Freshwater Biology

Freshwater Biology (2014) **59,** 1970–1981 doi:10.1111/fwb.12400

Managing microcystin: identifying national-scale thresholds for total nitrogen and chlorophyll *a*

LESTER L. YUAN*, AMINA I. POLLARD[†], SANTHISKA PATHER*, JACQUES L. OLIVER* AND LESLEY D'ANGLADA*

**Office of Science and Technology, Office of Water, U.S. Environmental Protection Agency, Washington, DC, U.S.A.*

[†]Office of Wetlands, Oceans, and Watersheds, Office of Water, U.S. Environmental Protection Agency, Washington, DC, U.S.A.

SUMMARY

1. The occurrence of high cyanotoxin concentrations can severely impair the use of a waterbody for drinking water and recreational purposes. Cyanotoxins are likely to occur under specific environmental conditions, and so identifying these conditions can facilitate management of the waterbody to reduce the likelihood of high cyanotoxin concentrations.
2. We analysed data collected from lakes across the contiguous United States to identify environmental variables that are strongly associated with occurrence of high concentrations of a common cyanotoxin, microcystin (MC).
3. Since many different environmental variables covary and are associated with high MC, we used least absolute shrinkage and selection operator (LASSO) regression to identify a few variables that provided accurate predictions of high MC ($\geq$1 μg L^{-1}).
4. Our analysis indicated that total nitrogen (TN) and chlorophyll *a* (chl *a*) concentrations yielded a parsimonious model that accurately predicted the occurrence of high MC. Based on this model, we identified management thresholds for TN and chl *a* that would maintain the probability of high MC at or below 10 and 5%.

Keywords: chlorophyll *a*, cyanotoxins, LASSO regression, microcystin, nutrients

Introduction

The ecological balance in freshwater ecosystems can be affected by human activities in the landscape. For example, excess nutrient loading from anthropogenic sources is one of the main threats to freshwater ecosystems worldwide (Smith, 2003). Increased nutrient concentrations in the water column can increase algal abundance, which, in turn, can cause undesirable ecological effects such as reductions in dissolved oxygen concentrations and changes in algal assemblage composition. Shifts from diatom to cyanobacteria-dominated algal assemblages are often associated with high nutrient concentrations (Downing, Watson & McCauley, 2001), and these changes can be accompanied by noxious odours and unsightly scums on the water surface. Furthermore, certain species of cyanobacteria produce cyanotoxins, which, when ingested in high enough concentrations, are poisonous to animals and people (Pilotto, 2008; Stewart, Seawright & Shaw, 2008). Among the cyanotoxins, microcystins (MCs) are commonly observed in lakes during warmer months. Microcystins are hepatotoxins and tumour promoters and have been associated with adverse effects, including deaths of fish, birds and domestic and wild animals, as well as illness and deaths in humans (Azevedo *et al*., 2002). Because of these potential adverse health effects, the occurrence of cyanotoxins in a waterbody can negatively affect the use of that waterbody for recreation or as a source of drinking water.

Cyanotoxins and cyanobacteria occur naturally in lakes and reservoirs, but the frequency of high concentration cyanotoxin events has increased in recent years due to human activities (Carmichael, 2008). These human activities have increased nutrient concentrations (Davis *et al*., 2010), altered ratios between nitrogen and

Correspondence: Lester L. Yuan, Office of Science and Technology, Office of Water, U.S. Environmental Protection Agency, 1200 Pennsylvania Ave NW, Washington, DC 20460, U.S.A. E-mail: yuan.lester@epa.gov

phosphorus (Downing *et al.*, 2001), increased temperatures (Davidson & Bradshaw, 1967) and altered flow regimes (Poff *et al.*, 1997; Carlisle, Wolock & Meador, 2010), all of which are factors that have been linked with the increased frequency of high concentration cyanotoxin events. Many other environmental factors have also been associated with the occurrence of cyanotoxin events, including high alkalinity (Carvalho *et al.*, 2011) and increased, but not saturating, light availability (Wiedner *et al.*, 2003).

One approach for managing and reducing the frequency of high cyanotoxin events is to identify environmental conditions under which these events are likely to occur and to manage waterbodies to avoid these conditions. More specifically, one can identify values for different anthropogenic pollutants that are associated with a low probability of cyanotoxin events. Once these protective numeric thresholds are identified, they can help facilitate management of different waterbodies by providing numerical targets for mitigation activities (e.g. discharge permits, non-point source best practices) and for routine monitoring. Protective thresholds can be derived by statistically modelling the relationship between the frequency of cyanotoxin occurrence and different pollutants, and then, using the model to determine protective levels of each pollutant. Ideally, when ambient levels of different pollutants are maintained below the established thresholds, managers can be reasonably confident that undesirable conditions (i.e. high cyanotoxin concentrations) will be infrequent.

Appropriate data for estimating widely applicable relationships between high cyanotoxin concentrations and different pollutants are limited. Manipulative studies provide site-specific insights into different factors that affect cyanotoxin production, but can be difficult to generalise across large areas. Observations of cyanotoxin concentrations across large spatial scales are relatively rare due to the expense of the laboratory analyses and the inherently complex logistics of coordinating large-scale monitoring programs. The U.S. National Lakes Assessment (NLA) provides one possible data set for estimating relationships between cyanotoxin occurrence and anthropogenic pollutants and natural factors. In this study, MC concentrations were measured in lakes across the contiguous United States, and hence, the data represent very broad environmental gradients (US EPA, 2010). Furthermore, a comprehensive suite of other environmental variables was quantified using consistent protocols at each of the sampled lakes.

Here, we ask whether the analysis of broad, national-scale data can inform the management of cyanotoxin occurrence in lakes. More specifically, we examined U.S. lakes in which sampled concentrations of one common cyanotoxin, MC, exceeded the World Health Organization (WHO) drinking water provisional guideline of 1 $\mu g\ L^{-1}$ (WHO, 1998) and estimated the relationship between these exceedances and different environmental factors using data collected by the NLA. Based on the analysis, we considered whether protective thresholds could be specified for nutrient variables with respect to MC occurrence.

Methods

Data

Data used for this analysis were collected by the NLA in the summer (May–September) of 2007 (US EPA, 2010). Lakes >4 ha were selected from the contiguous United States using a stratified random sampling design. At each of the 1077 lakes sampled, an extensive suite of environmental variables was measured, but here, we only provide sampling details regarding the variables used in the present analysis.

