Predicting microcystin occurrence in freshwater lakes and reservoirs: assessing environmental variables

Riley P. Buley, Hannah E. Correia, Ash Abebe, Tahir B. Issa & Alan E. Wilson

To cite this article: Riley P. Buley, Hannah E. Correia, Ash Abebe, Tahir B. Issa & Alan E. Wilson (2021): Predicting microcystin occurrence in freshwater lakes and reservoirs: assessing environmental variables, Inland Waters, DOI: 10.1080/20442041.2021.1938491

To link to this article: https://doi.org/10.1080/20442041.2021.1938491
Predicting microcystin occurrence in freshwater lakes and reservoirs: assessing environmental variables

Riley P. Buley, Hannah E. Correia, Ash Abebe, Tahir B. Issa, and Alan E. Wilson

ABSTRACT
Determining the environmental conditions that influence the occurrence and concentration of the cyanobacterial toxin microcystin (MC) is a critical step for predicting cases in which the toxin will adversely affect drinking water sources, recreational waterbodies, and other freshwater ecosystems. Although widely studied, little consensus exists regarding the factors that influence MC on a global scale. The objective of this study was to identify the environmental variables most strongly associated with MC concentrations using observational data from lakes and reservoirs around the world while also addressing the substantial proportions of missing values that a large aggregated dataset often involves. A total of 124 studies containing data from an estimated 2040 lakes and reservoirs in 22 countries was used to construct a global dataset. Variables including <35% of non-missing observations were removed prior to analysis. Missing values for the remaining 12 predictors of MC were imputed using an iterative imputation algorithm based on a random forest approach. Variable selection was performed with generalized additive modeling on the complete case and imputed datasets. Models applied to the imputed data produced lower prediction errors than those fit to the complete dataset. Variables of greatest significance to MC concentration included location (longitude–latitude pairs), total nitrogen, turbidity, and pH. Total phosphorus was not found to be a strong predictor of MC. In addition to assisting water resource managers in protecting their waterbodies against MC, the presented methodologies may provide a useful framework for future water quality modeling while accounting for varying proportions of missing data.

ARTICLE HISTORY
Received 19 August 2020
Accepted 25 May 2021

KEYWORDS
cyanobacteria; cyanotoxin; harmful algal bloom; modeling; water quality

Introduction
Cyanobacterial blooms are a growing threat to aquatic ecosystems worldwide due to increases in eutrophication and climate change (Paerl and Otten 2013). Blooms can impose numerous adverse effects on these systems (e.g., hypoxia, decreased light penetration; Paerl et al. 2001, Paerl and Otten 2013), including the production of secondary metabolites potentially toxic to animals and humans (i.e., cyanotoxins; Paerl et al. 2001, Graham et al. 2004, Malbruck and Kestemont 2006). Numerous cyanotoxins have been identified, including but not limited to dermatoxins (e.g., aplysiatoxin), neurotoxins (e.g., anatoxin, saxitoxin), and hepatotoxins (e.g., microcystin, nodularin; Sivonen 2009). Of these cyanobacterial toxins, microcystin (MC) is routinely observed in freshwater systems experiencing cyanobacterial blooms. MC is produced through non-ribosomal peptide-polyketide synthesis in cyanobacterial strains possessing the mcy gene cluster (Graham et al. 2004, Hotto et al. 2008, Joung et al. 2011, Li et al. 2012) and acts as an inhibitor to type 1 or 2A protein phosphatase (Sivonen 2009, Mankiewicz-Boczek et al. 2015). Although classified as a hepatotoxin, MC is known to affect the kidneys, intestines, and muscle tissues of fish and other aquatic organisms (Malbruck and Kestemont 2006, Martins and Vasconcelos 2009). More than 246 variants of MC are known, and toxicity between variants differs substantially (Hu et al. 2016, Li et al. 2017, Meriluoto et al. 2017). Because of its toxicity, the World Health Organization has set a recommended guideline of 1 µg/L of MC in drinking water (WHO 2003, Li et al. 2017). Although significant progress has been made in the management of MC levels in freshwater systems, MC remains a consequential topic in research, especially because the frequency and severity of toxic cyanobacterial blooms are expected to increase in the following decades (Paerl and Otten 2013).

Extensive field surveys relating common water quality measurements to MC occurrence have been
performed during the past 2 decades (Graham et al. 2004, Giani et al. 2005, US Environmental Protection Agency 2010, 2016, Mowe et al. 2014, Francy et al. 2016). Despite this, wide disparities in the environmental conditions that most influence MC production exist in the literature. For example, the documented optimal temperature range for toxigenic cyanobacteria varies widely (15–20 °C: Billam et al. 2006; ~23 °C: Li et al. 2017; 18–35 °C: Gagala et al. 2012; and >25 °C: Boutte et al. 2008). The difficulty in determining the precise relationships of environmental parameters to the production of MC stems from the complex dynamics of harmful cyanobacterial blooms. For instance, multiple cyanobacterial species produce MC (e.g., Microcystis, Anabaena/Dolichospermum, Nostoc, Planktothrix, and Nodularia; Hotto et al. 2008), and these species have various environmental preferences and competitive traits, such as gas vesicles, nitrogen-fixation capabilities, thermal tolerances, seasonal preferences, and wide nitrogen to phosphorus ratio tolerances that allow MC production in a wide range of environmental situations (Paerl et al. 2001, Fastner et al. 2016, Shan et al. 2020). The amount of MC produced per cell can also be influenced by a number of factors, including nutrient concentration (Horst et al. 2014), temperature (Mowe et al. 2014), and presence of toxigenic strains (i.e., a strain that possesses the genes for toxin production; Graham et al. 2004, Joung et al. 2011). Further, a bloom dominated by a toxigenic cyanobacterial species capable of producing MC will not always produce toxins, a factor that can often generate water quality data-sets with periods of low MC concentrations contrasted by periods of high MC concentrations. These numerous issues contribute to the difficulty of creating meaningful models that relate environmental factors to MC occurrence.

