Received: 29 December 2015 | Revised: 2 December 2016 | Accepted: 15 December 2016

DOI: 10.1111/geb.12569

RESEARCH PAPER

WILEY Global Ecology and Biogeography A Journal of Macroecology

Predicting microcystin concentrations in lakes and reservoirs at a continental scale: A new framework for modelling an important health risk factor

Zofia E. Taranu[1] | Irene Gregory-Eaves[2] | Russell J. Steele[3] | Marieke Beaulieu[4] | Pierre Legendre[1]

[1]Department of Biological Sciences, University of Montréal, Montréal, Québec H2V 2S9, Canada

[2]Department of Biology, McGill University, Montréal, Québec H3A 1B1, Canada

[3]Department of Mathematics and Statistics, McGill University, Montréal, Québec H3A 0B9, Canada

[4]Department of Civil Engineering, University of Sherbrooke, Sherbrooke, Québec J1K 2R1, Canada

Correspondence
Zofia Ecaterina Taranu, Department of Biological Sciences, University of Montréal, Montréal, Québec H2V 2S9, Canada.
Email: zofia.taranu@gmail.com

Editor: Martin Sykes

Abstract

Aim: Scientists, governments and non-governmental organizations are increasingly moving towards the collection of large, open-access data. In aquatic sciences, this effort is expanding the scope of questions and analyses that can be performed to further our knowledge of the global drivers of water quality. Cyanotoxin concentration is one variable that has received considerable attention, and although strong local-scale models have been described in the literature, modelling cyanotoxin concentrations across broader spatial scales has been more difficult. Commonly used statistical frameworks have not fully captured the complex response of toxic algal blooms to global change, limiting our ability to predict and mitigate the impairment of freshwaters by toxic algae. Here, we advance our understanding of emergent drivers of cyanotoxins across a structured landscape by applying a hierarchical "hurdle" model.

Location: Lakes and reservoirs in the conterminous United States [$n = 1127$].

Methods: We studied cyanobacteria and their toxins [microcystins] during the 2007 summer period. We applied a hierarchical zero-altered model to test the importance of multi-scale interactions among environmental features in driving microcystin concentrations above the limit of detection. We then used boosted regression trees [BRTs] to identify environmental thresholds associated with severe impairment by microcystins.

Results: Accounting for numerous non-detections, spatial heterogeneity and cross-scale interactions substantially improved continental-scale predictions of bloom toxicity. Our model accounted for 55% of the variance in the probability of detecting microcystins across the United States, and 26% of the variability in microcystin concentrations once detected. BRTs further showed that although both local and regional drivers were associated with microcystin concentrations at low to intermediate provisional guidelines, only local drivers came into play when predicting higher limits.

Main conclusions: Identifying the interaction between local and regional processes is key to understanding the heterogeneous responses of microcystins to environmental change. Our framework could increase the effectiveness of continental-scale analyses for many different water variables.

KEYWORDS

boosted regression tree [BRT], cross-scale interaction [CSI], cyanobacteria, ecoregion, eutrophication, lake, land use, microcystins, reservoir, zero-altered hurdle model

1 | INTRODUCTION

Human-induced global change is emerging as one of the greatest challenges of the 21st century with significant consequences for human health and ecosystem services. In particular, the stability and function of freshwater ecosystems, which provide critical water supplies and other services, are being increasingly threatened by climate warming, eutrophication and their symptomatic cyanobacterial blooms (Paerl & Paul, 2012). This on-going expansion of cyanobacteria in lakes across the globe (Taranu et al., 2015) is problematic because some genera produce neuro- and hepatotoxins ranging in their effects on human health from relatively minor to severe (Carmichael et al., 2001; Codd, Morrison, & Metcalf, 2005; Jonasson et al., 2010; Lévesque et al., 2014). In addition, cyanobacterial blooms can cause substantial economic losses, with management costs in areas with fully developed market economies reaching up to billions of dollars (Hunter et al., 2012). As a result, predicting when and where cyanotoxins will occur, and if concentrations will exceed guidelines for drinking water and recreational activities, is of increasing concern (Carvalho et al., 2013; Yuan, Pollard, Pather, Oliver, & D'Anglada, 2014).

Currently, important discrepancies in the predictive strengths of cyanotoxin models exist in the literature, which may be attributed to the timing of sampling, landscape heterogeneity or issues with messy data and the predominance of below-detection measurements. For instance, the rapid production of cyanobacteria and their dynamic response to changes in nutrient concentrations and ratios, light availability and weather conditions make it difficult to provide robust predictions of toxic bloom events (Kardinaal et al., 2007; Pimentel & Giani, 2013). Toxigenic cells are in some cases more abundant and/or toxic at the onset of a bloom (Davis, Berry, Boyer, & Gobler, 2009; Romo, Soria, Fernández, Ouahid, & Baró-Solá, 2013) or vary depending on the pool of potentially toxigenic genera in a given site (Monchamp, Pick, Beisner, & Maranger, 2014; Rolland, Bourget, Warren, Laurion, & Vincent, 2013). In addition to these sampling effects, the predictive strength of cyanotoxin models may vary with the spatial extent of the study. For instance, in a regional analysis [Canadian lakes situated within a spatial extent of 700 km; $n = 22$ lakes], Giani, Bird, Prairie, and Lawrence (2005) found that nutrient concentrations [total nitrogen] explained as much as 56% of the variance in microcystin [MC] concentrations, MCs being one of the most prevalent classes of cyanotoxins in the environment (Blàha, Babica, & Maršálek, 2009). In contrast, in a meta-analysis conducted across Canada [spatial extent *c.* 5000 km; $n = 246$ lakes], Orihel et al. (2012) found that nutrient concentrations [total nitrogen] explained far less variance in MC concentrations [15% explained]. Such differences in model fit when expanding the spatial extent from regional to continental scale may be due to greater heterogeneities in geology and climate, which could in turn blur the response of cyanobacteria to any one predictor [e.g., a confounding effect of low alkalinity; Carvalho et al., 2013]. Cross-scale interactions [CSIs], defined as patterns or processes at one scale that affect driver–response relationships taking place at a different scale (Peters, Bestelmeyer, & Turner, 2007; Soranno et al., 2014), may account for spatial heterogeneity in model performance. For instance, Fergus, Soranno, Cheruvelil, and Bremigan (2011) showed that CSIs between landscape features [regional land use and hydro-geomorphology] accentuated the effect of local land cover on nutrient concentrations in certain lakes. Similarly, cyanobacterial dominance may be driven by local changes in catchment land use or water residence time, which may in turn be modified by regional variations in precipitation or temperature. Failure to account for these different sources of variability and spatial dependence may bias model estimates and fit (Finley, 2011). Lastly, cyanotoxin models can be further biased by a skewed distribution (e.g., Beaver et al., 2014; Orihel et al., 2012), whereby a large proportion of observations fall below the detection limit.

At present it is unclear whether variability in model prediction across studies is the result of spatial hierarchy and/or the presence of a high proportion of values below the detection limit. This knowledge gap is likely to derive from the scattered nature of current knowledge, which largely focusing on narrow ranges in terms of spatial scale or environmental gradients. To expand from local to macro-scales, however, it is necessary to apply multi-scale research and novel statistical approaches to identify emergent properties and better characterize landscape structures (Cheruvelil, Soranno, Webster, & Bremigan, 2013; Heffernan et al., 2014). Here, we used a continental-scale data set to align heterogeneous sites along common environmental gradients and identify prominent local to regional drivers of toxic algal blooms across the conterminous United States. The framework presented herein also allowed us to control for the inflated number of observations below the detection limit, which in turn provides a clear understanding of cyanotoxin occurrence.

2 | METHODS

2.1 | Study site information

We analysed 1127 lakes, ponds and reservoirs from the 48 contiguous United States randomly selected by the U.S. Environmental Protection Agency (U.S. EPA, 2009) as part of their 2007 nationwide survey [National Lake Assessment, NLA]. To ensure an accurate representation of the reference population [*c.* 50 000 lakes in the conterminous United States], the NLA based their site selection on a state by lake size stratified probabilistic sampling design (Olsen, Stahl, Snyder, & Pitt, 2009; Peck et al., 2013). Lakes included in the survey met the criteria that the maximum depth exceed 1 m and that the surface area be at least 10^{-3} km^2 for lakes and 0.04 km^2 for reservoirs. The 2007 monitoring programme represents a complete lake trophic gradient [27% oligotrophic, 26% mesotrophic, 24% eutrophic and 24% hypereutrophic lakes and reservoirs] that spanned 11 U.S. EPA national nutrient water-quality ecoregions [aggregates of the original 84 Omernik level III ecoregions] (Herlihy et al., 2013; Omernik, 1987; Rohm, Omernik, Woods, & Stoddard, 2002) [see Figure S1(a) in the Supporting Information]. Ecoregions varied in soils, vegetation, climate and geology, and thus represent broad-scale patterns in ecosystem state and anthropogenic impact on water quality.