Chemical concentrations were quantified in water samples collected in open water at the deepest point of each lake (up to 50 m and in the mid-point of reservoirs). At this location, water was sampled using a vertical, depth-integrated methodology that collected 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-litre (L) cubitainer. When full, the cubitainer was gently inverted to mix the water, and subsamples were poured off to obtain a MC sample and a water chemistry sample. After collection, water samples for chemical analytes were placed on ice and shipped overnight to the Willamette Research Station in Corvallis, Oregon, which quantified TN, total P (TP), true colour, dissolved organic carbon (DOC), acid-neutralising capacity (ANC), specific conductivity (Cond) and chlorophyll *a* (chl *a*) concentrations at pre-specified levels of precision and accuracy (US EPA, 2006). Microcystin samples were placed on ice and shipped overnight to the U.S. Geological Survey Organic Geochemistry Research Laboratory in Lawrence, Kansas.

Microcystin sample processing began with three sequential freeze/thaw cycles to lyse cyanobacteria (Loftin *et al.*, 2008). Processed samples were filtered using 0.45-μm polyvinylidene diflouride membrane syringe filters and stored frozen until analysis. The concentration of MC in the filtered water sample was measured with a polyclonal enzyme-linked immuno-sorbent assay

(ELISA) using an Abraxis kit (Abraxis LLC, Warminster, PA, USA) for microcystin-ADDA. The binding mechanism of the microcystin-ADDA assay is specific to the MCs, nodularins and their congeners; therefore, results from this assay may include contributions from any compound within the ADDA functional group (Fischer *et al.*, 2001). The minimum reporting level for the assay was 0.1 μg L^{-1} as microcystin-LR. Owing to the open water sample location and the single sample collection, the NLA data set is likely to be a conservative representation of the occurrence of high MC concentrations in different lakes (Chorus, 2001).

We examined whether different environmental factors strongly affected the occurrence of MC by including variables in the analysis that quantified each factor (Table 1). Nutrient concentrations and ratios were quantified as TN, TP and the mass ratio between the two (TN : TP). Algal biomass was quantified as chl *a* concentration. Light availability was approximated using Secchi depth (SD) and water colour (quantified as platinum cobalt units). The latitude and longitude of each lake were also included as variables to account for unmeasured factors that were associated with geographic position. Latitude also potentially provided an alternate, albeit coarse, means of quantifying light availability. Sampling day of the year (DOY) was included as a possible explanatory variable because the day of sampling could potentially provide a means of quantifying seasonal trends in MC occurrence (Cronberg, Annadotter & Lawton, 1999) and seasonal variations in light availability and temperature. Maximum temperature along a vertical profile in the lake at the time of sampling (Tmax) was included to account for the possible effects of water temperature, while the maximum difference between temperatures along the profile (Tdiff) was used to quantify the strength of stratification. Alkalinity was quantified as ANC and conductivity. Depth was included as a candidate variable to characterize lake morphology and its potential effects on lake stratification and MC occurrence. Finally, DOC was included because of its role in controlling lake biological assemblages (Drakare *et al.*, 2002).

Statistical analysis

All water chemistry variables (TP, TN, ANC, DOC, Cond, Colour and TN : TP), chl *a*, depth and SD were log-transformed prior to further analysis to reduce the skewness of their distributions.

To explore the bivariate relationship between the exceedance of the WHO provisional guideline of MC ≥ 1 μg L^{-1} and each environmental variable, we computed the mean difference in the environmental variable value between lakes in which MC ≥ 1 μg L^{-1} and lakes in which MC < 1 μg L^{-1} and assessed the statistical significance of this difference using a two-sample *t*-test. We further examined the degree to which a linear model accurately characterised the relationship between each environmental variable and the probability of MC ≥ 1 μg L^{-1} using logistic regressions and nonparametric curves (Wood & Augustin, 2002) to estimate the effect of each environmental variable on the probability MC ≥ 1 μg L^{-1}. The proportion of deviance explained by each linear logistic regression was compared to the proportion of deviance explained by the nonparametric model using the same explanatory variable, and the difference between the two indicated the degree to which a linear model accurately represented the observed relationship between the explanatory variable and the probability of MC ≥ 1 μg L^{-1}.

Table 1 Summary of explanatory variables

Variable	Abbreviation	Units	Min	Max	Mean	Median
Acid-neutralising capacity	ANC	μeq L^{-1}	−63	91 630	2686	1797
Chlorophyll *a*	Chl *a*	μg L^{-1}	0.1	936	29.6	7.7
Colour	Colour	PCU	0	165	16.3	12
Conductivity	Cond	μS cm^{-1}	4	50 590	701	250
Day of year	DOY	NA	127	290	214	212
Depth	Depth	m	0.5	97	9.5	6.0
Dissolved organic carbon	DOC	mg L^{-1}	0.3	290	8.8	5.6
Nitrogen, total	TN	μg L^{-1}	5	26 100	1175	586
TN : TP ratio	TN : TP	NA	0.23	1200	34.1	21.1
Phosphorus, total	TP	μg L^{-1}	1	4679	109	29
Secchi depth	SD	m	0.04	36.7	2.2	1.4
Maximum profile temperature	Tmax	°C	10	38	24.3	24.7
Maximum difference in profile temperature	Tdiff	°C	0	25.1	9.5	6

We identified the subset of environmental variables that best predicted the probability of MC $\geq$ 1 μg L^{-1} using a least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996). Least absolute shrinkage and selection operator regression identifies parsimonious predictive models by gradually reducing (i.e. 'shrinking') the absolute value of regression coefficients such that the sum of all coefficients is less than a pre-specified threshold. After shrinking, coefficients for some explanatory variables are set to zero, and these variables no longer contribute to the model. Least absolute shrinkage and selection operator regression has been shown to yield models with less bias than stepwise regression when a few, moderately valued regression coefficients are expected (Tibshirani, 1996). It is used infrequently in ecology (Dahlgren, 2010), but has been applied extensively in other fields such as bioinformatics (Laaksonen *et al.*, 2006; Lu *et al.*, 2011) and medicine (Steyerberg, Eijkemans & Habbema, 2001).

In LASSO regression, explanatory variables were all standardised to a mean value of zero and a standard deviation of 1, and all variables were initially included in a logistic regression model predicting the probability of MC $\geq$ 1 μg L^{-1}. Then, a constraint was imposed on the sum of all regression coefficients at a predetermined threshold, t, and the regression model was refit subject to this constraint. Different values of t produced different models, and as t decreased in value, fewer explanatory variables contributed to the model. Cross-validation was used to identify the value of t that yielded a model that was both accurate and parsimonious. We used a 30-fold cross-validation procedure to select the best value of t, using the R library, glmnet (Friedman, Hastie & Tibshirani, 2010).

Since our response variable is binary (i.e. MC is either $\geq$1 μg L^{-1} or MC is <1), we quantified the predictive accuracy of our model by computing the area under the receiver operating characteristic curve (AUC). Values of AUC range from a low value of 0.5 (i.e. the model is no better than a random guess) to a high value of 1.0 (i.e. the model predicts the binary response perfectly).