Although difficult, the need to predict environmental variables most likely to influence the occurrence of MC is crucial for both the management of drinking water reservoirs and other freshwater systems. Previous studies have used multiparameter predictive and forecasting models to monitor MC occurrence (Giani et al. 2005, Otten et al. 2012, Francy et al. 2016, Yuan and Pollard 2017, 2019, Shan et al. 2019, 2020). Such research often uses survey data from a single body of water or region, thereby reducing the generality of models to be used in other areas. Effective MC models have been developed utilizing both national and local waterbody datasets (Yuan and Pollard 2019), but such research is uncommon. Contributing to the complexity of developing prediction models at larger spatial scales is securing large and spatially expansive datasets in the field of water quality because heterogeneous studies only collect subsets of potential variables of interest and produce an aggregated dataset with varying amounts of missing values. Instances of missing data are a persistent issue observed in all fields of science, a problem readily addressed in the fields of ecology, physiology, health, and social sciences (Bennett 2001, Wisz et al. 2008). Suggestions as to the acceptable limits of missing data within a dataset is a topic of debate, with the allowable limits given on a case-by-case basis (e.g., 10% allowable missingness: Bennett 2001; 30%: Taugourdeau et al. 2014; and 60%: Penone et al. 2014). Moreover, statements indicating the importance of the heterogeneity of missing data rather than the total amount missingness have been observed (Tabachnick et al. 2019). Although removing variables with large amounts of missingness may be perceived as a logical method to circumvent this issue, doing so may reduce power and introduce new bias into the developed model (Penone et al. 2014, Taugourdeau et al. 2014). Limited global models using aggregated data from various publications have been employed for inference in ecology and water quality because of these issues.

We present novel statistical analyses to determine the water quality variables that best predict MC concentra-tion in freshwater lakes and reservoirs for a global range of waterbodies while addressing large amounts of missing values within an aggregated dataset. We first address the issue of missing data using random forest (RF) imputation, a machine learning technique used to impute missing data without a regression model being specified (Tang and Ishwaran 2017). RF imputation efficiently inputs unknown values within a data matrix without the use or influence of the response variable and has been used successfully in human health and biological studies (Stekhoven and Buhlmann 2012, Penone et al. 2014). Furthermore, RF-based imputation has been shown to outperform other imputation methods for missing data (Shah et al. 2014, Kokla et al. 2019).

The effect of imputation on model fit and prediction was assessed by fitting generalized additive models (GAMs) with variable selection to the complete case and imputed datasets. GAMs account for nonnormal and spatially autocorrelated data, and they accommodate nonlinear predictors and response variables by means of nonparametric smooth functions fit using regression splines (Lehmann 1998, Colón-González et al. 2013, Brabec et al. 2014, Wood 2017). GAMs do not have predefined functions to which the model has to conform, allowing the data to determine the best-fit functions of a model (Suárez-Seoane et al. 2002). After the GAMs were fit, we then compared the predictive performance of each of the selected models on both the complete case and imputed data using 10-fold cross-validation.
Lastly, the GAM with the best prediction performance, a key benchmark to derive effective inference for management, was used to develop predictions about MC concentration from highly nonlinear data.

GAMs have been used in prior water quality and (Carvalho et al. 2013) and ecological studies (Lehmann 1998, Suárez-Seoane et al. 2002), and RF techniques have been used for modeling of cyanobacterial secondary metabolites (Kehoe et al. 2015, Harris and Graham 2017), but RF imputation has not been used in tandem with GAMs to predict MC concentration on such a spatial scale (to the authors’ knowledge). The use of RF imputation with GAM for inference of nonlinear relationships may provide researchers and resource managers with meaningful insights to the production of MC in freshwater systems, even with data typically constrained by considerable missingness. The objective of this study was to identify the environmental variables most strongly associated with MC concentration using a dataset containing observations from a wide spatial scale. Building on the insights observed in prior MC research (e.g., Graham et al. 2004), we hypothesized that many environmental parameters would have nonlinear relationships with MC.

**Methods**

**Data accumulation**

Articles were retrieved in February 2018 using the Web of Science database by combining “microcystin” with the keywords lake, reservoir, environment, parameters, nutrients, variables, environmental parameters, and environmental variables. Searches returned 3332 articles. Studies were included in this analysis if they fit the following criteria: (1) were observational field studies (i.e., not experimental in nature); (2) were from a freshwater reservoir or lake, defined as a system with little-to-no-flow (i.e., not including impeded rivers or streams); and (3) provided numerical sampling data in figures, tables, or supplementary files. We determined that 42 articles met these criteria (see references in Supplemental Materials). Data were taken directly from text or supplementary material in each article when possible, but data were also obtained using the *metaDitigise* package in R, which allowed the extraction of data from article figures (Pick et al. 2019). Additionally, the National Water Information System of the US Geological Survey was used to obtain real-time data containing MC values and respective environmental parameters from 80 sites around the United States (see Supplemental Material 1). Less than (<) symbols observed throughout the USGS dataset were assigned half values. USGS dataset values that were affected by contamination (V symbol) were removed. Data were also taken from the 2007 and 2012 US National Lake Assessments (NLA; US Environmental Protection Agency 2010, 2016).

