Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset

Ted D. Harris & Jennifer L. Graham

Taylor & Francis
Taylor & Francis Group

Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset

Ted D. Harris[a] and Jennifer L. Graham[b]

[a]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS; [b]US Geological Survey, Lawrence, KS

ABSTRACT

Harris TD, Graham JL. 2017. Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset. Lake Reserve Manage. 33:32–48.

Cyanobacterial blooms degrade water quality in drinking water supply reservoirs by producing toxic and taste-and-odor causing secondary metabolites, which ultimately cause public health concerns and lead to increased treatment costs for water utilities. There have been numerous attempts to create models that predict cyanobacteria and their secondary metabolites, most using linear models; however, linear models are limited by assumptions about the data and have had limited success as predictive tools. Thus, lake and reservoir managers need improved modeling techniques that can accurately predict large bloom events that have the highest impact on recreational activities and drinking-water treatment processes. In this study, we compared 12 unique linear and nonlinear regression modeling techniques to predict cyanobacterial abundance and the cyanobacterial secondary metabolites microcystin and geosmin using 14 years of physiochemical water quality data collected from Cheney Reservoir, Kansas. Support vector machine (SVM), random forest (RF), boosted tree (BT), and Cubist modeling techniques were the most predictive of the compared modeling approaches. SVM, RF, and BT modeling techniques were able to successfully predict cyanobacterial abundance, microcystin, and geosmin concentrations <60,000 cells/mL, 2.5 µg/L, and 20 ng/L, respectively. Only Cubist modeling predicted maxima concentrations of cyanobacteria and geosmin; no modeling technique was able to predict maxima microcystin concentrations. Because maxima concentrations are a primary concern for lake and reservoir managers, Cubist modeling may help predict the largest and most noxious concentrations of cyanobacteria and their secondary metabolites.

Cyanobacteria are photosynthetic bacteria capable of forming large-scale harmful algal blooms (CyanoHABs) in aquatic systems. CyanoHABs are likely increasing in frequency, duration, intensity, and geographical extent worldwide (Paerl 2014, Otten and Paerl 2015). The apparent global increase in CyanoHABs has been linked to a multitude of factors including increased water temperature, longer periods of water stratification, and nutrient (nitrogen [N] and phosphorus [P]) and persistent organic pollutant loading (Paerl and Huisman 2008, Paerl and Otten 2013, Harris and Smith 2016).

CyanoHABs pose a serious problem for water users because they can produce cyanotoxins, a suite of potent neurotoxins and hepatotoxins (e.g., microcystin; Otten and Paerl 2015) that can adversely affect human health. Direct contact with blooms may cause asthma and skin irritations, whereas ingestion may cause vomiting, muscle weakness, and in rare cases death (Chorus and Bartram 1999, Otten and Paerl 2015). Additionally, CyanoHABs are the primary producers of metabolites that cause taste-and-odor (e.g., geosmin and 2-methylisoborneol) events in drinking water supply reservoirs (Jüttner and Watson 2007). Taste-and-odor metabolites impart unpalatable tastes and earthy and/or musty odors to raw drinking water supplies, which ultimately lead to increases in customer complaints to the water utilities that supply the tainted finished drinking water (Dietrich 2006). In response, water utilities must use expensive advanced treatment options (e.g., activated carbon and/or ozone) to remove cyanotoxins and taste-and-odor compounds (Dunlap et al. 2015). Thus, CyanoHABs are an increasingly expensive problem

for recreational concessionaires and drinking water utilities.

Given the health hazards and taste-and-odor problems that CyanoHABs cause, there have been numerous attempts to create models that predict cyanobacteria bloom formation and their noxious secondary metabolite production. Most studies have used linear models to predict cyanobacteria and their secondary metabolites (e.g., Smith et al. 2002, Dzialowski et al. 2009, Beaulieu et al. 2014 and references therein), but recent studies have used more extensive, nonlinear modeling techniques, including neural networks (Nnet; Recknagel et al. 2006, Ahn et al. 2011, Parinet et al. 2013, Millie et al. 2014), support vector machines (SVM; Xie et al. 2012), and random forest (RF; Jacoby et al. 2015), among others (e.g., mixed effect modeling; Taranu et al. 2012). Although these nonlinear modeling techniques have been successful at predicting cyanobacteria and their secondary metabolites in the lakes and reservoirs studied, relatively few studies have investigated whether other nonlinear modeling techniques such as partial least squares, boosted tree (BT), multivariate adaptive regression splines (MARS), and Cubist can accurately predict cyanobacteria and their secondary metabolites.

Several studies have developed predictive models for cyanobacteria and their secondary metabolites in Cheney Reservoir, Kansas. Smith et al. (2002), the first study to create predictive models for Cheney Reservoir, used linear models to show that chlorophyll-*a* (Chl-*a*) was related to total P (TP), and that geosmin concentrations were related to Chl-*a*; however, only a small ($n = 6$) amount of data were collected. Christensen et al. (2006) used ordinary least-squares linear regression modeling on a slightly larger dataset and found that the cyanobacterium *Anabaena* (currently *Dolichospermum*) and geosmin ($n = 16$ and 18, respectively) were related to turbidity and specific conductance. Although these models initially were effective at predicting cyanobacterial-related events in Cheney Reservoir, they were not robust over time due to the relatively small amount of data collected. In contrast to earlier studies on Cheney reservoir that found relations between environmental variables, cyanobacteria, and their secondary metabolites, more recent studies were either unable to develop significant linear regression models (Dzialowski et al. 2009) or linear regression models that explained >46% of the variation within the collected data (Stone et al. 2013). Given

that linear models have performed poorly at predicting cyanobacteria and their secondary metabolites in Cheney Reservoir and many other lakes and reservoirs worldwide, studies investigating nonlinear modeling techniques are needed for more accurate predictions.

In this study, we developed and compared 12 unique linear and nonlinear regression modeling techniques to predict cyanobacterial abundance and the cyanobacterial secondary metabolites microcystin and geosmin using 14 years of physiochemical water quality data collected from Cheney Reservoir. Our primary study objectives were to (1) examine the temporal trends related to cyanobacterial blooms, (2) develop and compare modeling techniques, and (3) build the best predictive models for cyanobacterial abundance, microcystin, and geosmin occurrence. The 3 best modeling techniques for each response variable were chosen by lowest root mean square error (RMSE) and investigated further by comparing observed and predicted values. Additionally, the most important predictor variables of each modeling technique were examined to better understand the underlying factors that cause CyanoHABs in Cheney Reservoir.

Methods

Study site

Cheney Reservoir (97°50′16.11″W, 37°45′32.99″N) is a large (surface area = 31 km^2), shallow (average depth = 6.1 m), eutrophic (average TP = 100 µg/L) impoundment located in south-central Kansas (Stone et al. 2013). The reservoir rarely thermally stratifies because of persistent winds and shallow depths. Cheney Reservoir supplies 51–69% of the municipal water supply for the city of Wichita, Kansas (Ziegler et al. 2010). The reservoir has had cyanobacteria-caused taste-and-odor and toxin events since 1990 (Smith et al. 2002, Christensen et al. 2006), resulting in recreational advisories and increased water treatment costs.

Sample collection and data analysis

Since April 2001, the US Geological Survey (USGS) has routinely collected discrete water quality samples at Cheney Reservoir near the dam (USGS station 07144790; see Supplemental Fig. S1 for map of watershed and reservoir). All sample collection and analyses were conducted using USGS protocols as described in

Stone et al. (2013). Briefly, samples were collected at 2-week or monthly intervals from May 2001 to June 2015. Samples were collected at the surface (0.5 m) with a Kemmerer sampler from May 2001 to July 2004; vertical integrated photic zone samples were collected from August 2004 to June 2015. No significant differences existed between surface and vertical integrated photic zone samples (Stone et al. 2013). With the exception of geosmin, microcystin, and phytoplankton, all samples were analyzed by the USGS National Water Quality Laboratory. Geosmin was analyzed using gas chromatography-mass spectrometry (GC-MS) by Engineering Performance Solutions (Zimmerman et al. 2002). Microcystin was analyzed via the congener independent enzyme-linked immunosorbent assays (ELISA) by the USGS Organic Geochemistry Research Laboratory. Phytoplankton analyses were conducted by BSA Environmental Services, Inc., using membrane-filtered slides (McNabb 1960); a minimum of 400 natural units were counted per sample.

More than 100 physiochemical water-quality variables were measured at least once between April 2001 and November 2015 on Cheney Reservoir. All water quality data are available through the USGS National Water Information System at http://dx.doi.org/10.5066/F7P55KJN and in Graham and Harris (2016; data used for this study in Supplemental Tables S1–3). To avoid collinearity between potential explanatory variables, all explanatory variables with correlation coefficients $>|0.75|$ were removed from further analyses as per Kuhn and Johnson (2013). Additionally, any potential explanatory variable with $>5\%$ of the observations missing was removed from further analyses. Response and explanatory data with concentrations less than the analytical limit of detection were substituted with a value half of the limit of detection (see Supplemental Table 1 and Supplemental Tables 1–3 in Harris et al. 2016 for limits of detection). Because past studies (Christensen et al. 2006, Stone et al. 2013) on Cheney Reservoir have noted seasonality as a strong explanatory variable, all models used Fourier transformed variables (i.e., sin and cos) as potential explanatory variables (Helsel and Hirsch 2002), leaving 24 potential explanatory variables for the models (Table 1). Additionally, because we wanted to include the effects of antecedent weather conditions on physiochemical conditions at the sampling site, reservoir elevation was used as an explanatory variable and served as

surrogate for extreme precipitation events. Data for cyanobacterial abundance, microcystin, and geosmin included 185, 176, and 185 observations, respectively (Supplemental Tables S1–3).