An important caveat of using the 2007 NLA lake set to predict the occurrence of toxic blooms is that most lakes examined in this study [91.5%] were sampled only once during the open-water season [June to October; Figure S1(b)]. With a single water sample collected from one pelagic station, we were unable to accurately characterize within-lake seasonal dynamics, and large cyanobacterial blooms or peaks in cyanotoxin production might have been missed in some instances (Håkanson, Bryhn, & Hytteborn, 2007). To test for temporal variability in detection of MCs, we examined lakes with two visits per summer [8.5% of the sites used in this study were visited twice; $n = 95$] and compared the proportion of sites with MC concentrations above the detection limit and whether concentrations above the detection limit were correlated between both sampling events.

2.2 | Limnological and landscape variables

To address our study objective, we selected local, lake-specific explanatory variables previously identified in the literature as potential drivers of pelagic MC concentrations (Beaver et al., 2014; Yuan et al., 2014). In particular, we tested the importance of variables measured within the water column (total nitrogen [TN], total phosphorus [TP], TN:TP ratio, chlorophyll *a* [Chl *a*], cyanobacterial biomass, alkalinity [acid-neutralizing capacity, ANC; conductivity], dissolved organic carbon [DOC], turbidity, colour, and surface water temperature), as well as site and catchment characteristics [maximum depth, lake origin, drainage ratio, percentage agricultural land cover]. We also evaluated the effect of time of sampling [day of the year] to control for any seasonality effect. Cyanobacterial biomass was estimated from cell density data according to the conversion model of Beaulieu, Pick, and Gregory-Eaves (2013). Lake origin is a binary classification indicating whether water bodies are natural lakes or human-made reservoirs, as determined by visual inspection of maps by the NLA field technicians.

To test for the presence of multi-scale effects, we evaluated the importance of regional explanatory variables. In particular, MCs were found to frequently exceed the WHO drinking water provisional guideline [1 μg L^{-1}] in agriculturally productive ecoregions of the north-central United States (Beaver et al., 2014). This suggests that differences in land use among ecoregions may have an overarching effect on local-scale dynamics. To quantify this, we delineated ecoregion polygons [ArcGIS©] (ESRI 2011) and measured land-cover percentages in each using data from the 1992 U.S. Geological Survey National Land Cover data set [USGS] (Homer, Huang, Yang, Wylie, & Coan, 2004) to test for interactions among ecoregion land-cover and local-scale variables [cross-scale interaction]. To control for unmeasured heterogeneity across the landscape [e.g., due to geology or climate], we also evaluated the importance of site location [latitude and longitude] and random effects that grouped lakes according to ecoregions and/or major USGS hydrological unit [HUC-2 watershed]. The ecoregion classification is as described above and the hydrological unit consisted of large drainage basins that divided the conterminous United States into 18 major geographical areas.

2.3 | Statistical analysis

2.3.1 | Hierarchical zero-altered model

Approximately two-thirds [68%] of the lakes sampled by the NLA in 2007 had MC concentrations below the detection limit [DL], resulting in a strongly right-skewed distribution [Figure 1]. To develop predictive models for a variable with this right-skewed distribution, we applied a hierarchical two-stage model (Brilleman et al., 2016; Thorson, Shelton, Ward, & Skaug, 2015; Zuur & Ieno, 2016) [Figure 2], which consisted of two parts: (1) a binomial generalized linear mixed model [binomial GLMM] based on the full data set modelling the "presence" [MC concentrations at or above the detection limit; $n = 362$] and "absence" [MC concentrations below the detection limit; $n = 787$] of MC in lakes and reservoirs; and (2) a DL-truncated log-link Gamma model or lognormal model, where all observations with MC concentrations below the detection limit were removed and MC concentrations in the remaining sites [$n = 362$] were modelled as a function of the selected environmental variables. This two-part model is commonly referred to as a "hurdle" model because, irrespective of the mechanisms causing an increase in the response variable, a hurdle must first be crossed before it is observed (Zuur & Ieno, 2016; Zuur, Ieno, Walker, Saveliev, & Smith, 2009). Ecologically, it is relevant to consider these two processes separately because predictors that determine the presence–absence of MCs can be different from those predicting its linear dynamics once observed. Statistically, ignoring the large number of observations below the detection limit is problematic as it could result in exaggerated estimates of the variance [over-dispersed data] and biased estimates of standard errors and other parameters (Lachenbruch, 2001, 2002; Moulton, Curriero, & Barroso, 2002; Moulton & Halsey, 1995, 1996; Zuur & Ieno, 2016).

It is important to note that we choose a zero-altered hurdle model (Zuur & Ieno, 2016), as opposed to a zero-inflated mixture model, to deal with the high occurrence of non-detections because we were interested in the probability of not measuring any detectable quantity of MCs versus measuring any quantity of MCs. In contrast, the aim of the zero-inflated mixture model is to discriminate between false and true zeros [i.e., the count process allows for zeros], which we feel is a question better suited to studies interested in the efficacy of current laboratory techniques and/or sampling designs. Furthermore, when dealing with continuous biomass data that has too many zeros [or too many non-detections], a distribution with inflated error such as the lognormal or Gamma is needed. However, these distributions do not allow for zero values [strictly positive]. Therefore, modelling the zeros separately from the non-zeros in a binomial–lognormal [ZALN] or binomial–Gamma [ZAG] hurdle model (e.g., Brilleman et al., 2016; Thorson et al., 2015) is recommended.

For each part of the hurdle model, we tested the importance of local, lake-specific variables and regional-level environmental variables as fixed effect drivers of MCs across the conterminous United States [see Appendix S1 for details on the model selection procedure]. To account for non-independence among lakes from geographically similar locations, we tested random effects [intercept and slope] for each level of the ecoregion and hydrological unit, respectively. The significance of

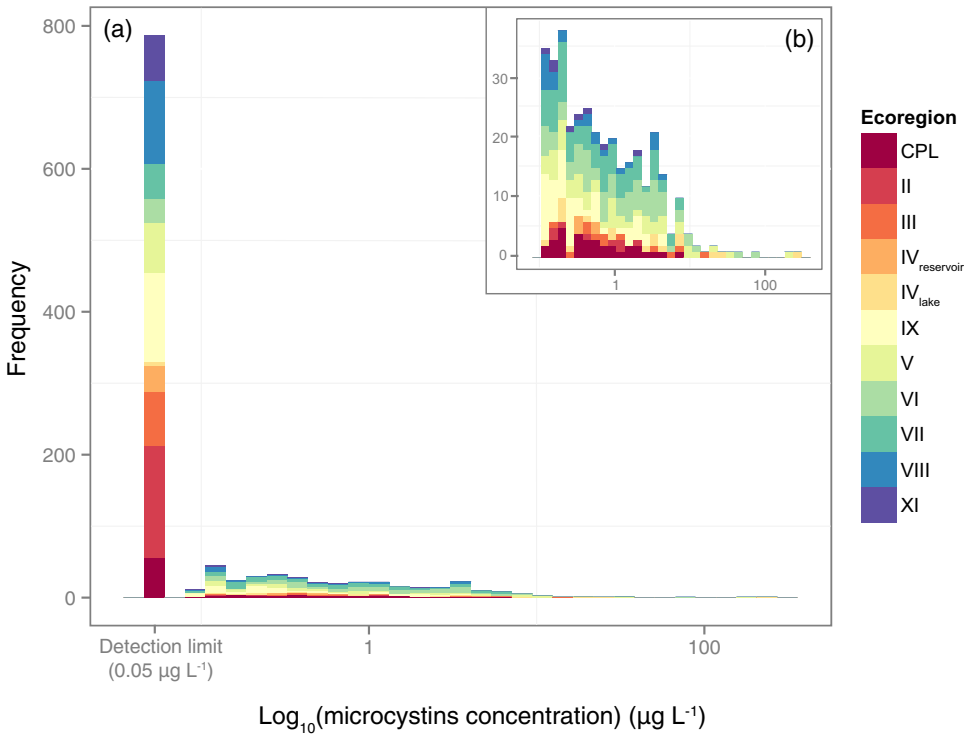

FIGURE 1 Histograms of microcystin concentration. (a) Observations across all NLA lakes [$n = 1127$], where colour coding represents the 11 U.S. EPA nutrient–water quality ecoregions of the contiguous United States. (b) Observations above the detection limit [>0.05 µg L^{-1}; $n = 362$ or 32% of sites]

