ABSTRACT

Baird, Brian Wray. Refining Chlorophyll-Nutrient Relationships across the U.S. Considering the Effect of Depth, Nutrient Enrichment, and Temperature. (Under the direction of Dr. Daniel Obenour).

Lake eutrophication is a significant challenge for water-quality management across the world. Many management strategies focus on reducing nutrient inputs, but other environmental factors can substantially influence nutrient-algal relationships. While there have been various large-scale studies assessing how different lake characteristics impact eutrophication, there remains a need to integrate these factors into a comprehensive model. In this study, we refine chlorophyll-nutrient relationships across the conterminous United States (CONUS) by considering auxiliary variables (e.g., temperature, lake depth, nutrient enrichment) within a Bayesian hierarchical framework. We leverage over 4000 sampling events of 2755 different lakes from the National Lakes Assessments (US EPA, 2007-2022) to inform model development. We first consider auxiliary variables independently, exploring how they influence the intercept, slope, and critical nutrient ratio (nitrogen:phosphorus) in a regression to predict chlorophyll based on the limiting nutrient. Next, we integrate significant auxiliary variables into a comprehensive model for chlorophyll prediction. Results indicate an increase in the critical nutrient ratio in relation to increased lake depth, and an increase in the slope of the regression in relation to temperature. We apply the model to map summer nutrient limitation and find similar percentages of CONUS lakes experience strong nitrogen and strong phosphorus limitation, with slightly more lakes indicating phosphorus limitation. These proportions, however, varied substantially across different regions. Our model can serve as an effective tool to aid water resources management, particularly at large spatial scales.

Refining Chlorophyll-Nutrient Relationships across the U.S. Considering the Effect of Depth, Nutrient Enrichment, and Temperature

by
Brian Wray Baird

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Civil Engineering

Raleigh, North Carolina
2025

APPROVED BY:

_____	_____
Dr. Daniel Obenour Committee Chair	Dr. Jacelyn Rice-Boayue

_____
Dr. Casey Dietrich

DEDICATION

To the love of my family, the joy of my friends, and the humble pursuit of knowledge.

BIOGRAPHY

Brian Baird was born in Louisiana and raised in the pine forests of Northeast Texas. He received his Bachelor of Science in Civil Engineering from the University of Arkansas in 2023. It was there, in the beauty of the Ozarks, where his passion for water resources took root. Inspired by his studies, he headed east to pursue a Master of Science in Civil Engineering with a focus in Water Resources Engineering under the esteemed guidance of Dr. Daniel Obenour at North Carolina State University in 2023. As his time at NC State drew to a close, he looked with anticipation to the next steps in his career as a surface water consultant back in his home state of Texas.

ACKNOWLEDGMENTS

This work was supported by the Science and Technologies for Phosphorus Sustainability (STEPS) Center, a NSF Science and Technology Center (CBET-2019435). The assistance of Dr. James Elser is especially appreciated.

The National Lakes Assessment 2007, 2012, 2017, and 2022 data were a result of the collective efforts of dedicated field crews, laboratory staff, data management and quality control staff, analysts and many others from EPA, states, tribes, federal agencies, universities, and other organizations. Please contact nars-hq@epa.gov with any questions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 Introduction

Algal blooms in aquatic systems cause many adverse effects. Harmful algal blooms (HAB) impact the ecosystem by increasing animal morbidity and mortality rates both directly (e.g., producing harmful toxins that lead to poisonings) and indirectly (e.g., dissolved oxygen depletion leading to large fish kills) (Anderson et al., 2002; Backer et al., 2015; Carmichael & Boyer, 2016). Algal blooms can also cause many negative health effects in humans such as dermal, gastrointestinal, or respiratory illnesses through the consumption of poisoned organisms, inhalation of toxins, or direct contact with water affected by the blooms (Backer et al., 2015; Grattan et al., 2016; Lad et al., 2022). The direct impacts from harmful algal blooms often lead to economic losses as well (Sharma et al., 2014). Not only are the effects adverse, but the occurrence of harmful algal blooms globally have been increasing in recent decades due to various anthropogenic and climatic factors (Fang et al., 2022; Huisman et al., 2018).

The development of national and international databases on algae and water quality have made it possible to explore potential HAB drivers over large spatial scales. Large-scale models have been developed to quantify underlying relationships between algal abundance and driving factors, assess spatio-temporal trends in algal bloom occurrence, and inform water management strategies in lakes across the world (Qin et al., 2020; Quinlan et al., 2021; Zhao et al., 2023). Within the United States, large-scale models have been developed to determine the effects nutrients, lake and watershed characteristics, meteorological conditions, and other potentially influential factors have on algal concentrations. These studies have found that nutrient concentrations (primarily nitrogen and phosphorus) are the driving factors behind algal production, but other factors such as temperature, lake depth, trophic status, and precipitation can

be used to further enhance the predictive power of models (Brehob et al., 2024; Ho & Michalak, 2020; Liang et al., 2023; Yuan & Pollard, 2014). Large-scale studies have also been used to show the convergence of US lakes towards a eutrophic condition over recent decades (Fernandez-Figueroa et al., 2024), enhance the prediction of harmful toxins produced by algal blooms across the country (Taranu et al., 2017; Yuan & Pollard, 2017), and depict spatiotemporal variation in chlorophyll production and nutrient limitation (Rock & Collins, 2024; Wagner et al., 2011).