Statistical analyses

To examine temporal trends in cyanobacterial abundance, microcystin, and geosmin, all discrete samples were normalized to the standard deviation of each variable, x, using the formula:

$$\frac{x - \text{Avg}}{\text{Stdev}}, \qquad (1)$$

where $x =$ a single discrete sample of a response variable, Avg $=$ the average of a response variable, and Stdev $=$ the standard deviation of a response variable. One-way analysis of variance (ANOVA) was used to compare monthly means of normalized cyanobacterial abundance, microcystin, and geosmin. If an ANOVA had a significant result ($P < 0.05$), post-hoc Tukey tests were used to distinguish differences among normalized monthly means.

Predictive models were run in combination using the train function in the caret package in R 3.2.2, as per Kuhn and Johnson (2013, see chapter 10). Data were split into training and test datasets using the createDataPartion function in R; 75% of the response variable data were used in training. The createDataPartion function selects 75% of the data at random; the set.seed function in R was used so that the random data selection was consistent throughout the modeling procedures. Data were centered and scaled using the center and scale functions in the caret package in R prior to predictive modeling (Kuhn and Johnson 2013). Each model used repeated (repeats $=$ 5) 10-fold cross-validation using the trainControl function in R and was tuned as per Kuhn and Johnson (2013).

Twelve different predictive models were trained using the training dataset and were compared by RMSE performance on the test dataset using the resamples function in the caret package in R. Models included linear and nonlinear regression models (Table 2); this suite of models was selected because they are commonly found (e.g., ordinary linear regression, Nnet, SVM, RF) or are absent from the current literature (e.g., elastic net, Cubist). For each response variable (i.e., cyanobacterial abundance, microcystin, and geosmin),

Table 1. List of variables used in model development. With the exception of the 3 response variables (cyanobacterial abundance, microcystin, and geosmin), all variables were used as explanatory variables. P-code represents USGS parameter code for the explanatory variable. NGVD 1929 = National Geodetic Vertical Datum of 1929. All data used in model development are available in Supplemental Tables S1–3.

Variable	Abbreviation	Units	P-code
Fourier transformed date	sin	unitless	—
Fourier transformed date	cos	unitless	—
Dissolved oxygen	DO	mg/L	P00300
Reservoir surface elevation above NGVD 1929	elev	ft	P62614
pH	pH	unitless	P00400
Specific conductance	Spc	µS/cm	P00095
Temperature	Temp	°C	P00010
Turbidity	Turb	FNU	P63680
Bicarbonate	Bicarb	mg/L	P29806
Bromide	Brom	mg/L	P71870
Silica	Si	mg/L as SiO_2	P00956
Total Kjeldahl nitrogen	TKN	mg/L as N	P00625
Ammonia	NH	mg/L as N	P00608
Nitrate plus nitrite	NO	mg/L as N	P00631
Orthophosphate	OP	mg/L as P	P00671
Dissolved phosphorus	DP	mg/L as P	P00666
Total Phosphorus	TP	mg/L as P	P00665
Total Nitrogen	TN	mg/L as N	P00600
Fecal coliforms	FC	colonies per 100 mL	P31625
Chlorophyll *a*	Chl-*a*	µg/L	P70953
Iron	Fe	µg/L	P01045
Suspended sediment concentration	SSC	mg/L	P80154
Total nitrogen to total phosphorus ratio	TNTP	unitless	—
Nitrate plus nitrite to ammonia ratio	NONH	unitless	—
Cyanobacterial abundance	Cyano	cells/mL	—
Microcystin	MC	µg/L	P65210
Geosmin	Geo	ng/L	P51285

the lowest 3 average RMSE predictive models (Supplemental Fig. S2a–c) were compared by examining predicted and observed response variable concentrations using temporal plots created with Sigmaplot 11.0. Additionally, predicted and observed values were extensively compared in bivariate plots by regressing observed values on predicted data using ordinary least squares linear regression (results from all developed models in Supplemental Tables S4–6).

Variable importance

Variable importance for each modeling technique and the varimp function from the caret package in R are

Table 2. List of compared modeling approaches and their abbreviation (Abbrev.).

Model	Abbrev.
Ordinary Linear Regression	Linear
Partial Least Squares	PLS
Elastic Net	Enet
Neural Networks	Nnet
Multivariate Adaptive Regression Splines	MARS
Support Vector Machines	SVM
Single Trees	CART
Bagged Trees	BagT
Boosted Trees	BT
Conditional Inference Trees	CI Tree
Random Forest	RF
Cubist	—

explained in (Kuhn and Johnson 2013); scale was set to "TRUE" for each varimp function used. Briefly, regardless of the modeling technique, each variable importance score represents how relatively important each explanatory variable is in predicting the response variable, scaled from 0 to 100, with 100 representing the most important predictor variable.

Results

Temporal patterns in cyanobacterial abundance, microcystin, and geosmin

Cyanobacterial abundance ranged from 1 to 129,836 cells/mL (median = 1861, average = 7532 cells/mL). Cyanobacterial blooms were evident in late summer from 2002 to 2015. Normalized cyanobacterial abundance was highest in August, September, and October (Fig. 1a). May, August, September, and October had standard deviations >1. March and June had the lowest normalized means and standard deviations compared to other months. Despite the apparent seasonal pattern, normalized means were not significantly different among months (ANOVA; $P = 0.39$).

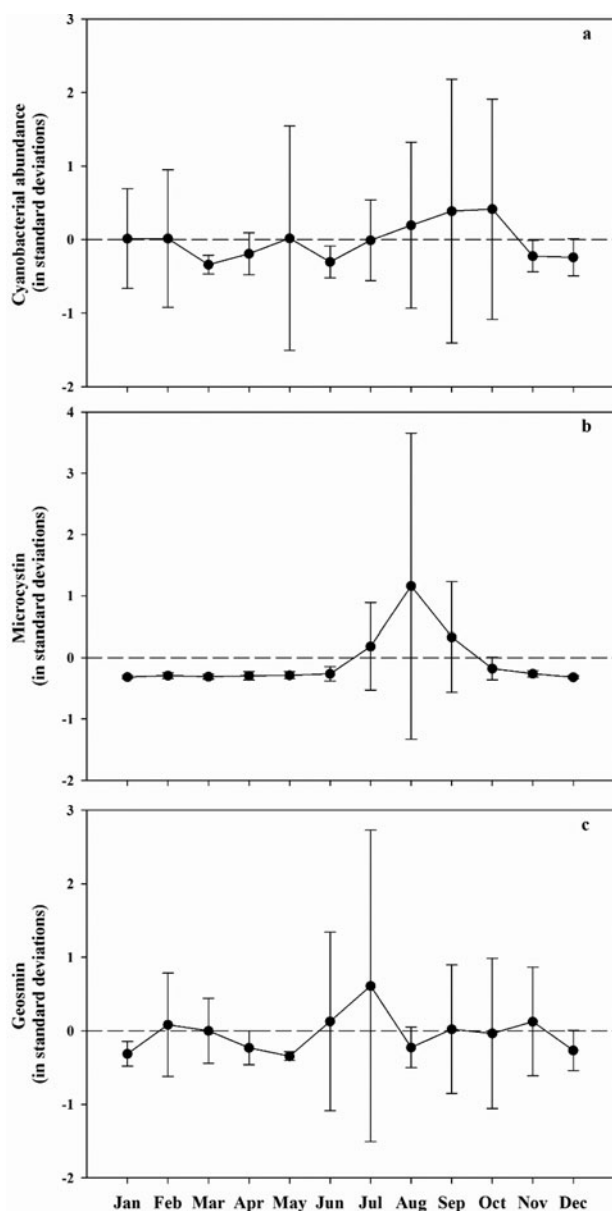Microcystin ranged from 0.1 to 9.0 µg/L (median = 0.1, average = 0.37 µg/L). Similar to

Figure 1. Normalized temporal trends of (a) cyanobacterial abundance, (b) microcystin, (c) and geosmin concentrations. Each data point represents the average normalized value of observations collected from a specific month from 2002 to 2015. Error bars represent standard deviations from the normalized mean.

cyanobacterial abundance, microcystin was also constrained to late summer; however, the highest microcystin concentrations occurred in August, whereas the highest cyanobacterial abundance concentrations occurred in September and October. Normalized microcystin was greater than the mean in only 3 of the 12 months (i.e., Jul, Aug, and Sep; Fig. 1b), and those months also had much higher (>0.2 normalized standard deviations) variability than other months. With the exception of September, the normalized microcystin mean for August was significantly (ANOVA;

$P < 0.001$) higher than all other months and also had the largest standard deviation. Thus, the highest and most variable microcystin concentrations seemed to precede the highest and most variable cyanobacterial abundances in Cheney Reservoir.

Geosmin ranged from 1 to 113 ng/L (median = 2.5, average = 6.3 ng/L). Normalized geosmin showed more inter- and intra-annual variation than cyanobacterial abundance or microcystin (Fig. 1c). Although geosmin was highest in June and July, which temporally proceeded the highest microcystin and cyanobacterial abundance concentrations, geosmin also had smaller peak concentrations in February and November. June, July, and October had standard deviations >1, indicating substantial inter-annual variation. Therefore, geosmin was highly variable throughout the year but seemed to have the highest concentrations before the highest microcystin concentrations (Aug) and cyanobacterial abundances (Sep and Oct). Normalized means were not significantly different among months (ANOVA; $P = 0.24$).

Predicting cyanobacterial abundance

SVM, RF, and BT models had the 3 lowest RMSE for cyanobacterial abundance (Supplemental Fig. S2a; Fig. 2a–c). The SVM model had the lowest RMSE of any of the cyanobacterial abundance models and performed best (by RMSE) on cyanobacterial abundances <60,000 cells/mL (Fig. 3a). For abundances >60,000 cells/mL, RF and BT models slightly outperformed SVM (Fig. 3b and c). The top 3 models by RMSE were unable to predict the highest cyanobacterial abundances in the dataset. Although the Cubist model was outperformed by multiple models on cyanobacterial abundances <60,000 cells/mL, it outperformed all models on cyanobacterial abundances >60,000 cells/mL and had the highest R^2 value between predicted and observed abundances of all cyanobacteria models (Fig. 3d and Supplemental Fig. S3a). Additionally, the Cubist model had a slope of observed versus predicted of 0.70, whereas other models had slopes ≤0.44, indicating the Cubist model had a more robust fit on larger cyanobacterial abundances compared to the other modeling techniques.

