Freshwater Biology

Freshwater Biology (2015) **60,** 1901–1916 doi:10.1111/fwb.12620

Deriving nutrient targets to prevent excessive cyanobacterial densities in U.S. lakes and reservoirs

LESTER L. YUAN AND AMINA I. POLLARD
Office of Water, U.S. Environmental Protection Agency, Washington, DC, U.S.A

SUMMARY

1. High densities of cyanobacteria can interfere with the use of lakes and reservoirs for recreation and as sources for drinking water, and one approach for reducing the amount of cyanobacteria is to reduce nutrient concentrations in the waterbody.
2. An approach is described for deriving numeric targets for concentrations of total phosphorus (TP) and total nitrogen (TN) that are associated with a pre-specified probability of cyanobacterial biovolume that exceeds the recommended World Health Organization thresholds for recreation in the water. The analysis consisted of two phases. First, a divisive tree algorithm was used to identify groups of lakes in which the relationship between nutrients and cyanobacterial biovolume was similar. Second, hierarchical Bayesian models were used to estimate relationships between cyanobacterial biovolume, TP and TN, while partitioning the observed variance in biovolume into components associated with sampling variability, temporal variability, and among-lake differences.
3. The final model accounted for 91% of the variance in cyanobacterial biovolume among different lakes and was used to identify nutrient concentrations that maintain a low probability of excessively high cyanobacterial biovolumes.
4. When no classes of lakes were specified and the relationship between cyanobacterial biovolume and nutrient concentrations was modelled using a national data set, mean targets of 87 and 1100 μg L^{-1} were derived for TP and TN, respectively, to maintain cyanobacterial biovolume below moderate risk levels as defined by the World Health Organization. After classification, mean nutrient targets in lakes that were found to be most susceptible to high biovolumes of cyanobacteria (i.e. deep lakes) were 61 and 800 μg L^{-1} for TP and TN, while higher nutrient thresholds were observed for other classes of lakes.

Keywords: cyanobacteria, nutrients, hierarchical Bayesian models, classification

Introduction

Cyanobacteria are important natural components of lake biological communities, but, under certain environmental conditions, their abundance can increase to levels that interfere with the use of the waterbody for recreation and as a source of drinking water. During these periods of high abundance and biovolume, cyanobacteria can form unsightly and odorous surface scums and substantially increase concentrations of cyanotoxins in the water. These toxins, in turn, restrict the use of the lake as both a source of drinking water and for recreation. One well-known health advisory has established cyanobacterial abundances of 20 000 and 100 000 cells mL^{-1} as the thresholds at which low and moderate human health risks exist for recreational users of a waterbody (WHO 2003). Other epidemiological studies have identified similar or lower thresholds of concern (Pilotto *et al.*, 1997; Stewart *et al.*, 2006; Lévesque *et al.*, 2014).

Many different environmental factors have been associated with increased amounts of cyanobacteria and increased occurrence of algal blooms (Paerl & Otten, 2013). Among anthropogenic pollutants, increased nutrient concentrations (i.e. nitrogen and phosphorus) have been identified as one of the main causes of increased

Correspondence: Lester L. Yuan, Office of Water, U.S. Environmental Protection Agency, Mail code 4304T Washington, DC 20460, U.S.A.
E-mail: yuan.lester@epa.gov

densities of cyanobacteria (Downing, Watson & McCauley, 2001; Ahn, Oh & Park, 2011). Other environmental factors that have been observed to increase the density of cyanobacteria include increased water temperature (Paerl & Huisman, 2008), alkalinity (Carvalho *et al*., 2011), water colour (Carvalho *et al*., 2011), light intensity (Huisman *et al*., 1999) and stratification strength (Wagner & Adrian, 2009). Lack of wind stress and the resulting increased stability of the water column have also been noted as conditions that are favourable for the formation of cyanobacterial blooms (Wynne *et al*., 2010).

Of the aforementioned, concentrations of nitrogen and phosphorus in a waterbody typically are the environmental factors that are more controllable by management actions, and hence, efforts to reduce the amount of cyanobacteria often focus on reductions in nutrient loads (Paerl & Otten, 2013). To facilitate these efforts, target concentrations for these nutrients, below which high densities of cyanobacteria occur with an acceptably low frequency, would be particularly useful. From a broader perspective, an empirical relationship relating ambient nutrient concentrations to the likelihood of high densities of cyanobacteria (e.g. >100 000 cells mL^{-1}) would allow water quality managers to specify nutrient concentrations that confer protection of source water and recreational activities.

Here, we ask whether analysis of large-scale synoptic data can yield nutrient thresholds that are useful for managing cyanobacteria. To this end, we describe an analysis of a national data set to estimate empirical relationships between the concentrations of phosphorus, nitrogen and cyanobacterial biovolume in lakes and reservoirs of the United States. Because many other environmental factors in addition to nutrient concentrations can potentially influence cyanobacteria, we describe an approach for classifying lakes to account for the effects of these other factors and to increase the precision of estimated relationships between nutrients and cyanobacteria. Relationships estimated between cyanobacteria and nutrients within each of the classes are then used to derive management targets for nitrogen and phosphorus.

Methods

Data

Data used for this analysis were collected by the U.S. Environmental Protection Agency's National Lake Assessment (NLA) in the summer (May–September) of 2012 (US EPA 2011). Lakes >1 ha were selected from the contiguous United States using a combination of a stratified random sampling design and a small number of hand-picked lakes and reservoirs. At each of the sampled lakes, an extensive suite of abiotic and biological variables was measured, but here we only provide sampling details regarding the parameters used in the present analysis.

At each lake, two sampling locations were established: one in open water at the deepest point of each lake (up to a maximum depth of 50 m) or in the midpoint of reservoirs, and one littoral zone sampling location approximately 10 m out from a randomly selected point on the shoreline. At the open water site, a vertical, depth-integrated methodology was used to collect a water sample from the photic zone of the lake (to a maximum depth of 2 m). Multiple sample draws were combined in a rinsed, 4-L cubitainer. When full, the cubitainer was gently inverted to mix the water, and a subsample was poured off to obtain a water chemistry sample. This subsample was placed on ice and shipped overnight to the Willamette Research Station in Corvallis, Oregon, U.S.A. which quantified total nitrogen (TN), total phosphorus (TP), true colour and acid neutralising capacity at prespecified levels of precision and accuracy (US EPA 2012). A second subsample for characterising the phytoplankton community was poured off and preserved with a small amount of Lugol's solution. At the littoral zone site, a grab water sample was collected 0.3 m below the surface at a depth of at least 1 m and also preserved with a small amount of Lugol's solution.

Cyanobacterial biovolume was quantified from the field samples in the laboratory. Samples collected from both open water and littoral zone locations where examined by taxonomists, who identified at least 400 natural algal units to species under 1000× magnification. Cyanobacteria were aggregated, and abundance was calculated as cells per mL. In each sample, the dimensions of the taxa that accounted for the largest proportions of the observed assemblage were measured and used to estimate biovolume. Biovolumes of the most abundant taxa were based on the average of measurements from at least 10 individuals, while biovolumes for less abundant taxa were based on somewhat fewer measurements. The overall biovolume of cyanobacteria was reported as $\mu m^3\ mL^{-1}$ (US EPA 2012), which we converted to $mm^3\ L^{-1}$. We focussed our analysis on cyanobacterial biovolume as this measure most accurately reflected the cyanobacterial density in the water column.

Physical characteristics of lake were estimated from mapped data (NHD+ version 2). These characteristics included lake surface area (Area), geographic location

(latitude and longitude), altitude (Alt) and lake perimeter. From these characteristics, we calculated the following composite variables: (i) the shoreline development, which is defined as the ratio between the perimeter of the lake and the perimeter of circle with the same area as the lake and characterises the geometric complexity of the lake shore, and (ii) the lake geometry ratio, which is defined as $Area^{0.25}$/Depth, or the ratio between fetch and lake maximum depth, and has been shown to differentiate lakes that stratify seasonally (low values of the geometry ratio) from lakes that are polymictic (Gorham & Boyce, 1989; Stefan *et al.*, 1996).

Variables quantifying the mean annual precipitation (Precip) and maximum monthly average air temperature (Temp) at the lake location were extracted from 30-year averaged climatic data (Daly *et al.*, 2008). Lakes were also noted as being man-made (i.e. reservoirs) or natural.

