SUPPLEMENTARY INFORMATION

Appendix S1. Hierarchical zero-altered model selection procedure and results

To test the significance of fixed and random effects, and whether cross-scale interactions modulated the response of microcystins to local-scale variables, we compared the five following nested models (sensu Fergus *et al*., 2011). The variance explained by the fixed and random components (where applicable) are presented in Tables 1 and 2 by the marginal and conditional coefficients of determination (Nakagawa & Schielzeth, 2013), respectively, where:

$$\text{Marginal-}R^2 = \frac{fixed\ effects}{total\ variance} = \frac{var(fixed)}{var(fixed) + var(rand) + var(resid)}$$

$$\text{Conditional-}R^2 = \frac{fixed\ and\ random\ effects}{total\ variance} = \frac{var(fixed) + var(rand)}{var(fixed) + var(rand) + var(resid)}$$

Model 1 – Local-scale variables

Here, we tested the significance of local, lake-specific variables by comparing a set of candidate models with different combinations of explanatory variables (Table S1). To determine which local-scale variables best explained the variation in the presence-absence (i.e. above versus below detection) and concentration of microcystins across the conterminous U.S., we ranked each competing model according to their Akaike Information Criterion (AICc; with a small sample size bias-correction) and Bayesian Information Criterion (BIC) using the ICtab function in R (*bbmle* package; Burnham & Anderson, 2002). The BIC penalizes more strongly for the number of parameters included in the model and is thus a more conservative approach than the AICc (Nakagawa & Schielzeth, 2013; Steele, 2013).

Model 2 – Random intercepts

To test for regional-scale variation in the microcystin presence-absence and concentration data, we then developed a set of unconditional models (random intercepts only) that allowed the response variable to vary among different regions (i.e. different baselines for each level of a

regional factor). These unconditional models partitioned the variance of the response in two components: the within-region variance (residual variance; σ^2) and the among-region variance (τ_{00}). The proportion of among-region variance (τ_{00}) relative to the total variance provides the degree of intra-class correlation (ICC), where a higher ICC indicates more similar observations within levels of a regional factor (i.e. larger among-group variance relative to within-group variance), and thus non-independence among sites from the same region. The larger the ICC, the more support there is to include the random effect in the model.

The most commonly used ICC metric in the literature is the proportion of among-region variance relative to residual variance:

$$\text{ICC}_{\text{literature}} = \frac{var(rand)}{var(rand) + var(resid)} = \frac{\tau_{00}}{(\tau_{00}+\sigma^2)}$$

However, a shortcoming of this metric is that the fixed effects must be kept constant in order to track changes in the ICC as random effects are added or deleted from the model (i.e. the residual variance changes as variables are added or removed from the model). When the fixed effects vary among models (as is the case here), using the total variance of the response in the denominator (as opposed to the residual variance) provides a more robust value of the intra-class correlation:

$$\text{ICC} = \frac{var(rand)}{var(rand) + var(y)} = \frac{\tau_{00}}{(\tau_{00} + var(MC))}$$

We used this latter metric throughout our analyses.

To identify the best-fit unconditional model, model likelihoods were compared and the model with the lowest AICc/ BIC was selected as the better model. Ideally, when comparing models that differ only in their random effects, the restricted maximum likelihood (REML) estimation must be used as it provides an unbiased estimate of the variance terms. However,

REML is not currently available for GLMMs (i.e. the glmer function always used the ML method), thus all models comparisons were made using ML.

Model 3 – Local-scale variables and random intercepts

To evaluate whether the fixed effects identified in Model 1 accounted for a substantial part of the random intercept variance identified in Model 2, we quantified the change in the random intercept variance (or equivalently, the change in ICC) when the best-fit local-scale variables identified in Model 1 were added to the unconditional Model 2. To test for further changes in within-region dependence, we re-evaluated the different random factors explored in Model 2. Lastly, we accounted for the geographic location of each site to test for any remaining spatial bias due to the latitudinal or longitudinal gradient. As mentioned, all models were fit using ML since REML is not available for glmer. That said, using the ML estimation is recommended when comparing models whose fixed effects differ (i.e. because the REML correction for biased variance estimates depends on the fixed effect structure, REML cannot be used when the fixed structure changes among models). Furthermore, when comparing the outputs of glmer and glm (i.e. with and without random effects), the ML method must also be used since the maximum likelihood approximation and the least squares methods are the same.

Model 4 – Local-scale variables, random intercepts and random slopes

To test whether local-scale relationships differed among regions, the fixed effect relationships were then allowed to vary among regions (i.e. adding random slopes to Model 3). Among-group differences in slopes were quantified using the random slope variance term (t_{11}). The best model was evaluated by comparing the information theoretic criteria.

Model 5 – Cross-scale interaction

To determine whether regional variables could account for any of the among-region

variability picked up by the random effects, we tested the importance of regional variables, and any cross-scale interaction (CSI) with local-scale variables. In particular, we tested the CSI between the proportion of agricultural land in each ecoregion and the best-fit local-scale variables.

Detailed model results

Hurdle model – Part 1

For the first part of the *hurdle* model, we applied a binomial GLMM using the full dataset to predict the probability of detecting microcystins in the NLA lakes and reservoirs. To do so, microcystins concentrations were transformed into presence (all values above the 0.05 μg L^{-1} detection limit) and absence (all values at or below the detection limit). The best local-scale model (Model 1) included TN, cyanobacteria biomass, Chl *a*, % basin agriculture, DOC and maximum depth.

With no *a priori* expectation on the preferred random intercept structure, we conducted an exploratory analysis (Model 2) by comparing the performance of a suite of unconditional models that differed only in their regional grouping variables (EPA-REG, HUC-2, WSA-ECO3, WSA-ECO9, NUT-REG, ECO-NUTA). The best unconditional structure accounted for 87% of the variance (ICC) in presence-absence of microcystins and was given by the U.S. EPA's national nutrient-water quality ecoregions (ECO-NUTA; n = 11 levels representing the NLA aggregate of the original Omernik level III ecoregions; Herlihy *et al.*, 2013). Lakes within the same nutrient-water quality ecoregion were thus strongly correlated to one-another and ignoring this structure would violate the statistical assumptions of independence.

The inclusion of both random intercepts and fixed effects (Model 3) decreased the variance in random intercepts (ICC decreased from ρ = 0.87 to ρ = 0.67), and significantly

reduced the effect size of one of the fixed effects; % basin agriculture (ΔAIC = -18 when removed). The loss of significance of % basin agriculture when conditioned on each ecoregion cluster (random intercept) suggests that within a given ecoregion, agricultural development was comparable among sites. Moreover, ecoregions with the highest probability of detecting microcystins (Fig. 3(a); VI and VII) had the highest median % basin agriculture (59% and 36%, respectively), while ecoregions with lower agricultural development (median <10%) had a lower than average probability of microcystin detection. By testing the effect of lake latitude and longitude, however, we appreciated that part of this variability was due to broad scale geographical patterns (ICC decreased from $\rho = 0.67$ to $\rho = 0.39$).

The relationships between microcystins presence-absence and the selected local-scale variables did not vary among ecoregions (Model 4; Table 1); all random slope models were within 3 AICc and BIC units of each other, and none substantially improve the model fit.

Lastly, to evaluate whether the variability in random intercepts among the different ecoregions (Model 3) was attributed to an overarching regional effect, we tested for any cross-scale interaction between % ecoregion agriculture and the local-scale variables (Model 5). Significant cross-scale interactions were detected between % regional agriculture and Chl *a*, as well as between % regional agriculture and TN, although the former interaction slightly outcompeted the latter (ΔIC = 2.7). This negative interaction tracked the saturating relationship between microcystins and Chl *a* (or TN) at more elevated values, combined with a stronger effect of agriculture in oligotrophic lakes (Figs. S2(g) & S10). Including the proportion of agricultural land at the ecoregion level and its cross-scale interaction with local-scale variables substantially decreased the random intercept variance and ICC (from $\rho = 0.39$ to $\rho = 0.04$). Thus, a large part of the random intercept variance was explained by regional agriculture.