SVM, RF, and BT models all identified reservoir elevation and Chl-a as relatively important predictor variables for cyanobacterial abundance (Fig. 4a–c). Orthophosphate/phosphorus species, iron,
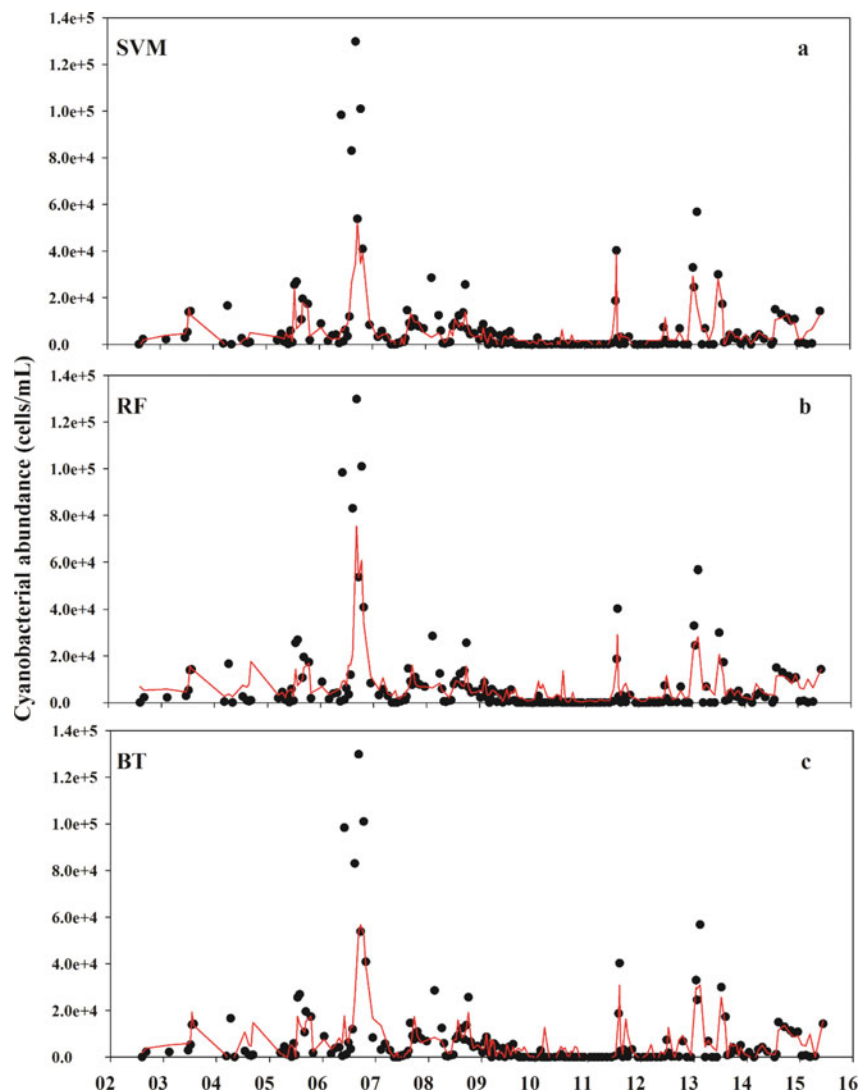
Figure 2. Observed (dots) and predicted (line) cyanobacterial abundance from 2002 to 2015 in Cheney Reservoir using (a) Support Vector Machine (SVM), (b) Random Forest (RF), and (c) Boosted Tree (BT) modeling approaches.

temperature, and time of year (i.e., sin) were also important predictor variables to the SVM, RF, and BT models. In contrast to the SVM, RF, and BT models, Cubist modeling identified specific conductance as an important predictor variable (Fig. 4d). The most important variables for BT indicated that reservoir elevation and Chl-*a* had the most impact on predictive BT modeling. Variable importance plots for SVM, RF, and Cubist modeling techniques had more variables with scaled scores >40 compared to BT, indicating that BT model performance is influenced by a smaller number of predictors than the other modeling approaches.

Predicting microcystin

Similar to cyanobacterial abundance, the SVM model for microcystin had the lowest RMSE compared to

the other models (Supplemental Fig. S2b; Fig. 5a). Cubist and BT models also had relatively low RMSE for predicting microcystin (Fig. 5b and c). The Cubist model outperformed SVM and BT models on the 2 largest microcystin concentrations in the dataset and explained nearly double the variance compared to the SVM and BT models in bivariate plots of predicted and observed concentrations (Fig. 6a–c). Yet, none of the top 3 models predicted microcystin concentrations >6 µg/L. With the exception of one predicted microcystin concentration of 6.1 µg/L by Nnet modeling (observed concentration = 7.3 µg/L; Supplemental Table S5), no other modeling technique predicted microcystin concentrations >2.5 µg/L, irrespective of the observed concentration.

The SVM, Cubist, and BT models showed that temperature and time of year (i.e., sin) were the most
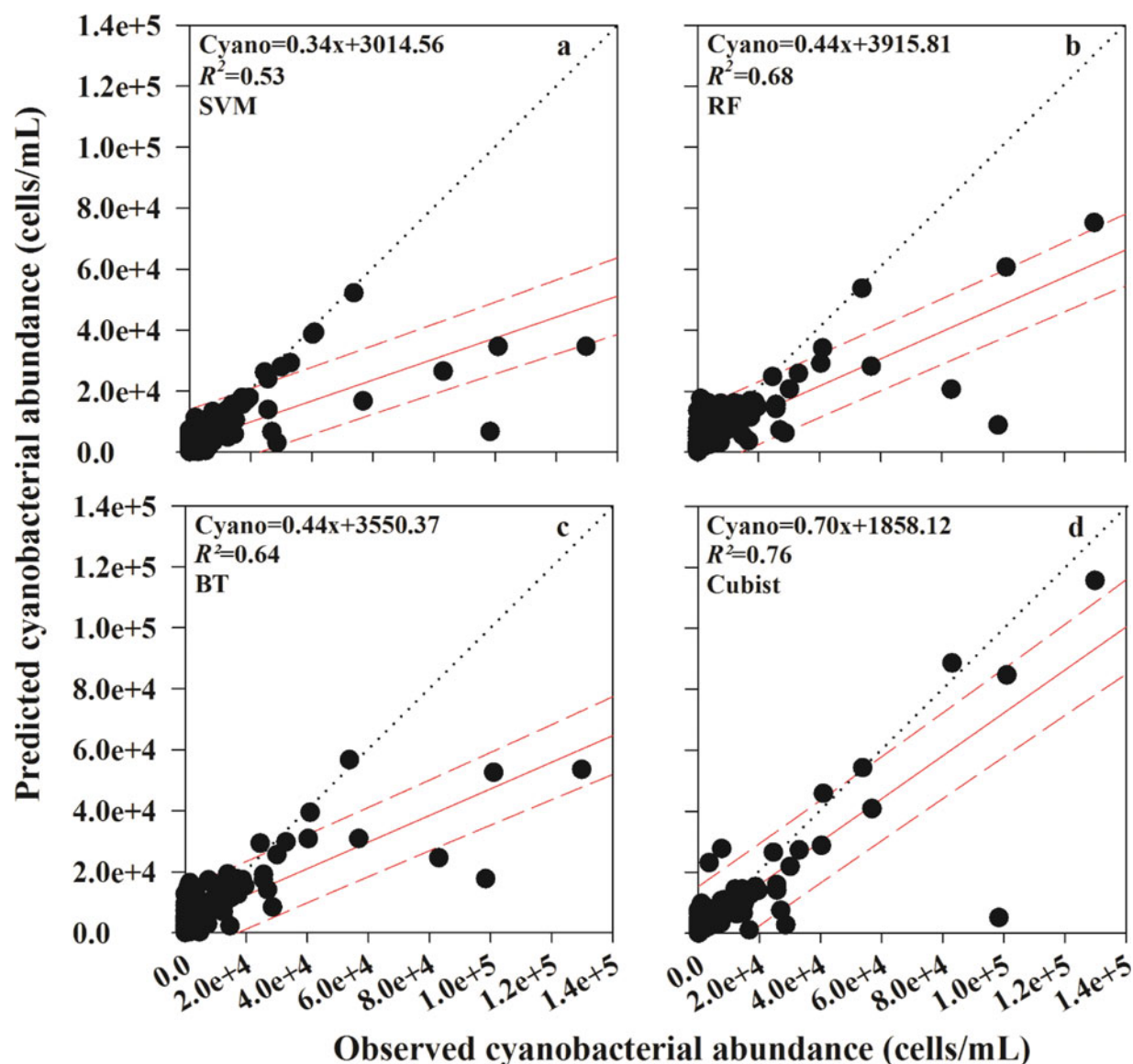
Figure 3. Observed compared to predicted cyanobacterial abundance using (a) Support Vector Machine (SVM), (b) Random Forest (RF), (c) Boosted Tree (BT), and (d) Cubist modeling approaches. Solid line represents linear regression line, dashed lines represent 95% prediction intervals, and dotted line represents 1:1 line.

important variables for predicting microcystin concentrations (Fig. 7a–c). Chl-*a*, iron, and dissolved oxygen also were important predictor variables for microcystin. Cubist and BT plots had fewer variables with scaled scores >40 compared to SVM, indicating that temperature, sin, and Chl-*a* variables had a substantially larger impact on modeling efforts compared to other predictor variables.

