining a hierarchical structure in the model can help ensure that each data set exerts the correct influence on the model results (Gelman and Hill, 2007).

As discussed above, a second issue that arises from combining different data sets is that different measurements are often collected in different data sets. This problem was addressed in the current analysis by modeling a comprehensive network of relationships between different parameters to take advantage of the many different measurements that were available in the national data (Qian and Miltner, 2015). Then, state data sets in which only a subset of measurements were collected could still be feasibly modeled by informing specific aspects of the network. Modeling a network of relationships also provided the means of specifying relatively simple relationships between pairs of measurements, that, when combined, yielded a complex range of functional forms.

For the present study, a network of relationships was specified that reflects current understanding of the linkage between lake eutrophication (as represented by chl *a*) to increased concentrations of microcystin in individual samples (Figure 1). At the bottom of the diagram, cyanobacterial biovolume is directly associated with microcystin concentration. Cyanobacterial biovolume is then expressed as the product of total phytoplankton biovolume and the proportion of that biovolume that is cyanobacteria (i.e., the relative biovolume of cyanobacteria), which clarifies the nature of the relationship between chl *a* and cyanobacterial biovolume. More specifically, chl *a* is directly proportional to phytoplankton biovolume (Kasprzak et al., 2008), and, as chl *a* increases, the relative biovolume of cyanobacteria has been observed to increase (Downing et al., 2001).

Each of the relationships in the network described above is expressed mathematically in the Bayesian network. Initially, this network model was developed using the national data. First, phytoplankton biovolume, B_i , is modeled as being directly proportional to chlorophyll concentration (chl_i), in sample i :

$$B_i = k_{c,i} chl_i \quad (1)$$

The reciprocal of the parameter $k_{c,i}$ is the average amount of chlorophyll per unit biovolume of phytoplankton. Because the chl a content of phytoplankton can vary with environmental conditions and assemblage composition, different values of this parameter are estimated for each sample, i . The overall distribution of the set of values for $k_{c,i}$ is assumed to be log-normal with mean value of μ_k and a standard deviation of σ_k .

Exploratory analysis suggested that a quadratic function provided a reasonable representation of the relationship between the relative biovolume of cyanobacteria, p_c , and chl a :

$$\text{logit}(p_{c,i}) = f_1 + f_2 \text{chl}_i + f_3 \text{chl}_i^2 + e_{1,i} \quad (2)$$

Where cyanobacterial relative biovolume has been logit transformed, f_1 , f_2 , and f_3 are coefficients estimated from the data, and the error term $e_{1,i}$ is a normally distributed random variable with a mean of zero and a standard deviation of σ_1 .

Because laboratory replicates of B_i and $p_{c,i}$ were available, we estimated uncertainty associated with measuring phytoplankton and cyanobacterial biovolume as follows,

$$\log(Bm_j) = \log(B_{i[j]}) + e_{2,j} \quad (3)$$

$$\text{logit}(pm_{c,j}) = \text{logit}(p_{c,i[j]}) + e_{3,j} \quad (4)$$

Where Bm_j and $pm_{c,j}$ are the laboratory measurements of phytoplankton biovolume and proportion cyanobacteria, and the index j maps replicate measurements to the corresponding estimate of the true value of the measurement for sample i . The error terms e_2 and e_3 are again assumed to be normally distributed with means of zero and standard deviations of σ_2 and σ_3 , respectively.

Cyanobacterial biovolume (C) can then be expressed as the product of the relative biovolume of cyanobacteria and total phytoplankton biovolume. After log-transforming, the expression is as follows:

$$\log(C_i) = \log(k_{c,i}) + \log(p_{c,i}) + \log(\text{chl}_i) \quad (5)$$

Where cyanobacterial biovolume in sample i is the sum of a log-transformed parameter k_c , the log-transformed cyanobacterial relative biovolume in the sample, and the log-transformed chlorophyll concentration.

The final component of the model relates cyanobacteria biovolume to microcystin concentration. Initial exploration of the data suggested that microcystin concentration increases at a rapid rate relative to cyanobacterial biovolume at high levels of cyanobacteria. At low levels of cyanobacteria, though, microcystin increases at a somewhat lower rate. To account for this change in rate, we modeled microcystin with a piecewise linear model:

$$\log(\mu_{MC,i}) = g(\log(C_i)) \quad (6)$$

Where the response variable in this relationship is $\mu_{MC,i}$ the estimated mean concentration of microcystin in sample i . The function $g(\cdot)$ is the piecewise linear function, which is characterized by four parameters: the intercept, d_1 , and slope, d_2 , of the first segment, the point along the gradient at which the slope changes, c_p , and the slope of the second segment, d_3 .

The distribution of observed microcystin concentrations about the mean value was then modeled as a negative binomial distribution,

$$MC_i \sim NB(\mu_{MC,i}, \phi) \quad (7)$$

Where MC_i is the microcystin concentration observed in sample i , $NB(\cdot)$ is a negative binomial distribution with overdispersion parameter, ϕ . Because the negative binomial distribution specifies only non-negative integer outcomes, before fitting the model, we multiplied microcystin measurements by 10 and rounded to the nearest integer. Microcystin measurements below the detection limit of 0.1 $\mu\text{g/L}$ were set to zero (Yuan and Pollard, 2017).