With the exception of cyanobacterial biovolume, TP and TN, the measurements described above were selected as candidate classification variables because of their potential influence on the amount of cyanobacteria. For example, as noted earlier, the amount of cyanobacteria has been observed to increase with increased stratification strength (Wagner & Adrian, 2009), and so we included lake geometry ratio as a candidate classification variable. Similarly, many studies have shown relationships between temperature and the amount of cyanobacteria (Paerl & Huisman, 2008; Beaulieu, Pick & Gregory-Eaves, 2013), so we included maximum monthly air temperature at the lake location (Temp) as a candidate classification variable. Similar rationales exist for the other selected classification variables.

Statistical analysis

Statistical analyses to estimate relationships between TN, TP and cyanobacterial biovolume consisted of two phases. First, TREED regression was used to classify lakes into groups in which the relationships between increased nutrient concentrations and cyanobacterial biovolume were similar (see below). Second, hierarchical Bayesian models were used to partition the variance in cyanobacterial biovolume into sampling variability, temporal variability and among-lake components.

Observed values of cyanobacterial biovolume were first Box–Cox-transformed (with the power parameter, lambda, set to 0.05) such that repeat samples of biovolume at each lake approximated a normal distribution. Concentrations of TN and TP were log-transformed to reduce the skewness of their distributions. Cyanobacterial biovolume was also corrected *a priori* for small differences attributed to the identity of the taxonomist that processed the sample (see Supplemental Information). Candidate classification variables were not transformed because the TREED algorithm is not sensitive to the distribution of each variable.

TREED analysis

To increase the precision of the estimated relationships between nutrient concentrations and cyanobacterial biovolume, we grouped lakes by applying a variant of classification and regression trees (Breiman *et al.*, 1984) known as TREED analysis (Alexander & Grimshaw, 1996). In classification and regression trees, the data set is partitioned in a stepwise manner to minimise the residual deviance about mean values of the response variable in each group. In contrast, in TREED analysis, the data set is partitioned to minimise the residual deviance about a functional relationship within each group. In the present analysis, each end node of the tree consisted of a multiple linear regression model that modelled cyanobacterial biovolume as a function of TN and TP, and the data set was partitioned to minimise the residual deviance in these estimated relationships across all end nodes. By using a function in the end node rather than a single value, the classification tree can be less complex, and therefore more interpretable (Alexander & Grimshaw, 1996; Yuan & Pollard, 2014).

Selection of classification variables and specifying the classification tree proceeded as follows. We first computed mean values of TN, TP and cyanobacterial biovolume for all available samples from each lake because in this phase of the analysis, we were only interested in accounting for sources of variance that varied among different lakes. The classification tree was then built sequentially. For each level of the tree, each candidate classification variable was considered in turn, and approximately 50 values spanning the observed range of that variable were selected as possible splitting values. With categorical variables (e.g. lake origin), the variable itself explicitly defined discrete categories so no splitting values were needed. The data set was divided into two groups based on each of the possible splitting values, and a multiple linear regression model relating cyanobacterial biovolume to TN and TP was fit, specifying group membership as a dummy variable and allowing different values of the regression coefficients for each group. The residual deviance of the model was then computed and retained. This procedure was repeated for each of the classification variables, and the combination

of classification variable and splitting value that yielded the greatest reduction in residual deviance was saved. Splitting the data set and building the classification tree continued recursively using this procedure until any further splits would have reduced the number of samples in a group to <100 lakes or until a pre-specified maximum number of lake groups was defined (see below). The minimum number of samples required for each lake group was established to ensure that a sufficient number of independent samples were available to reliably fit the regression relationship in each of the end nodes (Harrell, 2001), and sufficient samples were available to minimise the possible errors associated with estimating separate effects for TN and TP, which were correlated (Mason & Perreault, 1991).

To select the maximum number of lake groups to specify in the classification tree, we applied a 10-fold cross-validation procedure. Available data were randomly assigned to 10 equally sized partitions, and then each partition was sequentially held out as independent validation data. The classification tree and associated regression model were then calibrated using the remaining 90% of the data. This model was then used to predict cyanobacterial biovolume in both the calibration and held-out validation data, and root-mean-square (RMS) predictive errors were computed for both data sets. The process was repeated for each of the 10 partitions, yielding average estimates of the overall RMS calibration and validation error for the entire data set. We repeated the 10-fold cross-validation calculation for 25 different random assignments to the partitions and for trees with the maximum number of lake groups ranging from two to five. We expected both calibration and validation RMS error to decrease with increased numbers of lake groups, but also expected decreases in validation error to cease as the model became overfit (i.e. when too many lake groups were specified). The point at which performance on validation data no longer improved was selected as the maximum number of lake groups for subsequent modelling.

An enormous number of different trees are possible because of the number of candidate classification variables and the number of possible values at which splits can be specified for each of the variables. The stepwise, or 'greedy' algorithm described above only minimises the residual deviance at each level of the tree, an approach that is locally optimal, but may not yield the best global model (Chipman, George & McCulloch, 1998). After specifying the maximum number of lake groups, we explored the space of possible classification trees more broadly using a 'bootstrap umbrella of model parameters', or 'bumped' trees (Tibshirani & Knight, 1999; Yuan & Pollard, 2014a). More specifically, we fit 500 classification trees to bootstrap replicates of the data set. For each bumped tree, we fit the linear regression model relating TN and TP concentrations to the amount of cyanobacteria using group membership as a dummy variable. We then evaluated the resulting trees in terms of the degree to which each tree accounted for variability in the response variable. We selected one final tree that best accounted for variability in cyanobacterial biovolume to examine the structure and performance in greater detail. We also examined all of the classification variables selected in the five trees with the lowest residual deviance to identify variables that were most frequently identified and therefore, potentially more important.

Hierarchical Bayesian models

Like most biological abundance measurements, cyanobacterial biovolume varies strongly over time and space within any particular lake. Blooms of extremely high levels of cyanobacteria can appear and then disappear quickly because of changes in the environmental conditions and because of differences in the life cycles of different species of cyanobacteria (Pinckney *et al.*, 1998). Estimates of cyanobacterial biovolume can also vary strongly between different locations in a lake because of its patchy distribution and because of the inherent sampling variability associated with estimating an aggregate measure of cyanobacterial biovolume from a relatively small (i.e. 400 algal units) number of individuals. The sampling design of the NLA was synoptic, but data were available to directly estimate both the temporal and sampling variability in cyanobacterial densities. More specifically, as part of the NLA sampling design, 10% of sampled lakes were randomly selected and resampled on 1–2 different days at least 6 weeks apart, and these repeat visits provided data that could be used to estimate temporal variability. Furthermore, preliminary exploratory analysis indicated that cyanobacterial biovolume in samples collected at littoral sites did not differ systematically from samples collected at open water sites. That is, the location at which a sample was collected was not a statistically significant predictor of cyanobacterial biovolume. So, these two different samples collected on each lake visit were used to estimate the variability attributed to the combined effects of sampling and within-lake spatial variability (hereafter referred to only as sampling variability). We included both types of repeat samples in the data set and used

hierarchical Bayesian models to partition the overall variance of cyanobacterial biovolume into contributions from temporal, spatial and among-lake sources (Gelman & Hill, 2007).

We fit three hierarchical Bayesian models with different fixed effects. Our base model was a simple intercept-only model, in which cyanobacterial biovolume was modelled as a constant term plus a random effect associated with the specific lake and a random effect associated with a particular lake visit. This model can be written as follows:

$$y_i = a + b_{j[i]} + c_{k[i]} + r_i$$

where y_i is cyanobacterial biovolume in sample i, a is the overall mean value of y_i, $b_{j[i]}$ is a normally distributed random effect of lake j, with each sample i assigned to a particular lake j, $c_{k[i]}$ is a normally distributed random effect of lake visit k, with each sample i assigned to a particular lake visit, k and where r_i is the random contribution of sampling variance to the observed y_i. The variables, b_j, c_k and r_i are normally distributed with $b_j \sim N(0, s^2_{\text{among}})$, $c_k \sim N(0, s^2_{\text{time}})$ and $r_i \sim N(0, s^2_{\text{sample}})$.This base model partitioned the overall variance of cyanobacterial biovolume into among-lake (s_{among}), temporal (s_{time}) and sampling (s_{sample}) variance components.

In our second model, we introduced nutrient concentrations, which accounted for a portion of the among-lake variance:

$$y_i = a + d_1\text{TP}_{j[i]} + d_2\text{TN}_{j[i]} + b_{j[i]} + c_{k[i]} + r_i$$

where the effects of differences in nutrient concentrations are modelled with the regression coefficients, d_1 and d_2, and the mean TN and TP concentrations over all samples available in lake j. TN and TP were also standardised by subtracting their overall mean value and dividing by their standard deviations prior to using them in the model. This standardisation improved the convergence of the hierarchical Bayesian models and facilitated interpretation of the relative effects of TN and TP (Gelman & Hill, 2007).