We used the results of the LASSO regression to explore approaches for setting thresholds for the environmental variables identified by the model (chl *a* and TN). Using this final model, threshold concentrations of chl *a* and TN can be derived by identifying a frequency with which MC exceeding 1 μg L^{-1} that is acceptable. Because the mean probability of MC $\geq$ 1 μg L^{-1} is simultaneously a function of both TN and chl *a*, the threshold value for one of these parameters depends on the threshold selected for the other, and an infinite number of combinations of values can yield the desired frequency of MC $\geq$ 1 μg L^{-1}. To illustrate one approach for selecting numeric threshold values from this infinite set of possibilities, we first selected two examples for desired frequencies of occurrence (0.1 and 0.05). We then used quantile regression to estimate the 5th percentile of observed chl *a* concentration as a linear function of TN. This relationship approximated the lower boundary of the distribution of sampled values of chl *a* and TN. The points along this lower boundary that corresponded with the desired frequencies of occurrence of MC $\geq$ 1 μg L^{-1} then yielded estimates of the largest possible TN threshold and the corresponding chl *a* threshold.

We computed a second set of possible thresholds by identifying the largest chl *a* concentration and associated TN concentration that would yield the desired frequency of occurrence. In this case, quantile regression was used to estimate the 95th percentile of observed chl *a* concentrations as a linear function of TN. This relationship approximated the upper boundary of the distribution of samples values of chl *a* and TN. The points along the upper boundary that corresponded with the desired frequencies of MC $\geq$ 1 μg L^{-1} then yielded another set of TN and chl *a* thresholds.

Table 2 Summary of mean and 95% confidence limits of the difference in the value of the indicated variable between sites with MC $\geq$ 1 μg L^{-1} and sites with MC < 1 μg L^{-1}

Variable	Mean	95% CL	Proportion deviance explained (linear logistic)	Proportion deviance explained (nonparametric model)
log(ANC)	0.44	0.36, 0.52	0.09	0.10
log(Chl *a*)	0.74	0.63, 0.84	0.18	0.18
log(Colour)	0.22	0.16, 0.27	0.05	0.06
log(Cond)	0.42	0.33, 0.51	0.08	0.09
log(Depth)	−0.29	−0.35, −0.22	0.07	0.07
log(DOC)	0.36	0.3, 0.43	0.14	0.16
DOY	3.83	−0.77, 8.42	0.00	0.01
Latitude	2.08	1.25, 2.92	0.03	0.04
Longitude	−0.57	−2.44, 1.29	0.00	0.05
log(SD)	−0.42	−0.49, −0.34	0.11	0.12
Tdiff	−3.48	−4.48, −2.49	0.05	0.06
Tmax	−0.52	−1.16, 0.13	0.00	0.02
log(TN)	0.58	0.51, 0.66	0.23	0.24
log(TN : TP)	−0.12	−0.19, −0.06	0.01	0.02
log(TP)	0.67	0.57, 0.78	0.15	0.16

Abbreviations as defined in Table 1.
Mean differences are significantly different from zero with $P < 0.05$ when the 95% confidence limits do not include zero. Proportion deviance explained: deviance explained by linear logistic and nonparametric regression fits to the relationship between the indicated variable and the probability of MC $\geq$ 1.

Results

A total of 1168 sampling events were carried out in 1077 different lakes (*c.* 10% of lakes were sampled more than once). Microcystin was greater than the detection limit of 0.1 μg L^{-1} in 32% of the 1168 samples and in 12% of the samples, MC was ≥1 μg L^{-1}. The maximum MC concentration measured was 225 μg L^{-1}.

The mean values for the majority of the environmental variables differed significantly across lakes with and without MC ≥ 1 μg L^{-1} (Table 2). The only variables that were not significantly different across the two groups of lakes were DOY, Tmax and longitude (i.e. the 95% confidence limits on the mean difference for these three variables included zero). Of the remaining variables, depth, SD and TN : TP were significantly lower in lakes with MC ≥ 1 μg L^{-1} compared to lakes with MC < 1 μg L^{-1}. All other variables were significantly greater in lakes with MC occurring above the WHO threshold.

Relationships between most of the explanatory variables and the probability of high MC were monotonically increasing or decreasing, except for sampling day, maximum temperature and longitude (see plots in Supporting information). The relationship between longitude and the probability of MC ≥ 1 μg L^{-1} was the only model for which the nonparametric curve substantially improved the proportion of deviance explained relative to the linear model (Table 2). Among the explanatory variables, only chl *a*, DOC, SD, TN and TP accounted for >10% of the observed deviance in the response variable.

Plots of the relationships between a subset of the explanatory variables and the probability of high MC