Hurdle model – Part 2

For the second part of the *hurdle* model, we applied zero-altered lognormal (ZALN) and zero-altered Gamma (ZAG) GLMMs to evaluate the best predictors of log-transformed microcystins concentrations (with a constant offset to set the DL value to zero) in the subset of lakes where this toxin was detected (Fig. 2(c)). The ZALN and ZAG models were compared using the AICc and BIC to determine which provided the better fit (Zuur & Ieno, 2016). The ZAG count process systematically outperformed the ZALN for all sub-models tested ($\Delta IC \geq 67.9$). The results of the former are presented in Table 2.

The local-scale variables that accounted for a significant proportion of the variation in $\log_{10}$ microcystins concentration were TN, cyanobacteria biomass, turbidity, lake origin (natural versus man-made), and surface water temperature (Model 1; Table 2). In our exploratory analysis of the unconditional models (Model 2), we found little support for a hierarchical structure; the ICC varied from $\rho = 0.03$ (WSA-ECO3; wadeable stream assessment aggregate of Omernik level III ecoregions; $n = 3$ levels) to $\rho = 0.06$ (HUC-2; USGS major hydrological unit code with 18 levels across the nation) indicating only ~5% of the variation in toxin concentration in this subset of lakes could be attributed to regional heterogeneity. The unconditional model with a random structure for the HUC-2 classification had the lowest information criteria values, but differed by <4 AICc/ BIC from the nutrient-based regional classifications (i.e., EPA-REG, NUT-REG, and ECO-NUTA). The comparison of Models 1, 2, and 3 (using the ML method since REML is not available for GLMM) showed that the regional heterogeneity explained by the random effect was captured by the local-scale variables, where the best-fit Model 1 outperformed the unconditional model (Model 2; $\Delta AICc = 83.2, \Delta BIC = 67.9$) as well as the

model that included both the random intercept and local-scale variables (Model 3; ΔAICc = 0.7, ΔBIC = 4.5, ICC decreased to $\rho = 0.02$). We did not proceed with testing for the random slope effects (Model 4) given this lack of regional heterogeneity. Similarly, we failed to detect any effect of regional agriculture (Table 2; Model 5).

Appendix S2. Selection of optimal parameters for boosted regression trees

For each provisional threshold, we identified the optimal BRT parameters (nt, lr and tc) by creating a training set (75% of the sample size; n = 861 lakes) and a test set (n = 288) and examining how changing the lr (while keeping tc and nt fixed) changed the predicted model deviance (using the *predict.gbm* function in R; Ridgeway *et al.*, 2015). Since slowing down the lr increases the number of trees, it is suggested to have a slow lr combined with a large nt (Elith *et al.,* 2008). We used the model developed over the training set to predict values for each observation in the test set, calculated the difference between observed and predicted (i.e. deviance) and examined how the deviance changed for different lr and nt values (up to a maximum of nt = 10 000 trees). The learning rates tested were lr = 0.01, 0.05, 0.001, 0.0005. Once we identified the lr and nt combination with the lowest predicted deviance, we tested for optimal tc, by keeping lr constant, and letting the tree complexity increase from 1 to 5 nodes.

REFERENCES

Burnham, K.P. &Anderson, D.R. (2002) Model selection and multimodel inference: A practical information-theoretic approach. 2n Edition. pp. 488. Springer-Verlag New York, Inc.

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

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. *Limnol Oceanogr*, **56**, 2127-2146.

Herlihy, A.T. *et al.* (2013) Using multiple approaches to develop nutrient criteria for lakes in the conterminous USA. *Freshw Sci,* **32**, 367-384.

Nakagawa, S. & Schielzeth, H. (2013) A general and simple method for obtaining R^2 from generalized linearmixed-effects models. *Methods Ecol Evol*, **4**, 133-142.

Steele, R. (2013). Model selection for multilevel models. In M. Scott, J. Simonoff, & B. Marx (eds.), The Sage Handbook of Multilevel Modeling, pp. 109–125. London: SAGE Publications.

Appendix S3. Supplemental figure legends

Figure S1 Spatio-temporal summary of ecoregion factor. (a) Map of NLA lakes colour-coded by ecoregion levels, where the size of each point corresponds to the $\log_{10}$-transformed concentration of microcystins. The mountainous ecoregion (Ecoregion II), which includes the Rocky Mountains, the Cascade Range and Sierra Nevada, divides the intermountain xeric ecoregion to the West (Ecoregion III) from the coastal and interior plain ecoregions to the East. The Appalachian mountain range (Ecoregion XI) further divides the coastal and interior plains. (b) May to October lake sampling date broken down by ecoregion.

Figure S2 Fitted curves of best-fit binomial GLM predicting the presence (above detection limit) and absence (below detection limit) of microcystins (MC) in NLA lakes (R^2_{marg} = 0.55, n = 1127), where TN = total nitrogen concentrations, CBB = cyanobacteria biomass, DOC = dissolved organic carbon, Depth = maximum lake depth, and Chl *a* = chlorophyll *a* concentrations. All variables appear in standardized values. Partial fits were plotted using the *effects* package in R.

Figure S3 Fitted curves for the best-fit ZAG model (R^2_{marg}= 0.29; n = 362). Abbreviations are as described in Fig. S2. Partial fits were plotted using the *effects* package in R.

Figure S4 Diagnostic plots of final zero-altered *hurdle* model. (a, b) QQ-plots using randomized quantile residuals, (c, d) scatterplots of quantile residuals vs. fitted values, (e, f) scatterplots of observed vs. fitted values. Inset panel: (g-h) randomized quantile residuals vs. ecoregions, and (i-j) fitted values vs. ecoregions (shown are the mean and standard error of residuals and fitted values, respectively).

Figure S5 Probability that the fitted values of the final *hurdle* model fall above each provisional guideline; where the probability corresponds to probability that the microcystins concentration in a given lake is above the detection limit times the probability that the Gamma distribution with the fitted values and dispersion is above the provisional guideline. Probabilities have been separated by ecoregions to showcase the spatial heterogeneity in predicted values.

Figure S6 Partial dependence plots for the most influential variables (relative influence ≥ 5%) in the BRT model for microcystin (MC) occurrence above the U.S. EPA drinking advisory for children. See Table S1 for variable units. Y-axes are shown on the logit scale and centered to have zero mean over the data distribution. Rug plots at top of plots indicate the decile distribution of sites across each variable.

Figure S7 Partial dependence plots for the most influential variables in the BRT model for microcystin (MC) occurrence above the WHO drinking advisory for. See Fig. S6 for plot description.

Figure S8 Partial dependence plots for the most influential variables in the BRT model for microcystin (MC) occurrence above the U.S. EPA drinking advisory for adults. See Fig. S6 for plot description.

Figure S9 Partial dependence plots for the most influential variables in the BRT model for microcystin (MC) occurrence above the WHO recreational, low probability of effect advisory. See Fig. S6 for plot description.

Figure S10 Three-dimensional partial dependence plots for the strongest interaction in the BRT model for microcystin (MC) occurrence above the U.S. EPA drinking advisory for children. All variables except those graphed are held at their means, where DR = drainage ratio and TN = total nitrogen concentration.

Figure S11 Three-dimensional partial dependence plots for the strongest interaction in the BRT model for microcystin (MC) presence (above detection limit) and absence (below detection limit) illustrating the interaction between total nitrogen (TN) and latitude, as well as the cross-scale interaction between % ecoregion agriculture and chlorophyll *a* (Chl *a*).