Much historical algal modeling literature focuses on Liebig's law of the minimum, which states that an organism's growth is limited by the scarcest essential resource rather than the total amount of resources available (de Baar, 1994; Saito et al., 2008). In reference to algae, this principle becomes most relevant when assessing chlorophyll-nutrient relationships (i.e., the effects of nitrogen and phosphorus availability on algal abundance). Intuitively, this suggests any excess of nitrogen (N) in a phosphorus (P)-limited system or excess of P in an N-limited system would have minimal marginal impact on algal accumulation. Nutrient limitation is identified in relation to a critical N:P ratio, which represents how much nitrogen is needed in relation to phosphorus to support algal accumulation (Dolman & Wiedner, 2015). When observed N:P ratios are higher than this critical N:P ratio, the waterbody experiences P-limitation, and when observed N:P ratios are lower than this critical N:P ratio, the waterbody experiences N-limitation. Knowing the limiting nutrient enables water resource managers to identify what nutrient reductions will be most beneficial (Liang et al., 2020; Moon et al., 2021). However, most large-scale studies do not incorporate the law of the minimum directly. When nutrient limitation is incorporated, many studies assume a constant critical N:P ratio that is based on

typical values observed in freshwater and marine systems worldwide (Guildford & Hecky, 2000; Naderian et al., 2025; Qin et al., 2020).

In contrast to many large-scale algal studies, Dolman and Wiedner (2015) utilized a limiting nutrient model that provides unique insights into chlorophyll-nutrient relationships over large spatial scales (Dolman & Wiedner, 2015). This study utilized Liebig's law of the minimum in a novel approach that included calibrating a critical N:P ratio to observed data. This enabled the defining of lakes as nitrogen or phosphorus limited based on comparing the observed N:P ratio to the model's calibrated critical N:P ratio. This approach found that different lake datasets yielded different critical N:P mass ratios, ranging from 16 to 25 (when using total nutrients). Furthermore, this model was able to determine the frequency and seasonal patterns of nutrient limitation for algal accumulation for a variety of different lake types (Dolman et al., 2016).While this study provided helpful insights into chlorophyll-nutrient relationships in lakes, it did not consider how other factors aside from nutrient levels could impact the critical N:P ratio and overall chlorophyll levels.

Despite the benefits offered by a limiting nutrient formulation, many large-scale models have not adopted this approach. A limiting nutrient formulation results in non-linarites that cannot be readily accommodated within conventional regression approaches, and it cannot generally be imposed by more flexible machine learning approaches. This is where the employment of a Bayesian modeling approach becomes potentially beneficial. Bayesian approaches allow for obtaining probabilistic parameter estimates (e.g., critical N:P ratio) and predictions for increasingly complex models by leveraging both informative datasets and prior knowledge

(Katin et al., 2021). Bayesian models have successfully been used for various purposes within the field of surface water quality such as predicting the mass N:P ratios of phytoplankton (Yuan & Jones, 2022), quantifying phosphorus sources in watersheds (Karimi & Obenour, 2024; Wei et al., 2024), and predicting seasonal nutrient and chlorophyll concentrations (Del Giudice et al., 2021; Jackson-Blake et al., 2022).

In this study, a Bayesian regression approach was used to model relationships between algae (chlorophyll) and total nitrogen and phosphorus in conterminous United States (CONUS) lakes based on Liebig's law of the minimum. In addition, we assessed how chlorophyll-nutrient relationships are potentially mediated through auxiliary factors (i.e., lake depth, nutrient enrichment, and temperature). First, models were developed to determine how regression parameters (i.e., intercept, slope, and critical N:P ratio) may vary across different bin ranges for the auxiliary variables. Then, a final model was developed to integrate the combined, apparent effects of multiple auxiliary variables on algal production. Model output was used to assess if the critical N:P ratio varies substantially across CONUS lakes. Furthermore, the model was used to investigate what CONUS lakes are most likely to experience strong N- or P-limitation.

2.0 Methodology

2.1 Data processing

Models were created using data from EPA's National Lakes Assessment (NLA) for the years 2007, 2012, 2017, and 2022 (U.S. Environmental Protection Agency, 2010, 2016, 2022a, 2024). Each NLA year includes an abundance of data on lakes, ponds, and reservoirs across the conterminous United States (CONUS), which were collected and analyzed using generally consistent methods (U.S. Environmental Protection Agency, 2022b). The primary NLA data used for model formulation were chlorophyll-a (chl-a) concentration, total nitrogen (TN) concentration, total phosphorus (TP) concentration, and water temperature in the surface layer (upper 2 m), as well as water depth at the sampling location. Sampling was generally targeted toward the center or deepest portion of each lake. The combined dataset contained over 4000 sampling events across 2755 CONUS waterbodies. Data were collected from May-October with a great majority of collection occurring from June-August (Section SI 1.1), and thus represent warm conditions, referred to as "summer" in this study, for brevity.