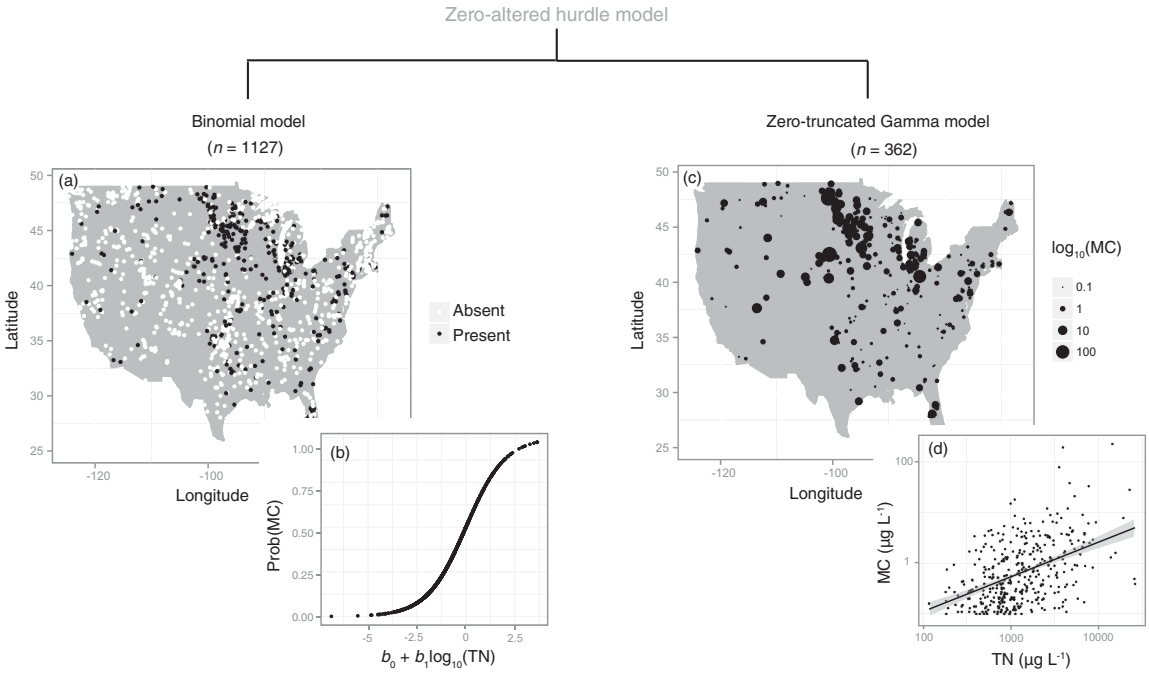

FIGURE 2 Schematic representation of the hierarchical zero-altered hurdle model. (a) Map of the sites where microcystin [MC] concentrations are below [Absent; white] and above [Present; black] the detection limit. (b) Example of a logistic regression [binomial model] of probability of MC detection versus total nitrogen as the explanatory variable. (c) Map of the sites where MCs fell above the detection limit, where circle size indicates the log-transformed concentration of MCs in each site. (d) Example of a Gamma regression of MC concentrations versus total nitrogen [TN]. The *invlogit()* function ["arm" package in R] was used to illustrate the logistic model

the fixed and random effects was evaluated in each part of the hurdle model by comparing nested fixed and random effect models [Appendix S1]. Each part of the hurdle model was calculated separately in R using the "lme4" package in R by applying a logistic regression to the full data set and a lognormal or Gamma model to the DL-truncated data. The ZALN and ZAG models were compared using information criteria to determine which provided the better fit [Appendix S1]. Local models without random effects were evaluated with a generalized linear model ["base stats" package in R].

2.3.2 | Transformations and centring

Log or square-root transformations were applied to reduce the skewness of most predictor variables [Table S1], with the exception of the land-cover variables describing the proportion of agriculture at the catchment and ecoregion scales; these variables could not be normalized and therefore raw values were used. To establish a zero point on all scales and help model convergence, all explanatory variables were then standardized (subtracting the overall mean value from each individual observation [$X_{ij} - \bar{X}$] and dividing by the standard deviation) using the *scale* base function in R. MC data above the detection limit were log-transformed to reduce the skewness. This resulted in log-transformed MC concentrations that were left-censored at -1.3 $\log_{10}$ μg MC L^{-1} [i.e., the equivalent of 0.05 μg MC L^{-1}]. Thus, to support the use of a Gamma or lognormal hurdle models [defined for a threshold value of zero], we used an additive constant offset so that the value of the detection limit became zero in the ZAG and ZALN models [i.e., we subtracted the $\log_{10}$ DL from the $\log_{10}$ MC concentrations, *sensu* Brilleman et al. (2016)]. The response for the ZAG and ZALN models therefore represents the amount by which the $\log_{10}$ MC exceeds the $\log_{10}$ DL.

2.3.3 | Validation of the hurdle model

Once we fitted the final hurdle model we assessed whether any patterns remained in the residuals by examining the diagnostic plots [i.e., QQ-plots, scatterplots of the randomized quantile residuals versus fitted values, and scatterplots of the observed versus fitted values] for both the presence–absence and the continuous distribution models. We used randomized quantile residuals as opposed to Pearson residuals because the former correct for the banding of zeros in the binomial model and adjust for the mean–variance relationship assumed by the Gamma distribution. The quantile residuals were calculated using the *qresiduals()* function in the "statmod" package in R (Dunn & Smyth, 1996). In addition, to examine how well the model captured the large-scale spatial patterns in the data, we plotted the quantile residuals and fitted values for each part of the hurdle models versus ecoregion; any substantive spatial patterns unaccounted for by the model would be highlighted in these diagnostic plots.

Lastly, to test the strength of the model predictions across the range of toxin concentrations, we explored how well the fitted hurdle model was able to predict whether a given lake passed different provisional MC guidelines. To do so, we multiplied the probability that a given lake is above the detection limit [fitted probabilities of the binomial component; π_i = probability of presence] (Zuur & Ieno, 2016) times the probability that the lognormal or Gamma distribution [with the fitted value and dispersion] is above each of the provisional guidelines.

2.3.4 | Boosted regression trees

In addition to improving our understanding of the environmental determinants of MCs, we were also interested in identifying predictors that could be used as indicators of whether lakes and reservoirs were susceptible to surpassing management targets for MCs. Here, our goal was to test which environmental conditions were associated with MC concentrations above key drinking water or recreational guidelines; namely, the U.S. EPA drinking water advisory for children [≥0.3 μg L^{-1}], the WHO drinking water advisory [≥1 μg L^{-1}], the U.S. EPA drinking water advisory for adults [≥1.6 μg L^{-1}] and the WHO recreational, low probability of effect advisory [≥2 μg L^{-1}] (Chorus & Bartram, 1999; Hollister & Kreakie, 2016; U.S. EPA, 2015). From a management perspective, we had an interest in what would split the population of data above or below each threshold as greater resources and energy are invested by governments to try to prevent lakes and their watersheds from surpassing these limits. From a statistical standpoint, however, it is a questionable practice to split a population of data along a continuous response gradient and treat them separately. As such, the rationale behind using the hurdle framework for non-detection versus detection could not be applied for this second objective. We thus used boosted regression trees [BRTs; "gbm" package in R] (Ridgeway, 2015) on the whole population of data to classify the target variable based on environmental variables and identify conditions associated with toxin levels exceeding each provisional guidelines.

Briefly, BRTs optimize the predictive model performance by building and merging results from multiple models [i.e., trees] using a stage-wise forward selection procedure that cumulatively combines numerous simple regression trees (Elith, Leathwick, & Hastie, 2008). The first tree maximally reduces the deviance in the response, and the following tree fits the residuals of the first tree. The process continues at each successive step, and the final model is a linear combination of trees, each fitted to the residuals of the previous tree in the stage-wise process. To improve accuracy and avoid over-fitting, randomness is introduced at each stage by randomly selecting and fitting 50% of the data without replacement [i.e., bag fraction = 0.5]. In addition, a regulation step is implemented that optimizes the number of trees [nt], the learning rate [lr; used to shrink the contribution of each tree added to the model] and the tree complexity [tc; used to control the maximum number of tree nodes]. Details on the procedure for selecting the optimal nt, lr and tc parameters are provided in Appendix S2.