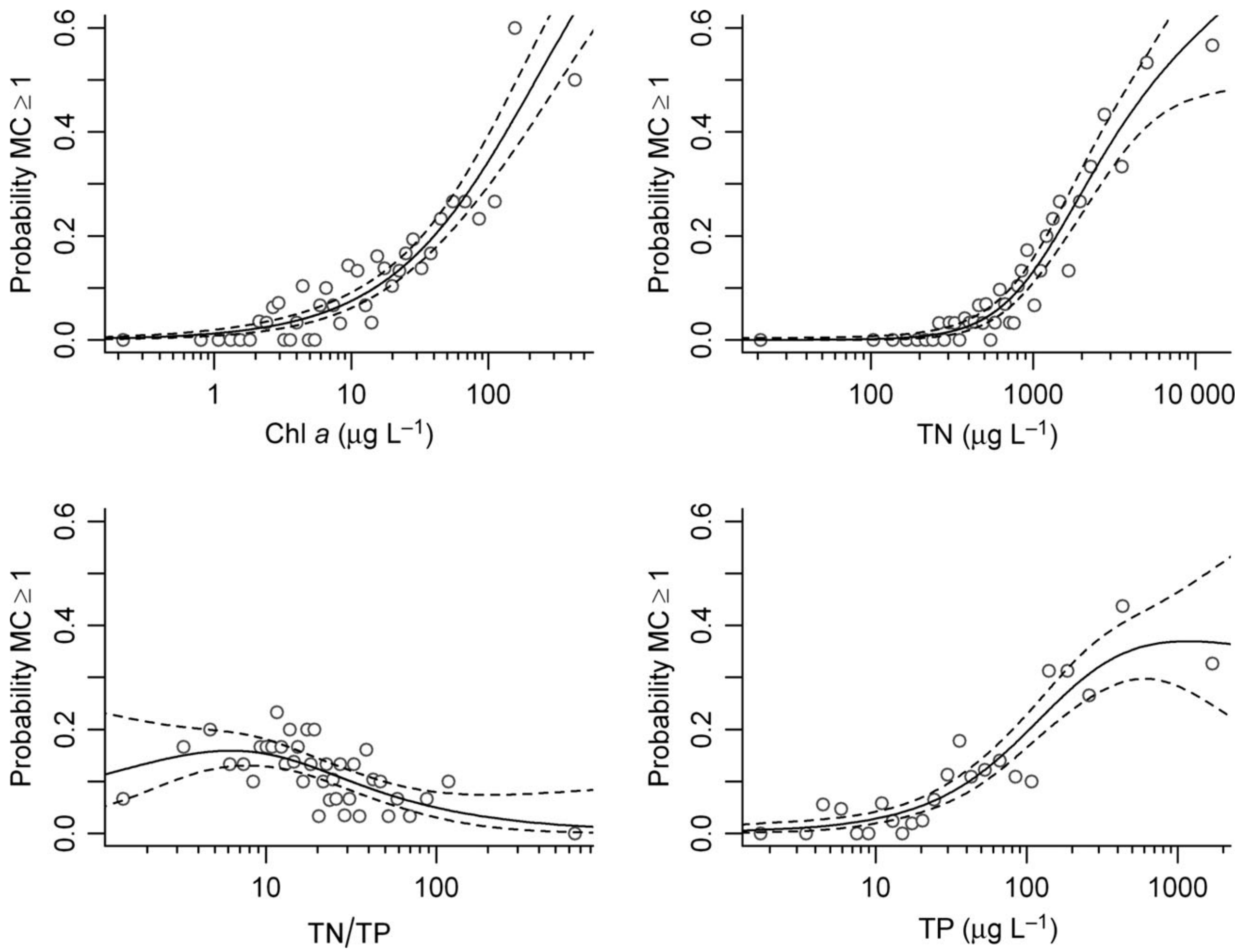

Fig. 1 Selected examples of nonparametric fit relating indicated explanatory variable and probability of microcystin ≥ 1 μg L^{-1}. Solid line: mean relationship; dashed line: estimated 90% confidence limits; open circles: average probability of occurrence in *c.* 30 samples in window centred about indicated value of the explanatory variable. Log scale shown on all *x* axes.

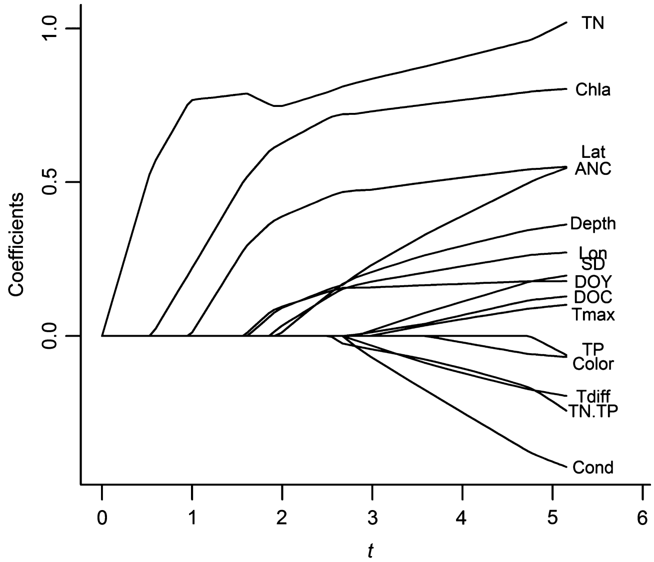

Fig. 2 Profiles of regression coefficient values for different explanatory variables and different values of constraint, *t*, during least absolute shrinkage and selection operator regression.

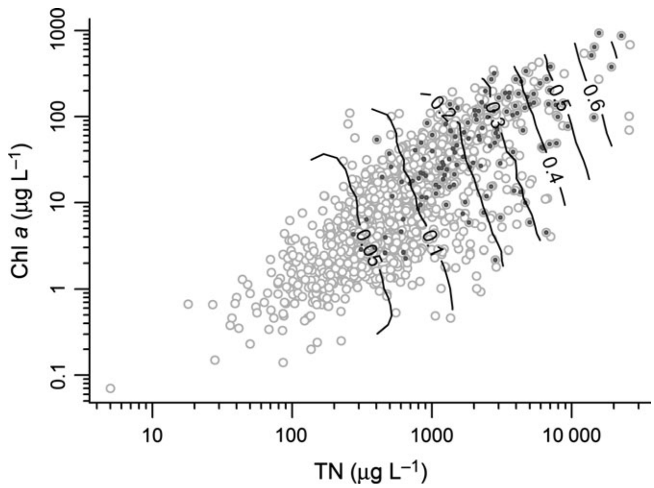

Fig. 3 Relationship between total nitrogen (TN) and chl *a*. Open circles show all sampled values of TN and chl *a*. Grey filled circles indicate lakes in which microcystin (MC) $\geq$1 μg L^{-1}. Solid lines indicate predicted mean probability of MC $\geq$ 1 μg L^{-1} from regression model.

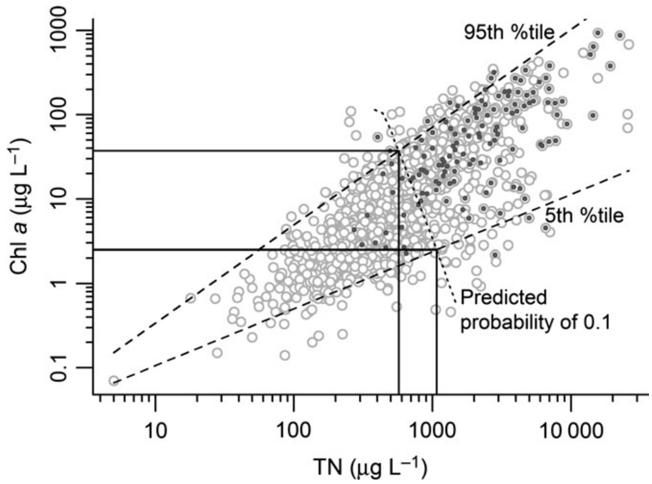

Fig. 4 Illustration of approach for deriving management thresholds for total nitrogen (TN) and chl *a* to maintain probability of microcystin (MC) $\geq$1 μg L^{-1} $\leq$0.1. Dashed line show quantile regressions for the 5th and 95th percentiles of chl *a* values, given TN concentrations. Diagonal dotted line shows combinations of TN and chl *a* for which mean predicted probability of MC $\geq$ 1 μg L^{-1} is 0.1. Horizontal and vertical solid lines show threshold values for TN and chl *a* corresponding with the estimated limits of the data. Open and filled circles as in Fig. 3.

concentrations showed strong increases in the probability with increased concentrations of TN, TP and chl *a*, while the effects of changes in TN : TP were comparatively weak (Fig. 1).