Predicting geosmin

RF, SVM, and BT had the 3 lowest RMSE values for predicting geosmin (Supplemental Fig. S2c; Fig. 8a–c). When observed and predicted geosmin

concentrations were compared, the RF model outperformed the SVM and BT models by RMSE overall, and especially on geosmin concentrations >20 ng/L (Fig. 8a and 9a–c). Similar to the cyanobacterial abundance and microcystin models, however, the Cubist model outperformed all models on the highest concentrations in the dataset (i.e., geosmin concentrations >20 ng/L; Fig. 9d; Supplemental Fig. 3b). Although the Cubist model had a lower R^2 than the RF model, the slope of the Cubist model was much closer to 1 (0.85 compared to 0.45, respectively), indicating a more robust fit when geosmin concentrations exceeded 20 ng/L. Additionally, the Cubist model for geosmin was the only model developed to accurately predict the highest
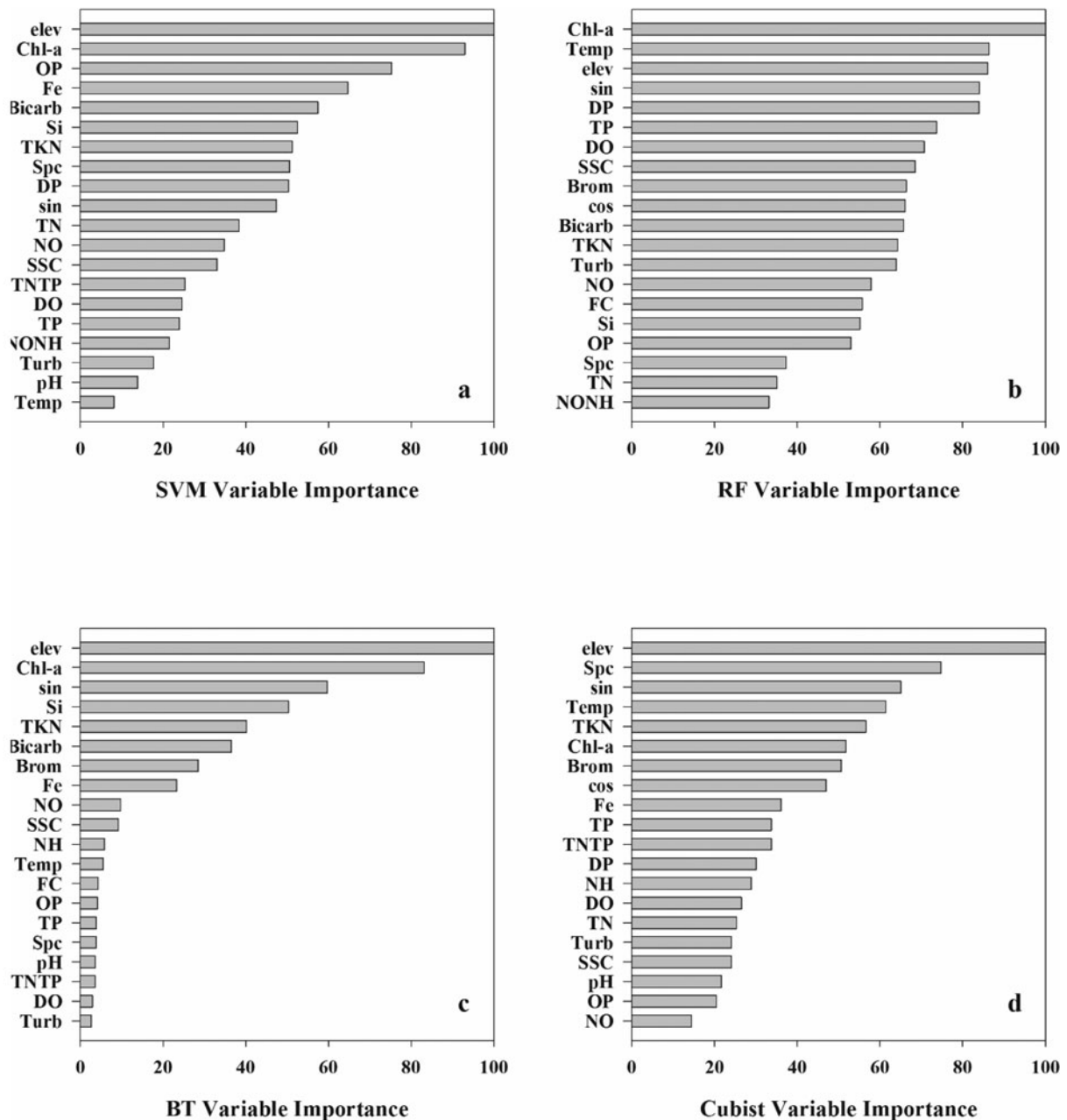
Figure 4. Top 20 most important variables for predicting cyanobacterial abundance for (a) Support Vector Machine (SVM), (b) Random Forest (RF), (c) Boosted Tree (BT), and (d) Cubist modeling approaches.

concentration of a response variable (Fig. 9d). The Cubist model indicated that when turbidity was >22.2 FNU and silica was <10.43 mg/L, maxima geosmin concentrations could be accurately predicted using Chl-*a* as an explanatory variable (data not shown).

The RF model identified suspended sediment, nitrate plus nitrite, and time of year (i.e., cos) as the 3 most important predictor variables. The SVM and BT models also identified light (i.e., suspended sediment concentration), N species and/or ratios (i.e.,

total Kjeldahl N [TKN], $NO_3:NH_3$ ratio), and Chl-*a* as important variables for geosmin prediction. In contrast to the other modeling techniques, the Cubist model used substantially fewer explanatory variables and identified silica as the most important predictor variable (Fig. 10a–d). Similar to microcystin variable importance scores, SVM, BT, and Cubist plots had few variables with scaled scores >40, indicating that TKN, cos, and silica, respectively, had a substantial impact on predicting geosmin, whereas RF modeling used
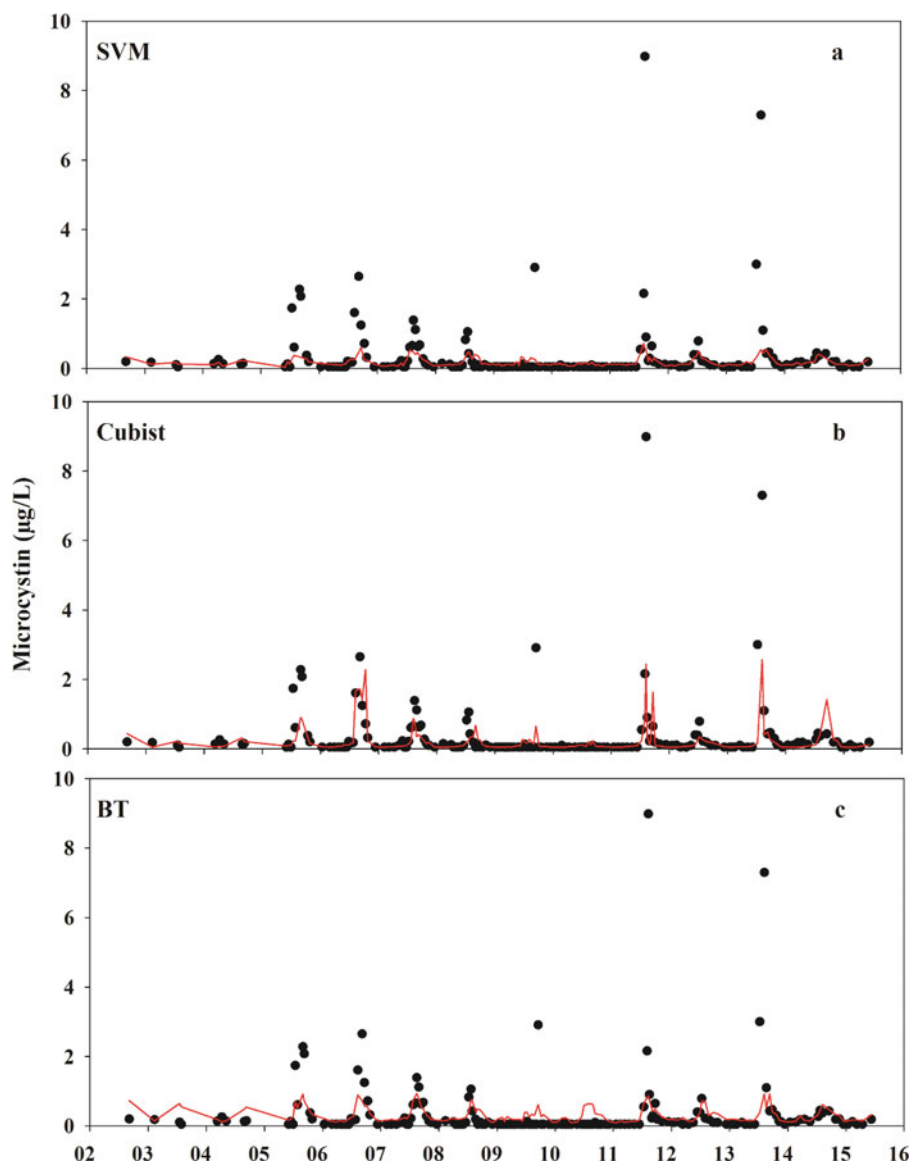
Figure 5. Observed (dots) and predicted (line) microcystin concentration from 2002 to 2015 in Cheney Reservoir using (a) Support Vector Machine (SVM), (b) Cubist, and (c) Boosted Tree (BT) modeling approaches.

substantially more explanatory variables within the model.