The data were processed and analyzed using R version 4.4.1 (R Core Team, 2023) in RStudio (RStudio Team, 2020) utilizing various R packages (Garnier et al., 2023; Koenker, 2023; Landis, 2022; Wickham et al., 2019). Exploratory assessment of the data included the development of histograms and other visualizations (Figures SI 1.1.1 – 1.1.3), and it was determined that outlier removal was appropriate, particularly for chl-a, TN, and TP. Outlier removal was primarily conducted using an interquartile range approach, such that 94 of the original 4242 observations (2.2%) were removed from the dataset (Section SI 1.1). Maps of the data used in this project were created using ArcGIS Pro (Esri, 2024) (Figure SI 1.1.4).

2.2 Algal models based on nutrients alone

Models were created using the structure shown in equations 2.2.1 and 2.2.2, where *a* is chl-a (µg/L), *TN* is total nitrogen (µg/L), *TP* is total phosphorus (µg/L), *LN* is the limiting nutrient in phosphorus equivalents (µg/L) for a waterbody, and β terms are parameters fitted through Bayesian inference. In most cases, variables were log-transformed, consistent with physical expectations and statistical considerations (Faraway, 2015; Ott, 1990).

$$\ln(a) = \beta_0 + \ln(LN) * \beta_{LN} + \epsilon_l + \epsilon_s \quad \text{Eq. 2.2.1}$$

$$LN = \min\left[TP, \frac{TN}{\beta_{CR}}\right] \quad \text{Eq. 2.2.2}$$

In this model, β_{CR} represents the critical nitrogen to phosphorus (N:P) mass ratio described in the introduction, β_{LN} represents the slope, and β_0 represents the intercept. Note that because nutrients and chl-a share the same units (µg/L), slope coefficients are unitless in this study. Finally, ϵ_l is a normally-distributed (on the log scale) lake-specific random effect (RE), $\epsilon_l \sim N(0, \sigma_l)$, and ϵ_s is the sample-specific model residual, $\epsilon_s \sim N(0, \sigma_s)$. This hierarchical formulation (Gelman & Hill, 2006; Qian et al., 2010) addresses the fact that 949 (34%) lakes include multiple samples across the four NLA years.

To compare the limiting nutrient approach to a more conventional regression approach, additional models were constructed where nutrient contributions are additive on the log scale. When back-transformed to the original scale, the nutrient effects become multiplicative, and so we refer to this as the "multiplicative" nutrient formulation, for brevity. The multiplicative model has the following structure:

$$\ln(a) = \beta_0 + \ln(TN) * \beta_{TN} + \ln(TP) * \beta_{TP} + \epsilon_l + \epsilon_s \quad \text{Eq. 2.2.3}$$

2.3 Algal models including individual auxiliary variable

To assess how algal-nutrient relationships may vary with other lake characteristics, the coefficients described in equations 2.2.1-2.2.3 (e.g., $\beta_0, \beta_{LN}, \beta_{CR}$ for Eq. 2.2.1.) were allowed to vary across different ranges (i.e., 'bins') of auxiliary variables. The auxiliary variables were motivated by previous studies (see Introduction) and included depth, temperature, and nutrient enrichment, all from the NLA datasets. In this study, nutrient enrichment was expressed as a "nutrient enrichment index" (NEI) determined by averaging the standardized natural logs of TN and TP concentrations for each observation (i.e., averaging the z-scores of TN and TP on the natural log scale for each observation). The means and standard deviations associated with this standardization process were 3.7 ln(µg/L) and 1.3 ln(µg/L) for TP, and 6.5 ln(µg/L) and 0.9 ln(µg/L) for TN, respectively.

We initially developed an individual model for each auxiliary variable, where all coefficients were allowed to vary hierarchically. Specifically, each model coefficient (β_X) was allowed to vary across five bins (j), where each bin contained a quintile of the dataset, ordered based on the auxiliary variable under consideration. Bin-level coefficients were constrained by a hyperdistribution with mean μ_X and standard deviation τ_X, such that $\beta_{X,j} \sim \mathrm{N}(\mu_X, \tau_X)$. In the case of Equations 2.2.1 and 2.2.2, the revised hierarchical formulation is as follows:

$$\ln(a) = \beta_{0,j} + \ln(LN) * \beta_{LN,j} + \epsilon_l + \epsilon_s \qquad \text{Eq. 2.3.1}$$

$$LN = \min\left[TP, \frac{TN}{\beta_{CR,j}}\right] \qquad \text{Eq. 2.3.2}$$

For each individual auxiliary variable (IAV) model, we examined patterns in the bin-level parameters (relative to the auxiliary variable under consideration). Furthermore, we statistically assessed the distinguishability of the bin-level parameters of the two end bins (i.e., $\beta_{X,1}$ and $\beta_{X,5}$) by determining the probability that the bin-5 coefficient exceeds the bin-1 coefficient (based on Bayesian posterior samples, described below). Probabilities greater than 97.5% and less than 2.5% indicate that the 95% credible interval for the difference between the two end-bin coefficients does not include zero. In this case, the auxiliary variable was generally considered to be significant in mediating algal-nutrient relationships.

2.4 Algal model with multiple auxiliary variables

Finally, we constructed a comprehensive model including multiple auxiliary variables (i.e., the MAV model). To avoid excessive binning and over-parameterization, we developed relationships between the auxiliary variables and the primary model coefficients (e.g., $\beta_0, \beta_{LN}, \beta_{CR}$) using linear formulations:

$$\beta_X = \gamma_0 + \sum_{k=1}^{m} w_k * \gamma_k \qquad \text{Eq. 2.4.1}$$

Here, each β_X is represented as the linear combination of *m* auxiliary variables (w_k) with coefficients, γ_k and γ_0 as an intercept. We initially included all of the significant auxiliary variables identified in the IAV models (described above). Most auxiliary variables were log-transformed to help linearize these relationships. Finally, for parsimony, any auxiliary variables that were found to be both insignificant (i.e., 95% credible interval for associated γ_k included zero) and without a clear mechanistic foundation were pruned from the MAV model using backward selection (Faraway, 2015).