Figure S12 Map illustrating spatial distribution of natural lakes versus man-made reservoirs, where size of circles represents the log-transformed microcystin (MC) concentration in each lake.

Table S1. Summary and transformations of explanatory variables used in *hurdle* and BRT models. The categorical variable coding for lake origin is not included in table; there were 631 man-made reservoirs and 518 natural lakes sampled.

Variable	Abbreviation	Mean	Median	Max	Min	Units	Transformation
Local-scale							
Total nitrogen	TN	1167	568	26 100	5.29	$\mu g\ L^{-1}$	Log_{10}
Total Phosphorus	TP	107.5	24	4 679	0.03	$\mu g\ L^{-1}$	Log_{10}
TN: TP ratio	TN: TP	54.7	21.33	12 090	0.23	––	Log_{10}
Chlorophyll *a*	Chl *a*	28.28	7.54	936	0.07	$\mu g\ L^{-1}$	Log_{10}
Cyanobacteria biomass	CBB	2.15×10^6	0.35×10^6	200×10^6	0	$\mu g\ L^{-1}$	Log_{10}
Acid-neutralizing capacity	ANC	2624	1711	91 630	-62.96	$\mu eq\ L^{-1}$	Log_{10}*
Specific conductivity	Cond	664.9	242	50 590	4.35	$\mu S\ cm^{-1}$	Log_{10}
Dissolved organic carbon	DOC	8.88	5.45	290.6	0.34	$mg\ L^{-1}$	Log_{10}
Turbidity	––	12.93	3.54	574	0.15	NTU	Log_{10}
Colour	––	16.23	12	165	0	PCU	Log_{10}
Surface water temperature	Temperature	24.3	24.7	37.6	10	ºC	––
Maximum depth	Depth	9.5	5.8	97	0.5	m	Log_{10}
Drainage ratio	DR	351	31.7	52 110	0.74	––	Log_{10}
Day of the year	DOY	213	212	291	128	––	––
Regional-scale							
Latitude	––	40.64	41.39	48.98	26.94	DD	––
Longitude	––	-84.61	-94.68	-67.79	-124.60	DD	––
% ecoregion agriculture	––	26.29	29.3	74.8	3.2	––	––
% basin agriculture	––	19.98	7.70	88.1	0	––	––

* The minimum value was added to all ANC values prior to log-transformation as the variable had negative values along its range.

Figure S1

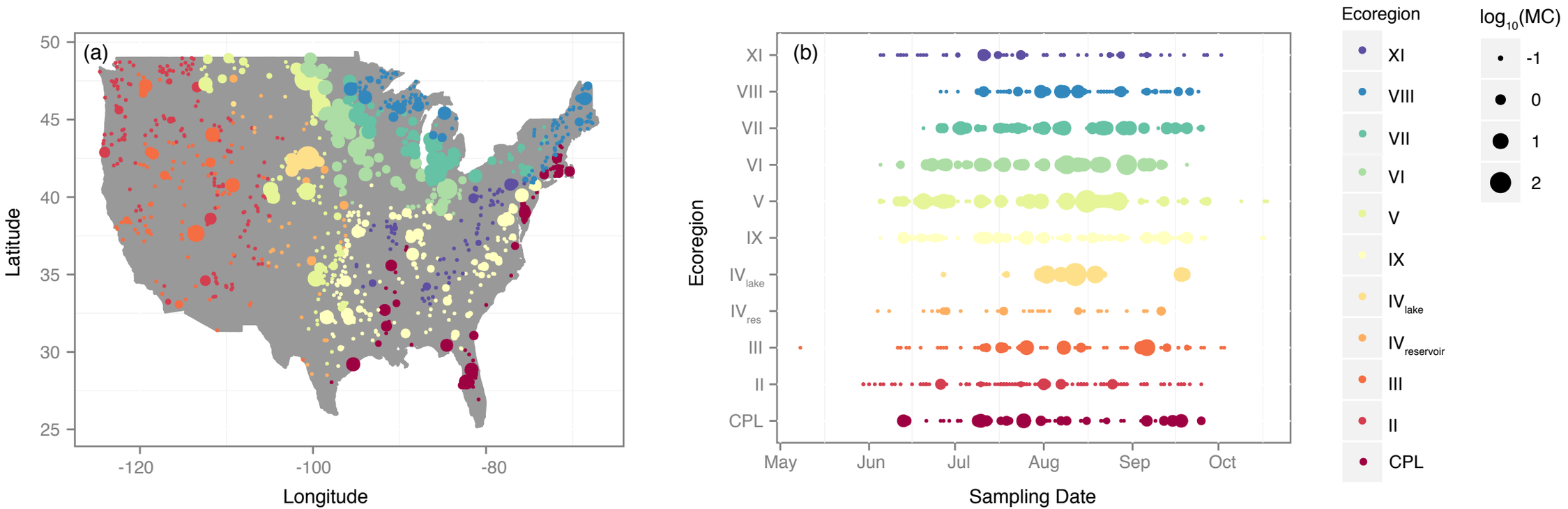

Figure S1 Spatio-temporal summary of ecoregion factor. (a) Map of NLA lakes colour-coded by ecoregion levels, where the size of each point corresponds to the $\log_{10}$-transformed concentration of microcystins. The mountainous ecoregion (Ecoregion II), which includes the Rocky Mountains, the Cascade Range and Sierra Nevada, divides the intermountain xeric ecoregion to the West (Ecoregion III) from the coastal and interior plain ecoregions to the East. The Appalachian mountain range (Ecoregion XI) further divides the coastal and interior plains. (b) May to October lake sampling date broken down by ecoregion.