State data from IA were included in the national model and inform estimates of relationships in the same network. As described above, though, only chl a and MC measurements were available in IA state data set, so to prevent over-specifying the model, we selected one of the relationships in the network that could be refined with data from the state. To that end, the relationship between chl a and the relative biovolume of cyanobacteria perhaps has the weakest support in theory and relies most heavily on empirical calibration, so it was selected for refinement with state data. Data at the scale of an individual state could then be introduced by first modifying the national model to allow the coefficients f_1 , f_2 , and f_3 to vary among states. That is, the relationship between chl a and the relative biovolume of cyanobacteria was re-expressed as,

$$\text{logit}(p_{c,i}) = f_{1,k[i]} + f_{2,k[i]}chl_i + f_{3,k[i]}chl_i^2 + e_{1,i} \quad (8)$$

Where different values of each of the coefficients were estimated for each U.S. state, k , based on the national data. The values of the coefficients for each state were constrained by normal distributions defined by the parameters, μ_f and σ_f . For example, the set of state-specific coefficients for f_1 were drawn from a single normal distribution:

$$f_1 \sim \text{Normal}(\mu_{f_1}, \sigma_{f_1}) \quad (9)$$

Identical expressions can be written for the set of f_2 values and f_3 values. These distributions constrained the range of possible values such that estimates of these parameters computed

with relatively small sample sizes within individual states can “borrow” information from estimates computed from other states (Gelman and Hill, 2007).

After specifying state-specific coefficients for f_1 , f_2 , and f_3 in the national model, IA state data can be naturally included in the model by noting that these data should only inform estimates of the coefficients in the state of Iowa. That is, estimates of f_1 , f_2 , and f_3 in IA are based on both the IA state data set and NLA data collected in Iowa. In other states, estimates of these coefficients are based only on NLA data. The influence of IA state data on the national distributions of these coefficients (as characterized by μ_f and σ_f) remains weak because these data only affect one element of the overall distributions of coefficients. However, within the state of IA, these coefficients can be fit to maximize the predictive accuracy of the overall relationship linking chl a to MC for both IA data and NLA data collected in IA, while remaining consistent with the range of possible values observed among all states.

One final difference in fitting the IA state data is that several sources of variability that are modeled separately in the national model (σ_1 , σ_2 , σ_3 , and σ_k) are combined into one, combined estimate of residual variability. This combination of error terms reflects the data that are available from the IA, in which no lab replicates and no direct measurements of cyanobacterial biovolume were available. Hence, measurements were only sufficient to estimate one lumped source of variability.

Two different Bayesian models were fit in this study. First, a national model was fit using only NLA data. Then, a model including IA data with the NLA data (referred to as the “state-national model”) was fit. The results of the state-national model were compared to a simple frequentist model, in which the relationship between chl a and MC concentration was estimated as a quadratic function using only IA state data and assuming that MC concentration could be modeled as a negative binomial distribution.

3. Results

3.1. National model

A total of 2352 observations of MC, cyanobacterial and phytoplankton biovolume, and chl a were available from the NLA for analysis. These measurements were collected from 1116 different lakes spanning the conterminous United States and two lakes in Hawaii. An additional 112 samples laboratory replicates of phytoplankton and cyanobacterial biovolume measurements were available to quantify measurement variability.

Three different relationships were estimated in the national model: (1) chl a and phytoplankton biovolume, (2) chl a and cyanobacterial relative biovolume, and (3) cyanobacterial biovolume and MC. The observed relationship between chl a and phytoplankton biovolume was accurately represented as a line with a slope equal to 1 on log-log axes (left panel, Figure 2), conforming with the initial hypothesis that chl a was directly proportional to phytoplankton biovolume. Based on the assumption that the density of phytoplankton is 1 g/mL, similar to a recommended value of 1.1 (Holmes et al., 1969), the estimate for the mean chlorophyll content of phytoplankton was 0.42%.

Cyanobacterial relative biovolume exhibited an increasing relationship with chl *a* (middle panel, Figure 2). The quadratic functional form allowed the model to represent the steepening of the relationship at higher concentrations of chl *a*. Mean MC concentration increased with cyanobacterial biovolume (right panel, Figure 2). The slope of the relationship increased at a cyanobacterial biovolume of 1.9 mm³/L, but the 90% credible intervals on the location of this changepoint ranged from 0.5 to 5 mm³/L. At cyanobacterial biovolumes greater than the changepoint, the slope of the mean relationship was statistically indistinguishable from 1, whereas at cyanobacterial biovolumes less than the changepoint, the slope was 0.61, with 90% credible intervals ranging from 0.51 to 0.69. Overall, the credible intervals about the cyanobacteria-MC relationship were narrow compared to those estimated for the chl *a*-cyanobacterial relative biovolume relationship.

3.2. State-national model

State-specific coefficients for the relationship between chl *a* and cyanobacterial relative biovolume varied substantially among states. Because coefficient values for quadratic relationships are not easily interpreted, we computed the predicted mean cyanobacterial relative biovolume at a chl *a* concentration of 20 µg/L to visualize the range of variation among states (Figure 3). Among all the national data, mean cyanobacterial relative biovolume was 0.18 at chl *a* = 20 µg/L. Systematic changes in cyanobacterial relative biovolume with latitude or longitude were not evident, but some regional differences were observed. For example, cyanobacterial relative biovolume at chl *a* = 20 µg/L in Northeast states was generally lower than elsewhere, whereas in Midwest states, it was somewhat higher.

A total of 556 samples of chl *a* were measured at 28 lakes in IA. In some lakes, MC concentrations were sampled at different beaches, so 686 observations of MC were matched to the chl *a* measurements. As described previously, the relationship between chl *a* and cyanobacterial relative biovolume in IA was adjusted to maximize the accuracy of the predicted MC. Inclusion of IA data reduced the magnitude of the slope of the relationship between chl *a* and cyanobacterial relative biovolume, but increased the intercept (Figure 4). So, greater values of cyanobacterial relative biovolume were observed at chl *a* concentrations less than about 10 µg/L. At higher chl *a* concentrations, inclusion of IA state data did not substantively change the predicted cyanobacterial relative biovolume. Overall, in IA, the estimated relationship between cyanobacterial relative biovolume and chl *a* was weak and was statistically indistinguishable from a constant value. The addition of the state data also narrowed the range of the credible intervals, as would be expected.