Third, we added the effect of lake classification, as specified by TREED regression:

$$y_i = a_{n[i]} + d_{1,n[i]}\,\text{TP}_{j[i]} + d_{2,n[i]}\,\text{TN}_{j[i]} + b_{j[i]} + c_{k[i]} + r_i$$

where each group of lakes was allowed different regression coefficients, a_n, $d_{1,n}$ and $d_{2,n}$, indexed by n for different groups of lakes.

After fitting the models, we compared the magnitude of among-lake variance across the different models to quantify the degree to which each model accounted for differences among lakes in the data. We calculated the proportion of among-lake variance explained by the 2nd and 3rd models (i.e. models including nutrient concentrations as fixed effects) as the difference between 1 and the ratio between the among-lake variance of each of these models and the among-lake variance of the intercept-only model. This value is comparable to a conventional R^2 statistic, except that it excludes the contribution of sampling and temporal variance (which models of among-lake differences cannot be expected to explain).

For the 2nd and 3rd models, we also estimated management target values for nutrient concentrations as follows. We first converted WHO management thresholds expressed as cyanobacterial cells per unit volume to thresholds expressed in terms cyanobacterial biovolume per unit volume using a major axis regression between cyanobacterial abundance and biovolume. Then, for each model and each class of lakes (in the case of the 3rd model), we calculated the mean cyanobacterial biovolume that yielded a 10% chance of exceeding each management threshold. The selection of a 10% chance is only illustrative in this analysis, but one would generally expect that a low probability would be selected when deriving a management threshold. The 10% chance of exceeding the cyanobacterial threshold corresponds to the 90th percentile of the distribution of possible cyanobacterial biovolumes within a lake over the course of a sampling season. This distribution was computed from the estimated mean biovolume of cyanobacteria in each lake and the temporal variance estimated by the hierarchical Bayesian model. Sampling variability was not considered in this computation because the random effects of sampling variability generally should not influence management decisions for a particular lake.

Once the targeted mean cyanobacterial biovolume was computed, the relationships between mean TP, mean TN and cyanobacterial biovolume were used to estimate target concentrations for TP and TN. Because the effects of TN and TP were modelled simultaneously, an infinite combination of TP and TN concentrations could possibly be selected to achieve the targeted amount of cyanobacteria. We selected unique target values from this set of possibilities by identifying the combination of TP and TN that corresponded with a major axis regression through TP and TN. These values represent the mean expected combination of TP and TN, given the observations of TP and TN in the data set, or within a group of lakes (Fig. 1).

A range of possible values was computed for each TP and TN target by combining the effects of uncertainty in the estimates of the regression parameters (i.e. traditional

confidence limits) and the effects of among-lake variability, as quantified by the value of s_{among} estimated from the hierarchical Bayesian models. Confidence limits quantify uncertainty in estimates of the mean relationships between TN, TP and cyanobacterial biovolume among all lakes, whereas among-lake variability represents differences in the relationships between TN, TP and cyanobacterial biovolume that one might observe in different individual lakes within a lake group (or within the entire data set in the case of the national model). Hence, selection of the mean value of the among-lake distribution (i.e. the mean reported TN and TP targets) would achieve the desired frequency of excessive cyanobacterial biovolumes in approximately half of the lakes within the lake group. Selection of a lower value within the distribution would yield desired frequencies in a larger proportion of lakes. Since the full distributions of sampled values were available for all parameters estimated in the hierarchical Bayesian models, calculating the combined effects of among-lake variability and mean confidence limits was straightforward.

Finally, generalised additive models (GAM) (Wood & Augustin, 2002) were used to explore functional relationships between TN, TP and cyanobacterial biovolume that were more complex than the linear relationships used in the hierarchical Bayesian models.

All statistical calculations were performed with R (R Core Team 2013). Regression trees were fit by adapting scripts provided in the partykit library (Hothorn and Zeileis 2013). Parameters for the hierarchical Bayesian models were estimated using the rstan library (http://mc-stan.org/rstan.html).

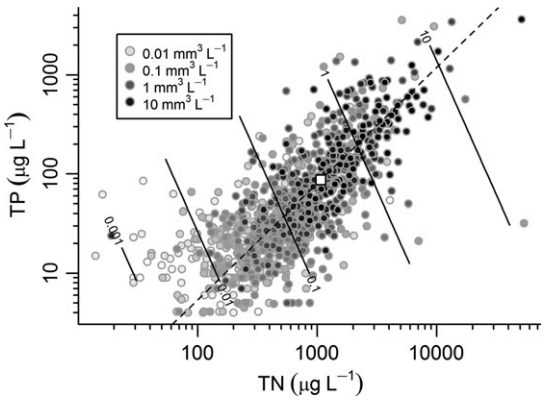

Fig. 1 Example showing calculation of total phosphorus (TP) and total nitrogen (TN) targets using predicted cyanobacterial biovolume and major axis regression relationship between TN and TP. Contour lines: predicted mean cyanobacterial biovolume, dashed line: estimated major axis regression between TN and TP, circles show the observed values of TN and TP with the observed values of cyanobacterial biovolume indicated by the colour of the circle, white square: example of unique TN and TP target for maintaining mean cyanobacterial biovolume at 0.29 mm^3 L^{-1}.

Results

A total of 2418 samples with complete environmental and cyanobacterial biovolume measurements were available from the NLA data set. These samples were collected from 1109 distinct sites. Of these sites, 100 were sampled on at least two different days during the sampling season.

A total of 58 different cyanobacteria genera were observed in the data set. *Chroococcus* and *Anabaena* were the most commonly observed genera (Table 1). *Chroococcus*, *Aphanocapsa* and *Planktolyngbya* each occurred in densities exceeding 100 000 cells mL^{-1} in more than 90 samples in the data set. Cyanobacterial genera that are known to produce toxins such as *Anabaena*, *Aphanizomenon*, and *Cylindrospermopsis* were also commonly observed in many samples and in high densities.

Median measured biovolume for cyanobacterial cells was 16 µm^3 per cell, but varied greatly among different cyanobacterial species. Management thresholds for cyanobacterial abundance have been specified as 20 000 and 100 000 cells mL^{-1} (WHO 2003), and based on the relationship between cyanobacterial abundance and biovolume, we converted these thresholds to biovolumes of 0.45 and 2.3 mm^3 L^{-1} (Fig. 2).

Median cyanobacterial biovolume in different lakes in the final data set was 0.18 mm^3 L^{-1}, and exceeded 2.3 mm^3 L^{-1} in 17% of the samples (Table 2). A total of 607 of the sampled lakes were designated as man-made reservoirs, while the remaining 502 were designated as natural in origin.

Table 1 Frequently occurring cyanobacteria genera. Genera occurring in >10% of samples shown or exceeding 100 000 cells mL^{-1} in at least 5 samples

Genus	Proportion of samples	Number of samples in which >100 000 cells mL^{-1}
Chroococcus	0.65	104
Anabaena	0.47	18
Aphanocapsa	0.41	95
Aphanizomenon	0.32	52
Pseudanabaena	0.29	41
Merismopedia	0.27	20
Planktolynbya	0.24	96
Cylindrospermopsis	0.16	62
Synechocystis	0.11	4
Planktothrix	0.09	41
Phormidium	0.06	19
Microcystis	0.05	5

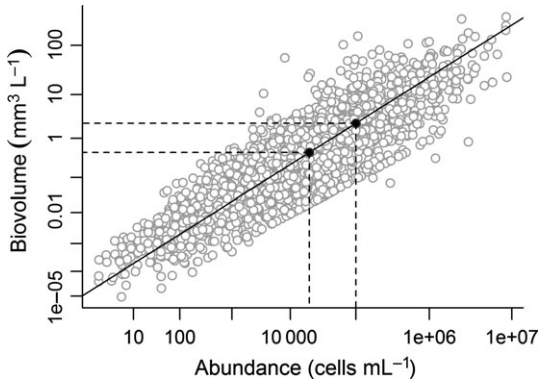

Fig. 2 Relationship between cyanobacterial abundance and biovolume in National Lakes Assessment samples. Solid line: estimated major axis regression relationship, dashed line segments: WHO threshold of 20 000 and 100 000 cells mL^{-1} and the associated biovolume thresholds.