Once the optimal parameters are identified, the final step is to simplify the BRT using an automatic rule that drops explanatory variables until the average change in predictive deviance exceeds its original standard error [*gbm.simplify* function]. To present the final model of each provisional guideline BRT, we provide a graphical summary of the relative influences of the explanatory variables as well as the marginal effects of the most influential variables [on a logit scale]. To provide information on tree complexity [number of nodes] we used the *gbm.interactions* function to illustrate important interactions among

explanatory variables. Lastly, the predictive performance of the BRTs [deviance explained and area under the receiver operating characteristic curve, ROC] is presented for each provisional guideline model.

3 | RESULTS

3.1 | Predicting the occurrence and concentration of MCs across 1127 U.S. lakes and reservoirs

Although most of the data represent measurements from a single sampling date per lake, our analyses showed that the repeated sampling of *c.* 9% of 1127 lakes was highly consistent with the results from the first sampling date and that there was no bias among ecoregions [or latitudes] with respect to the timing of sampling [Figure S1(b)]. The proportion of lakes with MC concentrations above the detection limit was comparable between the two visits [28% and 37% of the lakes sampled on the first and second visit, respectively] and comparable with the full data set [32% of the 1127 lakes sampled]. Once detected, total concentrations were significantly correlated between the first and second sampling dates [$r = 0.85$, $p < .0001$ on log-transformed data; $r = 0.42$, $p < .0001$ on raw data].

The first part of our zero-altered model, which involved the application of a binomial GLMM to the presence–absence MC data from all NLA lakes, indicated that the probability of detecting MC depended on both local and regional features. At the local scale, MCs were associated with deeper [maximum depth], more coloured [DOC] and productive lakes [TN, Chl *a* and cyanobacterial biomass; Table 1, Model 1]. However, substantial regional heterogeneity remained, and the model with both fixed local effects and a random intercept testing for an ecoregion effect [Table 1, Model 3.1] showed that toxic blooms were more likely to be detected in temperate and southern glaciated plains than along the arid and mountainous western coast [Figure 3a]. When we included geographical coordinates as fixed variables [Table 1, Model 3.2], we accounted for an additional 30% of the variance in random intercepts [decrease in random effect variance from $\tau_{00} = 0.44$ to $\tau_{00} = 0.14$; Figure 3b]. Considering the cross-scale interaction between percentage agriculture at the ecoregion scale and local drivers brought about further improvements. In particular, the best-fit CSI between percentage ecoregion agriculture and Chl *a* concentrations accounted for the remaining random intercept variance [from $\tau_{00} = 0.14$ to $\tau_{00} = 0.01$; Table 1, Model 5.1]. The probability of detecting MCs increased with percentage ecoregion agriculture, and the sign of the CSI between Chl *a* and percentage ecoregion agriculture was negative, indicating that high Chl *a* limited the effect of regional agriculture [i.e., when Chl *a* concentrations were very high, shading may limit algal growth]. The importance of this cross-scale interaction is clearly evident in Figure 3(c, d), where the random effect intercepts of each ecoregion are plotted versus percentage ecoregion agriculture. Overall, our best-fit binomial GLMM explained 55% of the variance in presence–absence of MCs across the NLA sites and showed that the occurrence of MCs was best explained by nitrogen enrichment, changes in the phytoplankton community [total cyanobacterial biomass and Chl *a* concentration], light availability [DOC], maximum depth, latitude, longitude and an overarching effect of ecoregion agriculture [Table 1, Model 5.2, Figure S2].

For the second part of the hurdle model, we found that the ZAG model systematically outperformed the ZALN model for all sub-models tested [ΔAIC or ΔBIC ≥ 67.9, where AIC is the Akaike information criterion and BIC the Bayes information criterion; Appendix S1]. The ZAG model indicated that once the data set was restricted to sites where MCs were detected, only local-scale variables were identified as significant predictors of MC concentration. Specifically, we found that MC was more abundant in natural lakes versus human-made reservoirs and increased as total nitrogen, cyanobacterial biomass and turbidity increased [Table 2, Model 1]. The best-fit model also included a negative relationship between MC concentration and surface water temperature, although the amount of variance explained by temperature was very weak and its effect bordered zero. The relationship between MC concentration and local-scale variables was largely homogeneous across the landscape of lakes where MCs were detected [i.e., random effects did not vary among the different nutrient ecoregion classifications or the HUC-2 hydrological units]. In addition, we failed to detect any effect of regional agriculture. Together, the best-fit local-scale variables explained 26% of the variance in MC concentration [Table 2; Figure S3].

The diagnostic plots of the final zero-altered Gamma hurdle model did not show strong patterns in the quantile residuals [Figure S4a–d, g, h]. We detected some pattern in the Gamma model fitted values, where the model tended to underestimate concentrations at low predicted values and overestimate concentrations at high predicted values [Figure S4f]. The plot of fitted Gamma values versus ecoregions showed that certain regions, with higher fitted values, also had greater variance in fitted values [Figure S4j].

To test how well the hurdle model fits the data across the range of toxin concentrations, we plotted the probability that the fitted values of the final hurdle model will fall above each provisional guideline. We separated these probabilities by ecoregion to showcase the spatial heterogeneity in predicted values. For the subset of data above the detection limit, the probability that predicted MC concentrations fell above the U.S. EPA drinking advisory for children [>0.3 μg L^{-1}], WHO drinking water advisory [>1 μg L^{-1}], U.S. EPA drinking advisory for adults [>1.6 μg L^{-1}] and U.S. EPA recreation advisory [>2 μg L^{-1}] were 77%, 22%, 6% and 2%, respectively. In comparison, the proportion of lakes with observed MC concentrations above each guideline [excluding lakes where MCs were not detected] were 63%, 35%, 27% and 22%; the model thus had greater difficulty predicting extreme MC values. The sites that were predicted to most likely fall below the U.S. EPA child advisory guideline [>0.3 μg L^{-1}] are located in the intermountain xeric ecoregions [Figure S5a], whereas sites predicted to exceed the highest provisional guideline [>2 μg L^{-1}] are situated in the interior agricultural plains [Figure S5d].

3.2 | Thresholds of severe lake impairment

Building on our hurdle model, which differentiated factors that restrict the distribution of cyanotoxins across the conterminous United States,

TABLE 1 Summary statistics for part 1 of the zero-altered hurdle model [binomial GLMM]

	β_0	β_{TN}	β_{CBB}	$\beta_{Chl\ a}$	β_{DOC}	β_{Depth}	β_{Lat}	β_{Long}	$\beta_{Eco\ Agric}$	$\beta_{Eco\ agric \times Chl\ a}$	$\beta_{Basin\ agric}$	R^2_{marg}	R^2_{cond}	ΔAICc	ΔBIC	ICC	σ^2	τ_{00}	τ_{11}
Model 1: local	−1.05 (0.08)	0.61 (0.16)	0.43 (0.09)	0.37 (0.13)	0.44 (0.14)	0.30 (0.10)	–	–	–	–	0.32 (0.08)	0.38	–	65.9	50.8	–	0.92	–	–
Model 2: random intercept	−0.87 (0.38)	–	–	–	–	–	–	–	–	–	–	0.00	0.31	215.6	170.6	0.87	1.01	1.47	–
Model 3.1: local + random intercept	−1.21 (0.23)	0.70 (0.18)	0.41 (0.10)	0.47 (0.14)	0.33 (0.14)	0.24 (0.11)	–	–	–	–	–	0.37	0.45	32.7	17.6	0.67	0.87	0.44	–
Model 3.2: . . . + site coordinates	−1.19 (0.15)	0.70 (0.18)	0.41 (0.10)	0.59 (0.14)	0.40 (0.14)	0.32 (0.11)	0.37 (0.10)	0.51 (0.14)	–	–	–	0.46	0.49	13.7	3.7	0.39	0.86	0.14	–
Model 4: . . . + random slope[a]	−1.12 (0.16)	0.72 (0.2)	0.42 (0.1)	0.61 (0.14)	0.41 (0.15)	0.36 (0.12)	0.44 (0.11)	0.53 (0.15)	–	–	–	0.47	0.50	13.4	13.4	0.54	0.85	0.16	0.09^{ns}
Model 5.1: . – random slope + CSI	−1.7 (0.17)	0.57 (0.18)	0.42 (0.10)	1.13 (0.23)	0.42 (0.14)	0.35 (0.11)	0.41 (0.10)	0.50 (0.12)	2.21 (0.49)	−1.65 (0.49)	–	0.55	0.55	1.9	6.9	0.04	0.86	0.01	–
Model 5.2: . . . – random intercept	**−1.7 (0.17)**	**0.57 (0.18)**	**0.42 (0.10)**	**1.14 (0.23)**	**0.42 (0.14)**	**0.35 (0.11)**	**0.41 (0.10)**	**0.49 (0.12)**	**2.19 (0.49)**	**−1.66 (0.50)**	**–**	**0.55**	–	**0.0**	**0.0**	–	**0.86**	–	**–**