In total, data from 124 studies or sites included in our analyses contained physicochemical factors (e.g., temperature; pH; Secchi disk depth, a measure of water clarity; and conductivity), nutrients (e.g., nitrogen, nitrate, nitrite, phosphorus, and phosphate), phytoplankton biomass (measured as the concentration of chlorophyll, both total and α), and/or concentrations of MC. Many USGS sites also contained a wide array of measured variables, such as heavy metals, organic or inorganic chemicals, physicochemical measurements, and nutrients. Variables not also reported in the obtained published studies were largely removed, resulting in 41 predictors. Lastly, the reporting of MC and its variants differed between studies and datasets. The 5 most occurring MC measurements observed, including total MC, total MC-LR, total MC + nodularins (largely Environmental Protection Agency [EPA] reported data), intracellular MC, and intracellular MC-LR, were kept for analysis and the others discarded. To maintain observation numbers and the global scale of the data, all MC concentration responses were treated as one variable. Disparate scales were not a concern because MC levels from all studies were measured in µg/L. MC values >500 µg/L were considered outliers and removed (n = 14, 0.3% of total data).

An estimated 2040 lakes in 22 countries were represented in the global dataset (Fig. 1). Because of the range of survey studies with differing sampling methodologies incorporated into this global dataset, varying rates of missingness were observed among the 41 predictor variables (Table 1). Missingness ranged from 0% (latitude and longitude) to 99.3% (soluble reactive phosphorus). Variables not measured in at least 35% of the observations were removed prior to statistical analysis because these variables could be considered as not regularly collected in relation to MC. This process left 12 remaining environmental predictor variables, including Secchi depth, pH, ammonium as nitrogen, nitrite as nitrogen, nitrate as nitrogen, nitrate + nitrite as nitrogen, total phosphorus, dissolved organic carbon, chlorophyll, total nitrogen, turbidity, and specific conductivity (Table 2). Observations without both latitude and longitude coordinates were also removed, resulting in 4316 observations.

**Statistical analysis overview**

Statistical analysis of the aggregated global dataset occurred in 4 steps: (1) impute missing data using RF
machine learning, (2) fit GAMs with variable selection on complete case and imputed data, (3) compare models fit in step 2 to both complete case and RF-imputed data using 10-fold cross-validation, and (4) fit GAM with lowest prediction errors from step 3 for inference (each step is described in detail). R 4.0.2 was used to perform all statistical analyses (R Core Team 2020).

**Imputation of missing data using random forest machine learning**

To address missingness, RF-based imputation was applied to impute missing values for the 12 remaining predictor variables. Imputation in this study was performed by RF machine learning using the statistical package `missForest` in R, which allows the imputation of both continuous and categorical values with limited assumptions of the data by utilizing RF machine learning (Stekhoven and Buhlmann 2012). Briefly, the RF algorithm in `missForest` first imputes all missing values with the mean for each variable of interest in a data matrix $X = (X_1, X_2, \ldots, X_p)$ with $p$ variables, then sorts the variables $X_m, m = 1, \ldots, p$ from least to highest proportions of missingness. Starting with the variable with the least amount of missingness $X_m$, an RF is fit on the observed data values in the data matrix and is used to predict the missing values in $X_m$. These RF predictions are then used as the new imputed values for the missing values of variable $X_m$. The algorithm proceeds through the remaining predictor variables from least to highest missingness, and the algorithm repeats until a user-specified maximum number of iterations is reached or a stopping criteria is met, typically when the difference between the new imputations and previously imputed values of the data matrix $X$ increases (Stekhoven and Buhlmann 2012).

The response value MC was removed from the data set before RF-based imputation to avoid biased estimates. Latitude and longitude values were also removed because RF-based imputation methods are not suitable for imputing geolocational data. The global data matrix was then imputed, after which MC, latitude,
and longitude were returned to the dataset in their respective rows. Out-of-bag mean squared error (OOBMSE) estimates of imputation error were assessed for each of the imputed predictor variables.

### Table 2. Out-of-bag mean squared error (OOBMSE) for random forest imputation for 12 predictor variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>OOBMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitrate as nitrogen</td>
<td>0.00</td>
</tr>
<tr>
<td>Ammonium as nitrogen</td>
<td>0.01</td>
</tr>
<tr>
<td>Total phosphorus</td>
<td>0.07</td>
</tr>
<tr>
<td>pH</td>
<td>0.22</td>
</tr>
<tr>
<td>Nitrate as nitrogen</td>
<td>0.70</td>
</tr>
<tr>
<td>Nitrate and nitrite as nitrogen</td>
<td>0.72</td>
</tr>
<tr>
<td>Secchi disk depth</td>
<td>1.31</td>
</tr>
<tr>
<td>Total nitrogen</td>
<td>1.46</td>
</tr>
<tr>
<td>Organic carbon filtered</td>
<td>142.48</td>
</tr>
<tr>
<td>Turbidity (NTU)</td>
<td>686.39</td>
</tr>
<tr>
<td>Chlorophyll</td>
<td>1897.74</td>
</tr>
<tr>
<td>Specific conductivity</td>
<td>2959.980</td>
</tr>
</tbody>
</table>