Discussion

Overall, predictive models for cyanobacterial abundance, microcystin, and geosmin performed poorly at predicting maxima concentrations in Cheney Reservoir. With the exception of the geosmin Cubist model, no model predicted the highest 3% (i.e., maxima) of cyanobacterial abundance, microcystin, or geosmin concentrations in the dataset. There are several probable reasons for the underestimation of maxima concentrations by developed models. With the exception of drought in 2006 and drought followed by extreme

rainfall events in 2006 and 2011–2013, the most important variables (i.e., reservoir elevation, nutrient concentrations, temperature; Fig. 4a–d) for cyanobacterial prediction exhibited seasonal patterns throughout the study period, whereas maxima cyanobacterial abundances during 2002–2015 were not constrained to a specific season (Fig. 1a). There were substantial intra- and inter-annual variations in maximum cyanobacterial abundance, timing, and dominant taxa during blooms. The underperformance of models for maxima cyanobacterial abundance likely indicates that models attempting to predict cyanobacteria using seasonally changing variables could not differentiate bloom forming conditions between seasons and/or years because cyanobacterial blooms
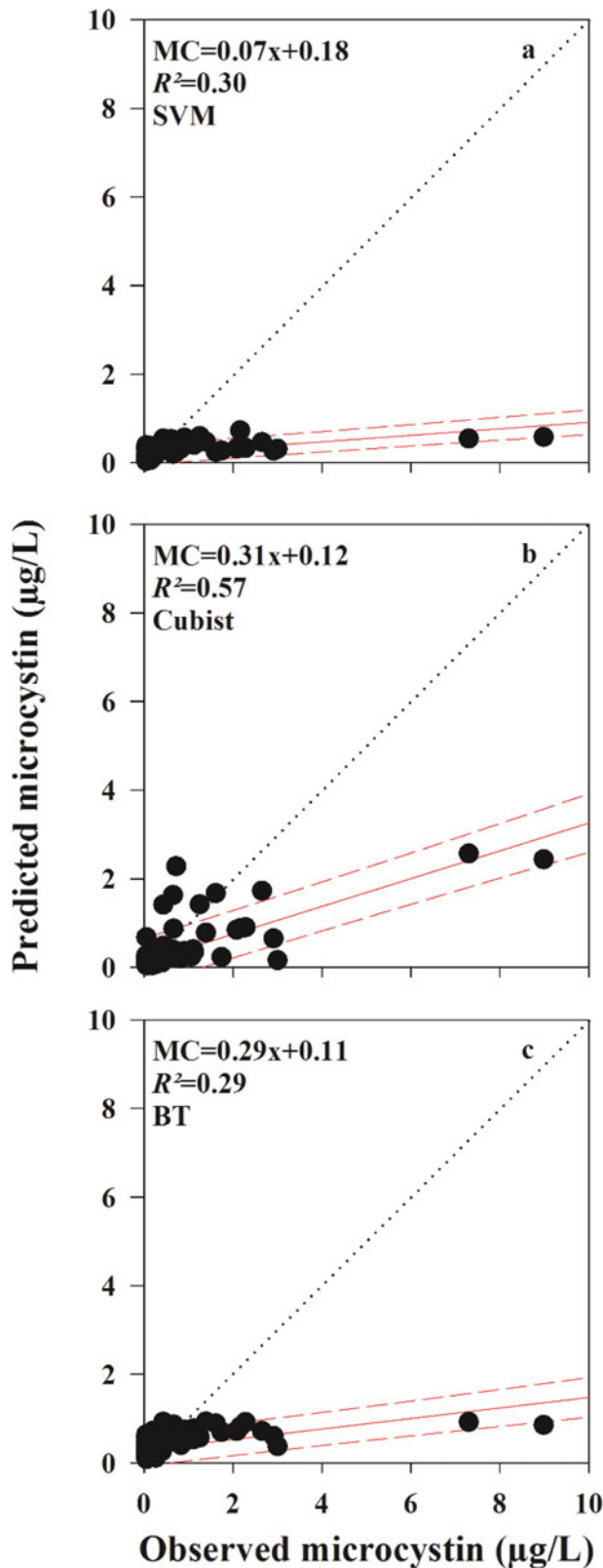
Figure 6. Observed compared to predicted microcystin concentration using (a) Support Vector Machine (SVM), (b) Cubist, and (c) Boosted Tree (BT) modeling approaches. Solid line represents linear regression line, dashed lines represent 95% prediction intervals, and dotted line represents 1:1 line.

occurred across a wide range of environmental conditions.

Although microcystin exhibited a clear seasonal pattern (Fig. 1b), no modeling technique predicted maxima concentrations. All modeling techniques recognized the seasonal pattern (i.e., by temperature and sin predictor variables; Fig. 7a–c) and accurately predicted that microcystin concentrations would occur, but no modeling approach discerned substantial inter-annual differences in the magnitude of late summer (Jul–Sep) microcystin concentrations. *Microcystis* is likely the dominant microcystin producer in Cheney Reservoir (Otten et al. 2016), and observed inter-annual variation in microcystin concentrations are likely linked to the overall abundance of *Microcystis*. Additionally, each bloom, including those dominated by *Microcystis*, is unique in (1) the percentage of cells capable of producing microcystin and (2) the amount of microcystin produced per cell (i.e., cell quota; Pearson et al. 2016). Thus, although the maxima microcystin concentrations were confined to late summer, not all late summer cyanobacterial blooms were capable of producing microcystin, which in turn caused predictive models to underperform on maxima concentrations.

Similar to cyanobacterial abundance, geosmin exhibited substantial intra- and inter-annual variation (Fig. 1c), which caused all modeling techniques except Cubist to underestimate the highest concentrations observed in the dataset. In contrast to all other modeling approaches, Cubist modeling recognized that when turbidity was relatively high (>22.2 NTU) and silica was relatively low (<10.43 mg/L of SiO_2), the highest concentrations were related to the Chl-*a* concentration. The recognition of this pattern by Cubist was likely due to its unique modeling structure. Similar to other tree-based modeling approaches, Cubist models are constructed by creating a set of rules that split the data at terminal nodes; each of these nodes uses a linear equation to predict response variables. Cubist modeling differs from other approaches because it uses (1) a unique smoothing process for linear models created at each terminal node, (2) a boosting-like procedure called committees, and (3) finalized committees adjusted to increase prediction performance using a K nearest neighbors-like procedure (further details on Cubist modeling construction and procedures are in Kuhn and Johnson 2013). Overall, these differences allowed the Cubist modeling technique
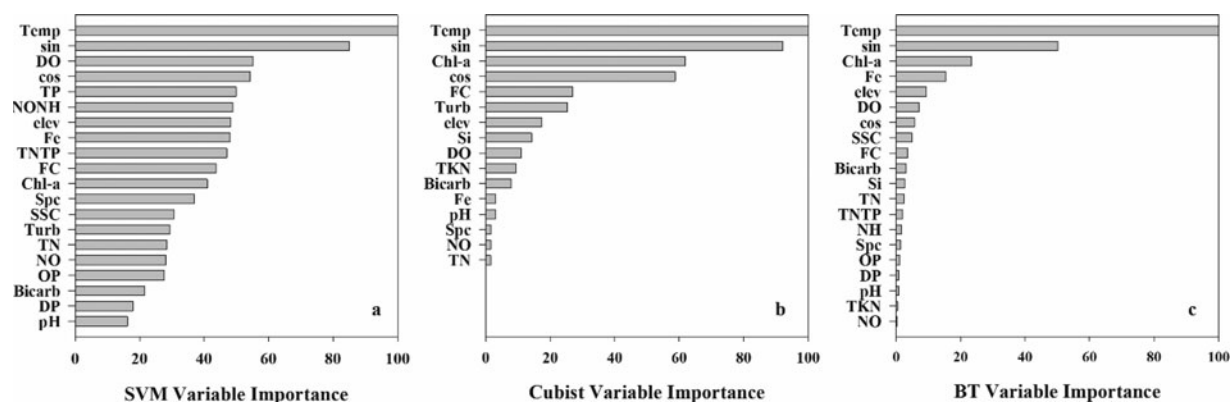
Figure 7. Top 20 (where applicable) most important variables for predicting microcystin concentration for (a) Support Vector Machine (SVM), (b) Cubist, and (c) Boosted Tree (BT) modeling approaches.
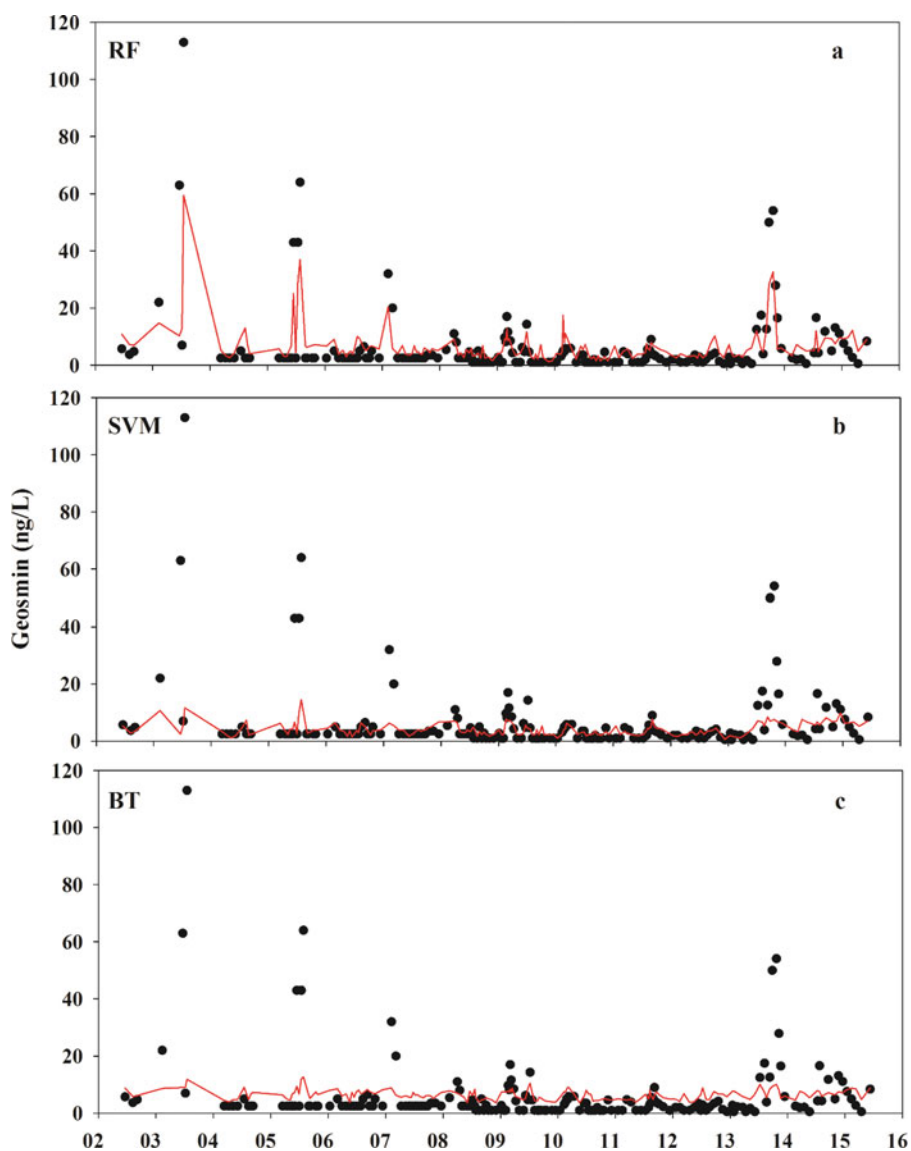