Profiles of the value of different coefficients with a range of values of *t* identified roughly three groups of variables (Fig. 2). The first group of variables (Cond, TN : TP, Tdiff, Tmax, TP, DOC, colour and SD) was reduced to zero quickly by the LASSO regression, indicating that they exerted very weak effects on the probability of high MC concentrations. A second group of explanatory variables (ANC, depth, longitude and DOY) was reduced to zero values somewhat more slowly. The results of cross-validation indicated that the best model, balancing predictive accuracy and parsimony, was achieved with only three variables (TN, chl *a* and latitude). Relative to TN and chl *a*, the magnitude of the regression coefficient for latitude was very small, and its effect was nearly negligible.

The final model provided accurate predictions of lakes in which MC $\geq$ 1 μg L^{-1} (AUC = 0.85). The frequency of occurrence of high MC concentrations depended most strongly on TN, with the greatest change in frequencies associated with increases in TN (Fig. 3). A weaker association between MC $\geq$ 1 μg L^{-1} and chl *a* was also observed, and the effects of latitude are barely perceptible in the plotted contours, manifested as small 'wiggles' in the otherwise straight contour lines.

Quantile regression estimates of the 5th and 95th percentiles of chl *a* as a function of TN quantified the range of possible combinations of TN and chl *a* (Fig. 4). Then, thresholds associated with the limits of this range and a desired frequency of occurrence of 0.1 for MC $\geq$ 1 μg L^{-1} were computed as TN concentrations of 570 or 1100 μg L^{-1}. These TN concentrations were paired with chl *a* concentrations of 37 and 3 μg L^{-1}, respectively. Repeating the same analysis for a desired frequency of occurrence of 0.05 gave TN concentrations of 250 or 400 μg L^{-1}, which were paired with chl *a* concentrations of 14 and 1 μg L^{-1}.

Discussion

The linkages between different environmental variables and increased concentrations of cyanotoxins are not well characterised but known to be complex. Different cyanobacteria species and strains of species favour different environmental conditions (Reynolds, 1998; Imai *et al.*, 2009), and each of these taxa differs in whether or not they produce toxins (Zurawell *et al.*, 2005). Furthermore, conditions that stimulate production of toxins often differ from conditions that support high cyanobacteria abundance (Sinang, Reichwaldt & Ghadouani, 2013). Hence, accurate predictions of the occurrence of high cyanotoxin concentrations in a particular lake would seem to require knowledge of the dominant cyanobacteria species and an understanding of conditions that stimulate those species to produce toxin. Despite these challenges, the present analysis has demonstrated that over the contiguous U.S., distinct patterns in the relationship between high MC concentrations and environmental conditions can be detected.

Other observational studies of MC occurrence conducted over large spatial areas provide findings that are comparable to the present work, and results from these studies generally conform to the results from the current analysis. In our analysis, we observed that the probability of occurrence of high MC concentrations increased with TN, TP and chl *a*, decreased with SD and decreased weakly with TN : TP (Table 1). Similar trends were observed in a survey of 30 shallow lakes in China, where changes in MC concentration were negatively associated with changes in SD and TN : TP and initially increased with TN and TP (Wu *et al.*, 2006). A different analysis of the same NLA data also suggested a strong association between high TN and high MC concentrations (Beaver *et al.*, 2014). Our finding of a positive association between MC occurrence and latitude was also observed in analysis of data collected from 241 lakes in the mid-west U.S. (Graham *et al.*, 2004). In this study, MC concentration also varied unimodally with TN and decreased with increasing SD (Graham *et al.*, 2004). Analysis of data collected from 246 lakes and reservoirs across Canada indicated that MC concentrations increased with TN and TP and tended to be higher at low values of TN : TP (Orihel *et al.*, 2012), findings similar to those of the present analysis.

Insights gained from simple bivariate associations can be sharpened by considering the simultaneous effects of multiple environmental variables. Bivariate analyses, as described above, do not provide a means of assessing whether observed associations exist because a particular predictor variable is correlated with another, stronger predictor. In the present analysis, LASSO regression indicated that when all environmental variables were considered together, TN and chl *a* were the best variables for predicting occurrences of high MC concentrations. Examples of comparable multivariate analyses of MC data across large regions are limited, but in Canadian lakes, TN was also identified from a suite of variables to be the strongest predictor of MC concentrations using classification and regression trees (Orihel *et al.*, 2012; Scott *et al.*, 2013).

The trends from multivariate analysis observed in the present study are also generally consistent with studies that provide stronger evidence of causal relationships between different environmental factors and MC occurrence than analyses of regional observational data. These studies include observational studies conducted within a single lake or reservoir, and field and laboratory manipulative studies. First, associations between MC concentrations and environmental conditions based on observations collected from a single lake or reservoir are less likely to be confounded by other variables because of the absence of across-lake differences in environmental conditions. In many of these studies, nitrogen (expressed in a variety of forms) and chl *a* have been strongly associated with MC concentration (Graham *et al.*, 2006; Yoshida *et al.*, 2007; Wang *et al.*, 2010; Te & Gin, 2011; Monchamp *et al.*, 2014). As one would expect, though, relationships vary across different lakes. For example, in some lakes, TP and temperature were most strongly associated with increased MC (Izydorczyk *et al.*, 2008; Rinta-Kanto *et al.*, 2009; Ni *et al.*, 2012), trends that we did not observe when we considered the effects of TP and temperature in the context of other predictor variables.

Manipulative studies conducted in the field and in the laboratory provide the strongest evidence of causal relationships, and findings from many of these studies are similar to those of the present analysis. For example, in a New York lake, nitrogen amendments were observed to increase cyanobacteria abundance and toxin concentrations in the late summer and early autumn (Gobler *et al.*, 2007). In the laboratory, evidence of a strong relationship between nitrogen availability and MC production by microcystis is emerging. For example, continuous and batch culture studies have shown strong positive correlations between MC concentration and both nitrate uptake and intracellular N concentrations (Orr & Jones, 1998; Downing *et al.*, 2005). Similarly, in fixed P or fixed N batch cultures, MC concentrations in *Microcystis aeruginosa* were found to be highly correlated with N concentrations and slightly negatively correlated with P concentrations (Lee *et al.*, 2000). Under conditions of high N and P concentrations, toxic strains of *Microcystis* grew better than non-toxic strains (Vézie *et al.*, 2002). Several other culture studies have documented positive correlations between cell MC concentration and chl *a* concentration (Lee *et al.*, 2000; Lyck, 2004). The results of our study echo these experimentally determined correlations between nitrogen, chl *a* and MC. As with the single lake studies, manipulative experiments have yielded some results that do not fully agree with those of the present analysis. For example, some studies found that increased P concentrations caused the greatest increases in MC production although concentrations of MC in the growth media were associated more strongly with N concentration (Rapala *et al.*, 1997).