The predicted mean relationship between chl *a* and MC in Iowa from the state-national model closely followed the observed data (left panel, Figure 5), exhibiting a slight increase in slope as chl *a* concentration increased. The 90% prediction intervals shown in the plot were based on the mean values of repeated random draws of 15 samples from the predicted distribution, to replicate the plotted observed data. These intervals were broad and included most the estimated mean values. The curvature observed in simple bivariate fit was opposite that observed from the state-national model, predicting that the rate of increase in MC was lower at high chl *a* concentrations than at low chl *a* concentrations (right panel, Figure 5).

The 90% prediction intervals of this fit also included most of the observed mean values, but qualitatively, the simple bivariate model did not match the observed data as closely as the state-national model.

The value of overdispersion parameter, ϕ , was estimated as 0.41 in the state-national model, whereas in the bivariate fit to IA state data, it was estimated as 0.24. Because the inverse of ϕ quantifies the degree to which observations are more dispersed than expected for a simple Poisson distribution, the difference in the estimates of ϕ indicate that the state-national model attributes less of the variability in MC observations to the negative binomial distribution. More specifically, the state-national model represents variability in MC as originating from two components: (1) variability in the value of cyanobacterial biovolume that is associated with a particular chl *a*, and (2) variability inherent in MC concentrations, as characterized by the negative binomial distribution. In contrast, in the simple bivariate model, the dispersion of MC observations is attributed entirely to the negative binomial distribution.

Quantile-quantile plots provide a more detailed comparison of the distribution of model predicted values of MC to the observed distribution of MC. To compare predicted distributions of the state-national model and the simple bivariate model, ten bins of the available IA state data were defined, each with comparable chl *a* concentrations and containing approximately 70 observations. For each bin, the distribution of MC concentrations was predicted using the state-national model and using the bivariate model based on the median chl *a* concentration in the bin. Quantiles of predicted and observed distributions were then compared with one another. If the predicted distribution is identical to the observed distribution, we would expect that the plotted pairs of quantiles would fall on the 1:1 line. In the example shown (Figure 6), the distribution predicted by the state-national model more closely matches the observed distribution of MC, whereas the bivariate model overpredicts MC concentrations in the middle of the distribution. These differences may stem from the quadratic functional form assumed in the bivariate model or by the limitation of requiring the distribution of MC to strictly adhere to a negative binomial distribution. Root mean square (RMS) errors computed between observed and predicted distributions for each of the bins quantify the difference in predictive accuracy. For the data shown in Figure 6 the RMS error for the state-national model was 0.32, and the RMS error for the bivariate fit was 0.50, so the state-national model predicted the observed distribution of MC with 36% more accuracy than the bivariate fit. Across all 10 bins, the state-national data improved on the fit of the simple bivariate model by an average of 13%.

4. Discussion

Combining national and state data improved the accuracy and precision of predictions of MC concentrations in state data. In addition to the simple advantage of increasing the number of samples available to analyze, modeling comparatively limited state data in the context of the network of relationships specified for the national data set allows one to effectively leverage the knowledge inherent in the national-scale data. The method described here can potentially provide an approach for combining many distinct state monitoring data sets to improve estimates of ecological relationships in lakes.

4.1. National model

The first step in the analysis was to develop a broadly applicable statistical model linking eutrophication to increased microcystin. To that end, microcystin concentration was predicted using cyanobacterial biovolume, a measurement that is proximal to the response of interest. Other statistical models have been described associating different predictors (e.g., nutrient concentrations) and microcystin concentration (Graham et al., 2004; Taranu et al., 2017; Yuan and Pollard, 2017). Models that directly associate nutrient concentrations with microcystin (Yuan and Pollard, 2017) would seem to be more useful for directing management activities because they directly provide information regarding the environmental variable that can most easily be manipulated by management actions. However, such models omit steps in the causal path, and relationships estimated between increased nutrients and microcystin may vary substantially among sites due to the influence of other environmental factors (Shipley, 2016; Vanreusel et al., 2007; Vaughan and Ormerod, 2003). The effects of these covariates can be further exacerbated and cause systematic errors when data sets from different locations are combined. Conversely, cyanobacteria biovolume is a predictor variable that is proximal to the microcystin response and therefore, relationships estimated between cyanobacteria biovolume and microcystin may be less sensitive to the influence of other environmental factors. However, using models based on this relationship for management decisions are complicated by the need to collect relatively costly cyanobacterial biovolume measurements. Therefore, we developed a set of relationships that quantitatively link cyanobacterial biovolume to chl *a*: a measurement that is more commonly collected in routine monitoring. Then, existing models relating nutrient concentrations to chl *a* (Dillon and Rigler, 1974; Dolman and Wiedner, 2014; Droop, 1973) can be applied to identify management targets for phosphorus and nitrogen.

By modeling several distinct relationships, the present results can be more easily compared with findings from other studies that have specifically examined each of these processes. For example, one common assumption in other studies is that chl *a* provides a direct measurement of phytoplankton biovolume (Hessen et al., 2006). In the current model, this assumption dictates that the functional form for the relationship between chl *a* and phytoplankton biovolume is one of direct proportionality, and the distribution of observed values of chl *a* and total phytoplankton biovolume was consistent with this assumption (Figure 2). Furthermore, the estimate of the average chl *a* content of phytoplankton (0.42%) was nearly identical with the median value of chl *a* content computed from a collection of past studies (Kasprzak et al., 2008). This consistency suggests that estimates of average phytoplankton chl *a* content from different locations converge to a single value.