### Fitting of GAMs with variable selection on complete case and imputed data

Because relationships between MC concentrations and environmental predictors were expected to be nonlinear, 2 GAMs of the form:

\[
g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_{12}(x_{12}) + s(u, v),
\]

were fit for the complete case data and RF-imputed data, where \(\beta_0\) is the parametric intercept; \(f_1, f_2, \ldots, f_{12}\) are smoothing functions describing the nonlinear relationships between the 12 predictors \(x_1, x_2, \ldots, x_{12}\) and the response \(Y\); and \(g(\cdot)\) represents the link function between the response and predictors (Wood 2017). The GAMs include a spatial interaction term \(s(u, v)\), where \(u, v\) are longitude-latitude pairs and \(s(\cdot)\) is a 2-dimensional smoothing function, allowing the response to vary over space and enabling the measurement of effects of other variables to be independent of location. Selection of terms significant to variation in MC levels was performed for each of the 2 GAMs using the mgcv package in R (Wood 2017). A Tweedie distribution was developed where mgcv estimated the distribution parameter, and a restricted maximum likelihood estimation was used for smoothing parameter estimation and variable selection. To assess fit of the models to the data, the reduced model selected for the complete case data (GAM\(_{CC}\)) and the reduced model selected using the RF-imputed data were each fit on both the complete case and RF-imputed data, and the percentage of deviance explained was recorded for each of the 4 fitted models. Deviance explained is approximately equivalent to unadjusted \(R^2\) as a measure of fit for GAMs with non-Gaussian families (Wood 2017).

### Comparison of selected GAMs using 10-fold cross-validation

A 10-fold cross-validation of models selected for the complete case data (GAM\(_{CC}\)) and RF-imputed data (GAM\(_{RF}\)) was performed to assess their prediction accuracy. Both datasets were randomly split into 10 parts with 9 parts used as training data and 1 part as testing data. GAM\(_{CC}\) was fit on the training partitions of both the complete case and RF-imputed data, then the model fits were used for prediction on the testing portions of both datasets. The same cross-validation procedure was implemented for GAM\(_{RF}\) fit on both the complete case and imputed data. The 4 GAMs were compared using median absolute deviation (MAD) to assess prediction accuracy (Davydenko and Fildes 2016).
Fitting of GAM with lowest prediction errors for inference

The GAM fit with lowest prediction errors determined by cross-validation was used for inference to determine significant predictors of MC concentration in lakes and reservoirs globally. Plots of the relationships between significant \((p < 0.05)\) predictor variables determined by GAM selection and MC were generated for interpretation. Proportion of deviance explained was determined for each significant predictor term in the final model (Wood 2017). The proportion of deviance explained for the variable of interest \(x\) was calculated as:

\[
D_x = \frac{D_F - D_R}{D_N},
\]

where \(D_F\) is the explained deviance of the full GAM, \(D_R\) is the deviance of the reduced GAM with the variable of interest removed, and \(D_N\) is the deviance of the intercept-only GAM. We allowed the Tweedie distribution parameters to be estimated by mgcv for the full model and then used the same distribution parameters to fit the reduced and null models. Percent deviance explained \((D_x \times 100)\) was reported for straight-forward interpretation.

Results

RF imputation of the aggregated dataset exclusive of latitude, longitude, and MC concentration generated OOBMSE for each of the 12 variables with missingness <65% (Table 2). OOBMSE varied substantially, with specific conductivity having the largest OOBMSE but also having a wide range \((0–2959.980)\); however, 8 of the 12 included variables had small OOBMSE \(\leq 1.46\).

Variable selection on the GAM fit with the complete dataset \((n = 986)\) removed ammonium as nitrogen and specific conductivity from the model, while ammonium as nitrogen, nitrate as nitrogen, and chlorophyll were dropped in the GAM using variable selection on the RF-imputed dataset \((n = 4316)\). The complete case dataset comprised data from the United States and Canada while the RF-imputed dataset utilized data from all 22 countries. When broken down by latitude into temperate \((30–60^\circ\) N and S), subtropical \((23–30^\circ\) N and S), and tropical \((0–23^\circ\) N and S) climate regions, the amount of data representing each region was 92.6%, 5.8%, and 1.4%, respectively. Both GAM\(_{cc}\) and GAM\(_{rf}\) had better fit (higher total deviance explained) when estimated using RF-imputed data than either of those selected models estimated using the complete case data (Table 3). GAM\(_{cc}\) and GAM\(_{rf}\) fit on RF-imputed data produced lower prediction errors than those models fit on complete data, with GAM\(_{rf}\) fit on imputed data performing the best \((MAD = 1.79; \text{Fig. 2})\).

The GAM\(_{rf}\) fit on the RF-imputed data, the preferred final model for inference, removed the variables ammonium as nitrogen, nitrate as nitrogen, and chlorophyll from the model \((p > 0.05; \text{Table 4})\). This final reduced model is given as:

\[
g(E(MC)) = \beta_0 + f_1(\text{Secchi}) + f_2(pH) + f_3(\text{nitrate} - N) + f_4(\text{nitrate} + \text{nitrite} - N) + f_5(\text{total phosphorus}) + f_6(\text{dissolved organic carbon}) + f_7(\text{total nitrogen}) + f_8(\text{turbidity}) + f_9(\text{specific conductivity}) + s(\text{longitude}, \text{latitude}),
\]

where \(\beta_0 = -0.76\) \((t = 33.97, p < 0.001)\), revealed highly nonlinear relationships between the selected predictor variables and MC concentrations (Fig. 3). It was observed that most variables had a negative relationship with MC in at least some portion of the range where the primary fraction of data were located. Nitrite as nitrogen, total phosphorus, and specific conductivity had largely negative relationships with MC in the sections where the majority of their data existed (nitrite as nitrogen \(0–0.1\) mg/L = 99.8%; total phosphorus \(0–2\) mg/L = 99.7%; specific conductivity \(0–5000\) µS/cm = 98.7%); however, the remaining variables had more nuanced relationships with MC (Fig. 3). The relationship between total nitrogen and MC went from negative to slightly positive from 0 to 10 mg/L, where 99.5% of the data was present. Nitrate + nitrite as nitrogen had 99.8% of its values between 0 and 5 mg/L where the relationship with MC was slightly negative. pH within the range of 6–10, where 99.0% of the data fell, changed from a negative to positive relationship with MC as pH values increased. Dissolved organic carbon between 0 and 50 mg/L, where 99.1% of its data were located, had a negligible relationship to MC. Turbidity measurements largely ranged from 0 to 200 NTU (99.1% of the data) and went from a negative to positive.