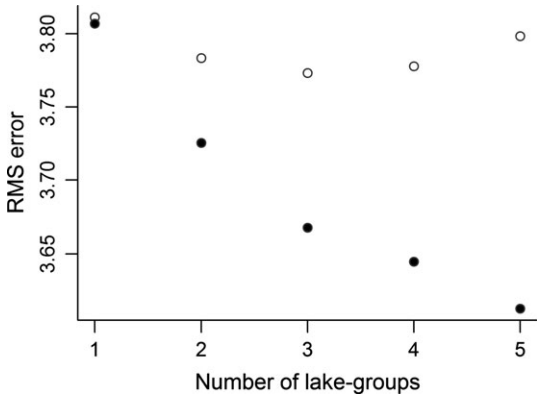

Fig. 3 Average root-mean-square prediction error (RMS) versus the number of lake groups in classification schemes with different numbers of groups. Open circles: mean RMS error in cross-validated data, filled circles: mean RMS error in calibration data.

Cross-validation of the classification trees indicated that the data could support up to three different lake groups (Fig. 3). Validation error increased when the maximum number of lake groups allowed was greater than three, whereas calibration error continued to decrease even up to five lake groups. Overall, use of the classification tree improved cross-validated RMS prediction error from 3.81 for the no-classification case (in units of Box–Cox-transformed cyanobacterial biovolume) to 3.77 for the three-group tree.

The best classification schemes identified by the bumped TREED analysis were very similar. Classification variables that were selected included lake depth and TN:TP (Fig. 4), and these same classification variables were identified in all five of the best-performing classification trees.

In the intercept-only hierarchical Bayesian model, we estimated the standard deviation of among-lake variability as 3.59, while the standard deviations of temporal and sampling variability were 2.94 and 2.18, respectively. When TN and TP were included as explanatory variables, the standard deviation of among-lake variability decreased to 1.71, so nutrient concentrations accounted for approximately 77% of among-lake variance in cyanobacterial biovolume. Incorporating the three-group classification scheme into the hierarchical Bayesian model further reduced the standard deviation of among-lake variability to 1.06. This third model accounted for approximately 91% of the among-lake variance in biovolume.

Estimated regression slopes and intercepts provided an indication of the sensitivity of cyanobacterial biovolume to increases in nutrient concentrations, and these slopes and intercepts varied across different lake groups.

Table 2 Summary statistics of nutrient concentrations, cyanobacterial abundance and candidate classification variables

	Minimum	25th percentile	Median	75th percentile	Maximum
Cyanobacterial biovolume (mm^3 L^{-1})	0	0.03	0.18	1.37	337
Cyanobacterial abundance (cells mL^{-1})	1.48	1.19×10^3	8.48×10^3	5.05×10^4	8.51×10^6
Total phosphorus (TP, μg L^{-1})	4	20	40	96	3640
Total nitrogen (TN, μg L^{-1})	14	318	623	1250	54000
Acid neutralising capacity (μeq L^{-1})	−3360	374	1570	2930	204 000
Lake surface area (ha)	1.0	10.9	31.6	110.0	167000.0
Colour (PCU)	0	12	19	29	840
Depth (m)	0.8	2.5	4.7	9.0	58.5
Altitude (m a.s.l.)	−53	185	331	683	3590
Lake ratio (m$^{-0.5}$)	0.49	2.81	5.04	9.43	88.00
Latitude	26.10	37.70	41.40	44.80	49.00
Longitude	−124.00	−107.00	−94.60	−85.00	−67.20
Mean annual precipitation (mm per year)	67	622	948	1180	3200
Shoreline development	1.01	1.28	1.63	2.27	28.60
Maximum monthly air temperature (Temp, °C)	4.11	12.20	15.20	19.40	31.50
TN:TP	0.3	8.7	14.2	22.5	1690.0

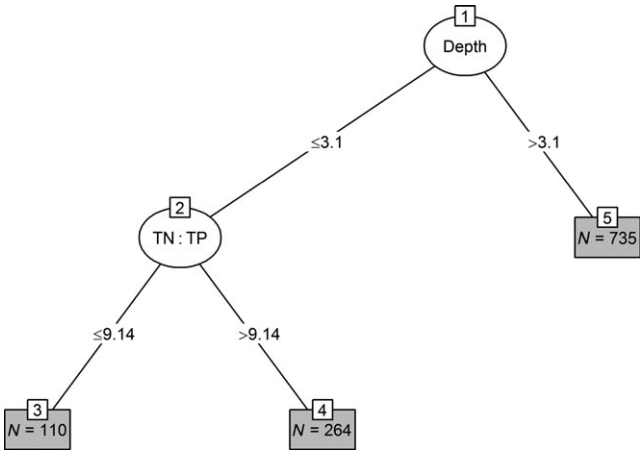

Fig. 4 Selected classification tree for cyanobacterial biovolume. Depth: lake maximum depth (m), total nitrogen (TN): total phosphorus (TP): mass ratio of TN to TP.

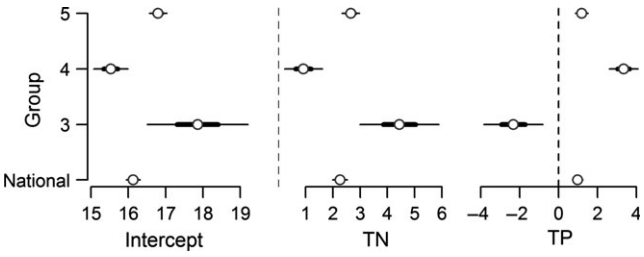

Fig. 5 Comparison of regression coefficients estimated for different lake groups. Coefficients labelled 'National' were estimated using full data set. Group numbers refer to groups defined in Fig. 4. Open circles: mean estimated coefficient value, thick line segment: 50th percentile confidence intervals, thin line segment: 90th percentile confidence intervals.

In the national, no-classification case, cyanobacterial biovolume increased by 2.26 (90% confidence intervals: 1.99–2.53) units for every unit increase in standardised TN concentration and by 0.98 (0.71–1.25) units for every unit increase in standardised TP concentration (Fig. 5). Within lake groups, regression slopes for TN varied from a low value of 0.92 (0.22–1.62) (lake group 4) to a high value of 4.45 (2.99–5.89) (lake group 3). Regression slopes for TP varied from a low value of −2.32 (−3.85 to −0.82) (lake group 3) to a high value of 3.34 (2.60–4.08) (lake group 4). The relative effects of TN and TP varied across lake groups. In lake group 4, which was identified as having relatively high values of TN:TP, the effects of TP were strong, while the effects of TN were relatively weak, whereas in shallow lakes with low values of TN:TP (lake group 3), the effects of TN were strong, while the effects of TP were weak.

Since nutrient concentrations were standardised, intercepts for the regression relationships provide an estimate of cyanobacterial abundance in different lake groups at the overall mean TP and TN concentrations. In the national, no-classification model, the intercept was 16.1 (0.14 $mm^3\ L^{-1}$), but after classification, the value of the intercept varied from 15.5 (0.10 $mm^3\ L^{-1}$) in lake group 4 to 17.9 (0.35 $mm^3\ L^{-1}$) in lake group 3.

Contour lines superimposed on the distributions of TN and TP within each of the lake groups provide another way to visualise the relationships between nutrient concentrations and cyanobacterial biovolumes (Fig. 6). The nearly vertical contour lines in lake group 5 indicated that TN concentration was a stronger predictor of cyanobacterial biovolume in this group of lakes than TP. Conversely, contour lines in lake group 4 indicated