Bayesian inference for model parameters was conducted using Markov Chain Monte Carlo (MCMC) implemented via the Nimble and Coda packages in R (Plummer et al., 2006; Valpine et al., 2024). Given the large dataset available for model fitting, we decided to primarily use uninformative priors (wide normal or uniform distributions). For the critical ratio, β_{CR}, a vaguely informative prior was selected considering previous research related to the critical ratio (Dolman & Wiedner, 2015; Guildford & Hecky, 2000). A list of priors used in the development of all models can be found in Table 2.4.1.

Table 2.4.1. Model parameters and priors. The second column indicates in which models each parameter is used. IAV_L and MAV refer to the individual and multiple auxiliary variable models, respectively, both using a limiting nutrient formulation. IAV_M denotes the IAV model with a multiplicative nutrient formulation. The MAV parameter rows include all coefficients used in Eq. 2.4.1. Units designated as "various" are either unitless or ln(µg/L), depending on the corresponding parameter. Note that SD indicates standard deviation. *For the critical ratio, we use a prior centered on 20 rather than zero.

Parameter	**Models**	**Units**	**Prior**
Intercept, β_0	IAV_L, IAV_M	ln(µg/L)	N(0, 10)
LN slope, β_{LN}	IAV_L	unitless	N(0, 10)
Critical ratio, β_{CR}	IAV_L	unitless	N(20, 5)
TN slope, β_{TN}	IAV_M	unitless	N(0, 10)
TP slope, β_{TP}	IAV_M	unitless	N(0, 10)
SD β_X, τ_x	IAV_L, IAV_M	various	U(0,10)
MAV parameters, γ_k	MAV	various	N(0, 10)
MAV parameters, γ_0	MAV	various	N(0*,10)
Lake RE SD, σ_l	all	ln(µg/L)	U(0,10)
Residual SD, σ_s	all	ln(µg/L)	U(0,10)

MCMC was conducted with 3 chains consisting of at least 40,000 samples each using default Nimble samplers (metropolis and conjugate samplers). At least the first 33% of samples were discarded as burn-in, allowing sufficient iterations to achieve convergence on the posterior parameter distribution. The Gelman-Rubin diagnostic (Gelman & Rubin, 1992) was used to assess chain convergence for each parameter (i.e., $\hat{R} < 1.1$ indicates acceptable convergence). For the MAV model, auxiliary variables were zero-centered to improve convergence diagnostics.

2.5 Model diagnostics and cross validation

Model performance was assessed based on the coefficient of determination (R^2), root mean squared error (RMSE), and residual plots of the data. The final MAV model was also subject to a four-fold cross-validation. In this process, the data were partitioned into four different folds, based on the four NLA years. The model was then trained (i.e., parameters were re-estimated) with one fold of data removed, and then the model was used to make predictions for the removed (hold-out) data. This process was repeated four times to generate hold-out predictions across the entire original dataset. The complete set of hold-out predictions could then be used to assess overall model performance (i.e., RMSE and R^2).

2.6 Model application and mapping

Due to the limiting nutrient structure in our final model, it was possible to determine the likelihood of whether the algal growth in each lake was limited by nitrogen or phosphorus availability. To do this, two uncertainties were accounted for: the uncertainty in the mean N:P ratio (R_o) and the uncertainty in the critical ratio identified by the model (β_{CR}). Both distributions were approximated for each individual waterbody using Bayesian posterior samples.

Each lake's mean summer N:P ratio (R_o) is imprecisely known from the limited samples available (R_i) in the NLA dataset. To account for this, R_o was expressed with a normal distribution, $R_0 \sim \mathrm{N}(\mu_X, \sigma_{l,R})$, where μ_X is the mean R_o across all lakes, and $\sigma_{l,R}$ represents inter-lake variability (i.e., μ_X and $\sigma_{l,R}$ are hyperparameters). R_i is then expressed as $R_i \sim \mathrm{N}(R_0, \sigma_{s,R})$ where $\sigma_{s,R}$ reflects within-lake variability. Using a similar MCMC sampling approach as outlined above, Bayesian posterior sample distributions for R_o were obtained for each lake.

The Bayesian posterior samples from the MAV model were used to obtain the distribution associated with the critical ratio, β_{CR}, for each lake. Specifically, the observed lake auxiliary variable conditions (w_k) and the posterior parameter samples (γ_k and γ_0) were used in the equation for β_{CR} (Equation 2.4.1). This yielded a distribution of the critical ratio for each lake that captures the uncertainty in β_{CR}.

A probabilistic assessment of nutrient limitation was then obtained by randomly sampling from the posterior distributions of R_o and β_{CR}. In particular, we obtained a distribution of the difference between R_o and β_{CR}. The percentage of samples where R_o exceeds β_{CR} indicates the probability of the waterbody being phosphorus limited in summer.