| Table 3. Total deviance explained (%) for variable selection on GAMs when estimated using complete and RF-imputed datasets. GAM\(_{cc}\) is the selected model from complete case data; GAM\(_{rf}\) is the selected model from RF-imputed data. |
|---------------------------------|-----------------|-----------------|
| **Data** | **Complete case** | **RF-imputed** |
| **Model** | **GAM\(_{cc}\)** | **GAM\(_{rf}\)** |
| | 60.7 | 62.5 |
| | 58.2 | 62.3 |
relationship to MC. Secchi depth had a slightly positive, but oscillating to negative, relationship to MC from 0 to 10 m, which included 99.1% of the data. Lastly, the 2-dimensional latitude–longitude smoother predicted MC to be much higher than the global average at locations within the grid of 25–50° longitude and 25–50° latitude (Fig. 4). Location, total nitrogen, turbidity, Secchi depth, nitrate + nitrite-nitrogen, and pH had a larger association with MC than the other variables within the GAM fit on the RF-imputed data (Table 5).

**Discussion**

**Modeling of MC in freshwater systems**

The development of prediction-based models to determine the occurrence of MC is an underutilized but growing practice in water resource management (Francy et al. 2016, Harris and Graham 2017, Yuan and Pollard 2017, 2019, Shan et al. 2019, 2020). Meaningful studies have been constructed assessing MC on large spatial areas and incorporating numerous waterbodies (Kotak et al. 2000, Graham et al. 2004, US Environmental Protection Agency 2016, 2010, Yuan and Pollard 2017, 2019, Shan et al. 2020) and have also been constructed for other cyanobacterial response variables like biomass (Carvalho et al. 2013, Shan et al. 2019, 2020, Vuorio et al. 2019).

This study is, to our knowledge, the largest determination of associative factors related to MC concentration in fresh waterbodies around the world. As such, the findings from this research may serve to support the often-anecdotal trends between MC concentration and select water quality variables and assist resource managers in determining the most relevant parameters to measure in a freshwater system experiencing MC issues.

The final model of this study utilized 12 variables that contained varying degrees of missingness up to 65%. Variables not used within our analyses merely illustrate that those variables are not commonly collected at freshwater sites, but their exclusion in our final model is not necessarily a reflection of their lack of significance to MC production in lab or region-specific studies. For instance, temperature was reported for <35% of the data collected for our analyses but is of noted importance to toxin-producing cyanobacteria (Billam et al. 2006, Boutte et al. 2008, Gągąla et al. 2012, Li et al. 2017) because select cyanobacteria have greater growth rates at higher temperatures and prefer a more stable water column brought on by thermally stratified systems (Paerl and Huisman 2008). This finding has been reflected in other modeling research, such as by Shan et al. (2019) who observed in a Bayesian network analysis that warmer water temperatures (≥24 °C) increased the probability of hazardous MC conditions (≥1 μg/L) occurring by 23.9%.
Limited availability is a prominent restraint in the accumulation of global data. In this work, limited relevant studies originated from the subtropical (30°–23° N and S) and tropical (23°–0° N and S) climate regions. A small number of studies in tropical regions compared to that of temperate regions is a known impediment to comprehensive inference in the field of limnology (Lewis 2002, Ramírez et al. 2020).

Continued assessments of water quality and MC toxicology in these areas will certainly improve our understanding of MC production in warmer climates and beyond. However, inclusion of the few currently available studies from these regions into our GAM framework, which accounts for spatial dependence, enables us to leverage information from these studies to understand average effects of environmental

Table 4. Summary of final GAM fit on RF-imputed data (n = 4316). Deviance explained = 62.5%. edf = estimated degrees of freedom. Ref. df = reference degrees of freedom.

<table>
<thead>
<tr>
<th>Parametric</th>
<th>edf</th>
<th>Ref. df</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitude, latitude</td>
<td>27.23</td>
<td>29</td>
<td>61.52</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Secchi</td>
<td>8.22</td>
<td>9</td>
<td>11.51</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>pH</td>
<td>5.07</td>
<td>9</td>
<td>16.07</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Ammonium as nitrogen</td>
<td>0.0001</td>
<td>9</td>
<td>0.00</td>
<td>0.563</td>
</tr>
<tr>
<td>Nitrite as nitrogen</td>
<td>4.84</td>
<td>9</td>
<td>2.59</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Nitrate as nitrogen</td>
<td>0.42</td>
<td>9</td>
<td>0.07</td>
<td>0.196</td>
</tr>
<tr>
<td>Nitrate + nitrite as nitrogen</td>
<td>5.59</td>
<td>9</td>
<td>10.95</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Total phosphorus</td>
<td>6.80</td>
<td>9</td>
<td>5.10</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dissolved organic carbon</td>
<td>7.17</td>
<td>9</td>
<td>5.66</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Fluorometric chlorophyll</td>
<td>1.02</td>
<td>9</td>
<td>0.22</td>
<td>0.123</td>
</tr>
<tr>
<td>Total nitrogen</td>
<td>6.22</td>
<td>9</td>
<td>34.69</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Turbidity (NTU)</td>
<td>8.43</td>
<td>9</td>
<td>22.51</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Specific conductivity</td>
<td>2.16</td>
<td>9</td>
<td>1.78</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Figure 3. Centered smooths of variables selected by the final GAM fit on RF-imputed data. Black hash marks represent the presence of data for the x-axis variable. Nitrite and nitrate + nitrite parameters reported as nitrogen.
factors on MC concentrations across several climate regions worldwide.