Figure S2

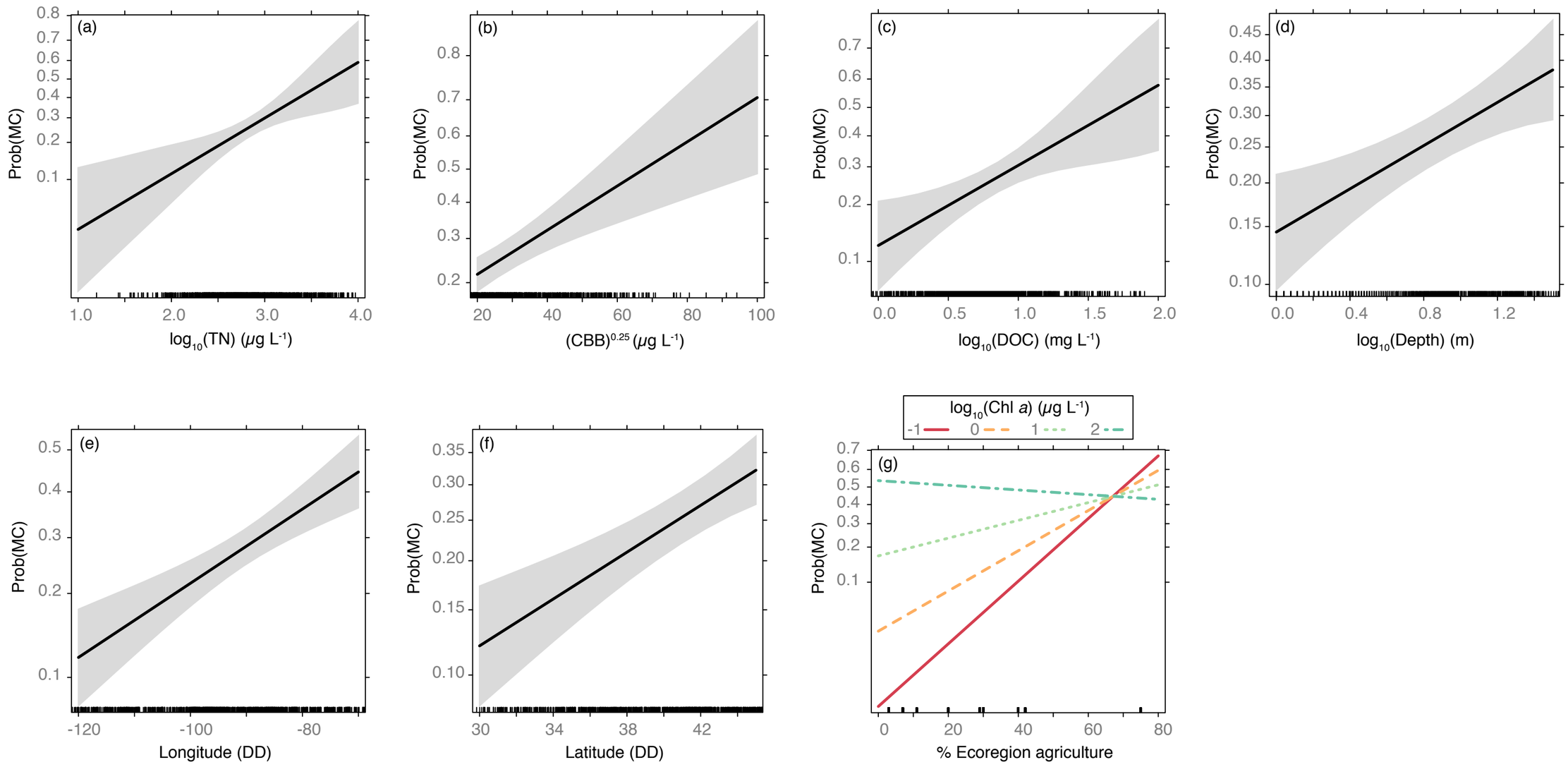

Figure S2 Fitted curves of best-fit binomial GLM predicting the presence (above detection limit) and absence (below detection limit) of microcystins (MC) in NLA lakes (R^2_{marg} = 0.55, n = 1127), where TN = total nitrogen concentrations, CBB = cyanobacteria biomass, DOC = dissolved organic carbon, Depth = maximum lake depth, and Chl *a* = chlorophyll *a* concentrations. All variables appear in standardized values. Partial fits were plotted using the *effects* package in R.

Figure S3

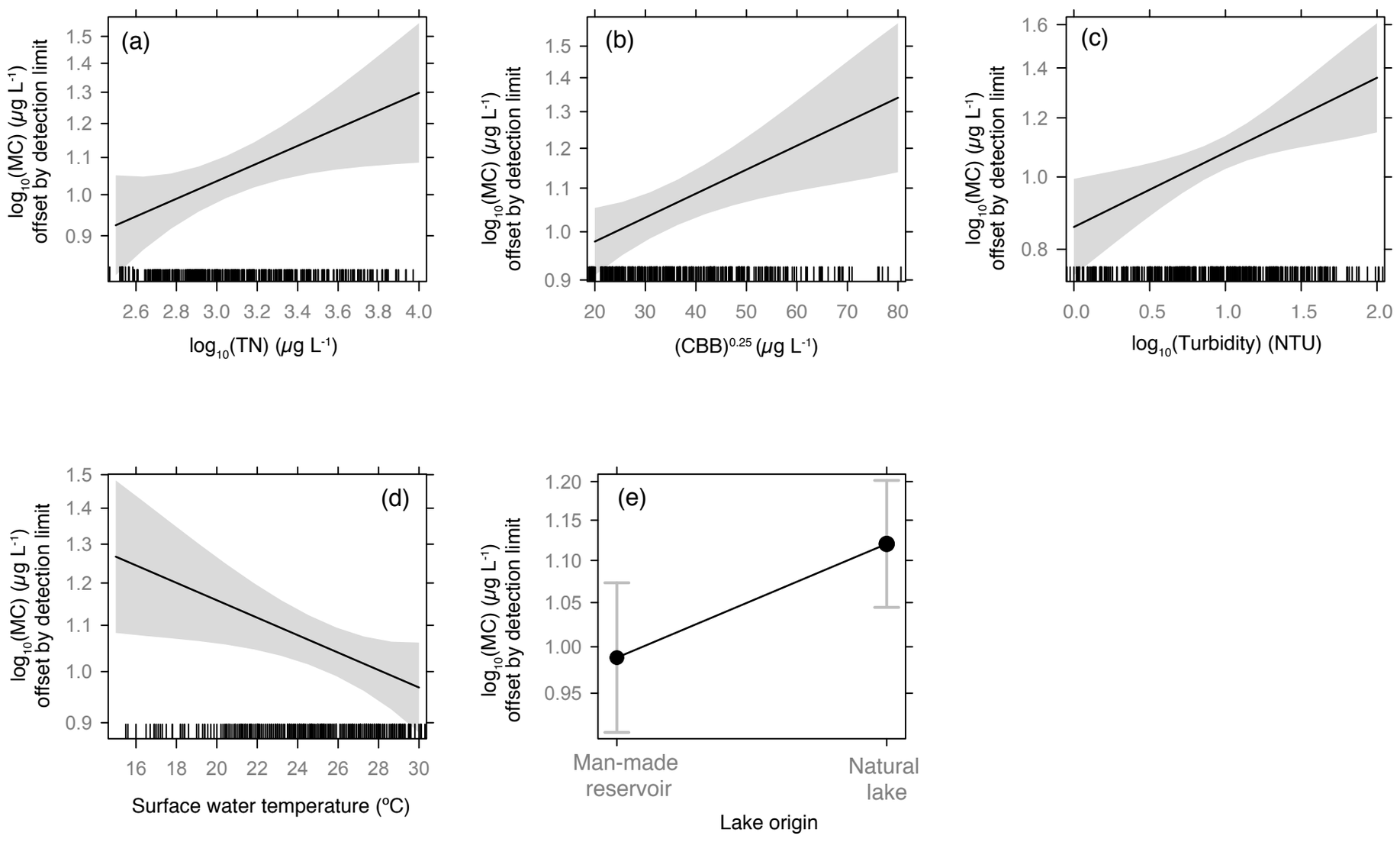

Figure S3 Fitted curves for the best-fit ZAG model (R^2_{marg} = 0.29; n = 362). Abbreviations are as described in Fig. S2. Partial fits were plotted using the *effects* package in R.