3.0 Results

3.1 Algal models including individual auxiliary variables

The IAV models capture the effect of adding auxiliary variables to a chl-a model based on nutrients alone. Focusing on the IAV models with a limiting nutrient formulation, goodness-of-fit metrics (i.e., R^2 and RMSE values) indicate modest improvements resulting from the inclusion of auxiliary variables when compared to a model without auxiliary variables (Table 3.1.1). Adding temperature resulted in the highest improvement in model fit (e.g., producing an R^2 of 0.64 vs. an R^2 of 0.59 for the model without auxiliary variables). Hyperparameter estimates are similar across the models, with relatively small standard errors. Residual plots of the models depict similar patterns, with residuals generally evenly distributed with slight differences in variability across the range of chl-a concentrations (Figure SI 2.1.1). Note, all IAV models achieve sufficient convergence as indicated by low Gelman-Rubin diagnostics (all $\hat{R} < 1.05$).

IAV models based on the multiplicative nutrient formulation (Eq. 2.2.3) perform worse than those based on the limiting nutrient formulation. The R^2 values for the NEI, depth, and temperature binned multiplicative models (0.58, 0.57, and 0.61) (Table SI 2.1.1) are 5-7% less than the R^2 values of the corresponding limiting nutrient models (Table 3.1.1). This suggests that a limiting nutrient structure is preferable over a multiplicative structure. For this reason, as well as mechanistic considerations (see Section 4.1), we focus on the limiting nutrient formulation throughout the rest of the model development.

Statistically significant auxiliary variables are identified based on the distinguishability of the bin-level parameters for the two end bins in each model. Recall that this is based on the 95%

credible interval for the difference between the two end-bin coefficients not including zero (Section 2.3), which translates to probabilities > 0.975 or < 0.025 in Table 3.1.1. Using this criterion, the IAV models indicate significant variations of the critical ratio in relation to all three auxiliary variables as well as significant variation of the slope parameter in relation to temperature. No significant variations were observed across the intercept parameters.

Table 3.1.1. Summary of Individual Auxiliary Variable (IAV) models using the limiting nutrient formulation. A model without any auxiliary variables is also included for comparison (last column). Model fit metrics include R^2 and RMSE. Parameters β_{CR}, β_{LN}, and β_0 are the hyperparameters for each model (or conventional parameters for the model with no auxiliary variables). Parameter standard errors are shown in parentheses. Bin-level parameters vary around the hyperparameters, within five bins determined by the auxiliary variables: the last three rows show the probability that the bin-5 parameter exceeds the bin-1 parameter.

Metric/parameter	**Units**	**Auxiliary variable included**			
		NEI	**Depth**	**Temperature**	**None**
R^2	unitless	0.61	0.61	0.64	0.59
RMSE	ln(µg/L)	0.88	0.88	0.84	0.90
β_{CR}	unitless	18.9 (2.9)	18.3 (2.6)	16.1 (1.7)	15.4 (0.5)
β_{LN}	unitless	0.90 (0.05)	0.94 (0.04)	0.96 (0.05)	0.99 (0.01)
β_0	ln(µg/L)	-0.72 (0.17)	-0.85 (0.11)	-0.94 (0.12)	-1.03 (0.05)
$P(\beta_{CR,5} - \beta_{CR,1} > 0)$	unitless	0.00	1.00	0.00	-
$P(\beta_{LN,5} - \beta_{LN,1} > 0)$	unitless	0.83	0.10	1.00	-
$P(\beta_{0,5} - \beta_{0,1} > 0)$	unitless	0.04	0.86	0.91	-

Assessing variations in the bin-level parameters for the IAV models suggests relationships between the auxiliary variables and chl-a production (Figure 3.1.1, Table SI 2.1.2). For the IAV model with bins based on NEI, there is significant variability in the bin-level critical ratio parameter (Table 3.1.1). The critical ratio appears to decrease nearly linearly with increasing NEI (Figure 3.1.1, r=0.82). There is no clear trend in the bin-level slope and intercept parameters, which is consistent with the lack of a significant difference between end-bin parameter estimates.

For the IAV model with bins based on depth, there is significant variability in the bin-level critical ratio parameters (Table 3.1.1). The critical ratio appears to increase nearly linearly with increasing log depth (Figure 3.1.1, r=0.98). There is no clear trend in the bin-level slope and intercept parameters, which is consistent with the lack of a significant difference between end-bin parameter estimates.

For the IAV model with bins based on temperature, there is significant variability in the bin-level slope and critical ratio parameters, but not for the bin-level intercept parameters (Table 3.1.1). The critical ratio and slope bin-level parameters follow roughly linear relationships (r=0.89 and 0.79) (Figure 3.1.1). Notably, the linear fit for the variation of the slope coefficient associated with temperature binning increases when modeled on the original scale as opposed to the log-scale (r=0.89). This indicates that there is no reason to log-transform temperature to capture this interaction in the MAV model (Section 3.2). There is no clear trend in the bin-level intercept parameters, which is consistent with the lack of a significant difference between end-bin parameter estimates.