Imputing missing values using RF machine learning from original data with variables having up to 65% missingness produced models with greater predictive ability than models fit only on complete data. RF imputation allowed us to include a greater number of locations in the final model, resulting in a model based on data from 22 countries. Complete case data on all the 41 variables were only available for 2 countries: the United States and Canada. Methods to address missing data are numerous, and limits on the total allowable amounts of missingness within each variable differs considerably even within fields, as previously described (Bennett 2001, Penone et al. 2014, Taugourdeau et al. 2014, Tabachnick et al. 2019). Effectiveness of imputation methods vary, but select methods, including RF-based imputation deployed by the R package missForest used in our analysis, can impute variables with up to 60% of their original values removed without significantly altering true relationships in the data (Penone et al. 2014, Taugourdeau et al. 2014). The use of machine learning imputation methods such as RF may therefore provide a useful way to address issues of missingness in global datasets and reduce personal bias from manual removal of incomplete data.

Nonlinear relationships between MC and environmental predictors were generated using GAMs in this study. The additive structure of GAMs contributes to the interpretability of the model and make it a preferable choice for real-world data, such as climatological and ecological research (Suárez-Seoane et al. 2002, Wisz et al. 2008, Brabec et al. 2014), and was an effective method for analyzing MC and water quality data across large spatial scales. Modest differences in the selected variables occurred in the GAMs fit on the complete case data and RF-imputed datasets. These dissimilarities are expected given the number of missing values imputed for some variables, but overall findings regarding potential contributors to MC in freshwater lakes and reservoirs are anticipated to be conserved (Penone et al. 2014). Note that the relationships found among the studied variables and MC are independent of location, which is included in the model. The interpretation is *ceteris paribus* (i.e., all other predictors held constant). For example, for a given lake, turbidity has a nonlinear relationship to MC concentration. Variation not captured by the selected environmental variables in the final model is captured through the spatial smooth, which is a nuisance variable in our model. This finding is a key strength of our approach because the effects shown in our results are average effects for any freshwater system within the spatial range of the data used in our analyses.

The concentration of data within the ranges of each of the 9 environmental variables reported (Fig. 3) should be noted (e.g., 99.5% of total nitrogen was from 0 to

![Figure 4. Centered MC concentrations (µg/L) estimated by the final GAM fit on RF-imputed data using 2-dimensional smoothing over longitude–latitude coordinates. Black points represent true locations from data.](image-url)

### Table 5. Percent deviance explained ($D_x \times 100$) for each of the selected predictor variables in the final GAM fit on RF-imputed data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Deviance explained (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude and longitude</td>
<td>9.01</td>
</tr>
<tr>
<td>Secchi</td>
<td>0.71</td>
</tr>
<tr>
<td>pH</td>
<td>1.13</td>
</tr>
<tr>
<td>Nitrite as nitrogen</td>
<td>0.10</td>
</tr>
<tr>
<td>Nitrate + nitrite as nitrogen</td>
<td>0.08</td>
</tr>
<tr>
<td>Total phosphorus</td>
<td>0.14</td>
</tr>
<tr>
<td>Dissolved organic carbon</td>
<td>0.31</td>
</tr>
<tr>
<td>Total nitrogen</td>
<td>1.80</td>
</tr>
<tr>
<td>Turbidity (NTU)</td>
<td>1.15</td>
</tr>
<tr>
<td>Specific conductivity</td>
<td>0.00</td>
</tr>
</tbody>
</table>
10 mg/L). The relationship between a select parameter and MC generated outside of these ranges has much larger confidence regions due to the limited range of data. Within these ranges, several trends in the data were observed and may relate to several different factors. Other than more obvious positive and negative relationships observed between variables and MC concentration, a flattened relationship between increasing variable values and MC concentration may be an indication of a saturation point. For instance, Dolman et al. (2012) found a saturation point between phosphorus content and cyanobacterial biovolume, whereas the relationship between nitrogen to cyanobacteria biovolume was not limited. Saturation of a toxic cyanobacterial bloom would eventually limit the amount of MC present within a system. Peaks in MC concentration may also reflect a limited preferable range for toxic cyanobacteria to thrive. For instance, Graham et al. (2004) found MC concentration and cyanobacterial biomass were highest between 1500 and 4000 µg/L total nitrogen. Such potential saturation values are revealed by using the GAM framework, which flexibly models these types of nonlinear relationships. Each selected variable in the final GAM and its relationship with MC is further discussed.