Figure S4

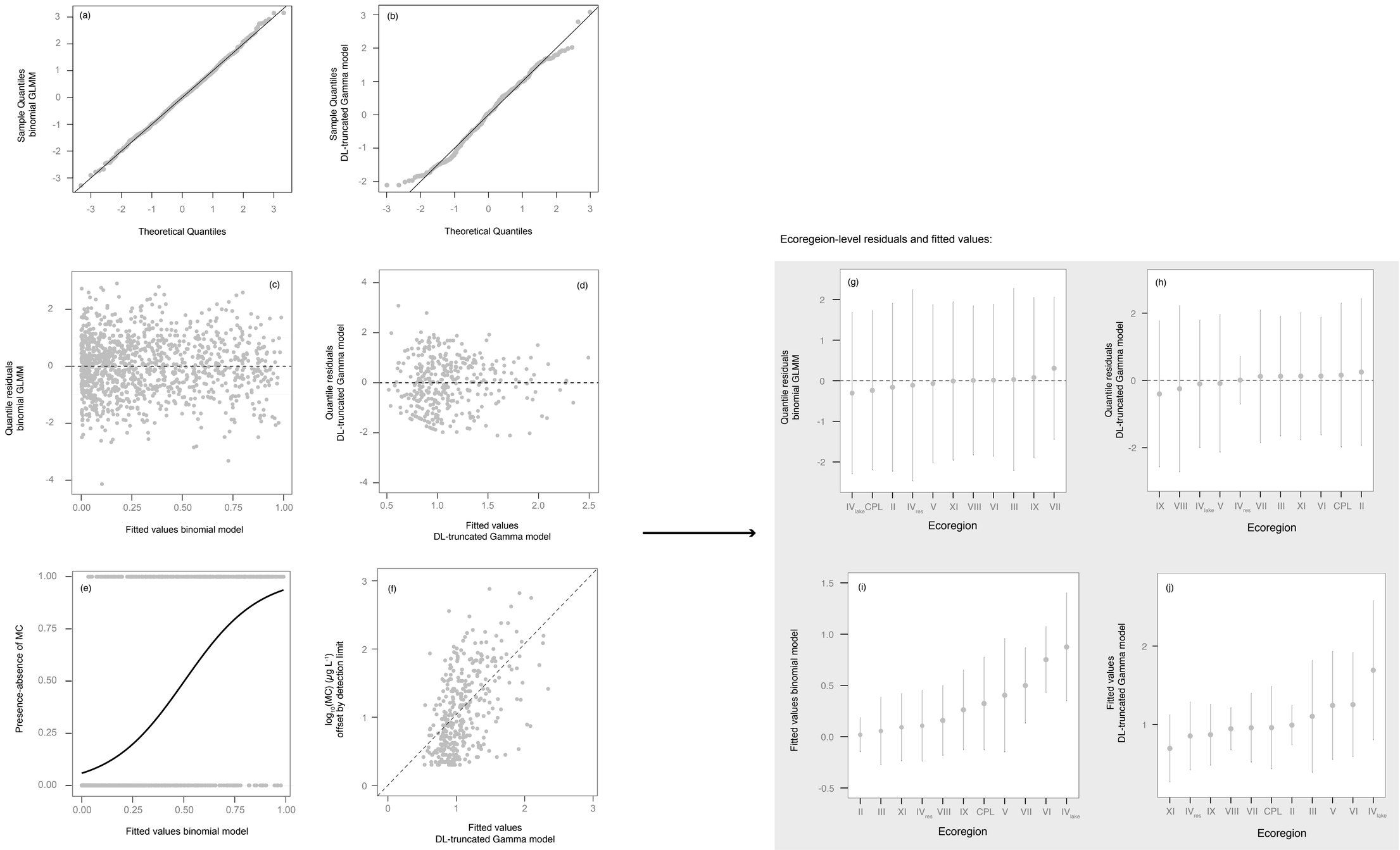

Figure S4 Diagnostic plots of final zero-altered *hurdle* model. (a, b) QQ-plots using randomized quantile residuals, (c, d) scatterplots of quantile residuals vs. fitted values, (e, f) scatterplots of observed vs. fitted values. Inset panel: (g-h) randomized quantile residuals vs. ecoregions, and (i-j) fitted values vs. ecoregions (shown are the mean and standard error of residuals and fitted values, respectively).

Figure S12

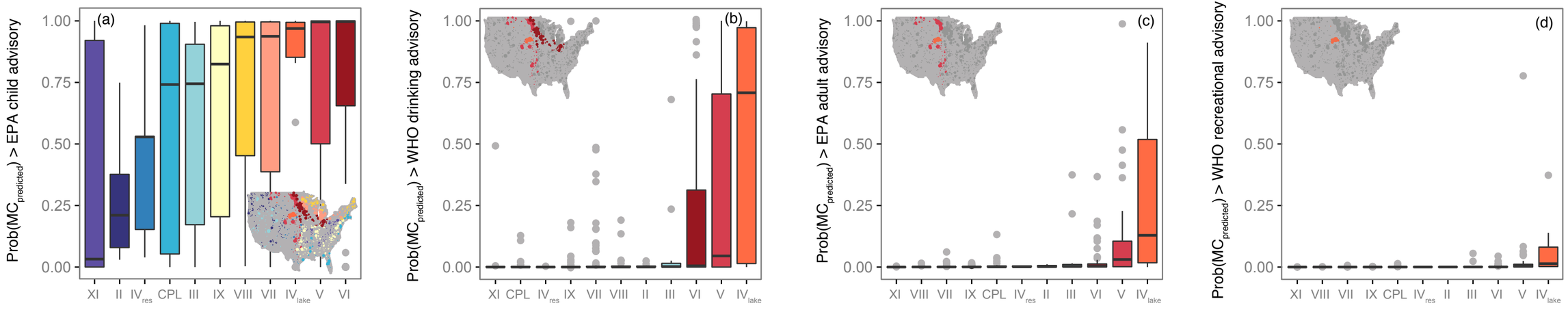

Figure S12 Probability that the fitted values of the final *hurdle* model fall above each provisional guideline; where the probability corresponds to probability that the microcystins concentration in a given lake is above the detection limit times the probability that the Gamma distribution with the fitted values and dispersion is above the provisional guideline. Probabilities have been separated by ecoregions to showcase the spatial heterogeneity in predicted values.

Figure S6

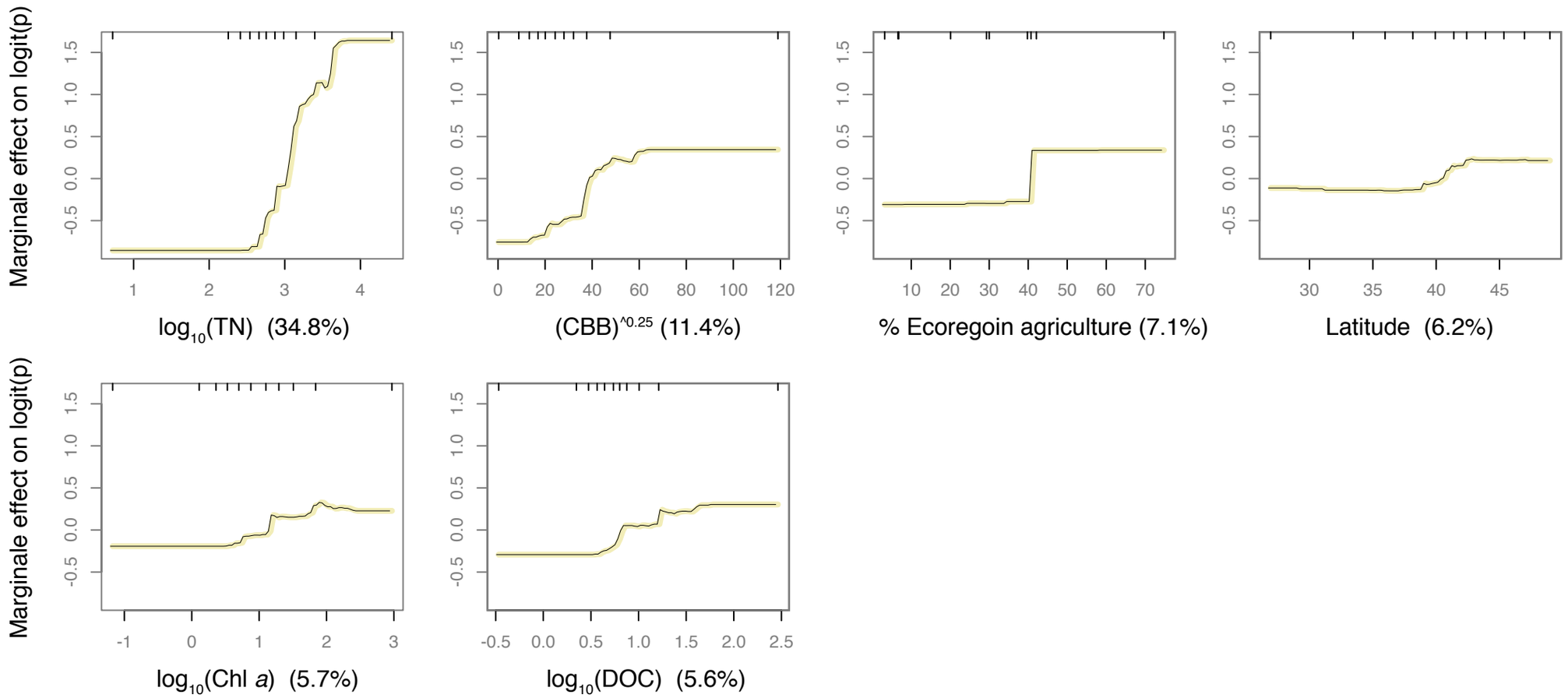

Figure S6 Partial dependence plots for the most influential variables (relative influence ≥ 5%) in the BRT model for microcystin (MC) occurrence above the U.S. EPA drinking advisory for children. See Table S1 for variable units. Y-axes are shown on the logit scale and centered to have zero mean over the data distribution. Rug plots at top of plots indicate the decile distribution of sites across each variable.