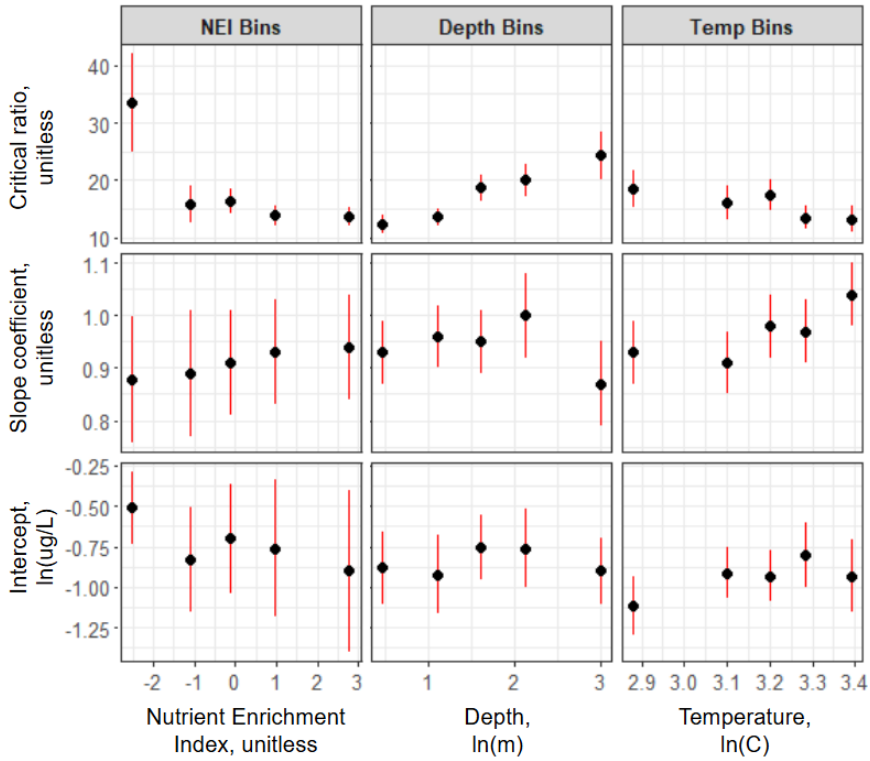


Figure 3.1.1. Variation in model parameters across auxiliary variable bins in Individual Auxiliary Variable (IAV) models with limiting nutrient formulation. Depth and temperature bin-level parameter values are plotted against the mean of the log-transformed auxiliary variables for each bin. NEI bin-level parameter values are plotted against the mean NEI for each bin. Red error bars indicate the 95% confidence intervals of the parameters.

3.2 Algal model with multiple auxiliary variables

The MAV model reflects the relationships identified from the IAV models (above), followed by backward variable selection to promote parsimony and prevent overfitting (Equations 3.1.1 and 3.1.2). In general, parameters converged well during the Bayesian inference process ($\hat{R} < 1.05$), and their 95% credible intervals did not include zero (Table SI 2.2.1), consistent with the backward selection process. Note that in Equations 3.1.1 and 3.1.2, the covariates representing

depth (ln(D)) and temperature (T) are centered by subtracting their respective means of 1.65 ln(m) and 24.2 °C.

$$\ln(a) = -0.85 + \ln(LN) * (0.018 * T + 0.94) \quad \text{Eq. 3.1.1}$$

$$LN = \min\left[TP, \frac{TN}{(3.3 * \ln(D) + 16.5)}\right] \quad \text{Eq. 3.1.2}$$

The parameters retained in the final model capture significant auxiliary variable relationships. Notably, the impacts of NEI and temperature on the critical ratio as well as the impacts of depth on the slope were removed during the backwards variable selection process. The impacts of auxiliary variables that remain in the MAV model are the impact of depth on the critical ratio and the impact of temperature on the slope parameter. Note that all of the relationships identified as insignificant in the IAV modeling results were not considered as potential predictors during backwards variable selection.

Both visual assessments and quantitative metrics of MAV model performance indicate robust model performance. No clear patterns were observed in the model residuals (Figure SI 2.3.1), suggesting that the variable transformations and parameterizations used in the MAV model were appropriate. The R^2 and RMSE of the model (0.65 and 0.84 ln(µg/L)) also shows improved performance compared to the IAV models.

The only auxiliary variable impacting the critical ratio in the MAV model is lake depth, which consequently causes the limiting nutrient of a lake to become a function of depth and nutrient concentrations (Equation 3.1.2). This formulation suggests that two lakes with the same nutrient concentrations can experience different limiting nutrient concentrations because of variations in

the critical ratio (Figure 3.2.1). As limiting nutrient concentration directly impacts chl-a levels, this also suggests that lakes with the same nutrient concentrations can produce different levels of algae.