Key variables thought to trigger cyanotoxin production, such as temperature and light availability, were not found to be important in predicting MC occurrences in the current study (Zurawell *et al.*, 2005; Davis *et al.*, 2009). This difference can likely be attributed to the broad spatial extent and synoptic sampling of the NLA, in which detection of conditions that trigger cyanotoxin production would be very difficult. That is, the current data set was collected for the purpose of understanding the distribution of conditions in space, and therefore, we would not expect to be able to estimate short-term temporal changes (i.e. daily or weekly) from the data. Coupling analyses of broad-scale data such as the NLA with intensive studies from selected lakes would likely yield deeper insights into factors that control MC concentrations that vary in space and time.

The comparisons between the current, broad-scale analysis, other similar regional scale observational studies, and smaller scale laboratory manipulations and local field studies support our finding that increased concentrations of nitrogen can increase the likelihood of observing high concentrations of MC. The additional predictive power of chl *a* concentration is also reasonable, given the presence of chl *a* in cyanobacteria. Overall, our results align with the current paradigm that nutrient pollution and cultural eutrophication in freshwater systems are strongly associated with MC occurrence. Based on our results, we assert that management of surface freshwater TN and chl *a* should be effective in managing cyanotoxin concentrations.

Least absolute shrinkage and selection operator regression provided a robust approach for specifying a predictive model from many candidate explanatory variables. Least absolute shrinkage and selection operator is known to be effective when explanatory variables are correlated (Dormann *et al.*, 2013; Hebiri & Lederer, 2013), and so is particularly well-suited to the present data set, in which a few pairs of variables (e.g. TP and TN, TN and DOC, and TN and chl *a*) are strongly correlated ($r > 0.7$). One limitation of LASSO, as applied here, is that it only allows variables to be modelled with linear relationships. Fortunately, in this data set, linear models reasonably represented the observed relationships for the majority of the candidate variables (Table 2). The one exception was the relationship between longitude and the probability of high MC. In this case, the relationship estimated using a nonparametric curve was strongly unimodal, and the omission of longitude from the selected list of variables may have been an artefact of the restriction to linear models. However, additional analysis (not shown here) suggests that the unimodal relationship between longitude and the probability of high MC arises from the fact that TN is higher in the middle longitudes of the U.S., corresponding to the agricultural mid-west region of the U.S. Indeed, when TN and longitude were included in the same regression model to predict the probability of high MC, longitude did not contribute significantly to the model.

The current analysis indicated that thresholds for TN and chl *a* may help protect against frequent occurrence of high MC. Thresholds associated with a desired frequency of occurrence of 0.1 for MC ≥ 1 μg L^{-1} were 570 or 1100 μg L^{-1} for TN. These TN concentrations were paired with chl *a* concentrations of 37 and 3 μg L^{-1}, respectively. As expected, reducing the desired frequency of occurrence to 0.05 also reduces the threshold concentrations for TN (250 and 400 μg L^{-1}) and chl *a* (14 and 1 μg L^{-1}). While this method for establishing thresholds is robust, three issues should be considered before applying these values.

First, the selection of the desired frequency of occurrence of MC ≥ 1 μg L^{-1} is a risk management decision. Our selections of 0.1 and 0.05 are illustrative, and for certain lakes that are the primary drinking water source for a community, exceedance of the 1 μg L^{-1} threshold 10 or 5% of the time would be too frequent. In other communities, drinking water treatment may reduce the concentrations of MC and allow for higher frequencies of high MC in the source water. The benefit of the current analysis resides in the strong connection between nutrient and chl *a* concentrations and an easily understood environmental outcome. Given this strong, quantitative relationship, the relative costs and benefits of controlling nutrient concentrations can be easily weighed against one another.

Second, our analysis yielded a model in which the response of interest (frequencies of occurrence of 0.1 or 0.05 for MC ≥ 1 μg L^{-1}) is predicted to be a function of two environmental variables, and hence, the threshold for one variable depends on the value selected for the other. We provided one example of an approach to illustrate a few of the challenges that arise when deriving thresholds from a multivariate relationship. The approach shown here uses the limits of the observed combinations of TN and chl *a* to select pairs of threshold values that are consistent with one another, and in general, the threshold values for each parameter should not be considered separately. For example, without knowledge of chl *a* concentration, the threshold value of 1100 μg L^{-1} for TN might not be stringent enough to achieve the desired frequency of high MC concentrations. Also, the parameters of the quantile regression used to approximate the upper and lower boundaries of the range of possible values of TN and chl *a* can be selected to suit different management goals. A combination of TN and chl *a* values that is more closely located to the centre of the range of observed values (e.g. the 50th percentile) may provide thresholds that are more broadly applicable to the region. Other methods for defining thresholds from multivariate models have been proposed as well (Malve & Qian, 2006; Lamon, Malve & Pietiläinen, 2008). Ultimately, as noted above, the selection of thresholds is a risk management decision, and the approach used for accounting for the combined effects of two variables can be incorporated into this decision.

Third, other environmental variables may be more or less important at local or regional spatial scales. As noted earlier, variables such as water temperature and light availability that were not found to be predictive at the broad spatial scale of the current study could very possibly be important at smaller spatial scales. The absence of TP from the final predictive equation, in particular, merits further discussion, as TP limits primary productivity in many freshwater systems, and increased concentrations of TP have been associated with increased cyanobacteria abundance and MC concentrations in other studies (e.g. Dolman *et al.*, 2012). However, as noted earlier, many other manipulative studies have found that increased nitrogen or combined increases of nitrogen and phosphorus cause increases in MC concentration (e.g. Rapala *et al.*, 1997; Paerl *et al.*, 2011). Interpretation of other analyses of large-scale field-collected MC data is currently inconclusive as these other studies have rarely considered the simultaneous effects of both TN and TP in multivariate models, hindering our ability to quantify the relative contributions of the two nutrients. Indeed, similar to other large-scale field studies, we also observed strong bivariate relationships between TP and the occurrence of high MC. So overall, scientific understanding of the relationships between nutrients and MC concentrations is still evolving, and the results of the current analysis further expand this knowledge base.

From a management perspective, the lack of an effect of TP does not imply that TP thresholds are not necessary. Instead, TP simply does not improve the predictive accuracy for high MC occurrences at the broad spatial scale of the current study. Furthermore, the presence of chl *a* as a predictor for high MC suggests that controlling TP and TN could reduce the frequency of high MC events by reducing the biomass of cyanobacteria in the system. Indeed, the results of the current analysis provide further support for the need for controlling both nitrogen and phosphorus loading to lakes and reservoirs (Bernhardt, 2013; Finlay, Small & Sterner, 2013). Further site- or region-specific studies may also help define an appropriate TP threshold specific to the control of the occurrence of high MC concentrations.