Figure S7

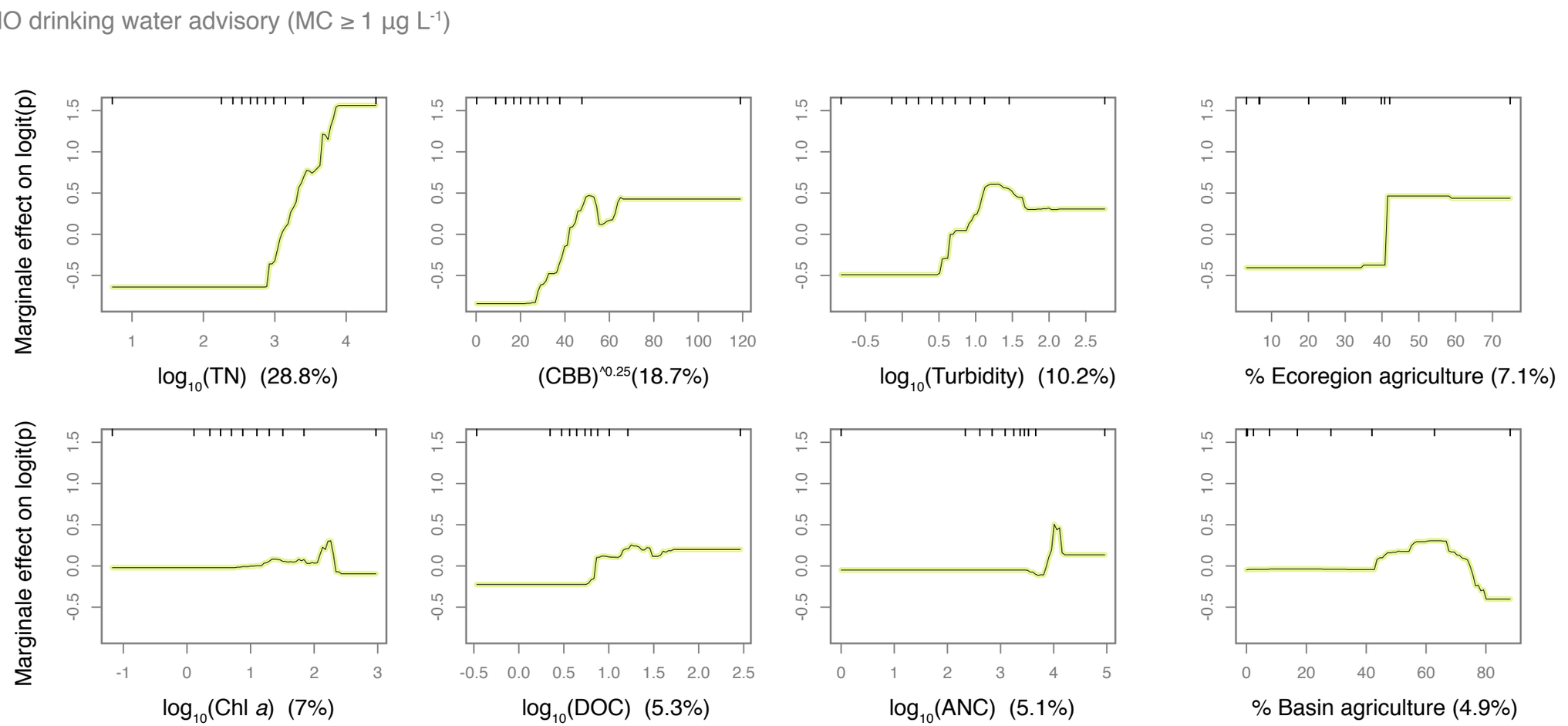

Figure S7 Partial dependence plots for the most influential variables in the BRT model for microcystin (MC) occurrence above the WHO drinking advisory for. See Fig. S6 for plot description.

Figure S8

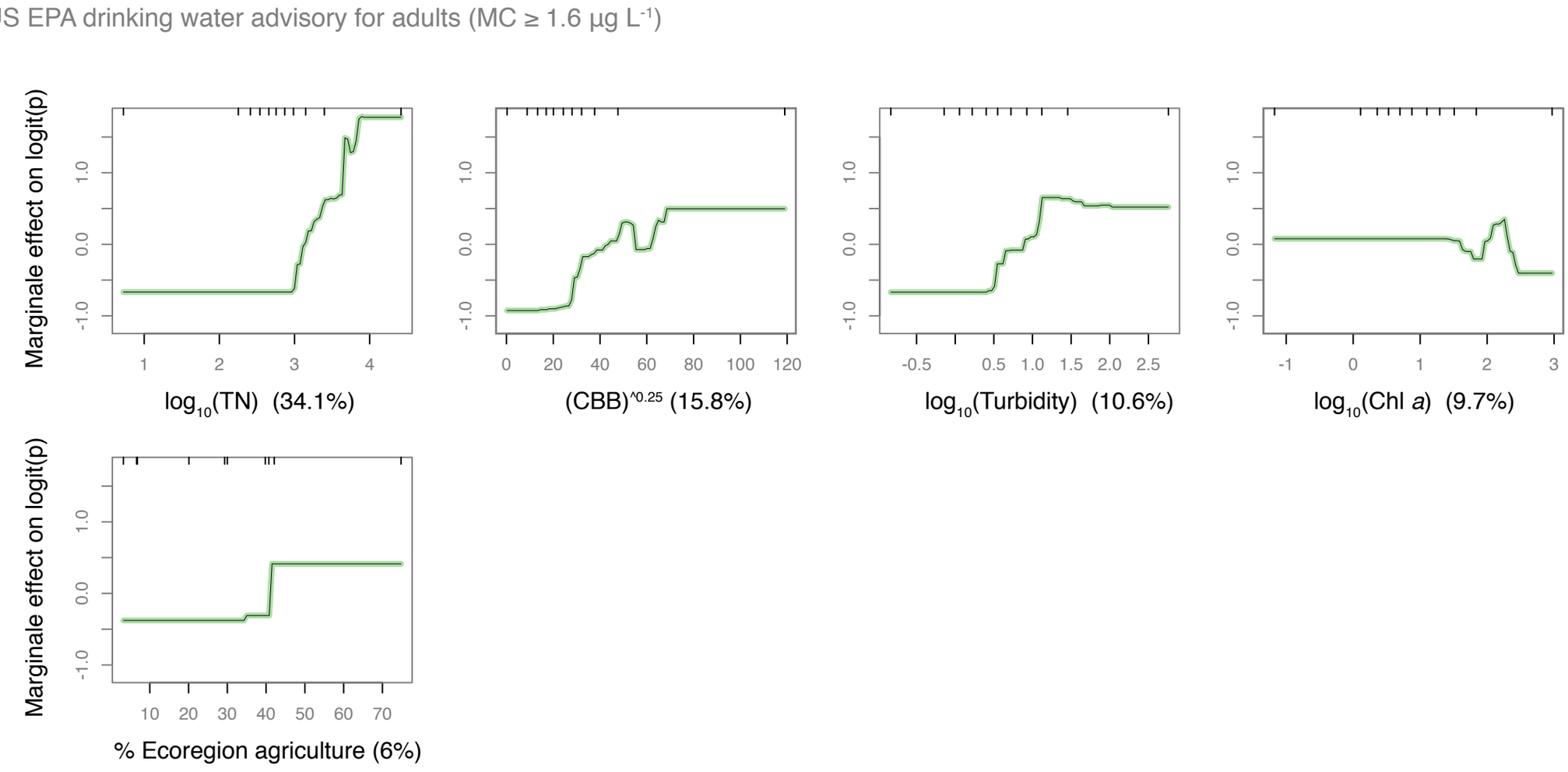

Figure S8 Partial dependence plots for the most influential variables in the BRT model for microcystin (MC) occurrence above the U.S. EPA drinking advisory for adults. See Fig. S6 for plot description.