We report logit-transformed regression coefficients, which are directly related to a linear predictor. The slope coefficient of each variable is shown by β_x with the standard error in parenthesis.
Key: 0=intercept; TN=$\log_{10}$[total nitrogen]; CBB=[cyanobacteria biomass]$^{0.25}$; Chl *a*=$\log_{10}$[chlorophyll *a*]; DOC=$\log_{10}$[dissolved organic carbon]; Depth=$\log_{10}$[maximum lake depth]; Lat=latitude; Long=longitude; Eco agric=percentage ecoregion agriculture; Basin agric=percentage basin agriculture; R^2_{marg}=marginal coefficient of determination; R^2_{cond}=conditional coefficient of determination; ΔAICc=difference in Akaike information criterion with a correction for finite sample sizes; ΔBIC=difference in Bayes information criterion; ICC=intra-class correlation; σ^2=residual variance; τ_{00}=random intercept variance; τ_{11}=random slope variance; n.s=non-significant; CSI=cross-scale interaction. The best-fit model appears in bold text. See Table S1 for variable units.
[a]Random slope for the relationship between microcystin presence–absence and TN is shown; however, all random slope models were within three AICc and BIC units of each other, thus τ_{11} was comparable throughout.

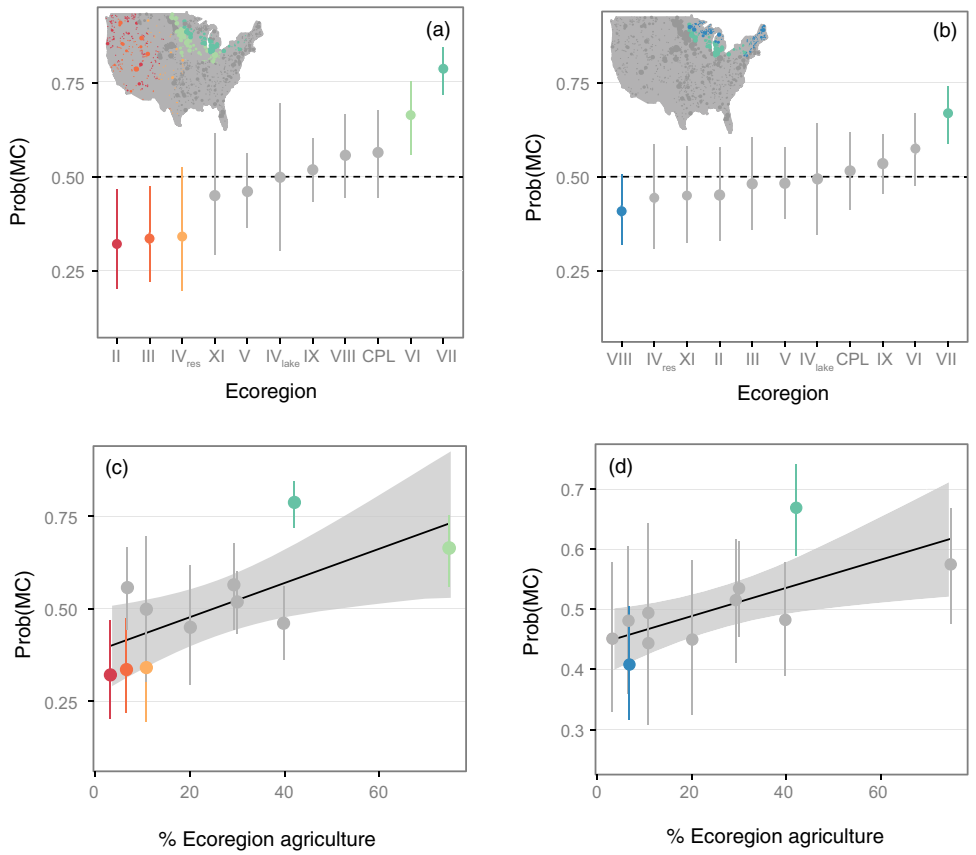

FIGURE 3 Ecoregion-level probability of detecting microcystins [MC]. (a), (b) Random intercepts derived from (a) the binomial generalized linear mixed model [GLMM] with local-scale variables and an ecoregion-level random intercept versus respective ecoregion codes [Table 1, Model 3.1] and (b) the binomial GLMM with local-scale variables, site coordinates and ecoregion-level random intercept versus ecoregion codes [Table 1, Model 3.2]. Ecoregions with significantly higher or lower than average probabilities of detecting microcystins are shown in colour and illustrated in the inset map. (c), (d) Partial regression of ecoregion-level random intercepts from the binomial GLMM shown in panels (a) and (b) versus percentage ecoregion agriculture

our BRT models further identified conditions that were associated with toxin levels falling above or below the WHO and U.S. EPA drinking water and recreational provisional guidelines. Elevated nitrogen concentrations and cyanobacterial biomass were consistently associated with MC concentrations above guideline limits, whereas other variables lost [latitude, longitude, percentage agricultural land use] or gained [turbidity] importance along the gradient of minor to severe lake impairment [Figures 4, S6–S9]. In particular, the probability of exceeding the lower provisional limits [0.3–1 μg MC L^{-1}] increased sharply in mesotrophic lakes [>350 μg TN L^{-1}; Figures S6 and S7], whereas more elevated provisional limits [1.6–2.0 μg MC L^{-1}] were associated with eutrophic to hypereutrophic conditions [>650 μg TN L^{-1}; Figures S8 & S9]. The probability of exceeding higher MC guidelines also increased with bloom size [cyanobacterial biomass above *c.* 160 mg L^{-1}], with maximum MC values observed at very high cyanobacterial densities [above *c.* 13,000 mg CBB L^{-1}]. Conditions of elevated percentage ecoregion agriculture [>40%] and DOC [>6 mg DOC L^{-1}] were important predictors of low to intermediate impairment [U.S. EPA children and WHO drinking advisories; Figures S6–S8]. However, the influence of both factors was much reduced when impairment was severe [WHO recreational advisory limit of 2 μg MC L^{-1}; Figures 4f and S9]. The highest provisional limit also tended to correspond with more turbid water columns [Figure 4f].

Tree complexity selected for each BRT was generally low [Table 3], whereby little to no interaction [number of nodes] was identified among environmental drivers. However, we detected an important interaction for the low impairment limit [U.S. EPA drinking advisory for children], where lakes were more likely to exceed this guideline when both TN and the catchment area [relative to lake area] were high [Figure S10]. Interestingly, our BRT on presence–absence data tracked the importance of the cross-scale interaction between percentage ecoregion agriculture and Chl *a* as well as a spatial heterogeneity of the TN effect [most pronounced for sites north of 40° latitude; Figure S11].

4 | DISCUSSION

Our application of a hierarchical zero-altered model to cyanotoxin data from a large population of lakes and reservoirs showed that significant advances could be made in understanding the impacts of local and regional factors on a critical lake-water quality metric by quantifying spatially structured environmental gradients. We showed that although local-scale variables explained a substantial portion [38%] of the

TABLE 2 Summary statistics of part 2 of the zero-altered hurdle model [i.e., detection limit-truncated Gamma model]

	β_0	β_{TN}	β_{CBB}	$\beta_{Turbidity}$	$\beta_{Lake\ origin}$	$\beta_{Surface\ water\ temperature}$	$\beta_{Eco\ agric}$	R^2_{marg}	R^2_{cond}	ΔAICc	ΔBIC	ICC	σ^2	τ_{00}	τ_{11}
Model 1: local	**−0.01 (0.04)**	**0.09 (0.04)**	**0.09 (0.03)**	**0.12 (0.04)**	**0.13 (0.06)**	**−0.07 (0.03)**	–	**0.26**	–	**0.0**	**0.0**	–	**0.23**	–	–
Model 2: random intercept	0.04 (0.07)	–	–	–	–	–	–	0.00	0.07	83.2	67.9	0.06	0.28	0.02	–
Model 3: local + random intercept	0.01 (0.06)	0.11 (0.04)	0.08 (0.03)	0.14 (0.04)	0.10 $(0.10)^{n.s.}$	−0.06 $(0.03)^{n.s.}$	–	0.28	0.31	0.07	4.5	0.02	0.22	$0.01^{n.s.}$	–
Model 5: . . . – random intercept + CSI	−0.01 (0.05)	0.08 (0.04)	0.10 (0.02)	0.12 (0.04)	0.13 (0.06)	−0.07 (0.03)	0.01 $(0.02)^{n.s.}$	0.30	–	69.6	73.4	–	0.27	–	–

The best-fit model appears in bold text. See footnote to Table 1 for all abbreviations and explanations and Table S1 for variable units.

occurrence [presence–absence] of MCs in U.S. lakes and reservoirs, geographical location and regional heterogeneity across the landscape accounted for an additional 17% of the residual variance [Table 1; comparison of R^2_{marg} of Models 1 and 5.2]. Furthermore, by modelling presence–absence and concentration data as a two-part hurdle model, we identified which factors were associated with the occurrence of a toxic bloom and its continued increase. We suggest that such a framework could be applied to a wide variety of contaminants that are measured in surface waters at continental to global scales [e.g., atrazine, arsenic and benzene] and could provide valuable insights into the drivers of such analytes in the presence of overarching spatial structures and a large number of missing values.