**Variables of significance**

**Nutrients: nitrogen and phosphorus**

Nitrogen and phosphorus have been identified as major contributors to cyanobacterial blooms and MC occurrence (Paerl et al. 2001, Yuan and Pollard 2017), with some models able to account for large amounts of MC variation in US lakes using only nitrogen and phosphorus (Yuan and Pollard 2017). In our study, total nitrogen had a largely positive relationship with MC at concentrations >1 mg/L and had the second strongest association with MC in the final GAM. Positive linear relationships between total nitrogen and MC have also been identified in past field studies (Downing et al. 2001, Pham et al. 2020; Graham et al. 2004 data not incorporated in this study). Because MC is a peptide structure, nitrogen is a key building-block in its production (Hotto et al. 2008). Nitrogen is also an integral nutrient for cyanobacterial growth and function (Hotto et al. 2008). The structure of MC comprises 14% nitrogen, and conditions in which the carbon to nitrogen molar ratio is <4.3 can reduce MC production in *Microcystis aeruginosa* cultures (Wagner et al. 2019). These factors contribute to nitrogen being noted as one of the primary drivers of MC production in freshwater systems (Otten et al. 2012, Beaulieu et al. 2013).

Although total nitrogen was observed to have a positive relationship with MC, differences were found in the relationships between the various forms of nitrogen incorporated into the final GAM. Despite nitrate being separately nonsignificant to the analysis, nitrite as nitrogen and nitrate + nitrite as nitrogen displayed negative or nearly-zero associations with MC, possibly suggesting a nuance in best forms of nitrogen for MC production in cyanobacteria. Findings may also suggest that if toxins are at high densities, a sizable bloom would require an ample amount of nitrogen, which could include nitrate as nitrogen, nitrite as nitrogen, or ammonium as nitrogen.

Total phosphorus, the only phosphorus derivative to be incorporated into the final GAM, had a negative relationship to MC. However, 98.8% of the data fell below 1 mg/L, where the relationship between total phosphorus and MC was nearly-zero to slightly negative. This limited relationship was not expected and may be attributed to several factors. First, such a relationship may indicate that although evidence suggests that phosphorus is a key nutrient to the bloom formations of phytoplankton, including cyanobacteria (Trimbee and Prepas 1987, Schindler et al. 2008, Paerl and Otten 2013), it is not a dominant factor in the production of MC. Laboratory observations indicate that MC production in *M. aeruginosa* requires a carbon to phosphorus molar ratio of <200 but a carbon to nitrogen molar ratio <8 (Wagner et al. 2019). Wagner et al. (2019) suggested that phosphorus is important for cell biomass, but nitrogen has a greater importance to the production of toxins, which may reflect the relatively weak association of phosphorus to MC compared to that of nitrogen observed in this analysis. Second, the sampled sites were possibly highly eutrophic with ample amounts of phosphorus, and therefore a strong relationship between phosphorus and MC would not be observed. The average phosphorus content of the sampled lakes before the imputation of the dataset was 120 ± 309 µg/L, which equates to conditions of possible hypereutrophy based on the Carlson trophic state index (Carlson and Simpson 1996). Such eutrophic, phosphorus-rich conditions have been shown to have other nutrients, such as nitrogen, as the dominant contributors to cyanobacterial bloom formation and/or the production of MC (Scott et al. 2019).

The reduction of both nitrogen and phosphorus will likely be needed to reduce cyanobacterial blooms and subsequent MC toxins. This possibility has been put forward in prior modeling studies such as Shan et al. (2020), whose Bayesian analysis identified that reducing MC risks in 3 lakes in China could be achieved by setting total phosphorus and nitrogen thresholds of 0.5
and 1.8 mg/L, respectively. Such findings solidify the growing call for dual nitrogen and phosphorus reductions in systems plagued by cyanobacterial blooms (Paerl and Otten 2013).

**pH**

The majority of pH values used in this study fell within 6–10, in which the relationship of pH to MC turned from negative to positive with the increase in pH value. The increase in MC production at more alkaline pH conditions has been documented (Song et al. 1998), and MC-producing species have been shown to out-compete other phytoplankton in prior laboratory studies (Yang et al. 2018). Because of the ease of measuring pH with handheld meters, it is a useful measurement to track cyanobacterial bloom formation and possibly MC occurrence, but note that bloom densities and sampling time may affect the value of pH within a system. Typically, pH values will be higher in the daylight hours as primary producers photosynthesize and lower at night when respiration occurs. The variation in pH values is also affected by the alkalinity (i.e., buffering capacity) in a waterbody and should be considered as well (Boyd et al. 2016).

**Specific conductivity**

Specific conductivity had a slightly negative relationship with MC from 0 to 5000 µS/cm, where 98.7% of the data fell. Past lab research has indicated that both the growth of cyanobacteria and MC concentration decrease with the increase in salinity (Georges des Aulnois et al. 2019). Moreover, the MC-producing species, *Microcystis*, has been found to have a lower salinity tolerance of up to 2 practical salinity units (PSU) when compared to other cyanobacterial species that can tolerate salinity of 35 PSU or greater (Paerl et al. 2001). However, *Microcystis* has also been found to tolerate a salinity of up 9.8 g/L, suggesting that the tolerances across MC-producing strains and species may vary (Orr et al. 2004). Specific conductivity measures a range of salts and inorganic capable of holding an electrical current; however, such compounds may hold differing importance to MC production. For instance, Cerasino and Salmaso (2012) found a positive correlation to MC (Spearman correlation = 0.50) in Italian subalpine lakes, whereas Aboal and Puig (2005) found a negative correlation (Pearson correlation = −0.31) in the reservoirs of the Segura river basin of southeastern Spain. Because of these understudied differences, further research is needed to understand the specific relationships that compounds have with MC occurrences, and specific conductivity’s usefulness in monitoring MC should be assessed on a case-by-case basis.