Chl *a* content does, however, vary with environmental conditions and among different phytoplankton species, and the present model provided a mechanism for chl *a* content to vary among samples because a distribution of values was allowed for chl *a* content, k_c . Some systematic variation in k_c may have been discernible in the data because lake characteristics, such as eutrophication status and light availability (Cloern et al., 1995; Riemann et al., 1989), can influence the chl *a* content of phytoplankton. However, at the spatial scales of the current analysis, residual variability attributed to this relationship was likely to be small

relative to the other relationships in the network. Hence, the effects of different predictors on k_c were not explored in this study.

At the national scale, cyanobacterial relative biovolume increased with lake eutrophication, as quantified by chl *a* concentration, similar to trends observed by others (Downing et al., 2001). Different hypotheses have been proposed to account for the increased dominance of cyanobacteria at the high levels of nutrient loading that accompany eutrophic conditions, but agreement on a single mechanism has been elusive. For example, some have suggested that the ability of certain cyanobacterial species to fix atmospheric nitrogen explains their dominance at high concentrations of phosphorus (O'Neil et al., 2012; Schindler et al., 2008). Others have observed that in stable water columns and high levels of nutrient loading, abundant cyanobacteria can outcompete other phytoplankton by restricting the availability of light to other species (Huisman et al., 1999). The applicability of these and other hypotheses depends in part on the identity of cyanobacteria in a lake (e.g., cyanobacteria species differ in their ability to fix nitrogen). Given that cyanobacterial assemblage composition can vary strongly among lakes, the broad confidence intervals associated with the estimated relationship (Figure 2 and Figure 4) and the variability in the estimated relationships among different states (Figure 3) seem reasonable. Indeed, increased cyanobacterial relative biovolume in eutrophic lakes is not consistently observed in all locations. Studies conducted at smaller spatial scales (among lakes with more similar cyanobacteria species) may be better able explain differences in the cyanobacterial relative biovolume among lakes. For example, in oligotrophic lakes sampled in a smaller region, iron was an important predictor of cyanobacteria dominance (Sorichetti et al., 2014).

Others have observed an association between cyanobacterial biovolume and cyanotoxin concentration in field data (Fahnenstiel et al., 2008), but the present analysis is the first to our knowledge to identify two distinct regimes in the relationship between cyanobacterial biovolume and microcystin. Identifying different regimes also revealed that, at high levels of cyanobacteria, the slope of the relationship between mean cyanobacteria biovolume and mean microcystin concentration was indistinguishable from 1 on log-log axes, indicating that mean microcystin was directly proportional to mean cyanobacterial biovolume. Because the relationship estimated here between cyanobacterial biovolume and microcystin is directly analogous to experimental studies that examine factors that influence cell quota of microcystin (Granéli and Flynn, 2006), this finding of direct proportionality confirms laboratory observations that microcystin production rate is directly comparable to cell division rate (Briand et al., 2005; Jähnichen et al., 2008; Millie et al., 2009; Orr and Jones, 1998).

A key difference between the laboratory studies and the present analysis of field data is that a fraction of the cyanobacteria species included in the measurement of cyanobacteria biovolume likely do not produce microcystin. That is, total cyanobacteria biovolume overestimates the biovolume of cyanobacteria in each sample that can produce microcystin. Hence, the present finding of a proportional relationship between cyanobacteria biovolume and microcystin depends on the assumption that the proportion of cyanobacteria that produce microcystin in each sample is relatively constant with respect to other sources of variability in the model. The need for this assumption primarily reflects the practical

constraints of measuring cyanobacteria biovolume. Total cyanobacteria biovolume is calculated as the sum of many individual measurements of biovolume with error associated with each measurement. The relative precision of the estimate of total cyanobacteria biovolume then increases with the number of individual measurements, and so, estimated cyanobacteria biovolume of only microcystin-producing taxa is substantially less precise than an estimate of total cyanobacteria biovolume. The precision of statistical models based on these two metrics reflect this difference, as preliminary analyses (not shown here) showed that total cyanobacteria biovolume accounted for more variability in microcystin concentration than biovolume based only on known microcystin-producing genera. Use of total cyanobacteria biovolume to characterize the potential for producing microcystin also reflects current understanding of the factors that influence toxin production. The proportion of cyanobacteria that can produce microcystin varies both among and within species, so shifts in cyanobacterial assemblage composition can change the overall relationship between cyanobacteria and microcystin. Indeed, different morphospecies of *Microcystis* have different potentials for producing toxin (Via-Ordorika et al., 2004), and the presence of the *mcy* gene and the associated ability to produce toxins varies within a given species (Bukowska et al., 2017; Hotto et al., 2008). Different cyanobacteria metapopulations also have been shown to have different cell quotas (Salmaso et al., 2014), and other environmental factors have also been shown to influence microcystin cell quota (Downing et al., 2005). Overall, at the scale typical monitoring programs, this present finding suggests that a simple constant of proportionality can provide useful predictions of mean microcystin concentrations from cyanobacteria biovolume.

At lower levels of cyanobacteria (biovolume < 1.9 mm³/L) a slower rate of increase in mean microcystin was observed. In terms of cell quota, this finding suggests that microcystin per unit of cyanobacteria biovolume increases as total cyanobacterial biovolume decreases. A variety of mechanisms may be responsible for this trend. For example, other environmental factors, including light availability (Rastogi et al., 2014; Wiedner et al., 2003), temperature (Davis et al., 2009), and nutrient availability or nutrient ratio (Kotak et al., 2000) have been shown to also affect toxin content of cyanobacterial cells. Furthermore, the assumption that microcystin producing cyanobacteria is a constant proportion of total cyanobacteria may be violated at these low biovolumes. Note, though, that these different factors exert only weak effects on MC concentrations when cyanobacterial biovolumes exceeded the change point.