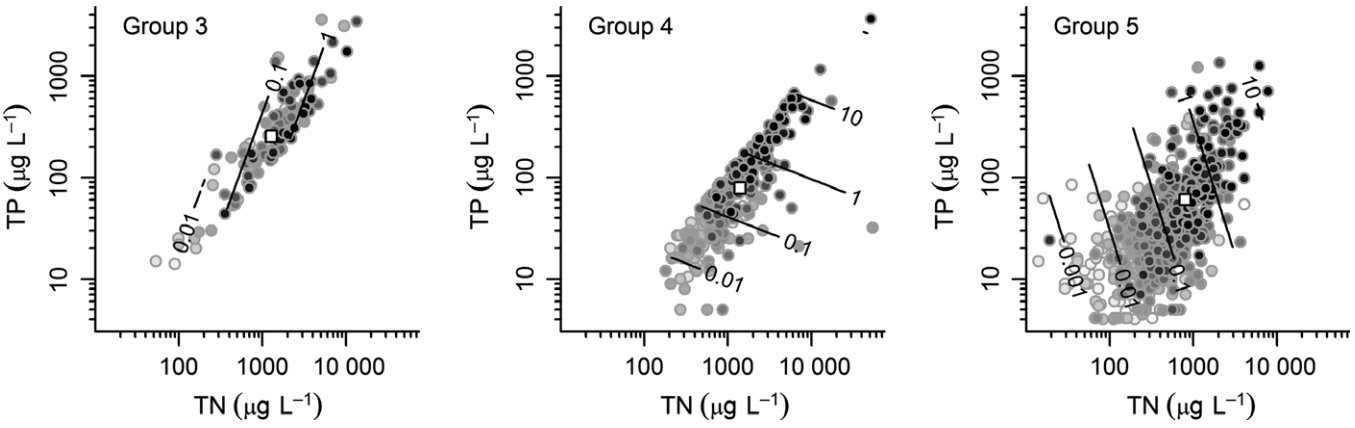

Fig. 6 Relationships between total nitrogen (TN) and total phosphorus (TP) within each lake group. Filled circles: observed values of TN and TP, contours: predicted mean cyanobacterial biovolume associated with each combination of TN and TP, white squares: mean TN and TP target values to achieve targeted mean cyanobacterial biovolume. Colours of circles same as shown in Fig. 1. Group numbers refer to groups defined in Fig. 4.

that changes in both TN and TP predicted cyanobacterial biovolume. These figures also provided a visualisation of the degree to which TP and TN were correlated within each of the lake groups, with lake groups 3 and 4 exhibiting strong correlations. Correlation coefficients (r) between TN and TP confirmed these qualitative observations, as r was 0.92 in lake group 3, 0.82 in lake group 4 and 0.63 in lake group 5.

We characterised the temporal distribution of cyanobacterial biovolumes within a given lake using a mean standard deviation of temporal variability of 3.12, as estimated by the hierarchical Bayesian model that included the lake classification scheme. Based on this distribution, we estimated that when mean cyanobacterial biovolume was 0.29 $mm^3\ L^{-1}$, the 90th percentile of the distribution of observed values during the sampling season was 2.3 $mm^3\ L^{-1}$, the WHO moderate risk threshold value. In other words, maintaining mean cyanobacterial biovolume below 0.29 $mm^3\ L^{-1}$ should also maintain the frequency of cyanobacterial biovolumes that exceed the WHO moderate risk threshold to <10% of samples. A similar calculation yields a threshold of 0.048 $mm^3\ L^{-1}$ to maintain the frequency of cyanobacterial biovolumes exceeding the WHO low risk threshold at 10%.

Targets for TN and TP that corresponded with mean cyanobacterial biovolumes at the WHO moderate risk threshold varied across different lake groups. In the base, no-classification model, mean management targets for TN and TP were 1100 and 87 μg L^{-1}, respectively (Table 3). After incorporating lake groups as selected by TREED regression, mean target nutrient concentrations varied from 800 to 1300 μg L^{-1} for TN and from 61 to 250 μg L^{-1} for TP (Table 3). Deeper lakes (>3.1 m deep) were associated with the lowest nutrient targets. These low targets arose from a relatively high value of the regression intercept and a relatively strong relationship between TN, TP and cyanobacterial biovolume (lake group 5 in Fig. 5). TN and TP targets for the WHO low risk threshold followed the same pattern as those for the moderate risk threshold (Table 4).

Table 4 Management targets for nutrient concentrations related to exceedance of WHO low risk threshold

	TN (μg L^{-1})			TP (μg L^{-1})		
Group	Mean	50% Conf Lim	25th and 75th %tile	Mean	50% Conf Lim	25th and 75th %tile
5	320	300–340	270–380	22	20–23	18–26
4	600	560–640	510–710	31	29–33	25–37
3	270	220–340	180–390	44	36–57	29–67
National	370	340–410	260–540	25	23–28	16–39

Mean: mean management target; 50% Conf Lim: 50% confidence limits on mean target; 25th and 75th %tile: 25th and 75th percentiles of the distribution of possible targets, including among-site variability; TP: total phosphorus; TN: total nitrogen.

Confidence limits on mean nutrient targets varied depending on the number of samples in each lake group and the confidence with which relationships between cyanobacterial biovolume and nutrient concentrations were estimated (Fig. 7). Mean TP and TN targets using the national model (with no lake groups) were estimated very precisely because of the size of the national database. Confidence limits on mean TN and TP targets for lake groups 3 and 5 were also narrow, reflecting the narrow confidence limits of the regression coefficients (Fig. 5).

Incorporating among-lake variability into the estimated range of possible TN and TP targets provides values that reflect the different outcomes that one might expect to observe in different lakes. For example, in the national model, after incorporating among-lake variability, the 25th percentile of the distribution of possible TN targets was 750 μg L^{-1} (Table 3, Fig. 7). That is, the model predicted that maintaining TN concentrations at 750 μg L^{-1} would ensure that at least 75% of lakes in the data set would achieve targeted cyanobacterial bio-

Table 3 Management targets for nutrient concentrations related to exceedance of WHO moderate risk threshold

	TN (μg L^{-1})			TP (μg L^{-1})		
Group	Mean	50% Conf Lim	25th and 75th %tile	Mean	50th Conf Lim	25th and 75th %tile
5	800	750–850	670–940	61	56–65	49–73
4	1400	1300–1500	1200–1600	79	73–84	65–93
3	1300	1100–1500	900–1700	250	220–300	170–360
National	1100	990–1200	750–1500	87	79–96	57–130

Mean: mean management target; 50% Conf Lim: 50% confidence limits on mean target; 25th and 75th %tile: 25th and 75th percentiles of the distribution of possible targets, including among-site variability; TP: total phosphorus; TN: total nitrogen.

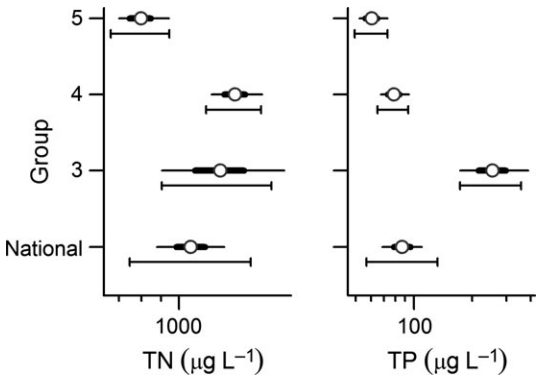

Fig. 7 Estimated nutrient targets. Open circle: mean target, thick line: 50th percentile confidence interval on mean value, thin line: 90th percentile confidence interval on mean values, line with end caps: 25th and 75th percentiles of distribution of targets, incorporating among-site variability. Group numbers refer to groups defined in Fig. 4.

volumes. In contrast, the mean TN target of 1100 μg L^{-1} would achieve targeted cyanobacterial biovolumes in approximately half of the lakes in the data set, and maintaining TN at 1500 μg L^{-1} would achieve cyanobacterial goals in 25% of lakes. The range of possible nutrient target values arises from both among-site variability and confidence limits on the mean target estimates. Hence, the ranges of nutrient targets in the national model were broad because of a large among-site variability, even though confidence limits for the mean targets were narrow. Conversely, ranges for lake groups 4 and 5 were relatively narrow, owing to a smaller among-site variability and precise estimates of the model parameters (Fig. 7).

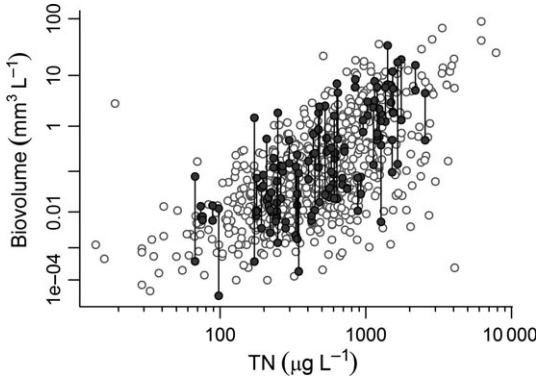

Fig. 8 Plot showing the magnitude of temporal variability relative to residual variability between total nitrogen and cyanobacterial biovolume in lake group 5. Circles: average cyanobacterial biovolume for a single lake visit, filled circles and vertical line segments: cyanobacterial biovolumes observed at the same lake at different times during the sampling season.