Figure 8. Observed (dots) and predicted (line) geosmin concentration from 2002 to 2015 in Cheney Reservoir using (a) Random Forest (RF), (b) Support Vector Machine (SVM), and (c) Boosted Tree (BT) modeling approaches.
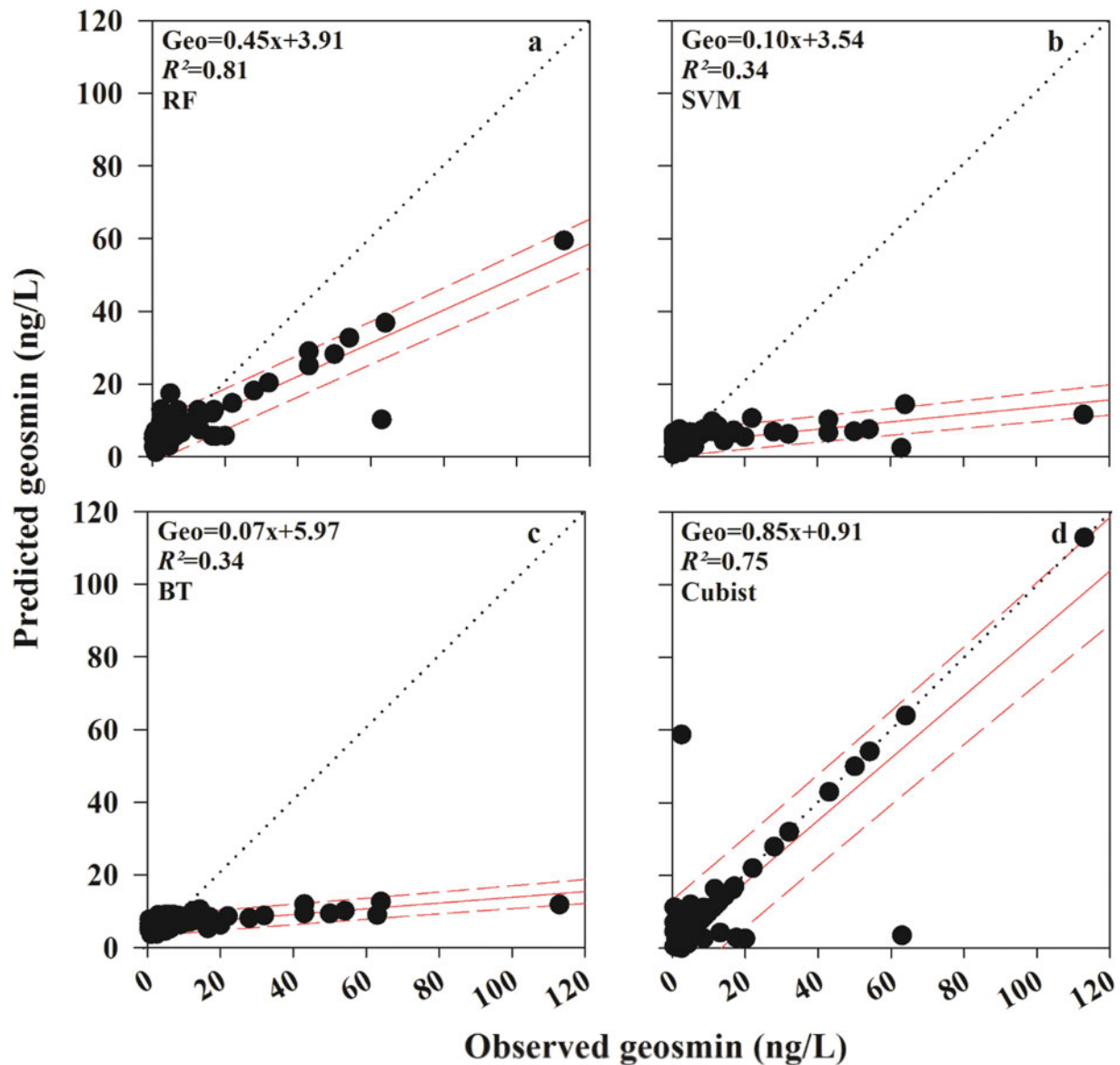
Figure 9. Observed compared to predicted geosmin concentration using (a) Random Forest (RF), (b) Support Vector Machine (SVM), (c) Boosted Tree (BT), and (d) Cubist modeling approaches. Solid line represents linear regression line, dashed lines represent 95% prediction intervals, and dotted line represents 1:1 line.

to effectively predict maxima geosmin concentrations compared to other modeling attempts; however, the reason the Cubist modeling technique could not accurately predict cyanobacterial abundance or microcystin concentration maxima is unknown. Therefore, although the Cubist modeling technique should be attempted in a variety of systems before widespread implementation, it has the potential to improve regression modeling efforts aimed at predicting maxima cyanobacterial metabolites.

In the earliest attempt to model cyanobacteria or their metabolites, Smith et al. (2002) found that geosmin concentration was linearly related to water-column Chl-*a* concentrations in a small ($n = 6$) dataset and explained 72% of the variation within the collected data. Christensen et al. (2006) was able develop a linear model to predict geosmin concentrations; log-transformed geosmin was related to log transformed turbidity and specific conductance and explained 71% of the variation in the data. Although these studies explained >70% of the variation in geosmin concentrations with linear models, Dzialowski et al. (2009) was unable to develop any significant regression models for geosmin in Cheney Reservoir. Additionally, Stone
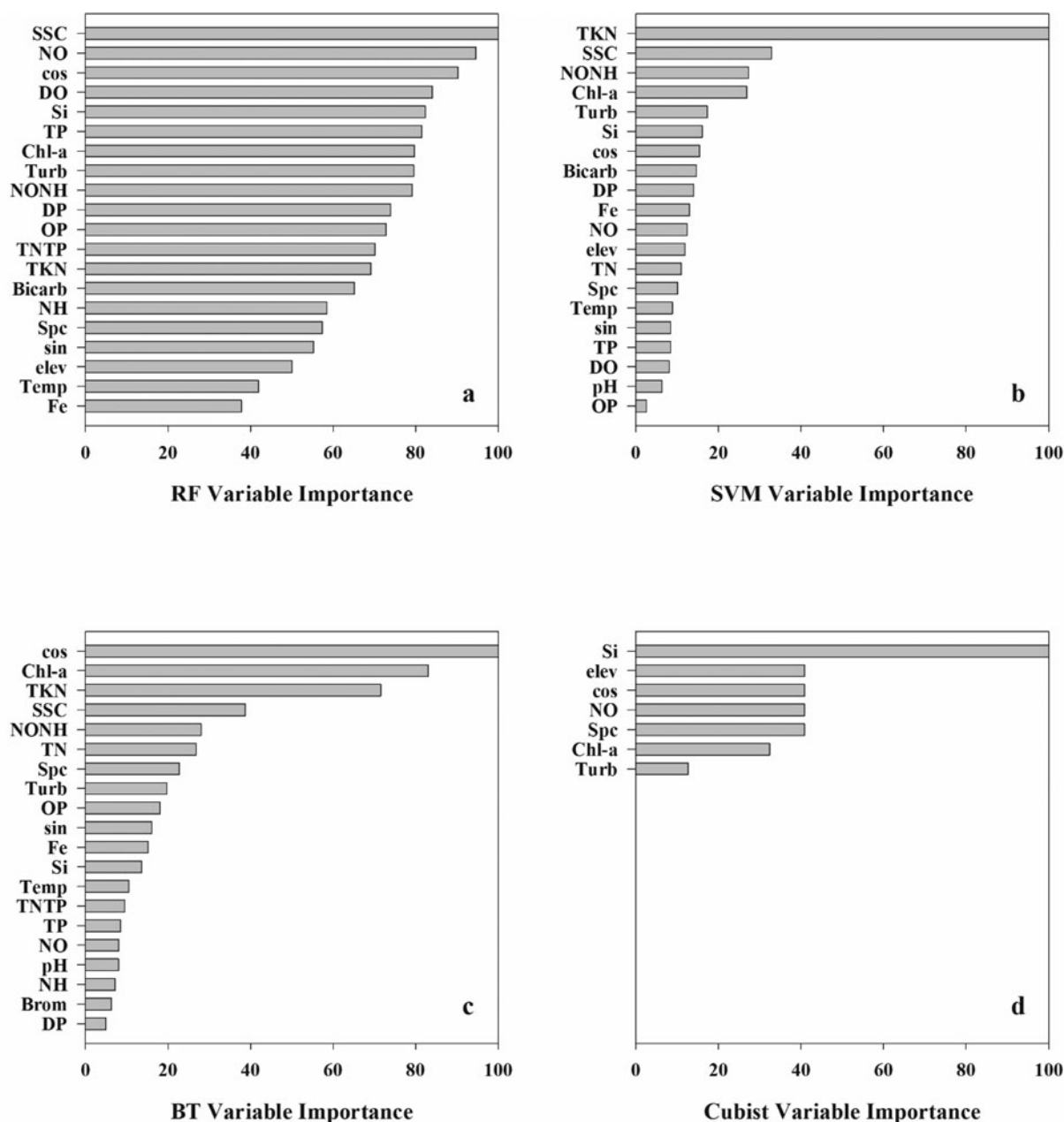
Figure 10. Top 20 (where applicable) most important variables for predicting geosmin concentration for (a) Random Forest (RF), (b) Support Vector Machine (SVM), (c) Boosted Tree (BT), and (d) Cubist modeling approaches.