4.2. State-national model

The state-national model predictions were substantially more accurate than a direct bivariate fit to the observed data (Figure 5 and Figure 6), and different reasons may account for this improved performance. At the simplest level, fitting a model to more data would seem always to improve model predictive performance. However, for this assumption to hold true, the data generating processes must be the same for the two data sets. If any systematic differences exist among different data sets, then a model fit to the combined data would identify a solution that averages the contributions of each data set, and predictive accuracy on any single data set would suffer. In the current analysis, the large differences in spatial scale between the national and the state data suggest that systematic differences among the

data sets are likely, and hence, this simple explanation for the improved predictions is not likely to be valid.

We argue instead that the *knowledge* represented in the network model supplements the IA data set to yield a model that more accurately predicts the observations. Three features inherent to the state-national model, in particular, are likely responsible for the improved predictions of observations in the IA data set. First, the network of relationships specified in the national model define a non-linear function linking chl *a* to MC that yielded the curved mean response we observed (Figure 5). In specifying the functional form for the bivariate fit or in any regression analysis of the relationship between chl *a* and MC, information regarding the form of this function is not available. Hence, it is difficult for a simple bivariate fit to identify the correct shape of the curve. Indeed, the curvature in the mean relationship identified by the bivariate model was opposite that estimated in the state-national model. Second, the network of relationships in the state-national model provided information regarding unobserved variables and relationships that could be used in lieu of direct observations. In this example, the relationships between chl *a* and total phytoplankton biovolume and between cyanobacterial biovolume and MC were supplied by the national model. The simple bivariate fit lacked the benefit of this additional information, and hence, for this model we had to estimate a direct relationship between chl *a* and MC that aggregated the distinct causal linkages. Finally, the hierarchical structure of the national model placed constraints on the range of possible values for parameters estimated within each state. These constraints limited model parameters for the state data set to values that were generally consistent with national parameters.

The finding that, among IA samples, cyanobacterial relative abundance was weakly associated with chl *a* concentration (Figure 4) suggests that increases in microcystin in the state are driven directly by increased eutrophication and the associated increased phytoplankton biovolume. Elsewhere, an increase in cyanobacterial relative biovolume with chl *a* would be expected to increase the amount of cyanobacteria biovolume per unit of chl *a* at higher levels of eutrophication. In IA, though, assuming a constant cyanobacterial relative biovolume likely would yield a model that performs nearly as well as the current model. Further analysis of state data would help further understand these trends (see below), but this finding further increases the uncertainty associated with the notion that cyanobacterial relative biovolume increases consistently with eutrophication.

The prediction intervals associated with the modeled relationship between chl *a* and MC from the state-national model were broad (Figure 5), and a portion of this uncertainty likely originated from the temporal and spatial separation of the chl *a* and MC samples in IA. Recall that IA MC samples were matched to chl *a* samples collected on the same day, as well one day prior and one day after the MC sample. Furthermore, chl *a* samples were collected near the middle of the lake, whereas MC samples were collected at swimming beaches. In contrast, NLA chl *a*, phytoplankton, and MC samples were co-located in time and space. MC in IA was also the average of several different measurements, compared with single grab samples in the NLA. Despite these sampling mismatches and protocol differences, the resulting mean predictions were still accurate, while the residual variability included in the model seemed to adequately represent the uncertainty in linking MC

measurements to chl *a* measurements that were not exactly co-located. In previous analysis of the NLA data, the variation in microcystin concentration attributed to differences between littoral and open-water samples was small relative to temporal variation and differences among lakes (Yuan and Pollard, 2017), and so, the lack of an effect of spatially mismatched samples in IA was not surprising.

We defined a hierarchical structure in this model to facilitate the incorporation of state data, but other classifications would likely be more ecologically meaningful (Wagner et al., 2011; Yuan and Pollard, 2014). Classification by state only provided a crude proxy for spatial location and did not account for other covariates that can affect the modeled relationships (e.g., lake depth). However, we focused here on combining state data sets with the national model, and because differences among standard sampling protocols in state monitoring programs likely also exerted a strong influence on observed data, structuring the model in terms of states was a logical first step. With more data at national and state scales, classification based on environmental covariates as well as state political boundaries may be possible and could yield improved models.

The current example of state data provided only chl *a* and MC data, but in other states, different measurements may be available. The basic process described here for combining state data with the national modeled network of relationships is generally applicable, though. Indeed, if more parameters are measured in a state, it may be possible to explore whether incorporating among-state variations in other relationships (e.g., the cyanobacteria-MC relationship) can improve the model, or whether our current assumption that certain relationships are invariant among states holds true. The ultimate goal for this model is to simultaneously represent data contributed from many different states within the same model network, such that both national data and data from many states contribute to updating the hyperparameter distributions of each of the key relationships. Such a system would realize the often-described vision of continuously updating a Bayesian model as new data and information become available (Clark, 2005).

From a broader perspective, monitoring data are routinely collected by many countries, states, and tribes, and robust methods for combining these data to extract inferences are needed to fully realize the potential for these data to inform environmental decision making. Environmental data from a variety of sources are frequently collected and pooled to estimate relationships of interest (Downing et al., 2001; Filstrup et al., 2014). This simple pooling approach can yield results that are informative, but assumptions regarding the comparability of different aspects of the data sets are unstated and unacknowledged, an oversight that can ultimately affect the accuracy of the inferences. The hierarchical Bayesian approach described here compels analysts to explicitly state assumptions regarding relationships between data sets, which can then be evaluated in conjunction with examining the model results. Furthermore, these assumptions provide hypotheses for future testing that can ultimately improve our understanding of the processes that underlie the characteristics that we observe in different systems.