Classification and modelling the effects of TN and TP reduced among-lake variability to nearly negligible levels relative to temporal and sampling variability. The contribution of temporal variability to overall variance of cyanobacterial biovolume can be visualised when cyanobacterial biovolume is plotted as a function of TN (Fig. 8). In this group of lakes, the effect of TN on cyanobacterial biovolume is stronger than TP, and so only TN is plotted. The effects of sampling variability on the plotted data were reduced by plotting the average cyanobacterial biovolume across littoral and open water sites for each lake visit. Then, the magnitude of temporal variability is displayed with vertical lines connecting samples collected at the same lake on different days. The lengths of the vertical lines relative to the spread of the data points illustrate that temporal variance accounted for the majority of the residual variability about the estimated TN-cyanobacterial biovolume relationship.

Discussion

Increased concentrations of phosphorus and nitrogen have been associated with increased occurrences of cyanobacterial blooms and increased cyanobacterial biovolume in studies conducted at a variety of locations, and the present analysis provides further support for these relationships using continental-scale data from the United States. Analysis of data collected from lakes in Florida (Canfield, Phlips & Duarte, 1989), from Europe (Carvalho *et al.*, 2013) and from northern temperate lakes (Downing *et al.*, 2001) all found strong associations between nutrient concentrations and cyanobacterial abundance. We observed strong associations between increases in cyanobacterial biovolume and increases in nutrient concentration in the vast majority of lakes and reservoirs of the United States, similar to other analyses of U.S. data (Beaulieu *et al.*, 2013). Also, the TP targets identified in the present study were very comparable to TP targets using the same WHO thresholds and data from Europe (Carvalho *et al.*, 2013).

The present study expands on previous analyses by providing results that are easily communicated to stakeholders and directly applicable to management decisions. The two key components of this analysis approach were (i) using classification to control for the effects of covariates and to allow the use of simple models to represent the relationships between nutrient concentrations and cyanobacterial biovolume, and (ii) using hierarchical Bayesian models to partition variability of cyanobacterial biovolume observations into among-lake,

temporal, and sampling components, which can then more accurately predict the probability of high cyanobacterial biovolume in different lakes.

Classification

To classify lakes, we used TREED analysis to identify groups of lakes that maximised the degree to which changes in nutrient concentrations accounted for variations in cyanobacterial biovolume. TREED analysis provides a means of screening many different candidate classification variables and identifying the most relevant variables and the appropriate cut points for those variables. As such, it yields inherent performance advantages over *a priori* classifications (e.g. Cheruvelil *et al.*, 2008). In our bumped search of possible classification schemes, we identified several similar classifications that improved the predictive accuracy of the model. Other approaches are also available for searching the space of possible trees that may identify a different set of classification trees (Gramacy, 2007). In cases in which the data support a greater numbers of discrete classes, many different trees representing different classification schemes may be identified with comparable predictive performance (Yuan & Pollard, 2014). However, in the present analysis, the three-group trees that best represented the data were generally similar.

Understanding the effects of other environmental factors on cyanobacterial biovolume was not the primary focus of this work. Indeed, as noted above, in some cases, different classification schemes, each using a different combination of classification variables, can provide a similar improvement in predictive performance, and so the selection of a particular suite of variables here should not be interpreted as an indication of their overall importance in predicting cyanobacterial abundance. Given those caveats, though, we did observe that the classification variables frequently identified by TREED analysis were consistent with environmental factors identified by other studies.

TN:TP was one of the frequently selected classification variables, and the strength with which this ratio influences cyanobacterial abundance and biovolume has long been the subject of debate (Smith, 1983; Downing *et al.*, 2001). Similar to our current analysis, recent work has suggested that TN and TP are independently correlated with cyanobacterial biovolume and that the ratio between the two may influence the respective relationships with TN and TP (Dolman *et al.*, 2012). In our model, high values of TN:TP were associated with lakes in which TP was a strong predictor of cyanobacterial biovolume (lake group 4), a finding that is consistent with an intuitive understanding of nutrient limitation. That is, in lakes with high TN:TP ratio, one might expect that TP is limiting and therefore the better predictor for cyanobacterial biovolume. Similarly, in lakes with low values of TN:TP, we found TN was a better predictor, as would be expected.

TN:TP is frequently selected as a classification variable likely because the linear relationships used to model the effects of TN and TP on cyanobacterial biovolume did not fully represent the interactions between the two nutrients in certain types of lakes. To further explore this possibility, we fit a nonparametric, GAM to the relationship between TN, TP and cyanobacterial biovolume in combined data from lake groups 3 and 4, and the resulting model predictions illustrate the changing effects of TN and TP with changes in the ratio between the two nutrients (Fig. 9). At low values of TN:TP (e.g. below the dashed line that indicates TN:TP = 9.1), the orientation of the contour lines suggests a combined effect of TN and TP. However, at higher values of TN:TP, and especially at high values of TN, the contour lines are nearly vertical, indicating that TN is the strongest predictor of cyanobacterial biovolume. The GAM represents the smooth transition between different nutrient regimes, and use of this model may provide more appropriate nutrient targets for shallow lakes (i.e. depth $\leq$3.1 m) than the separate targets for lake groups 3 and 4. In deeper lakes, the absence of TN:TP as a classification variable suggests that the linear models provide an adequate representation of the observed relationships.

Lake depth was the second variable that was selected frequently to classify lakes, and its selection is consistent

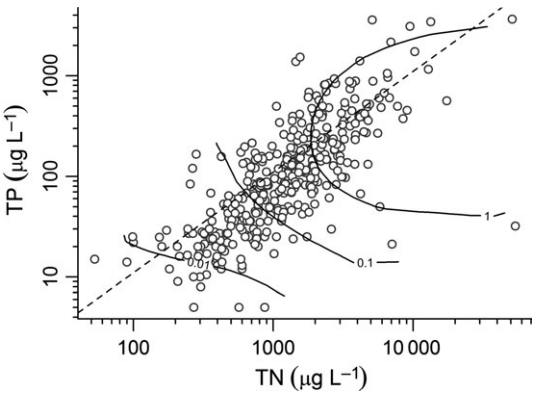

Fig. 9 Relationship between total nitrogen (TN), total phosphorus (TP) and cyanobacterial biovolume in lake groups 3 and 4. Open circles: observed values of TN and TP, contour lines: predicted mean cyanobacterial biovolume ($mm^3\ L^{-1}$) from generalised additive model, dashed line: threshold value of TN:TP = 9.1.

with emerging insights into fundamental differences between shallow and deep lakes. In shallow lakes, nutrient concentrations often exert weaker controls on algal assemblages as top-down effects can be more important (Jeppesen *et al.*, 1997). Conditions in shallow lakes also can settle in alternate turbid or clear stable states (Scheffer & van Nes, 2007), and growth of macrophytes can contribute substantially to the overall nutrient budget (van Donk *et al.*, 1993). Elucidating the precise mechanisms by which lake depth affects the relationships between nutrients and cyanobacterial biovolume is beyond the scope of the present analysis, but given the functional differences between shallow and deep lakes, the frequent selection of lake depth as a classification variable was not surprising.

Overall, TREED regression improved the precision of the estimated relationships and the predictive accuracy by accounting for the effects of other environmental factors. Furthermore, as shown in Fig. 6, the results of the analysis can be shown as simple contour plots that clearly illustrate the effects of increased nutrients on cyanobacterial biovolume. Other statistical models that simultaneously quantify the effects of nutrients and other environmental factors [e.g. random forests (Cutler *et al.*, 2007) and Bayesian trees (Chipman, George & McCulloch, 2010)] can provide more accurate predictions than the TREED regression approach shown here (Yuan & Pollard, 2014). However, the results from these types of analyses cannot be easily plotted and are therefore more difficult to interpret for management decisions. The TREED regression approach provides a useful compromise between prediction accuracy and interpretability.

Variance components

Using the hierarchical Bayesian model to partition variance in the amount of cyanobacteria into among-lake, temporal and sampling components enhanced our ability to interpret the data and analysis in at least two ways. First, we could more accurately assess model performance when we compared the amount of variability explained by the model to the amount of variability that the model could reasonably explain. Since the classification scheme and the relationships between nutrient concentration and cyanobacterial biovolume were based solely on differences among lakes, we can only expect this model to account for among-lake variance. Hence, estimates of the proportion of among-lake variance explained by the model provide a much more informative assessment of model performance. Our results suggest that differences in mean nutrient concentrations accounted for the vast majority of differences in cyanobacterial biovolume among U.S. lakes. Then, incorporating lake classification further increased the proportion of explained among-lake variance to 91%. A simple R^2 calculated for this same model was only 40%, which does not truly reflect the performance of the model.