**Dissolved organic carbon**

Dissolved organic carbon had an overall nearly-zero to negative relationship with MC from 0 to 100 mg/L, where most of the data were present. This relationship may be indicative of conditions not favorable for cyanobacterial growth or MC production, although a positive relationship between cyanobacterial growth and dissolved organic matter has been observed previously under laboratory conditions (Paerl et al. 2001, Zhao et al. 2019), possibly because of the ability of cyanobacterial cells to directly uptake this carbon source for growth or to be used by bacterial communities supporting the cyanobacterial blooms (Paerl et al. 2001, Znachor and Nedoma 2010). More dissolved organic matter is typically present in a system as cyanobacterial blooms decay and cells lyse (Tessarolli et al. 2018). The negative relationship observed between dissolved organic carbon and MC in this study may be indicative of the break-down of a bloom and therefore a reduction in MC being produced.

**Transparency: turbidity and Secchi depth**

Turbidity had a nonlinear relationship to MC, with a positive relationship occurring at ~50–200 NTU. The relationship between MC concentration and turbidity has been observed in past lake surveys. Kotak et al. (2000) suggested that MC was produced at low light intensities, thereby making turbid systems favorable. Increased MC production at low light intensities has been reported in laboratory culture studies using *Microcystis* (Wiedner et al. 2003), but toxic strains of *Microcystis* do not produce more MC at low-light versus high-light conditions. However, toxic strains will dominate nontoxic strains under both light conditions (LeBlanc et al. 2011). *Microcystis* colonies are equipped with gas vesicles, allowing them to regulate buoyancy (Paerl and Otten 2013). This physiological characteristic may allow them to form blooms on the surface, giving them preferential access to sunlight over other phytoplankton species and other cyanobacteria by circumventing the low-light conditions of highly turbid areas.

Secchi depth, a measure of water transparency, had an oscillating relationship with MC. Approximately 99.2% of the Secchi depth measurements in the data were <10 m, in which a peak of positive relationship was observed from ~2–4 m before returning to a negative or nearly-zero relationship on either side of this range. In general, MC was unrelated to Secchi depth.
as its value increased. This finding was expected because greater Secchi depths can be associated with oligotrophic conditions containing low phytoplankton densities (Carlson and Simpson 1996). Secchi depth is a useful tool to track the progression and density of the bloom in some situations (Joung et al. 2011). However, measurements are not always proportional to cyanobacterial or phytoplankton abundances in a waterbody because inorganic turbidity (e.g., sediment) or other pollution factors may affect Secchi depth values (Swift et al. 2006). Such disruptive factors may have contributed to the minimal relationship between Secchi depth and MC and may be the reason for the negative relationship between MC and Secchi depths <2 m.

**Location: latitude and longitude**

Location was the most important predictor of MC. In a study of 200 Midwestern United States lakes and reservoirs, particulate MC was significantly correlated to an increase in latitude, which was attributed to correlated changes in nutritional, physical, and chemical parameters (Graham et al. 2004). An assessment of MC concentration over the conterminous United States by grouping land patterns into 9 ecoregions also found substantial differences in the MC concentrations between regions (Beaver et al. 2014). In this study, MC was predicted to be highest at ~25–50° longitude and 25–50° latitude, corresponding to some of the highest MC concentrations documented in this study found in eastern Europe, including Serbia (Simeunovic et al. 2010, Taugourdeau et al. 2014) and Poland (Mankiewicz-Bozcek et al. 2006). Because data collected for this study largely originated in North America, increasing MC data collection in Asia, Europe, and Africa may better help determine global factors contributing to MC occurrence and accumulation.

**Conclusions**

The results of multiple statistical methods incorporated into this study revealed various environmental variables significantly related to the concentration of MC worldwide and included a reliable technique to impute missing data within an aggregated water quality dataset. To our knowledge, the aggregated dataset is the largest global accumulation of MC data. Not only did the fit of the model on imputed data produce more accurate predictions, an important metric for managers, but the imputed data also provided more spatial coverage for the model. Utilizing datasets with low overall missingness is recommended, but some missingness may be resolved using imputation and will likely need to be assessed on a case-by-case basis.

The final GAM indicated that environmental variables, such as location, total nitrogen, turbidity, and pH, are associated with MC concentrations in fresh waterbodies. Numerous nonlinear relationships were observed between the selected predictors and MC concentration. Such results reflect the usefulness of GAMs to assess nonlinear and nonnormal data.

These findings may serve as both reference and validation to often-anecdotal summaries of the trends in MC concentration worldwide. Moreover, the combination of machine learning methods for imputation and nonparametric modeling used in this analysis may serve as an effective procedure for utilizing global datasets containing large amounts of missingness, which is often the case when using data from many origins and sampling methodologies. Lastly, water resource managers can apply these methods to observational data to improve future water quality forecasting of toxic cyanobacterial blooms.

**Acknowledgements**

The authors thank the input and comments of Dr. F. Stephen Dobson to this research.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This project was supported by a grant from the National Science Foundation (DEB-1831094), the Alabama Agricultural Experiment Station, and the Hatch Program (ALA016-1-16007) and Aquaculture program (2017-70007-27132) of the National Institute of Food and Agriculture, US Department of Agriculture.

**ORCID**

Riley P. Buley  http://orcid.org/0000-0003-0721-3933
Hannah E. Correia  http://orcid.org/0000-0003-3476-3674
Ash Abebe  http://orcid.org/0000-0001-5759-2383
Tahir B. Issa  http://orcid.org/0000-0002-4322-5264
Alan E. Wilson  http://orcid.org/0000-0003-1080-0354

**References**

Beaulieu M, Pick F, Gregory-Eaves I. 2013. Nutrients and water temperature are significant predictors of