Acknowledgements

We thank the NLA and Iowa Department of Natural Resources for their field data collection efforts. The views expressed in this paper are those of the authors and do not reflect the official policies of the U.S. Environmental Protection Agency.

References

- Briand J-F, Jacquet S, Flinois C, Avois-Jacquet C, Maisonnnette C, Leberre B, Humbert J-F, 2005 Variations in the Microcystin Production of *Planktothrix rubescens* (Cyanobacteria) Assessed from a Four-Year Survey of Lac du Bourget (France) and from Laboratory Experiments. *Microb. Ecol* 50, 418–428. 10.1007/s00248-005-0186-z [PubMed: 16328657]
- Bukowska A, Kali ski T, Koper M, Kostrzewska-Szlakowska I, Kwiatowski J, Mazur-Marzec H, Jasser I, 2017 Predicting blooms of toxic cyanobacteria in eutrophic lakes with diverse cyanobacterial communities. *Sci. Rep* 7, 8342 10.1038/s41598-017-08701-8 [PubMed: 28827675]
- Clark JS, 2005 Why environmental scientists are becoming Bayesians. *Ecol. Lett* 8, 2–14. 10.1111/j.1461-0248.2004.00702.x
- Cloern JE, Grenz C, Vidregar-Lucas L, 1995 An Empirical Model of the Phytoplankton Chlorophyll: Carbon Ratio-the Conversion factor Between Productivity and Growth Rate. *Limnol. Oceanogr* 40, 1313–1321. 10.2307/2838689
- Davis TW, Berry DL, Boyer GL, Gobler CJ, 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, 715–725. 10.1016/j.hal.2009.02.004
- Dillon PJ, Rigler FH, 1974 The phosphorus-chlorophyll relationship in lakes. *Limnol. Oceanogr* 19, 767–773.
- Dolman A, Wiedner, 2014 Predicting phytoplankton biomass and estimating critical N:P ratios with piecewise models that conform to Liebig's law of the minimum. *Freshw. Biol* 60, 686–697. 10.1111/fwb.12518
- Downing JA, Watson SB, McCauley E, 2001 Predicting cyanobacteria dominance in lakes. *Can. J. Fish. Aquat. Sci* 58, 1905–1908. 10.1139/cjfas-58-10-1905
- Downing TG, Meyer C, Gehringer MM, van de Venter M, 2005 Microcystin content of *Microcystis aeruginosa* is modulated by nitrogen uptake rate relative to specific growth rate or carbon fixation rate. *Environ. Toxicol* 20, 257–262. 10.1002/tox.20106 [PubMed: 15892070]
- Droop MR, 1973 Some Thoughts on Nutrient Limitation in Algae. *J. Phycol* 9, 264–272. 10.1111/j.1529-8817.1973.tb04092.x
- Fahnenstiel GL, Millie DF, Dyble J, Litaker RW, Tester PA, McCormick MJ, Rediske R, Klarer D, 2008 Microcystin concentrations and cell quotas in Saginaw Bay, Lake Huron. *Aquat. Ecosyst. Health Manag* 11, 190–195. 10.1080/14634980802092757
- Filstrup CT, Wagner T, Soranno PA, Stanley EH, Stow CA, Webster KE, Downing JA, 2014 Regional variability among nonlinear chlorophyll—phosphorus relationships in lakes. *Limnol. Oceanogr* 59, 1691–1703. 10.4319/lo.2014.59.5.1691
- Fischer WJ, Garthwaite I, Miles CO, Ross KM, Aggen JB, Chamberlin AR, Towers NR, Dietrich DR, 2001 Congener-independent immunoassay for microcystins and nodularins. *Environ. Sci. Technol* 35, 4849–4856. 10.1021/es011182f [PubMed: 11775161]
- Gelman A, Hill J, 2007 Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York, NY.
- Graham JL, Jones JR, Jones SB, Downing JA, Clevenger TE, 2004 Environmental factors influencing microcystin distribution and concentration in the Midwestern United States. *Water Res.* 38, 4395–4404. 10.1016/j.watres.2004.08.004 [PubMed: 15556214]
- Granéli E, Flynn K, 2006 Chemical and physical factors influencing toxin content, in: *Ecology of Harmful Algae*. Springer, pp. 229–241.
- Hessen DO, Faafeng BA, Brettum P, Andersen T, 2006 Nutrient Enrichment and Planktonic Biomass Ratios in Lakes. *Ecosystems* 9, 516–527. 10.1007/s10021-005-0114-6