Second, quantifying temporal and sampling variability of cyanobacterial biovolume allows us to accurately account for temporal variability in the amount of cyanobacteria observed within a particular lake, and therefore, more accurately predict the probability of biovolumes that exceed a pre-determined threshold. Simple linear regression only estimates the overall residual variability, which combines all sources of variance. Using such a variance estimate that includes sampling, temporal and among-lake sources would generally overestimate the variability of the amount of cyanobacteria in a particular lake. From a practical perspective, overestimating variability may lead one to target a lower mean cyanobacterial biovolume and lower nutrient concentrations than would be necessary to reduce occurrences of high cyanobacterial biovolumes to a desired frequency.

Two issues with regard to the variance estimates require further discussion. First, we assumed that temporal variability for all lakes in the data set was the same, as our 100 repeat visits was only sufficient for reliably calculating an average temporal variance for the entire data set. However, different lakes possibly exhibit different levels of temporal variability in cyanobacterial biovolume, and additional repeat sample data, as it becomes available, would be useful to explore this issue. Second, to avoid adding another layer of complexity to the present analysis and again because of the limited number of repeat samples, we did not include predictor variables that were associated with within-lake, temporal changes. However, exploratory analysis (not shown here) suggested that factors such as the sampling day of the year may account for a substantial proportion of the variability in cyanobacterial biovolume over time, and with additional data, accounting for these known within-lake effects may further improve the precision of these models.

Nutrient effects and targets

We estimated separate effects of TN and TP on cyanobacterial biovolume and used these relationships to derive nutrient targets for different types of lakes in the United States. As with most observations of nutrient concentrations in lakes, we observed strong correlations

between TN and TP at the national scale, and within individual lake groups. Correlations in predictor variables such as observed here have often posed problems for statistical models of the relationships between nutrients and algal biomass. Confronted with strongly correlated TN and TP, some researchers have chosen to model only one of the nutrients (Carvalho *et al.*, 2011; Yuan & Pollard, 2014), reasoning that relationships between algal biomass and the other, unmodelled nutrient would be similar, given the strong correlations. Others have developed independent models for TN and TP and compared the results (Downing *et al.*, 2001; Beaulieu *et al.*, 2013). We assert here, as others have (Canfield *et al.*, 1989; Dolman *et al.*, 2012), that for large data sets such as the national-scale data used here, models that simultaneously estimate the effects of both TN and TP provide the most accurate predictions and illuminate differences in the relative effects of the two nutrients. Errors in the estimates of the effects of two correlated predictors are greatly reduced with increased sample size (i.e. >100 samples) and in models with high R^2 values, as is the case in the current analysis (Mason & Perreault, 1991), and so, we believe our estimates of TN and TP effects are minimally affected by the TN-TP correlation for the majority of lake groups. In lake groups 3 and 4, the correlations between TN and TP may be high enough to introduce some variability to the estimated effects, but even in those cases, the combination of a high R^2 and high sample size suggests that the errors are not large (Mason & Perreault, 1991).

Our finding that the best nutrient predictors of cyanobacterial biovolume varied across lake groups mirrors the diverse findings regarding nutrient limitation from experimental studies in lakes. Some manipulative studies have found that combined increases in TN and TP caused greater increases in primary productivity than either TN or TP alone (Maberly *et al.*, 2002; Dzialowski *et al.*, 2005), and nutrient colimitation has also been observed with cyanobacteria (Paerl & Otten, 2013). Other studies have found that either TN (Levine & Whalen, 2001) or TP is the limiting nutrient (Schindler *et al.*, 2008) and indeed, that the limiting nutrient can vary over time within the same lake (Maberly *et al.*, 2002). Our results based on data collected at a broad spatial scale provides support for the idea that nutrient limitation can vary among different lakes, and these analyses can potentially guide the selection of lakes for future comparative experiments.

Because our analyses were based on data collected within different lakes and did not include nutrient loading data, we could identify TN and TP as the best predictors of cyanobacterial biovolume in different lake groups, but our ability to inform appropriate remediation actions is limited. Some studies have strongly linked reductions in loadings of phosphorus to decreases in algal biomass (Effler & O'Donnell, 2010), and some have advocated that control only of phosphorus will achieve desired conditions in lakes (Schindler *et al.*, 2008). Others have advocated that reduced loadings of both nitrogen and phosphorus would be more effective (Lewis, Wurtsbaugh & Paerl, 2011). In some lakes, fixation of N_2 by cyanobacteria supplies the necessary nitrogen to fuel excess cyanobacterial growth (Beversdorf, Miller & McMahon, 2013), and in these lakes, reduction only in phosphorus loads may indeed result in decreases in observed concentrations of both TN and TP. However, other studies have found that even nitrogen-fixing cyanobacteria preferentially use environmental inorganic nitrogen when it is available (Ferber *et al.*, 2004), and in lakes with high levels of external nitrogen loading, reduction in both nitrogen and phosphorus loads would likely restore desired lake conditions most efficiently. In general, additional lake-specific consideration of the nature of external nutrient loads would help identify the management actions that will most effectively achieve nutrient targets and desired amounts of cyanobacteria in a particular lake.

Since both TN and TP are used to predict cyanobacterial biovolume, an additional assumption is required to identify individual management targets for TN and TP from the infinite set of possible values. A number of different approaches are possible, and we demonstrated the simple idea of calculating TN and TP targets based on major axis regression. This approach yields targets for TN and TP that, by definition, are located in the middle of the joint distribution of TN and TP concentrations. As such, the targets for TN and TP derived in this way represent the overall mean concentrations of TN and TP across the NLA data set, or within a particular lake group. Other approaches for selecting individual management targets are possible, including examining the upper and lower bounds of the possible combined values of TN and TP (Yuan *et al.*, 2014), and the final approach selected depends on the management objectives.

The analyses described here directly links numerical nutrient targets to a known threshold for cyanobacterial biovolume in an easily interpreted format. As such, these analyses can inform management decisions. The simple model, in which all available data are considered together, provides single target values for TP and TN for all lakes and reservoirs in the continental United

States. These target values maintain the *average* probability of cyanobacterial biovolume exceeding the WHO moderate risk threshold of 2.3 $mm^3 L^{-1}$ at 10%. However, differences among lakes, as quantified by the among-lake variance, indicate that for different lakes, these target nutrient concentrations value may yield cyanobacterial biovolumes that exceed the threshold much more frequently or much less frequently than 10%. The hierarchical Bayesian approach for estimating these relationships provided the means to explicitly incorporate different sources of variability into the estimates of nutrient targets, and the ranges provided for each nutrient target provide values that are associated with different risks of exceeding the WHO threshold. The lower end of each range corresponds with an estimated target that ensures that cyanobacterial biovolume in 75% of the lakes in the group remains below the WHO threshold at least 90% of the time. Conversely, the upper end of each range provides assurance that 25% of lakes in the group will achieve the desired condition. Hence, values within this range can be selected and weighed against other factors that inevitably enter into environmental decisions.

When using these nutrient targets to inform management decisions, one should consider all of the management objectives for a particular lake. The nutrient targets described here pertain only to the occurrence of excessive densities of cyanobacteria, and other considerations regarding the health of a lake likely yield different nutrient targets. For example, the extent of hypoxia in the lake (Yuan & Pollard, 2015) or the occurrence of high concentrations of microcystin (Yuan *et al.*, 2014) may be relevant endpoints to consider when making the management decisions.

Acknowledgments

The authors with to thank J. Oliver, S. Santell and D. Thomas for reviewing an earlier version of this paper. We also thank the many crews of the National Lakes Assessment for collecting the data used in this analysis. The views expressed in this paper are those of the authors and do not represent the official policy of the U.S. Environmental Protection Agency.

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. *Journal of Phycology*, **47**, 495–504.

Alexander W.P. & Grimshaw S.D. (1996) Treed regression. *Journal of Computational and Graphical Statistics*, **5**, 156–175.

Beaulieu M., Pick F. & Gregory-Eaves I. (2013) Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes data set. *Limnology and Oceanography*, **58**, 1736–1746.

Beversdorf L.J., Miller T.R. & McMahon K.D. (2013) The role of nitrogen fixation in cyanobacterial bloom toxicity in a temperate, eutrophic lake. *PLoS ONE*, **8**, e56103.

Breiman L., Friedman J., Stone C.J. & Olshen R.A. (1984) *Classification and Regression Trees*. Chapman and Hall/CRC, Boca Raton, FL.

Canfield D.E. Jr, Phlips E. & Duarte C.M. (1989) Factors influencing the abundance of blue-green algae in Florida lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **46**, 1232–1237.

Carvalho L., McDonald C., de Hoyos C., Mischke U., Phillips G., Borics G. *et al.* (2013) Sustaining recreational quality of European lakes: minimizing the health risks from algal blooms through phosphorus control. *Journal of Applied Ecology*, **50**, 315–323.