Figure S9

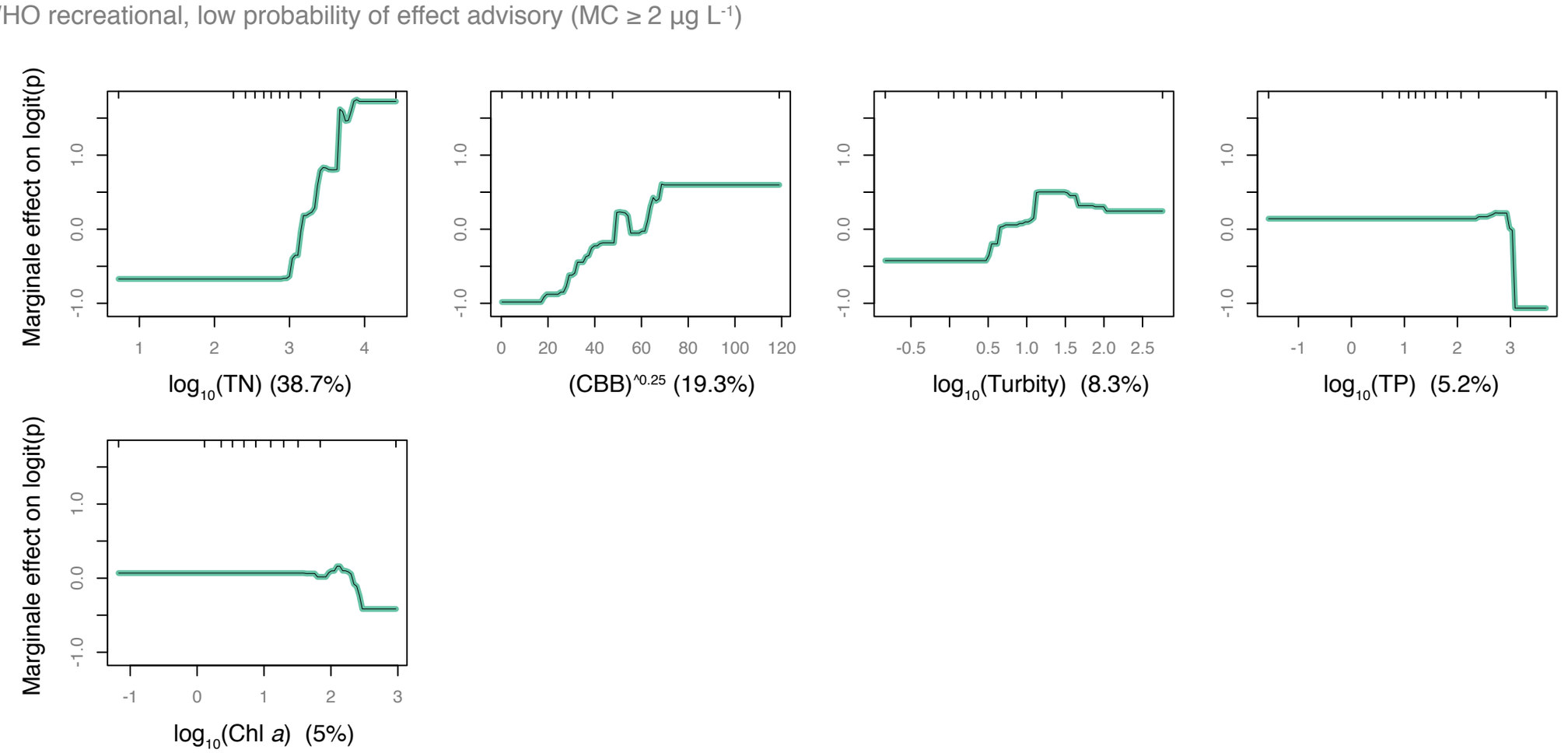

Figure S9 Partial dependence plots for the most influential variables in the BRT model for microcystin (MC) occurrence above the WHO recreational, low probability of effect advisory. See Fig. S6 for plot description.

Figure S10

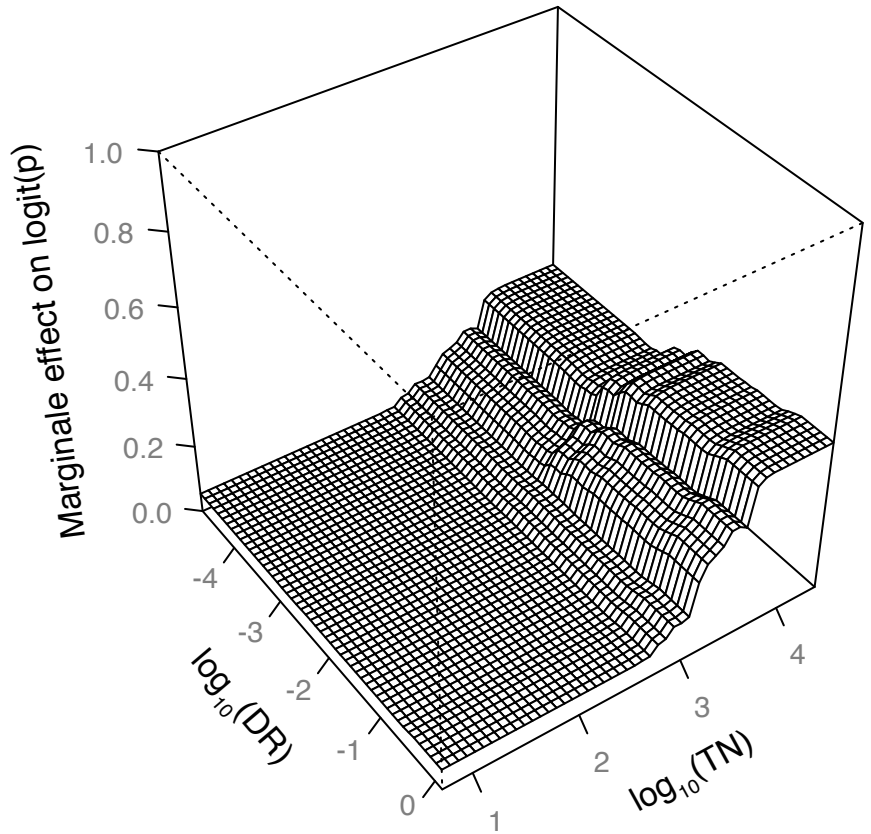

Figure S10 Three-dimensional partial dependence plots for the strongest interaction in the BRT model for microcystin (MC) occurrence above the U.S. EPA drinking advisory for children. All variables except those graphed are held at their means, where DR = drainage ratio and TN = total nitrogen concentration.