Interestingly, the hurdle model showed that certain combinations of variables helped explain the detection versus non-detection of MCs, whereas others were identified as drivers of MC abundance once detected. For instance, we noted an important relationship between DOC and presence–absence of MCs, but failed to detect a relationship between DOC and MC concentrations once above the detection limit. Similarly, our BRT models identified a strong effect of DOC at lower provisional limits, but its effect was weaker at higher management thresholds. This is in line with the growing evidence that suggests that DOC is beneficial for toxin-producing cyanobacteria as it allows them to outcompete other algae under high UV radiation. Under stratified conditions and higher UV radiation, DOC is photocatalysed into superoxide and hydrogen peroxide, but MC-producing cyanobacteria have several unique strategies to deal with these reactive oxygen species, including the formation of MC–protein complexes that prevent proteolytic degradation within the cyanobacterial cell (Paerl & Otten, 2013). Thus, MCs provide a competitive advantage to toxic cyanobacteria via their protective role under oxidative stress (Pimentel & Giani, 2014). We suggest that at higher MC limits, where algal concentrations are much higher, UV penetration into the water column is greatly limited and this effect is much diminished. At this extreme, MC production may also increase several fold due to quorum sensing and stress from nutrient limitation (Pereira & Giani, 2014; Pimentel & Giani, 2014; Van de Waal et al., 2009; Wood et al., 2011).

In contrast to the effect of DOC, MC concentrations [measured above the detection limit, not just presence–absence] differed between lakes and reservoirs. This finding echoes previous work that demonstrated significant differences in cyanobacterial biomass between lakes and reservoirs from the same data set, where the predictive strength of cyanobacterial biomass was consistently weaker in NLA reservoirs than natural lakes (Beaulieu et al., 2013). These findings suggest that differences between natural lakes and reservoirs play a key role in determining cyanobacterial and cyanotoxin concentrations in surface waters across the U.S. There are many known differences between lakes and reservoirs, including hydrological variability and lake connectivity (Read et al., 2015), as well as differences in their geographical distribution [reservoirs in the southern states; Figure S12]. These differences could, in turn, influence the dominance of toxic cyanobacteria in surface waters of the conterminous United States. For example, numerous authors have noted that water residence time is a significant

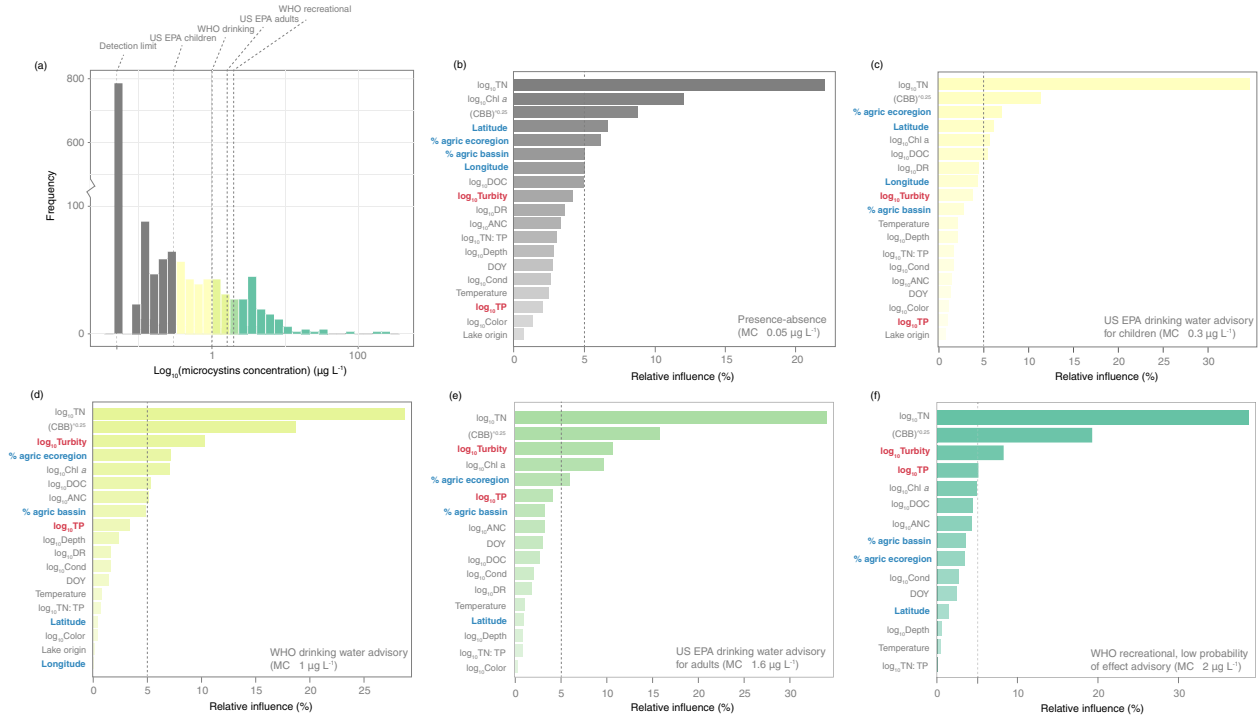

FIGURE 4 Summary of the relative contributions [%] of predictor variables for each boosted regression tree [BRT] model. (a) Histogram of microcystin [MC] concentration [*y*-axis truncated] illustrating the cut-off of each World Health Organization [WHO] and U.S. Environmental Protection Agency [EPA] provisional guidelines. Relative contributions of BRT predictor variables developed for MC occurrence above (b) the detection limit, (c) the U.S. EPA drinking water advisory for children, (d) the WHO drinking water advisory, (e) the U.S. EPA drinking water advisory for adults, and (f) the WHO recreational, low probability of effect advisory. Dashed lines in (b)–(f) indicate the most influential environmental variables [using an arbitrary cut-off of ≥5% relative influence]. Bold text and colour coding is used to highlight the gradual loss [blue] and gain [red] in the relative importance of regional versus local variables across the different provisional guidelines. $\log_{10}$TN = $\log_{10}$[total nitrogen], $\log_{10}$Chl a = $\log_{10}$[Chlorophyll a], $[CBB]^{0.25}$ = [cyanobacteria biomass]$^{0.25}$, % agric = percent agricultural land cover, $\log_{10}$DOC = $\log_{10}$[dissolved organic carbon], $\log_{10}$DR = $\log_{10}$[drainage ratio], $\log_{10}$ANC = $\log_{10}$[acid-neutralizing capacity], DOY = day of the year, $\log_{10}$Cond = $\log_{10}$[conductivity], $\log_{10}$TP = $\log_{10}$[total phosphorus].

factor influencing cyanobacterial communities, where blooms do not reach their full potential in systems that flush quickly, like reservoirs (Carvalho et al., 2011; Rolland et al., 2013; Romo et al., 2013). Using additional NLA lake and catchment data provided by Read et al. (2015), we noted that in addition to having shorter water residence times, reservoirs had higher catchment connectivity [i.e., typically stream or lake drainage systems, rather that isolated and headwater; $\chi^2 = 21.4$, $p < .0001$], and larger catchment to lake area ratios [drainage ratio; $F = 37.32$, $p < .0001$] than lakes. Given that systems with large catchments relative to lake area tend to have short water residence times (Kalff, 2002), MC concentrations may be related indirectly to drainage ratio via its effect on water residence times. The link between lake connectivity and cyanotoxin concentration is a relatively new observation, and will require further research. However, one might expect that, at least in some low-connectivity lakes, the nutrients supplied are retained for longer and may become more concentrated during drought events, which together could ensure an increased availability of nutrients to support cyanobacterial blooms and their toxins (O'Neil, Davis, Burford, & Gobler, 2012).