- Holmes R, Norris R, Smayda T, Wood EJJ, 1969 Collection, fixation, identification, and enumeration of phytoplankton standing stock. *Recomm. Proced. Meas. Product. Plankton Standing Stock Relat. Ocean. Prop. Wash. DC Natl. Acad. Sci* 17–46.
- Hotto AM, Satchwell MF, Berry DL, Gobler CJ, Boyer GL, 2008 Spatial and temporal diversity of microcystins and microcystin-producing genotypes in Oneida Lake, NY. *Harmful Algae* 7, 671–681. 10.1016/j.hal.2008.02.001
- Huisman J, Jonker RR, Zonneveld C, Weissing FJ, 1999 Competition for light between phytoplankton species: experimental tests of mechanistic theory. *Ecology* 80, 211–222. 10.1890/0012-9658(1999)080[0211:CFLBPS]2.0.CO;2
- Jähnichen S, Ihle T, Petzoldt T, 2008 Variability of microcystin cell quota: A small model explains dynamics and equilibria. *Limnol. - Ecol. Manag. Inland Waters, Integrating approaches in limnological research: from basic to applied and from empirical to theoretical* 38, 339–349. 10.1016/j.limno.2008.05.003
- Kasprzak P, Padišák J, Koschel R, Krienitz L, Gervais F, 2008 Chlorophyll a concentration across a trophic gradient of lakes: An estimator of phytoplankton biomass? *Limnol. - Ecol. Manag. Inland Waters, Integrating approaches in limnological research: from basic to applied and from empirical to theoretical* 38, 327–338. 10.1016/j.limno.2008.07.002
- Kotak BG, Lam AK-Y, Prepas EE, Hrudey SE, 2000 Role of chemical and physical variables in regulating microcystin-LR concentration in phytoplankton of eutrophic lakes. *Can. J. Fish. Aquat. Sci* 57, 1584–1593. 10.1139/f00-091
- Loftin KA, Meyer M, Rubio F, Kamp L, Humphries E, Whereat E, 2008 Comparison of two cell lysis procedures for recovery of microcystins in water samples from Silver Lake in Dover, Delaware, with microcystin producing cyanobacterial accumulations (Open-File Report No. 2008–1341). United States Geological Survey, Reston, VA.
- Millie DF, Fahnenstiel GL, Bressie JD, Pigg RJ, Rediske RR, Klarer DM, Tester PA, Litaker RW, 2009 Late-summer phytoplankton in western Lake Erie (Laurentian Great Lakes): bloom distributions, toxicity, and environmental influences. *Aquat. Ecol* 43, 915–934. 10.1007/s10452-009-9238-7
- O’Neil JM, Davis TW, Burford MA, Gobler CJ, 2012 The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. *Harmful Algae, Harmful Algae--The requirement for species-specific information* 14, 313–334. 10.1016/j.hal.2011.10.027
- Orr PT, Jones GJ, 1998 Relationship between microcystin production and cell division rates in nitrogen-limited *Microcystis aeruginosa* cultures. *Limnol. Oceanogr* 43, 1604–1614.
- Paerl HW, Huisman J, 2009 Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. *Environ. Microbiol. Rep* 1, 27–37. 10.1111/j.1758-2229.2008.00004.x [PubMed: 23765717]
- Paerl HW, Huisman J, 2008 Blooms like it hot. *Science* 320, 57–58. 10.1126/science.1155398 [PubMed: 18388279]
- Phillips G, Pietiläinen O-P, Carvalho L, Solimini A, Lyche Solheim A, Cardoso A, 2008 Chlorophyll–nutrient relationships of different lake types using a large European dataset. *Aquat. Ecol* 42, 213–226. 10.1007/s10452-008-9180-0
- Pollard AI, Hampton SE, Leech DM, 2018 The Promise and Potential of Continental-Scale Limnology Using the U.S. Environmental Protection Agency’s National Lakes Assessment. *Limnol. Oceanogr. Bull* 27, 36–41. 10.1002/lob.10238
- Qian SS, Miltner RJ, 2015 A continuous variable Bayesian networks model for water quality modeling: A case study of setting nitrogen criterion for small rivers and streams in Ohio, USA. *Environ. Model. Softw* 69, 14–22. 10.1016/j.envsoft.2015.03.001
- Rastogi RP, Sinha RP, Incharoensakdi A, 2014 The cyanotoxin-microcystins: current overview. *Rev. Environ. Sci. Biotechnol* 13, 215–249. 10.1007/s11157-014-9334-6
- Riemann B, Simonsen P, Stensgaard L, 1989 The carbon and chlorophyll content of phytoplankton from various nutrient regimes. *J. Plankton Res* 11, 1037–1045. 10.1093/plankt/11.5.1037
- Salmaso N, Copetti D, Cerasino L, Shams S, Capelli C, Boscaini A, Valsecchi L, Pozzoni F, Guzzella L, 2014 Variability of microcystin cell quota in metapopulations of *Planktothrix rubescens*: Causes and implications for water management. *Toxicon* 90, 82–96. 10.1016/j.toxicon.2014.07.022 [PubMed: 25108147]