Figure S11

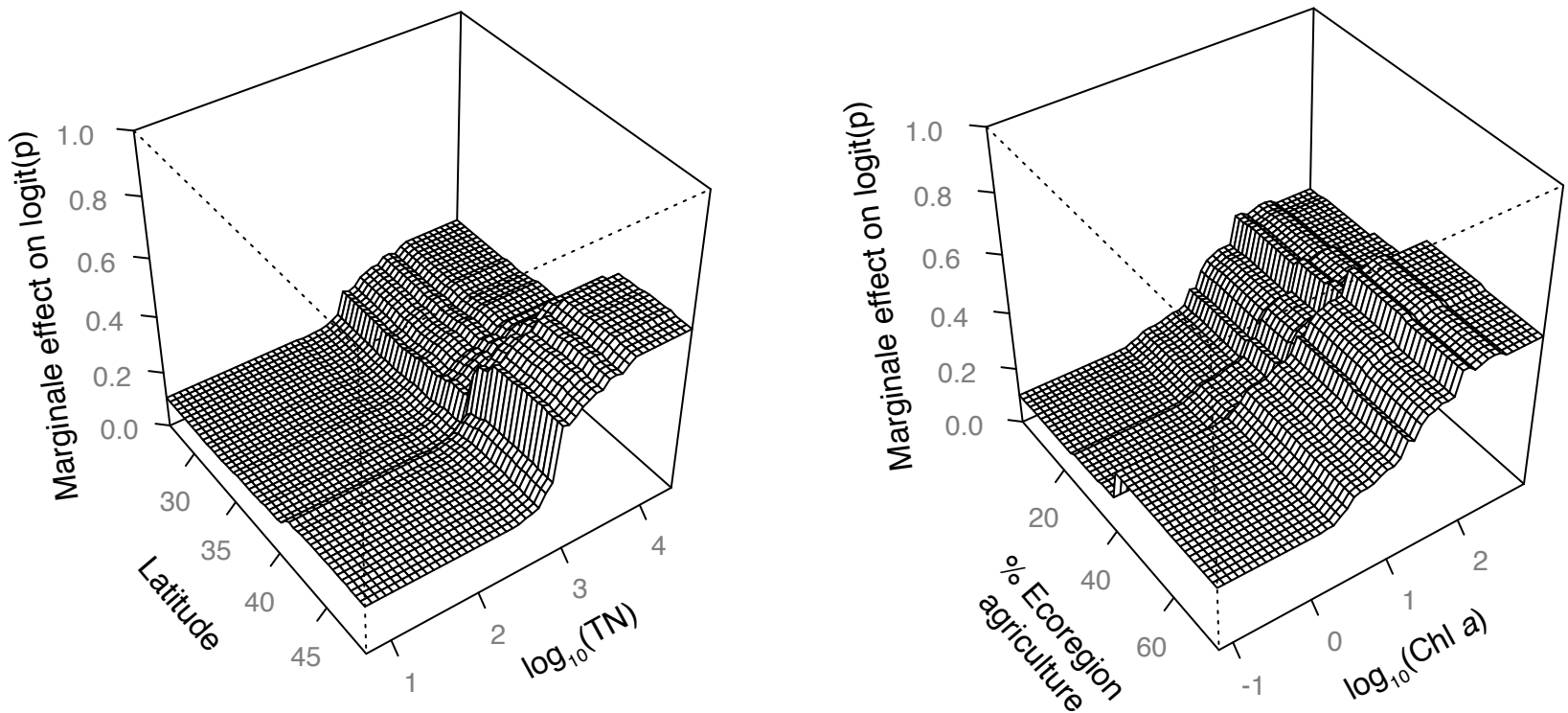

Figure S11 Three-dimensional partial dependence plots for the strongest interaction in the BRT model for microcystin (MC) presence (above detection limit) and absence (below detection limit) illustrating the interaction between total nitrogen (TN) and latitude, as well as the cross-scale interaction between % ecoregion agriculture and chlorophyll *a* (Chl *a*).

Figure S12

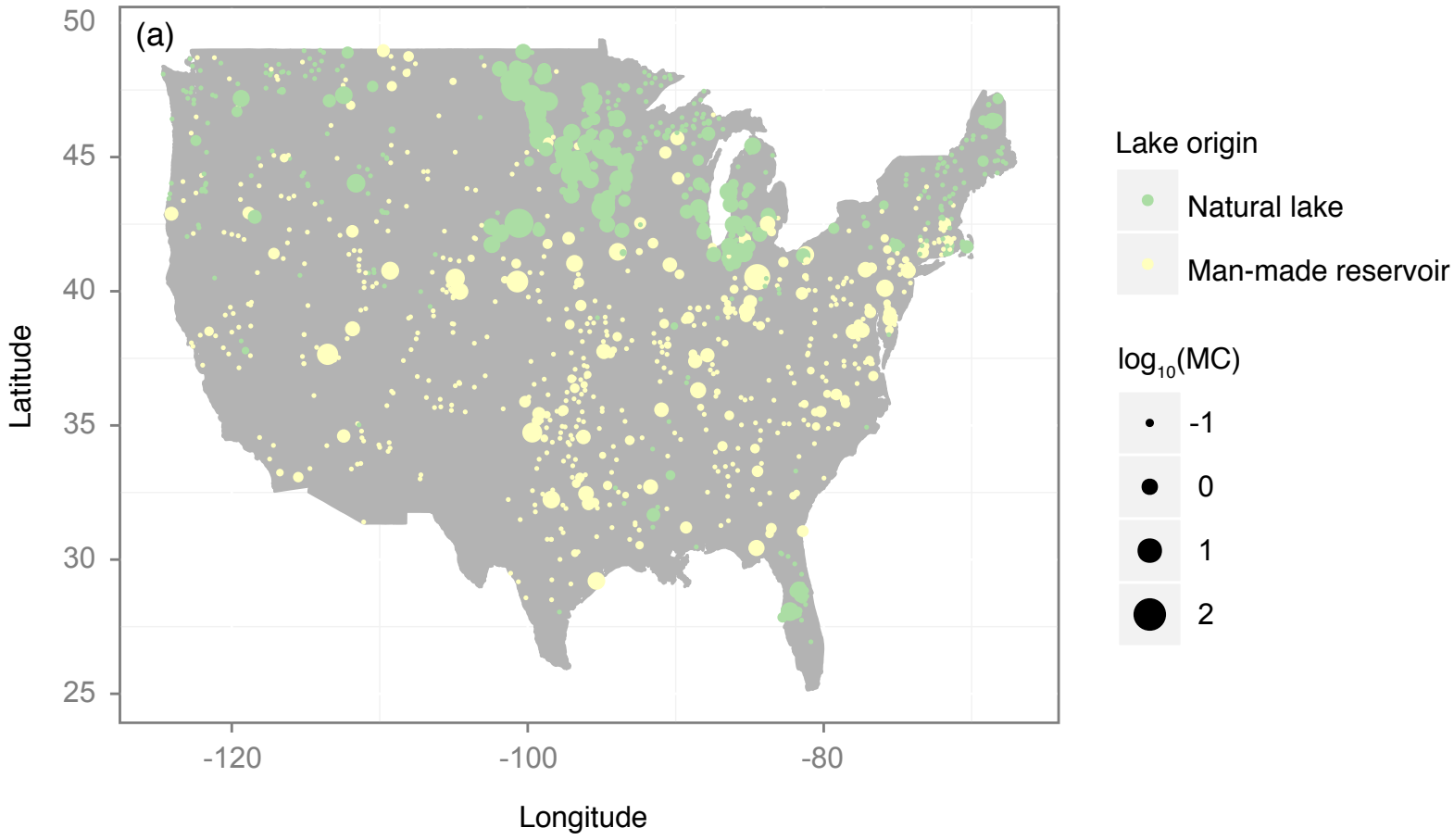

Figure S12 Map illustrating spatial distribution of natural lakes versus man-made reservoirs, where size of circles represents the log-transformed microcystin (MC) concentration in each lake.