The hurdle and BRT models also tracked a gradual decrease in the effect of regional variables along the MC gradient. Local and regional factors were strongly associated with MC values exceeding the lower MC guidelines, whereas there was a gradual loss of regional-level effects at higher impairment. Thus, multi-scale effects [percentage ecoregion agriculture] were important in predicting the detection and

TABLE 3 Predictive performance of boosted regression tree models for microcystin [MC] presence–absence [pa; above–below the detection limit] data and for concentrations above-below each provisional guideline

MC threshold	% Deviance explained	ROC	lr	nt	tc
pa (0.05 µg L^{-1})	13.4 (0.97)	0.86 (0.01)	0.01	500	5
U.S. EPA child drinking (0.3 µg L^{-1})	28.1 (0.98)	0.86 (0.01)	0.05	900	3
WHO drinking (1.0 µg L^{-1})	50.2 (0.98)	0.87 (0.02)	0.005	2550	1
U.S. EPA adult drinking (1.6 µg L^{-1})	57.4 (0.99)	0.87 (0.01)	0.01	1750	1
WHO recreational (2.0 µg L^{-1})	62.1 (0.98)	0.86 (0.02)	0.01	750	2

ROC=area under the receiver operating characteristic curve; lr=optimal learning rate; nt=number of trees; tc=optimal tree complexity [number of nodes]. Standard errors of model statistics are shown in parenthesis.

initial rise in MC concentrations, but once they occurred, MCs tended to increase in response to localized factors [e.g., higher turbidity]. The loss of regional drivers may also have been due to an increasingly constrained and localized distribution of lakes.

5 | CONCLUSION

Our study highlights the importance of multi-scale processes and how these represent sources of uncertainty and spatial dependence across national data sets. The continent-wide analysis allowed us to synthesize the results across all ecoregions, which spanned wide gradients of land use [e.g., agricultural land cover ranging from 0% to 75%], climate [e.g., mean water temperature ranging from 16°C to 26°C] and water quality [oligotrophic to hypereutrophic], thus providing generalizations about the continent as a whole and identifying relatively problematic regions where more severe mitigation is needed to curtail local water quality impairment. The framework used here allowed us to overcome major challenges in continental-scale analyses [e.g., high frequency of non-detections and spatial aggregation] and to correctly expand and analyse fine-scaled responses to broad-scale patterns. This approach echoes the arguments raised by Heffernan et al. (2014), who made the case that novel insights into environmental change will be acquired through a macro-ecological perspective.

The prevalence of toxic algal blooms is emerging as one of the most important water quality and health issues we face today. Nonetheless, provisional guidelines on toxic blooms vary greatly among countries, suggesting that more effort should be made to develop a comprehensive risk management framework (Ibelings, Backer, Kardinaal, & Chorus, 2014). Such a framework could help identify the importance of overarching processes and how lakes filter regional changes in land use or climate, both present and future, leading to heterogeneous patterns in algal response (Bleckner, 2005; Maheaux, Leavitt, & Jackson, 2015; Pennock, 2003). These conceptual frameworks, however, poorly apply to skewed distributions inherent to broad-scale, empirical data, thus limiting their usefulness in predicting bloom occurrence in other sites along heterogeneous landscapes. Here we showed how zero-altered mixed models and BRTs provide a broader representation of a potential health hazard by modelling zero inflation, the lack of independence among lakes and how the relative importance of predictors varied at different guideline limits, which thus helped set endpoints better adapted to geographical and environmental context.

ACKNOWLEDGEMENTS

This work was partially supported by a fellowship from Fonds Québécois de la Recherche sur la Nature et les Technologies [FQRNT] awarded to Z.E.T., a team grant from FQRNT allocated to I.G.-E. and P.L. and an NSERC discovery grant [NSERC RGPIN 261488-12] awarded to R.J.S. I.G.-E. also acknowledges funding from the Canada Research Chair Program. We greatly thank the Landscape Limnology Research Group [P. A. Soranno, K. Cheruvelil, J.-F. Lapierre, S. Oliver, E. Fergus, S. Collins, N. Skaff and J. Marino], F. Pick and A. Winegardner for fruitful discussions and valuable feedback on analyses and results. We thank the field crews and laboratory personnel of the 2007 National Lakes Assessment for the data used in this study. We also thank three anonymous referees for their helpful and constructive comments that contributed to improving the manuscript.

REFERENCES

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.

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.

Blàha, L., Babica, P., & Maršálek, B. (2009). Toxins produced in cyanobacterial water blooms – toxicity and risks. *Interdisciplinary Toxicology*, *2*, 36–41.

Bleckner, T. (2005). A conceptual model of climate-related effects on lake ecosystems. *Hydrobiology*, *533*, 1–14.

Brilleman, S. L., Crowther, M. J., May, M. T., Gompels, M., & Abrams, K. R. (2016). Joint longitudinal hurdle and time-to-event models: An application related to viral load and duration of the first treatment regimen in patients with HIV initiating therapy. *Statistics in Medicine*, *35*, 3583–3594.

Carmichael, W. W., Azevedo, S. M., An, J. S., Molica, R. J., Jochimsen, E. M., Lau, S., … Eaglesham, G. K. (2001). Human fatalities from cyanobacteria: Chemical and biological evidence for cyanotoxins. *Environmental Health Perspectives*, *109*, 663–668.

Carvalho, L., McDonald, C., de Hoyos, C., Mischke, U., Phillips, G., Borics, G., … Cardoso, A. C. (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., Webster, K. E., & Bremigan, M. T. (2013). Multi-scaled drivers of ecosystem state: Quantifying the importance of the regional spatial scale. *Ecological Applications*, *23*, 1603–1618.

Chorus, I., & Bartram, J. (1999). *Toxic cyanobacteria in water: A guide to their public health consequences, monitoring and management*. Bury St Edmunds, UK: St Edmundsbury Press.

Codd, G. A., Morrison, L. F., & Metcalf, J. S. (2005). Cyanobacterial toxins: Risk management for health protection. *Toxicology and Applied Pharmacology*, *203*, 264–272.

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.

Dunn, K. P., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*, 1–10. Retrieved from http://www.statsci.org/smyth/pubs/residual.html

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813.

ESRI. (2011). *ArcGIS desktop: Release 10*. Redlands, CA: Environmental Systems Research Institute.

Fergus, C. E., Soranno, P. A., Cheruvelil, K. S., & Bremigan, M. T. (2011). Multiscale landscape and wetland drivers of lake total phosphorus and water color. *Limnology and Oceanography*, *56*, 2127–2146.

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, *2*, 143–154.

Giani, A., Bird, D. F., Prairie, Y. T., & Lawrence, J. F. (2005). Empirical study of cyanobacterial toxicity along a trophic gradient of lakes. *Canadian Journal of Fisheries and Aquatic Science*, *62*, 2100–2109.

Håkanson, L., Bryhn, A. C., & Hytteborn, J. K. (2007). On the issue of limiting nutrient and predictions of cyanobacteria in aquatic systems. *Science of the Total Environment*, *379*, 89–108.

Heffernan, J. B., Soranno, P. A., Angilletta, M. J., Buckley, L. B., Gruner, D. S., Keitt, T. H., … Weathers, K. C. (2014). Macrosystems ecology: Understanding ecological patterns and processes at continental scales. *Macrosystems Ecology*, *12*, 5–14.

Herlihy, A. T., Kamman, N. C., Sifneos, J. C., Charles, D., Enache, M. D., & Stevenson, J. R. (2013). Using multiple approaches to develop nutrient criteria for lakes in the conterminous USA. *Freshwater Science*, *32*, 367–384.

Hollister, J. W., & Kreakie, B. J. (2016). Associations between chlorophyll *a* and various microcystin-LR health advisory concentrations. *F1000Research*, *5*, 151. doi:10.12688/f1000research.7955.1

Homer, C., Huang, C., Yang, L., Wylie, B., & Coan, M. (2004). Development of a 2001 national land cover database for the United States. *Photogrammetric Engineering Remote Sensing*, *70*, 829–840.