et al. (2013) was only able to explain 21% of the variation in geosmin concentrations with a linear model based on turbidity and pH; a linear model for microcystin was only able to explain 46% of the variation in the data for Cheney Reservoir. Otten et al. (2016) also attempted linear regression models for cyanobacterial secondary metabolites and found that a model with 4 explanatory variables was only able to explain 51% of the variation in geosmin concentrations in 2013 and 2014. By contrast, a microcystin model was able to explain 82% of the variation; however, the y-intercept

in the model was -17 µg/L, indicating poor predictive power. Our study also attempted linear regression models and found results consistent with Dzialowski et al. (2009) and Stone et al. (2013). Note that the R^2 values of prior studies were likely overly optimistic because they were based on the computed fit model, whereas our study computed R^2 values only on the validation (test) dataset.

The range in predictive ability of, and explanatory variables in, developed models over time is likely due to several reasons. The earliest models developed

(e.g., Smith et al. 2002, Christensen et al. 2006) used relatively small datasets, and data were primarily collected in spring and summer months. Because data were collected over a relatively short period and were constrained to relatively warm months, these datasets did not capture the full spectrum of intra- and inter-annual variability within the reservoir. Thus, these models were not robust over time and indicate the importance of using long-term datasets that capture multi-year variability within explanatory and response variables when developing predictive models. Second, climate change has caused environmental variability to increase over time throughout the Midwest, which also may affect model outcomes (Committee on Extreme Weather Events and Climate Change Attribution et al. 2016; discussed later, but see http://www.ncdc.noaa.gov/extremes/cei/graph/wn/4/04-09 for data and graphics on increases in the frequency of extremes in 1-day precipitation for the Midwest from 1910 to 2015). Thus, results of predictive modeling efforts have likely been so varied in Cheney Reservoir because (1) earlier models did not successfully capture the environmental variation occurring in the reservoir and (2) environmental variability caused by climate change is likely increasing, and may pose challenges to developing predictive models.

Effects of climate change on predictive models

Predictive models rely on the recognition of pattern-based processes; however, these processes have been and will continue to be altered by climate change. As recently shown by the Committee on Extreme Weather Events and Climate Change Attribution et al. (2016), climate change will likely cause more frequent extreme heat and rainfall events, droughts, and severe storms, resulting in ever more extreme environmental variability. These climate change driven patterns have also been recognized to make existing models less predictive over time, which led Gail (2016) to term the upcoming era the "dark age" of predictive modeling. These extreme events have also been shown to stimulate cyanobacterial blooms in temperate, subtropical, and tropical regions of the world (Kosten et al. 2012, Jeppesen et al. 2015, Brasil et al. 2016). In Cheney Reservoir, several extreme weather events, possibly exacerbated by climate change, seemed to fuel cyanobacterial blooms and relatively high secondary metabolite concentrations while potentially reducing

the predictive ability of models on maxima concentrations. For example, droughts from August 2006–April 2007 and August 2011–July 2013 caused extremes in reservoir elevation (Supplemental Fig. S4) and reduced the mean depth of the reservoir by 1 and 2.5 m, respectively. In both cases, lowered reservoir mean depth created conditions that resulted in the highest annual cyanobacterial abundances (*Microcystis*-dominated) and microcystin concentrations observed in the reservoir (Fig. 2a–c and 5a–c).

Extreme rainfall events have also led to cyanobacterial related events in Cheney Reservoir. Following a severe drought from 2011 to July 2013, a heavy precipitation event in the watershed caused the ninth largest inflow event in the history of the reservoir (Supplemental Fig. S4; Stone et al. 2015). This large inflow event stimulated an *Anabaena* bloom that caused geosmin concentrations to exceed 50 ng/L (Otten et al. 2016). Our results are consistent with Reichwaldt and Ghadouani (2012), who hypothesized that heavy rainfall events may cause sharp increases in cyanobacterial metabolites. Although our predictive models recognized that reservoir elevation (cyanobacterial abundance and geosmin; Fig. 4a–d and 10d) and temperature (microcystin; Fig. 7a–c) were important explanatory variables for cyanobacteria and secondary metabolite prediction in Cheney Reservoir, drought and extreme rainfall events likely caused predictive models to underestimate observed maxima cyanobacteria and secondary metabolite concentrations. If the frequency of extreme drought and inflow events continue to increase, we hypothesize that larger, more frequent cyanobacterial blooms will occur in the future. Consequently, pattern-reliant modeling approaches for cyanobacterial abundance, microcystin, and geosmin based on historical datasets may not be robust over time (Gail 2016).

Important explanatory variables for predicting cyanobacteria and cyanobacterial metabolites

Explanatory variables identified as important in Cheney Reservoir were consistent with those shown to be predictive of cyanobacteria in other studies. Similar to Cheney Reservoir, Taranu et al. (2012) found that water column TP, water temperature, and seasonality (similar to the sin and cos predictor variables) were predictive of cyanobacterial blooms in polymictic Canadian lakes, indicating that these factors likely

promote cyanobacterial blooms in well-mixed systems regardless of waterbody location (Fig. 4a–d). Beaulieu et al. (2013) and Millie et al. (2014) found that water column nutrient concentrations (TN and TP), water temperature, and chlorophyll were important factors in predicting cyanobacteria in 1147 US lakes and Lake Erie, respectively. By contrast, few studies aimed at predicting cyanobacterial blooms in the current literature identified reservoir elevation as an important predictor variable (but see Francy et al. 2016; Fig. 4a–d). Although not examined here, water column stability and spring P loads were also identified as an important explanatory variables for predicting cyanobacterial blooms (Wagner and Adrian 2009, Millie et al. 2014, Bertani et al. 2016).

Although Cheney Reservoir rarely stratifies, water column stability could be evaluated in future studies as changing environmental conditions may change patterns of stratification in the reservoir. Additionally, including spring P loading in predictive models was shown to substantially increase forecast accuracy of summer cyanobacterial blooms in Lake Erie (Bertani et al. 2016). Because spring nutrients loads are increased by heavy rainfall events within the lake watershed, and large inflow events seem to stimulate cyanobacterial related events in Cheney Reservoir (Otten et al. 2016), future studies on Cheney Reservoir may improve predictive models by including spring nutrient loading. Therefore, variables like reservoir elevation, water column stability, and spring nutrient loading may be important variables to consider in addition to water column nutrient concentrations and water temperature for cyanobacterial prediction in drinking water reservoirs.

Water temperature, nutrients, and nutrient ratios have been shown to be predictive of microcystin and geosmin concentrations. For instance, multiple studies have found that elevated water temperatures are predictive of, and favor, elevated microcystin concentrations (Davis et al. 2009, Dziallas and Grossart 2011, Joung et al. 2011, Beaver et al. 2014). Additionally, Jacoby et al. (2015) and Dzialowski et al. (2009) found relatively low TN:TP ratios were predictive of microcystin and geosmin concentrations, respectively, in Northwestern and Midwestern US reservoirs, respectively. In contrast to other studies in the literature, accurate prediction of maxima geosmin concentrations depended on silica and turbidity (i.e. light environment) in Cheney Reservoir (Fig. 10a, b, and d). Although Christensen

et al. (2006) showed that turbidity was a strong predictor variable for geosmin concentrations in Cheney Reservoir, few other studies have indicated that silica or turbidity are important predictors for geosmin. Thus, although some factors (e.g., nutrients, Chl-*a*, and water temperature) identified here seem to be predictive of cyanobacteria and their metabolites throughout North America, including Cheney Reservoir, other factors (e.g., reservoir elevation, suspended solid concentration/turbidity, and silica; Fig. 4a–d, 7a–c, and 10a–d) may be specific to Cheney Reservoir.

Given that climate change likely will cause predictive models developed using historical datasets to underperform, further research using recent analytical and technological advances are needed to accurately predict cyanobacterial blooms. For example, the incorporation of analytical techniques such as quantitative polymerase chain reaction (qPCR) have been shown to significantly improve data inputs for cyanobacterial abundance and cyanobacterial secondary metabolite models (Francy et al. 2016, Otten et al. 2016). Advances in sensor technology could also improve modeling efforts. Specifically, sensors that can accurately measure cyanobacteria and other algal taxa in real-time will help reservoir managers better understand the factors that lead to cyanobacterial or other harmful algal blooms and allow for real-time prediction of events. Additionally, real-time measurement of N species and P concentrations will allow (1) more frequent measurements of nutrient concentrations and (2) the development of real-time cyanobacterial management, which will ultimately lessen the reliance on predictive modeling. Programs currently using discrete sampling to monitor cyanobacteria and their metabolites may need to consider continuous real-time monitoring using new advanced sensor technologies to provide public and private stakeholders more accurate predictions of cyanobacterial related events. Therefore, although we may be entering a "dark age" (sensu Gail 2016) of predicting environmental biotic variables, emerging analytical and technological advances could be used to combat historical pattern-based information made erroneous by climate change.

Acknowledgments

References

Ahn C-Y, Oh H-M, Park Y-S. 2011. Evaluation of environmental factors on cyanobacterial bloom in eutrophic reservoir using artificial neural networks. J Phycol. 47:495–504.