Carvalho L., Miller (nee Ferguson) C.A., Scott E.M., Codd G.A., Davies P.S. & Tyler A.N. (2011) Cyanobacterial blooms: statistical models describing risk factors for national-scale lake assessment and lake management. *Science of the Total Environment* **409**, 5353–5358.

Cheruvelil K.S., Soranno P.A., Bremigan M.T., Wagner T. & Martin S.L. (2008) Grouping lakes for water quality assessment and monitoring: the roles of regionalization and spatial scale. *Environmental Management*, **41**, 425–440.

Chipman H.A., George E.I. & McCulloch R.E. (1998) Bayesian CART model search. *Journal of the American Statistical Association*, **93**, 935.

Chipman H.A., George E.I. & McCulloch R.E. (2010) BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, **4**, 266–298.

Cutler D.R., Edwards T.C., Beard K.H., Cutler A., Hess K.T., Gibson J. *et al.* (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Daly C., Halbleib M., Smith J.I., Gibson W.P., Doggett M.K., Taylor G.H. *et al.* (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, **28**, 2031–2064.

Dolman A.M., Rücker J., Pick F.R., Fastner J., Rohrlack T., Mischke U. *et al.* (2012) Cyanobacteria and cyanotoxins: the influence of nitrogen versus phosphorus. *PLoS ONE*, **7**, e38757.

van Donk E., Gulati R.D., Iedema A. & Meulemans J.T. (1993) Macrophyte-related shifts in the nitrogen and phosphorus contents of the different trophic levels in a biomanipulated shallow lake. *Hydrobiologia*, **251**, 19–26.

Downing J.A., Watson S.B. & McCauley E. (2001) Predicting cyanobacteria dominance in lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 1905–1908.

Dzialowski A.R., Wang S.-H., Lim N.-C., Spotts W.W. & Huggins D.G. (2005) Nutrient limitation of phytoplankton growth in central plains reservoirs, USA. *Journal of Plankton Research*, **27**, 587–595.

Effler S.W. & O'Donnell S.M. (2010) A long-term record of epilimnetic phosphorus patterns in recovering Onondaga Lake, New York. *Fundamental and Applied Limnology/Archiv für Hydrobiologie*, **177**, 1–18.

Ferber L.R., Levine S.N., Lini A. & Livingston G.P. (2004) Do cyanobacteria dominate in eutrophic lakes because they fix atmospheric nitrogen? *Freshwater Biology*, **49**, 690–708.

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

Gorham E. & Boyce F.M. (1989) Influence of lake surface area and depth upon thermal stratification and the depth of the summer thermocline. *Journal of Great Lakes Research*, **15**, 233–245.

Gramacy R.B. (2007) tgp: an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *Journal of Statistical Software*, **19**, 6.

Harrell F.E. (2001) *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer-Verlag, New York, NY.

Hothorn T. & Zeileis A. (2015) partykit: A Modular Toolkit for Recursive Partytioning in R. Journal of Machine Learning Research (accepted 2015-02-20). URL http://EconPapers.RePEc.org/RePEc:inn:wpaper:2014-10

Huisman J., Jonker R.R., Zonneveld C. & Weissing F.J. (1999) Competition for light between phytoplankton species: experimental tests of mechanistic theory. *Ecology*, **80**, 211–222.

Jeppesen E., Jensen J.P., Søndergaard M., Lauridsen T., Pedersen L.J. & Jensen L. (1997) Top-down control in freshwater lakes: the role of nutrient state, submerged macrophytes and water depth. In: *Shallow Lakes '95. Developments in Hydrobiology*. (eds Kufel L., Prejs A. & Rybak J.I.), pp. 151–164. Springer Netherlands, Rotterdam.

Lévesque B., Gervais M.-C., Chevalier P., Gauvin D., Anassour-Laouan-Sidi E., Gingras S. *et al.* (2014) Prospective study of acute health effects in relation to exposure to cyanobacteria. *Science of the Total Environment*, **466–467**, 397–403.

Levine M.A. & Whalen S.C. (2001) Nutrient limitation of phytoplankton production in Alaskan Arctic foothill lakes. *Hydrobiologia*, **455**, 189–201.

Lewis W.M., Wurtsbaugh W.A. & Paerl H.W. (2011) Rationale for control of anthropogenic nitrogen and phosphorus to reduce eutrophication of inland waters. *Environmental Science & Technology*, **45**, 10300–10305.

Maberly S.C., King L., Dent M.M., Jones R.I. & Gibson C.E. (2002) Nutrient limitation of phytoplankton and periphyton growth in upland lakes. *Freshwater Biology*, **47**, 2136–2152.

Mason C.H. & Perreault W.D. (1991) Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, **28**, 268.

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

Paerl H.W. & Otten T.G. (2013) Harmful cyanobacterial blooms: causes, consequences, and controls. *Microbial Ecology*, **65**, 995–1010.

Pilotto L.S., Douglas R.M., Burch M.D., Cameron S., Beers M., Rouch G.J. *et al.* (1997) Health effects of exposure to cyanobacteria (blue–green algae) during recreational water–related activities. *Australian and New Zealand Journal of Public Health*, **21**, 562–566.

Pinckney J.L., Paerl H.W., Harrington M.B. & Howe K.E. (1998) Annual cycles of phytoplankton community-structure and bloom dynamics in the Neuse River Estuary, North Carolina. *Marine Biology*, **131**, 371–381.

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Scheffer M. & van Nes E.H. (2007) Shallow lakes theory revisited: various alternative regimes driven by climate, nutrients, depth and lake size. *Hydrobiologia*, **584**, 455–466.

Schindler D.W., Hecky R.E., Findlay D.L., Stainton M.P., Parker B.R., Paterson M.J. *et al.* (2008) Eutrophication of lakes cannot be controlled by reducing nitrogen input: results of a 37-year whole-ecosystem experiment. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 11254–11258.

Smith V.H. (1983) Low nitrogen to phosphorus ratios favor dominance by blue-green algae in lake phytoplankton. *Science*, **221**, 669–671.

Stefan H.G., Hondzo M., Fang X., Eaton J.G. & McCormick J.H. (1996) Simulated long-term temperature and dissolved oxygen characteristics of lakes in the north-central United States and associated fish habitat limits. *Limnology and Oceanography*, **41**, 1124–1135.

Stewart I., Webb P.M., Schluter P.J. & Shaw G.R. (2006) Recreational and occupational field exposure to freshwater cyanobacteria – a review of anecdotal and case reports, epidemiological studies and the challenges for epidemiologic assessment. *Environmental Health*, **5**, 6.

Tibshirani R. & Knight K. (1999) Model search by bootstrap "bumping". *Journal of Computational and Graphical Statistics*, **8**, 671–686.

US EPA (2011) *2012 National Lakes Assessment. Field Operations Manual*. U.S. Environmental Protection Agency, Washington, DC.

US EPA (2012) *2012 National Lakes Assessment. Laboratory Operations Manual*. U.S. Environmental Protection Agency, Washington, DC.

Wagner C. & Adrian R. (2009) Cyanobacteria dominance: quantifying the effects of climate change. *Limnology and Oceanography*, **54**, 2460–2468.

WHO (2003) *Guidelines for Safe Recreational Water Environments. Volume 1. Coastal and Fresh Waters*. World Health Organization, Geneva.

Wood S.N. & Augustin N.H. (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, **157**, 157–177.

Wynne T.T., Stumpf R.P., Tomlinson M.C. & Dyble J. (2010) Characterizing a cyanobacterial bloom in Western Lake Erie using satellite imagery and meteorological data. *Limnology and Oceanography*, **55**, 2025–2036.

Yuan L.L. & Pollard A.I. (2014) Classifying lakes to improve precision of nutrient–chlorophyll relationships. *Freshwater Science*, **33**, 1184–1194.

Yuan L.L. & Pollard A.I. (2015) Classifying lakes to quantify relationships between epilimnetic chlorophyll a and hypoxia. *Environmental Management* **55**, 578–587.

Yuan L.L., Pollard A.I., Pather S., Oliver J.L. & D'Anglada L. (2014) Managing microcystin: identifying national-scale thresholds for total nitrogen and chlorophyll a. *Freshwater Biology*, **59**, 1970–1981.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Deriving nutrient targets to prevent excessive cyanobacterial densities in U.S. lakes and reservoirs.

Figure S1. Relationships between chl *a* and estimated total phytoplankton biovolume.

(Manuscript accepted 22 May 2015)