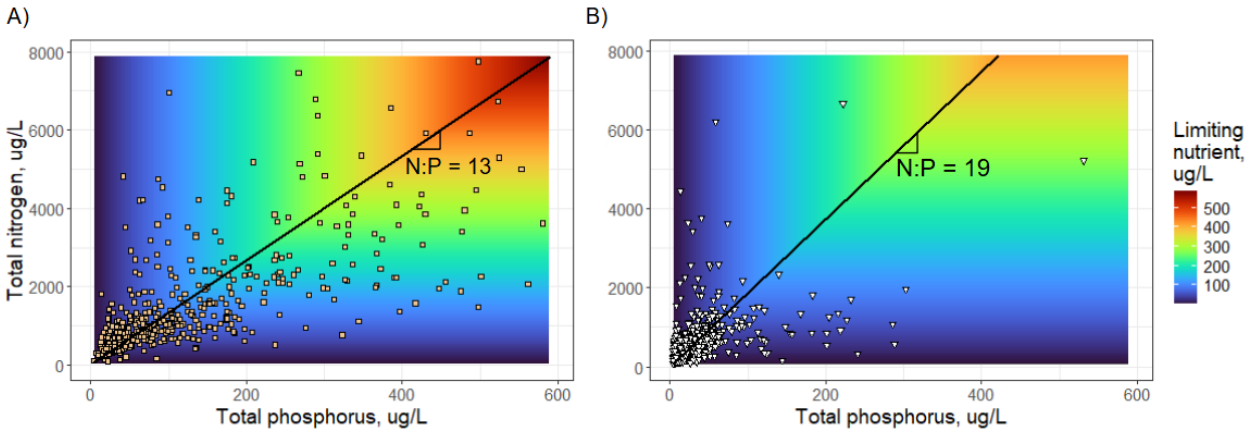


Figure 3.2.1. Variation in limiting nutrient concentration with respect to total phosphorus and total nitrogen concentrations based on the MAV model. Limiting nutrient predictions are shown for two lakes with depths of two (A) and ten (B) meters. Points correspond with observed data where depths range from 1.6 to 2.4 meters (n=490) (A) and 8 to 12 meters (n=486) (B). Note, the slopes of the black lines correspond with the critical ratios for each depth and indicate the approximate boundaries between nitrogen limitation and phosphorus limitation.

Once the limiting nutrient concentration is known, temperature is the only variable needed to predict chl-a in the MAV model (Equation 3.1.1). Chl-a levels are shown to primarily increase with an increase in limiting nutrient concentrations and secondarily increase with an increase in temperature (Figure 3.2.2). Shallow lakes are also shown to have a higher likelihood of elevated limiting nutrient concentrations compared to deeper lakes and subsequently have a higher likelihood of elevated chl-a levels.

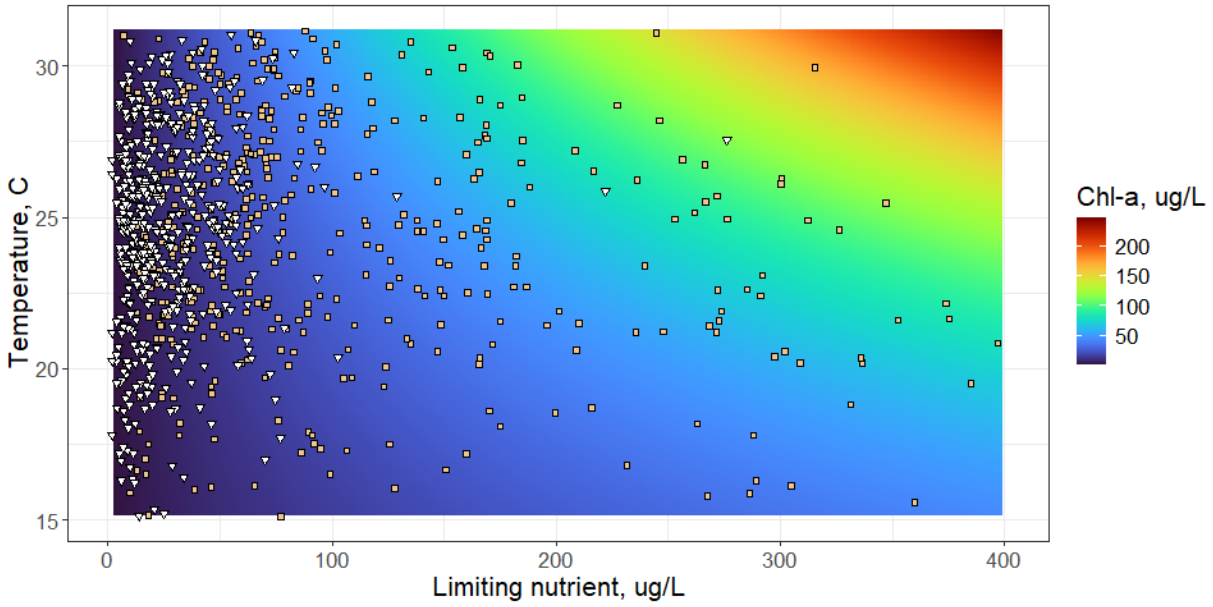


Figure 3.2.2. Chlorophyll-a predictions across typical ranges of temperature and limiting nutrient concentration for the MAV model. Brown squares and white triangles correspond with observations from approximately 2-m and 10-m deep lakes, respectively.

To further illustrate the relationships found in the MAV model, three example lakes were selected: a strongly P-limited Michigan lake, a strongly N-limited Nevada lake, and an Ohio lake with uncertain (or intermediate) nutrient limitation (additional lake details in Supporting Information Section SI 2.2). These lakes have depths of 4.1 meters, 3.0 meters, and 12.5 meters, respectively, such that the Ohio lake has the highest estimated critical ratio (19.4). The response of chl-a to changing phosphorus can be expressed assuming that TN levels remain constant, or by assuming that TN changes proportionally to TP (Figure 3.2.3a). The chlorophyll-nitrogen relationships can be expressed based on similar assumptions (Figure 3.2.3b). These lake examples highlight how an increase in the non-limiting nutrient alone has no effect on additional algal accumulation. Furthermore, these examples show that the chlorophyll-limiting nutrient relationship is approximately linear. The relatively cool Nevada lake departs most noticeably

from linearity, showing a slightly saturating (concave down) response of chlorophyll to additional nutrients (also, see Discussion Section 4.3).