While the considerations discussed above are important, the thresholds identified here are useful because they have been derived from a national-scale dataset and, as such, are applicable throughout the contiguous U.S. In other regions of the world, these thresholds can complement other thresholds that have been developed to facilitate the management of lakes and reservoirs (e.g. Solheim *et al.*, 2008). Lake-specific analysis will almost always yield thresholds that more accurately predict responses in a particular lake, but since MC data are only available from a small set of locations, lake-specific analyses are frequently not available for informing management decisions. In the absence of lake-specific data, the predictive model described here provides a useful approach for setting appropriate thresholds.

Acknowledgments

The authors are grateful to T. Crawford, D. Thomas, E. Behl, B. Walsh and two anonymous reviewers for comments that improved the manuscript. The authors also thank Charles P. Hawkins for providing catchment summaries of climatic variables. The views expressed in this paper are those of the authors and do not reflect the official policy of the U.S. Environmental Protection Agency.

References

Azevedo S.M.F., Carmichael W.W., Jochimsen E.M., Rinehart K.L., Lau S., Shaw G.R. *et al.* (2002) Human intoxication by microcystins during renal dialysis treatment in Caruaru—Brazil. *Toxicology*, 181–182, 441–446.

Beaver J.R., Manis E.E., Loftin K.A., Graham J.L., Pollard A.I. & Mitchell R.M. (2014) Land use patterns, ecoregion, and microcystin relationships in U.S. lakes and reservoirs: a preliminary evaluation. *Harmful Algae*, **36**, 57–62.

Bernhardt E.S. (2013) Cleaner lakes are dirtier lakes. *Science*, **342**, 205–206.

Carlisle D.M., Wolock D.M. & Meador M.R. (2010) Alteration of streamflow magnitudes and potential ecological consequences: a multiregional assessment. *Frontiers in Ecology and the Environment*, **9**, 264–270.

Carmichael W. (2008) A world overview — one-hundred-twenty-seven years of research on toxic cyanobacteria — where do we go from here? In: *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs*. Advances in Experimental Medicine and Biology (Ed. H.K. Hudnell), pp. 105–125. Springer, New York, NY.

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.

Chorus I. (2001) *Cyanotoxins: Occurrence, Causes, Consequences*. Springer, Berlin.

Cronberg G., Annadotter H. & Lawton L.A. (1999) The occurrence of toxic blue-green algae in Lake Ringsjön, southern Sweden, despite nutrient reduction and fish biomanipulation. *Hydrobiologia*, **404**, 123–129.

Dahlgren J.P. (2010) Alternative regression methods are not considered in Murtaugh (2009) or by ecologists in general. *Ecology Letters*, **13**, E7–E9.

Davidson B. & Bradshaw R.W. (1967) Thermal pollution of water systems. *Environmental Science & Technology*, **1**, 618–630.

Davis T.W., Berry D.L., Boyer G.L. & Gobler C.J. (2009) The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of *Microcystis* during cyanobacteria blooms. *Harmful Algae*, **8**, 715–725.

Davis T.W., Harke M.J., Marcoval M.A., Goleski J., OranoDawson C., Berry D.L. *et al.* (2010) Effects of nitrogenous compounds and phosphorus on the growth of toxic and non-toxic strains of *Microcystis* during cyanobacterial blooms. *Aquatic Microbial Ecology*, **61**, 149–162.

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.

Dormann C.F., Elith J., Bacher S., Buchmann C., Carl G., Carré G. *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

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

Downing T.G., Meyer C., Gehringer M.M. & van de Venter M. (2005) Microcystin content of *Microcystis aeruginosa* is modulated by nitrogen uptake rate relative to specific growth rate or carbon fixation rate. *Environmental Toxicology*, **20**, 257–262.

Drakare S., Blomqvist P., Bergström A.-K. & Jansson M. (2002) Primary production and phytoplankton composition in relation to DOC input and bacterioplankton production in humic Lake Örträsket. *Freshwater Biology*, **47**, 41–52.

Finlay J.C., Small G.E. & Sterner R.W. (2013) Human influences on nitrogen removal in lakes. *Science*, **342**, 247–250.

Fischer W.J., Garthwaite I., Miles C.O., Ross K.M., Aggen J.B., Chamberlin A.R. *et al.* (2001) Congener-independent immunoassay for microcystins and nodularins. *Environmental Science & Technology*, **35**, 4849–4856.

Friedman J., Hastie T. & Tibshirani R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Gobler C.J., Davis T.W., Coyne K.J. & Boyer G.L. (2007) Interactive influences of nutrient loading, zooplankton grazing, and microcystin synthetase gene expression on cyanobacterial bloom dynamics in a eutrophic New York lake. *Harmful Algae*, **6**, 119–133.

Graham J.L., Jones J.R., Jones S.B. & Clevenger T.E. (2006) Spatial and temporal dynamics of microcystin in a Missouri reservoir. *Lake and Reservoir Management*, **22**, 59–68.

Graham J.L., Jones J.R., Jones S.B., Downing J.A. & Clevenger T.E. (2004) Environmental factors influencing microcystin distribution and concentration in the Midwestern United States. *Water Research*, **38**, 4395–4404.

Hebiri M. & Lederer J. (2013) How correlations influence Lasso prediction. *IEEE Transactions on Information Theory*, **59**, 1846–1854.

Imai H., Chang K.-H., Kusaba M. & Nakano S. (2009) Temperature-dependent dominance of *Microcystis* (Cyanophyceae) species: *M. aeruginosa* and *M. wesenbergii*. *Journal of Plankton Research*, **31**, 171–178.

Izydorczyk K., Jurczak T., Wojtal-Frankiewicz A., Skowron A., Mankiewicz-Boczek J. & Tarczyńska M. (2008) Influence of abiotic and biotic factors on microcystin content in *Microcystis aeruginosa* cells in a eutrophic temperate reservoir. *Journal of Plankton Research*, **30**, 393–400.

Laaksonen R., Katajamaa M., Päivä H., Sysi-Aho M., Saarinen L., Junni P. *et al.* (2006) A systems biology strategy reveals biological pathways and plasma biomarker candidates for potentially toxic statin-induced changes in muscle. *PLoS One*, **1**, e97.

Lamon E.C. III, Malve O. & Pietiläinen O.-P. (2008) Lake classification to enhance prediction of eutrophication endpoints in Finnish lakes. *Environmental Modelling & Software*, **23**, 938–947.

Lee S.J., Jang M.-H., Kim H.-S., Yoon B.-D. & Oh H.-M. (2000) Variation of microcystin content of *Microcystis aeruginosa* relative to medium N:P ratio and growth stage. *Journal of Applied Microbiology*, **89**, 323–329.

Loftin K.A., Meyer M., Rubio F., Kamp L., Humphries E. & Whereat E. (2008) *Comparison of two cell lysis procedures for recovery of microcystins in water samples from Silver Lake in Dover, Delaware, with microcystin producing cyanobacterial accumulations*. United States Geological Survey.

Lu Y., Zhou Y., Qu W., Deng M. & Zhang C. (2011) A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, **27**, 2406–2413.