Hunter, P. D., Hanley, N., Czajkowski, M., Mearns, K., Tyler, A. N., Carvalho, L., & Codd, G. A. (2012). The effect of risk perception on public preferences and willingness to pay for reductions in the health risks posed by toxic cyanobacterial blooms. *Science of the Total Environment*, *426*, 32–44.

Ibelings, B. W., Backer, L. C., Kardinaal, W. E. A., & Chorus, I. (2014). Current approaches to cyanotoxin risk assessment and risk management around the globe. *Harmful Algae*, *40*, 63–74.

Jonasson, S., Eriksson, J., Berntzon, L., Spácil, Z., Ilag, L. L., Ronnevi, L. O., … Bergman, B. (2010). Transfer of a cyanobacterial neurotoxin within a temperate aquatic ecosystem suggests pathways for human exposure. *Proceedings of the National Academy of Sciences USA*, *107*, 9252–9257.

Kalff, J. (2002). *Limnology: Inland water ecosystems*. Upper Saddle River, New Jersey: Prentice Hall.

Kardinaal, W. E. A., Tonk, L., Janse, I., Hol, S., Slot, P., Huisman, J., & Visser, P. M. (2007). Competition for light between toxic and nontoxic strains of the harmful cyanobacterium *Microcystis*. *Applied and Environmental Microbiology*, *73*, 2939–2946.

Lachenbruch, P. A. (2001). Comparison of two-part models with competitors. *Statistics in Medicine*, *20*, 1215–1234.

Lachenbruch, P. A. (2002). Analysis of data with excess zeros. *Statistical Methods in Medical Research*, *11*, 297–302.

Lévesque, B., Gervais, M. C., Chevalier, P., Gauvin, D., Anassour-Laouan-Sidi, E., Gingras, S., … Bird, D. (2014). Prospective study of acute health effects in relation to exposure to cyanobacteria. *Science of the Total Environment*, *466–467*, 397–343.

Maheaux, H., Leavitt, P. R., & Jackson, L. J. (2015). Asynchronous onset of eutrophication among shallow prairie lakes of Nothern Great Plains, Alberta, Canada. *Global Change Biology*, *22*, 271–283.

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.

Moulton, L. H., Curriero, F. C., & Barroso, P. F. (2002). Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research*, *11*, 317–325.

Moulton, L. H., & Halsey, N. A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, *51*, 1570–1578.

Moulton, L. H., & Halsey, N. A. (1996). A mixed gamma model for regression analyses of quantitative assay data. *Vaccines*, *14*, 1154–1158.

Olsen, A. R., Stahl, L. L., Snyder, B. D., & Pitt, J. L. (2009). Survey design for lakes and reservoirs in the United States to assess contaminants in fish tissue. *Environmental Monitoring and Assessment*, *150*, 91–100.

Omernik, J. M. (1987). Ecoregions of the conterminous United States. *Annals of the Association of American Geographers*, *77*, 118–125.

O'Neil, J. M., Davis, T. W., Burford, M. A., & Gobler, C. J. (2012). The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. *Harmful Algae*, *14*, 313–334.

Orihel, D., Bird, D. F., Brylinsky, M., Chen, H., Donald, D. B., Huang, D. Y., … Vinebrooke, R. D. (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.

Paerl, H. W., & Otten, T. G. (2013). Blooms bite the hand that feeds them. *Science*, *342*, 433–434.

Paerl, H. W., & Paul, V. J. (2012). Climate change: Links to global expansion of harmful cyanobacteria. *Water Research*, *46*, 1349–1363.

Peck, D. V., Olsen, A. R., Weber, M. H., & Paulsen, S. G. (2013). Survey design and extent estimates for the National Lakes Assessment. *Freshwater Science*, *32*, 1231–1245.

Pennock, D. J. (2003). Terrain attributes, landform segmentation and soil redistribution. *Soil and Tillage Research*, *69*, 15–26.

Pereira, D. A., & Giani, A. (2014). Cell density-dependent oligopeptide production in cyanobacterial strains. *FEMS Microbiology Ecology*, *88*, 175–183.

Peters, D. P. C., Bestelmeyer, B. T., & Turner, M. G. (2007). Cross-scale interactions and changing pattern–process relationships: Consequences for system dynamics. *Ecosystems*, *10*, 790–796.

Pimentel, J. S. M., & Giani, A. (2013). Estimating toxic cyanobacteria in a Brazilian reservoir by quantitative real-time PCR, based on the microcystin synthetase D gene. *Journal of Applied Phycology*, *25*, 1545–1554.

Pimentel, J. S. M., & Giani, A. (2014). Microcystin production and regulation under nutrient stress conditions in toxic *Microcystis* strains. *Applied and Environmental Microbiology*, *80*, 5836–5843.

Read, E., Patil, V. P., Oliver, S. K., Hetherington, A. L., Brentrup, J. A., Zwart, J. A., … Weathers, K. C. (2015). The importance of lake-specific characteristics for water quality across the continental US. *Ecological Applications*, *25*, 943–955.

Ridgeway, G. (2015). gbm: Generalized Boosted Regression Models. R package version 2.1.1. Retrieved from http://CRAN.R-project.org/package=gbm

Rohm, C. M., Omernik, J. M., Woods, A. J., & Stoddard, J. L. (2002). Regional characteristics of nutrient concentrations in streams and their application to nutrient criteria development. *Journal of the American Water Resources Association*, *38*, 213–239.

Rolland, D. C., Bourget, S., Warren, A., Laurion, I., & Vincent, W. E. (2013). Extreme variability of cyanobacterial blooms in an urban drinking water supply. *Journal of Plankton Research*, *35*, 744–758.

Romo, S., Soria, J., Fernández, F., Ouahid, Y., & Baró-Solá, Á. (2013). Water residence time and the dynamics of toxic cyanobacteria. *Freshwater Biology*, *58*, 513–522.

Soranno, P. A., Cheruvelil, K. S., Bissell, E. G., Bremigan, M. T., Downing, J. A., Fergus, C. E., ... Webster, K. E. (2014). Cross-scale interactions: Quantifying multiscaled cause–effect relationships in macrosystems. *Frontiers in Ecology and the Environment*, *12*, 65–73.

Taranu, Z. E., Gregory-Eaves, I., Leavitt, P. R., Bunting, L., Buchaca, T., Catalan, J., ... Vinebrooke, R. D. (2015). Acceleration of cyanobacterial dominance in north temperate-subarctic lakes during the Anthropocene. *Ecology Letters*, *18*, 375–384.

Thorson, J. T., Shelton, A. O., Ward, E. J., & Skaug, H. J. (2015). Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. *ICES Journal of Marine Science*, *72*, 1297–1310.

U.S. EPA (2009). *National lakes assessment: A collaborative survey of the nation's lakes*. EPA 841-R-09-001. Washington, DC: U.S. Environmental Protection Agency, Office of Water and Office of Research and Development.

U.S. EPA (2015). *Drinking water health advisory for the cyanobacterial microcystin toxins*. EPA-820-R-15100. Washington, DC: U.S. Environmental Protection Agency, Office of Water and Office of Research and Development.

Van de Waal, D. B., Verspagen, J. M., Lürling, M., Van Donk, E., Visser, P. M., & Huisman, J. (2009). The ecological stoichiometry of toxins produced by harmful cyanobacteria: An experimental test of the carbon–nutrient balance hypothesis. *Ecology Letters*, *12*, 1326–1355.

Wood, S. A., Rueckert, A., Hamilton, D. P., Cary, S. C., & Dietrich, D. R. (2011). Switching toxin production on and off: Intermittent microcystin synthesis in a *Microcystis* bloom. *Environmental Microbiology Reports*, *3*, 118–124.

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.

Zuur, A. F., & Ieno, E. N. (2016). *Beginners guide to zero-inflated models with R*. Newburgh, UK: Highland Statistics Ltd.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Berlin: Springer.

BIOSKETCH

ZOFIA E. TARANU is a postdoctoral fellow with interests in advancing our understanding of water quality issues through rigorous quantitative analyses [http://zofiaecaterinataranu.weebly.com/]. More broadly, the research team is involved in the large-scale assessment of how anthropogenic activities severely alter and impair aquatic ecosystems. Through the application of key statistical approaches, the group aims to improve our knowledge of the effect of global environmental drivers on harmful algal blooms in lakes.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Taranu ZE, Gregory-Eaves I, Steele RJ, Beaulieu M, Legendre P. Predicting microcystin concentrations in lakes and reservoirs at a continental scale: A new framework for modelling an important health risk factor. *Global Ecol Biogeogr*. 2017;26:625–637. https://doi.org/10.1111/geb.12569