- Schindler DW, Hecky RE, Findlay DL, Stainton MP, Parker BR, Paterson MJ, Beaty KG, Lyng M, Kasian SEM, 2008 Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. *Proc. Natl. Acad. Sci* 105, 11254–11258. 10.1073/pnas.0805108105 [PubMed: 18667696]
- Shipley B, 2016 Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R. Cambridge University Press.
- Soranno PA, Schimel DS, 2014 Macrosystems ecology: big data, big ecology. *Front. Ecol. Environ* 12, 3–3. 10.1890/1540-9295-12.1.3
- Soranno PA, Spence Cheruvelil K, Webster KE, Bremigan MT, Wagner T, Stow CA, 2010 Using landscape limnology to classify freshwater ecosystems for multi-ecosystem management and conservation. *BioScience* 60, 440–454. 10.1525/bio.2010.60.6.8
- Sorichetti RJ, Creed IF, Trick CG, 2014 Evidence for iron-regulated cyanobacterial predominance in oligotrophic lakes. *Freshw. Biol* 59, 679–691. 10.1111/fwb.12295
- Sprague LA, Oelsner GP, Argue DM, 2017 Challenges with secondary use of multi-source water-quality data in the United States. *Water Res.* 110, 252–261. 10.1016/j.watres.2016.12.024 [PubMed: 28027524]
- Taranu ZE, Gregory-Eaves I, Leavitt PR, Bunting L, Buchaca T, Catalan J, Domaizon I, Guilizzoni P, Lami A, McGowan S, Moorhouse H, Morabito G, Pick FR, Stevenson MA, Thompson PL, Vinebrooke RD, 2015 Acceleration of cyanobacterial dominance in north temperate-subarctic lakes during the Anthropocene. *Ecol. Lett* 375–384. 10.1111/ele.12420 [PubMed: 25728551]
- Taranu ZE, Gregory-Eaves I, Steele RJ, Beaulieu M, Legendre P, 2017 Predicting microcystin concentrations in lakes and reservoirs at a continental scale: A new framework for modelling an important health risk factor. *Glob. Ecol. Biogeogr* 26, 625–637.
- US EPA, 2012a 2012 National Lakes Assessment Site Evaluation Guidelines (No. EPA 841-B-11–003). Office of Water, Washington, DC.
- US EPA, 2012b 2012 National Lakes Assessment. Laboratory Operations Manual (No. EPA-841-B-11–004). U.S. Environmental Protection Agency, Washington, DC.
- US EPA, 2011 2012 National Lakes Assessment. Field Operations Manual (No. EPA 841-B-11–003). Office of Water, US Environmental Protection Agency, Washington, DC.
- US EPA, 2010 National Lakes Assessment: A Collaborative Survey of the Nation's Lakes (No. EPA 841-R-09–001). Office of Water and Office of Research and Development, Washington, DC.
- Vanreusel W, Maes D, Dyck HV, 2007 Transferability of Species Distribution Models: a Functional Habitat Approach for Two Regionally Threatened Butterflies. *Conserv. Biol* 21, 201–212. 10.1111/j.1523-1739.2006.00577.x [PubMed: 17298526]
- Vaughan IP, Ormerod SJ, 2003 Improving the Quality of Distribution Models for Conservation by Addressing Shortcomings in the Field Collection of Training Data. *Conserv. Biol* 17, 1601–1611. 10.1111/j.1523-1739.2003.00359.x
- Via-Ordorika L, Fastner J, Kurmayer R, Hisbergues M, Dittmann E, Komarek J, Erhard M, Chorus I, 2004 Distribution of Microcystin-Producing and Non-Microcystin-Producing *Microcystis* sp. in European Freshwater Bodies: Detection of Microcystins and Microcystin Genes in Individual Colonies. *Syst. Appl. Microbiol* 27, 592–602. 10.1078/0723202041748163 [PubMed: 15490561]
- Wagner T, Soranno PA, Webster KE, Cheruvelil KS, 2011 Landscape drivers of regional variation in the relationship between total phosphorus and chlorophyll in lakes. *Freshw. Biol* 56, 1811–1824. 10.1111/j.1365-2427.2011.02621.x
- Wiedner C, Visser PM, Fastner J, Metcalf JS, Codd GA, Mur LR, 2003 Effects of Light on the Microcystin Content of *Microcystis* Strain PCC 7806. *Appl. Environ. Microbiol* 69, 1475–1481. 10.1128/AEM.69.3.1475-1481.2003 [PubMed: 12620831]
- Wood SN, 2006 Generalized additive models: An introduction with R. CRC Press, Boca Raton, FL.
- Yuan LL, Pollard AI, 2017 Using national-scale data to develop nutrient–microcystin relationships that guide management decisions. *Environ. Sci. Technol* 51, 6972–6980. 10.1021/acs.est.7b01410 [PubMed: 28561562]
- Yuan LL, Pollard AI, 2014 Classifying lakes to improve precision of nutrient–chlorophyll relationships. *Freshw. Sci* 33, 1184–1194. 10.1086/678465

Yuan LL, Pollard AI, Pather S, Oliver JL, D'Anglada L, 2014 Managing microcystin: identifying national-scale thresholds for total nitrogen and chlorophyll a. *Freshw. Biol* 59, 1970–1981. 10.1111/fwb.12400

EPA Author Manuscript

EPA Author Manuscript

EPA Author Manuscript

Figure 1. Schematic of model structure. Numbers in parentheses refer to equation for the labeled relationship.

Figure 2.

National model relationships. Left panel: relationship between chl *a* and phytoplankton biovolume, open circles: observed measurements of chl *a* and phytoplankton biovolume, solid line has a slope of 1. Middle panel: Relationship between chl *a* and cyanobacterial relative biovolume, open circles: average cyanobacterial relative biovolume in ~20 samples at the indicated chl *a* concentration (data binned to highlight mean relationships in the plot), solid line: estimated mean relationship, gray shading: 90% credible intervals about the mean relationship, vertical axis has been logit-transformed. Right panel: Relationship between cyanobacterial biovolume and microcystin (MC), open circles: average MC concentration in ~15 samples at the indicated cyanobacterial biovolume (data binned to highlight mean relationships in the plot), solid line: estimated mean relationship, gray shading: 90% credible intervals about the mean relationship, small filled circles: chl *a* bins in which MC in all samples was zero.

Figure 3. Variation in the relationship between chl *a* and cyanobacterial relative biovolume among states. PropCyano: predicted mean relative biovolume of cyanobacteria at chl *a* = 20 µg/L.

Figure 4.

Comparison of chl *a*/cyanobacterial relative biovolume relationships in Iowa. Filled gray: 90% credible intervals for estimate of relationship using only National Lake Assessment (NLA) data collected in Iowa. Solid and dashed lines: mean and 90% credible intervals for estimate of relationship using both Iowa (IA) state and NLA data.

Figure 5.

Comparison of predicted relationship between chl *a* and MC for the state-national model (left panel) and a model using only IA state data (right panel). In both panels: open circles: average MC concentration computed in 15 samples at the indicated chl *a*, solid lines: mean relationship, dashed lines: 5th to 95th percentiles of distribution of means of 15 samples drawn from predicted distribution. Mean values in bins of 15 samples shown in the plot to highlight mean relationship.

Figure 6. Quantile-quantile (Q-Q) plots of predicted and observed distribution of MC for chl *a* between 33 and 45 $\mu\text{g/L}$. Left panel: Q-Q plot for state-national model, right panel: Q-Q plot for simple bivariate model. Solid line: 1:1 relationship. 0.1 added to all values to allow 0 values to be log-transformed.