Lyck S. (2004) Simultaneous changes in cell quotas of microcystin, chlorophyll a, protein and carbohydrate during different growth phases of a batch culture experiment with *Microcystis aeruginosa*. *Journal of Plankton Research*, **26**, 727–736.

Malve O. & Qian S.S. (2006) Estimating nutrients and chlorophyll a relationships in Finnish Lakes. *Environmental Science & Technology*, **40**, 7848–7853.

Monchamp M.-E., Pick F.R., Beisner B.E. & Maranger R. (2014) Nitrogen forms influence microcystin concentration and composition via changes in cyanobacterial community structure. *PLoS One*, **9**, e85573.

Ni W., Zhang J., Ding T., Stevenson R.J. & Zhu Y. (2012) Environmental factors regulating cyanobacteria dominance and microcystin production in a subtropical lake within the Taihu watershed, China. *Journal of Zhejiang University Science A*, **13**, 311–322.

Orihel D.M., Bird D.F., Brylinsky M., Chen H., Donald D.B., Huang D.Y. *et al.* (2012) High microcystin concentrations occur only at low nitrogen-to-phosphorus ratios in nutrient-rich Canadian lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **69**, 1457–1462.

Orr P.T. & Jones G.J. (1998) Relationship between microcystin production and cell division rates in nitrogen-limited *Microcystis aeruginosa* cultures. *Limnology and oceanography*, **43**, 1604–1614.

Paerl H.W., Xu H., McCarthy M.J., Zhu G., Qin B., Li Y. *et al.* (2011) Controlling harmful cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): the need for a dual nutrient (N & P) management strategy. *Water Research*, **45**, 1973–1983.

Pilotto L.S. (2008) Epidemiology of cyanobacteria and their toxins. In: *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs* (Ed. H.K. Hudnell), pp. 639–649. Springer, New York, NY.

Poff N.L., Allan J.D., Bain M.B., Karr J.R., Prestegaard K.L., Richter B.D. *et al.* (1997) The natural flow regime. *BioScience*, **47**, 769–784.

Rapala J., Sivonen K., Lyra C. & Niemelä S.I. (1997) Variation of microcystins, cyanobacterial hepatotoxins, in *Anabaena* spp. as a function of growth stimuli. *Applied and Environmental Microbiology*, **63**, 2206–2212.

Reynolds C. (1998) What factors influence the species composition of phytoplankton in lakes of different trophic status? *Hydrobiologia*, **369–370**, 11–26.

Rinta-Kanto J.M., Konopko E.A., DeBruyn J.M., Bourbonniere R.A., Boyer G.L. & Wilhelm S.W. (2009) Lake Erie *Microcystis*: relationship between microcystin production, dynamics of genotypes and environmental parameters in a large lake. *Harmful Algae*, **8**, 665–673.

Scott J.T., McCarthy M.J., Otten T.G., Steffen M.M., Baker B.C., Grantz E.M. *et al.* (2013) Comment: an alternative interpretation of the relationship between TN:TP and microcystins in Canadian lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **70**, 1265–1268.

Sinang S.C., Reichwaldt E.S. & Ghadouani A. (2013) Spatial and temporal variability in the relationship between cyanobacterial biomass and microcystins. *Environmental Monitoring and Assessment*, **185**, 6379–6395.

Smith V.H. (2003) Eutrophication of freshwater and coastal marine ecosystems a global problem. *Environmental Science and Pollution Research*, **10**, 126–139.

Solheim A.L., Rekolainen S., Moe S.J., Carvalho L., Phillips G., Ptacnik R. *et al.* (2008) Ecological threshold responses in European lakes and their applicability for the Water Framework Directive (WFD) implementation: synthesis of lakes results from the REBECCA project. *Aquatic Ecology*, **42**, 317–334.

Stewart I., Seawright A.A. & Shaw G.R. (2008) Cyanobacterial poisoning in livestock, wild mammals and birds – an overview. In: *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs*. Advances in Experimental Medicine and Biology (Ed. H.K. Hudnell), pp. 613–637. Springer, New York, NY.

Steyerberg E.W., Eijkemans M.J.C. & Habbema J.D.F. (2001) Application of shrinkage techniques in logistic regression analysis: a case study. *Statistica Neerlandica*, **55**, 76–88.

Te S.H. & Gin K.Y.-H. (2011) The dynamics of cyanobacteria and microcystin production in a tropical reservoir of Singapore. *Harmful Algae*, **10**, 319–329.

Tibshirani R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

US EPA. (2006) *Survey of the Nation's Lakes. Laboratory Methods Manual*. Office of Water and Office of Research and Development, Washington, DC.

US EPA. (2010) *National Lakes Assessment: A Collaborative Survey of the Nation's Lakes*. Office of Water and Office of Research and Development, Washington, DC.

Vézie C., Rapala J., Vaitomaa J., Seitsonen J. & Sivonen K. (2002) Effect of nitrogen and phosphorus on growth of toxic and nontoxic Microcystis strains and on intracellular microcystin concentrations. *Microbial Ecology*, **43**, 443–454.

Wang Q., Niu Y., Xie P., Chen J., Ma Z., Tao M. *et al*. (2010) Factors affecting temporal and spatial variations of microcystins in Gonghu Bay of Lake Taihu, with potential risk of microcystin contamination to human health. *The Scientific World Journal*, **10**, 1795–1809.

WHO. (1998) *Guidelines for Drinking-Water Quality – Second Edition – Volume 2 – Health Criteria and Other Supporting Information – Addendum*. World Health Organization, Geneva.

Wiedner C., Visser P.M., Fastner J., Metcalf J.S., Codd G.A. & Mur L.R. (2003) Effects of light on the microcystin content of Microcystis strain PCC 7806. *Applied and Environmental Microbiology*, **69**, 1475–1481.

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.

Wu S.K., Xie P., Liang G.D., Wang S.B. & Liang X.M. (2006) Relationships between microcystins and environmental parameters in 30 subtropical shallow lakes along the Yangtze River, China. *Freshwater Biology*, **51**, 2309–2319.

Yoshida M., Yoshida T., Takashima Y., Hosoda N. & Hiroishi S. (2007) Dynamics of microcystin-producing and non-microcystin-producing Microcystis populations is correlated with nitrate concentration in a Japanese lake. *FEMS Microbiology Letters*, **266**, 49–53.

Zurawell R., Chen H., Burke J. & Prepas E. (2005) Hepatotoxic cyanobacteria: a review of the biological importance of microcystins in freshwater environments. *Journal of Toxicology and Environmental Health Part B: Critical Reviews*, **8**, 1–37.

Supporting Information

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

Figure S1. Nonparametric fits relating indicated explanatory variable and probability of MC ≥ 1 μg L^{-1}.

(Manuscript accepted 18 May 2014)