Beaulieu M, Pick F, Gregory-Eaves I. 2013. Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes dataset. Limnol Oceanogr. 58:1736–1746.

Beaulieu M, Pick F, Palmer M, Watson S, Winter J, Zurawell R, Gregory-Eaves I. 2014. Comparing predictive cyanobacterial models from temperate regions. Can J Fish Aquat Sci. 71:1830–1839.

Beaver JR, Manis EE, Loftin KA, Graham JL, Pollard AI, Mitchell RM. 2014. Land use patterns, ecoregion, and microcystin relationships in U.S. lakes and reservoirs: a preliminary evaluation. Harmful Algae. 36:57–62.

Bertani I, Obenour DR, Steger CE, Stow CA, Gronewold AD, Scavia D. 2016. Probabilistically assessing the role of nutrient loading in harmful algal bloom formation in western Lake Erie. J Great Lakes Res. [cited 2016 Jul 25]. Available from: http://www.sciencedirect.com/science/article/pii/S0380133016300405

Brasil J, Attayde JL, Vasconcelos FR, Dantas DDF, Huszar VLM. 2016. Drought-induced water-level reduction favors cyanobacteria blooms in tropical shallow lakes. Hydrobiologia. 770:145–164.

Chorus I, Bartram J, editors. 1999. Toxic Cyanobacteria in water: a guide to their public health consequences, monitoring and management. 1st ed. London (UK): E & FN Spoon.

Christensen VG, Graham JL, Milligan CR, Pope LM, Ziegler AC. 2006. Water quality and relation to taste-and-odor compounds in North Fork Ninnescah River and Cheney Reservoir, south-central Kansas, 1997–2003. Washington (DC): US Department of the Interior, US Geological Survey Scientific Investigations Report 2006–5095. Available from: http://pubs.usgs.gov/sir/2006/5095/

Committee on Extreme Weather Events and Climate Change Attribution, Board on Atmospheric Sciences and Climate, Division on Earth and Life Studies, National Academies of Sciences, Engineering, and Medicine. 2016. Attribution of extreme weather events in the context of climate change. Washington (DC): National Academies Press; [cited 2016 May 24]. Available from: http://www.nap.edu/catalog/21852

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.

Dietrich AM. 2006. Aesthetic issues for drinking water. J Water Health. 4(Suppl 1):11–16.

Dunlap CR, Sklenar K, Blake L. 2015. A costly endeavor: addressing algae problems in a water supply. J Am Water Works As. 107:E255–E262.

Dziallas C, Grossart H-P. 2011. Increasing oxygen radicals and water temperature select for toxic Microcystis sp. PLoS ONE. 6:e25569.

Dzialowski AR, Smith VH, Huggins DG, deNoyelles F, Lim N-C, Baker DS, Beury JH. 2009. Development of predictive models for geosmin-related taste and odor in Kansas, USA, drinking water reservoirs. Water Res. 43: 2829–2840.

Francy DS, Brady AMG, Ecker CD, Graham JL, Stelzer EA, Struffolino P, Dwyer, DF, Loftin KA. 2016. Estimating microcystin levels at recreational sites in western Lake Erie and Ohio. Harmful Algae 58:23–34.

Gail WB. 2016. A new dark age looms. New York Times [Internet]. [cited 2016 May 24]. Available from: http://www.nytimes.com/2016/04/19/opinion/a-new-dark-age-looms.html

Graham JL, Harris TD. 2016. Phytoplankton data for Cheney Reservoir near Cheney, Kansas, June 2001 through November 2015: US Geological Survey data release. Available from: http://dx.doi.org/10.5066/F71N7Z7V.

Harris T, Smith V, Graham J, Van de Waal D, Tedesco L, Clercin N. 2016. Combined effects of nitrogen to phosphorus and nitrate to ammonia ratios on cyanobacterial metabolite concentrations in eutrophic Midwestern USA reservoirs. Inland Waters. 6:199–210.

Harris TD, Smith VH. 2016. Do persistent organic pollutants stimulate cyanobacterial blooms? Inland Waters. 6:124–130.

Helsel DR, Hirsch RM. 2002. Statistical methods in water resources. Techniques of Water Resources Investigations, Book 4, Chapter A3. US Dept. of the Interior, US Geological Survey.

Jacoby JM, Burghdoff M, Williams G, Read L, Hardy J. 2015. Dominant factors associated with microcystins in nine midlatitude, maritime lakes. Inland Waters. 5: 187–202.

Jeppesen E, Brucet S, Naselli-Flores L, Papastergiadou E, Stefanidis K, Nõges T, Nõges P, Attayde JL, Zohary T, Coppens J, et al. 2015. Ecological impacts of global warming and water abstraction on lakes and reservoirs due to changes in water level and related changes in salinity. Hydrobiologia. 750:201–227.

Joung S-H, Oh H-M, Ko S-R, Ahn C-Y. 2011. Correlations between environmental factors and toxic and non-toxic Microcystis dynamics during bloom in Daechung Reservoir, Korea. Harmful Algae. 10:188–193.

Jüttner F, Watson SB. 2007. Biochemical and ecological control of geosmin and 2-methylisoborneol in source waters. Appl Environ Microbiol. 73:4395–4406.

Kosten S, Huszar VLM, Bécares E, Costa LS, van Donk E, Hansson L-A, Jeppesen E, Kruk C, Lacerot G, Mazzeo N, et al. 2012. Warmer climates boost cyanobacterial dominance in shallow lakes. Glob Change Biol. 18:118–126.

Kuhn M, Johnson K. 2013. Applied predictive modeling. New York (NY): Springer; [cited 2015 Sep 16]. Available from: http://link.springer.com/10.1007/978-1-4614-6849-3

McNabb CD. 1960. Enumeration of freshwater phytoplankton concentrated on the membrane filter. Limnol Oceanogr. 5:57–61.

Millie DF, Weckman GR, Fahnenstiel GL, Carrick HJ, Ardjmand E, Young WA, Sayers MJ, Shuchman RA. 2014. Using artificial intelligence for CyanoHAB niche modeling: discovery and visualization of *Microcystis*–environmental associations within western Lake Erie. Can J Fish Aquat Sci. 71:1642–1654.

Otten TG, Graham JL, Harris TD, Dreher TW. 2016. Elucidation of taste-and-odor producing bacteria and toxigenic cyanobacteria by shotgun metagenomics in a Midwestern drinking water supply reservoir. Appl Environ Microbiol. AEM.01334-16. 82:5410–5420.

Otten TG, Paerl HW. 2015. Health effects of toxic cyanobacteria in US drinking and recreational waters: our current understanding and proposed direction. Curr Environ Health Rep. p. 1–10.

Paerl HW. 2014. Mitigating harmful cyanobacterial blooms in a human- and climatically-impacted world. Life. 4:988–1012.

Paerl HW, Huisman J. 2008. Blooms like it hot. Science. 320:57–58.

Paerl HW, Otten TG. 2013. Harmful cyanobacterial blooms: causes, consequences, and controls. Microb Ecol. 65:995–1010.

Parinet J, Rodriguez MJ, Sérodes J-B. 2013. Modelling geosmin concentrations in three sources of raw water in Quebec, Canada. Environ Monit Assess. 185:95–111.

Pearson LA, Dittmann E, Mazmouz R, Ongley SE, D'Agostino PM, Neilan BA. 2016. The genetics, biosynthesis and regulation of toxic specialized metabolites of cyanobacteria. Harmful Algae. 54:98–111.

Recknagel F, Talib A, van der Molen D. 2006. Phytoplankton community dynamics of two adjacent Dutch lakes in response to seasons and eutrophication control unravelled by non-supervised artificial neural networks. Ecol Inform. 1:277–285.

Reichwaldt ES, Ghadouani A. 2012. Effects of rainfall patterns on toxic cyanobacterial blooms in a changing climate: between simplistic scenarios and complex dynamics. Water Res. 46:1372–1393.

Smith VH, Sieber-Denllinger J, DeNoyelles Jr. F, Campbell S, Pan S, Randtke S. 2002. Managing taste and odor problems in a eutrophic drinking water reservoir. Lake Reserv Manage. 18:319–323.

Stone ML, Graham JL, Gatotho J. 2013. Model documentation for relations between continuous real-time and discrete water-quality constituents in Cheney Reservoir near Cheney, Kansas, 2001–2009. Washington (DC): US Department of the Interior, US Geological Survey Scientific Investigations Report 2013–1123. Available from: http://pubs.usgs.gov/of/2013/1123/

Stone ML, Juracek KE, Graham JL, Foster GM. 2015. Quantifying suspended sediment loads delivered to Cheney Reservoir, Kansas: temporal patterns and management implications. J Soil Water Conserv. 70:91–100.

Taranu ZE, Zurawell RW, Pick F, Gregory-Eaves I. 2012. Predicting cyanobacterial dynamics in the face of global change: the importance of scale and environmental context. Glob Change Biol. 18:3477–3490.

Wagner C, Adrian R. 2009. Cyanobacteria dominance: quantifying the effects of climate change. Limnol Oceanogr. 54:2460–2468.

Xie Z, Lou I, Ung WK, Mok KM. 2012. Freshwater algal bloom prediction by support vector machine in Macau storage reservoirs. Math Probl Eng. 2012:e397473.

Ziegler AC, Hansen CV, Finn DA. 2010. Water quality in the Equus Beds aquifer and the Little Arkansas River before implementation of large-scale artificial recharge, south-central Kansas, 1995–2005. Washington (DC): US Department of the Interior, US Geological Survey Scientific Investigations Report 2010–5023.

Zimmerman LR, Ziegler AC, Thurman EM. 2002. Method of analysis and quality-assurance practices by US Geological Survey organic geochemistry research group-determination of geosmin and methylisoborneol in water using solid-phase microextraction and gas chromatography/mass spectrometry. US Department of the Interior, US Geological Survey Scientific Investigations Report 2002–337.