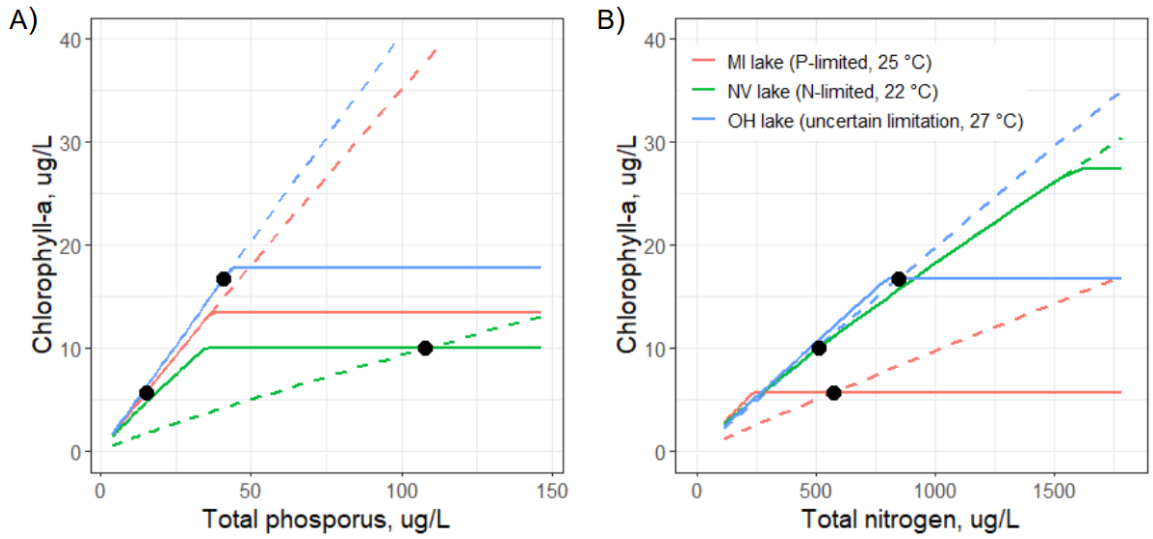


Figure 3.2.3. Variation in chlorophyll across a range of (A) TP and (B) TN concentrations. Results are shown for three example lakes with varying nutrient limitation. We include scenarios where the alternative nutrient levels are held constant (solid line) and where nutrient levels increase proportionally to one another in order to maintain a constant N:P ratio (dashed line). Black points indicate the observed nutrient concentrations for each lake.

3.3 Cross validation

The four-fold cross validation results indicate that the model performs well when making predictions for waterbodies outside of its calibration dataset (Table SI 2.3.1; Figure SI 2.3.2). The hold-out validation folds yield a combined R^2 and RMSE of 0.64 and 0.85 ln(µg/L), respectively. This represents a very small departure from the model calibrated to the entire dataset (R^2 = 0.65 and RMSE = 0.84 ln(µg/L)). In addition, across the five folds, all model parameters remained significant and had an average deviation of 6% (parameter deviation ranged

from 0.2% to 29%) relative to the model calibrated to the full dataset (Table SI 2.3.2). Together, these findings indicate that the MAV model is robust.

3.4 Model application and mapping

Our probabilistic estimates of waterbody-specific critical ratios and in-lake N:P ratios allow us to determine the probability of phosphorus (versus nitrogen) limitation in summer. The hyperparameters of the Bayesian model used to estimate the natural log of the mean summer N:P ratio of each lake (R_o) calibrated to have a mean (μ_X) of 2.85, an inter-lake standard deviation ($\sigma_{l,R}$) of 0.69, and a within-lake standard deviation ($\sigma_{s,R}$) of 0.45. While there can be notable variation within a region, in-lake summer N:P ratios tend to be higher in the Northeast and around the western Great Lakes regions, and lower in several western states, such as Nevada and Idaho (Figure 3.4.1a). On the other hand, estimated critical ratios (β_{CR}) tend to be higher in the southeast coastal plains and northern plains, lower in the Northeast and Northwest, and highly variable across most other regions (Figure 3.4.1b). When these results are combined, we find high concentrations of P-limited lakes in New England and the Great Lakes region, and substantially more N-limited lakes across the western US (Figure 3.4.1c). Out of the 2755 different lakes included in this study, 435 (16%) of lakes indicate nitrogen limitation with a probability greater than 90%, and 644 (23%) of lakes indicate phosphorus limitation with a probability greater than 90%.

Across CONUS, the overall (statistical) distribution of N versus P limitation is notably symmetric (Figure SI 2.4.1) suggesting approximately equal numbers of waterbodies being limited by each of the two nutrients, with slightly more P limited lakes. However, the situation

can be quite different for particular regions and states. At one extreme, 68% of waterbodies are found to have strong P limitation (i.e., >90% probability of P limitation) in Michigan, while only 2% have strong N limitation. In contrast, only 12% of waterbodies in Nevada suggest strong P limitation, with 64% indicating strong N limitation.

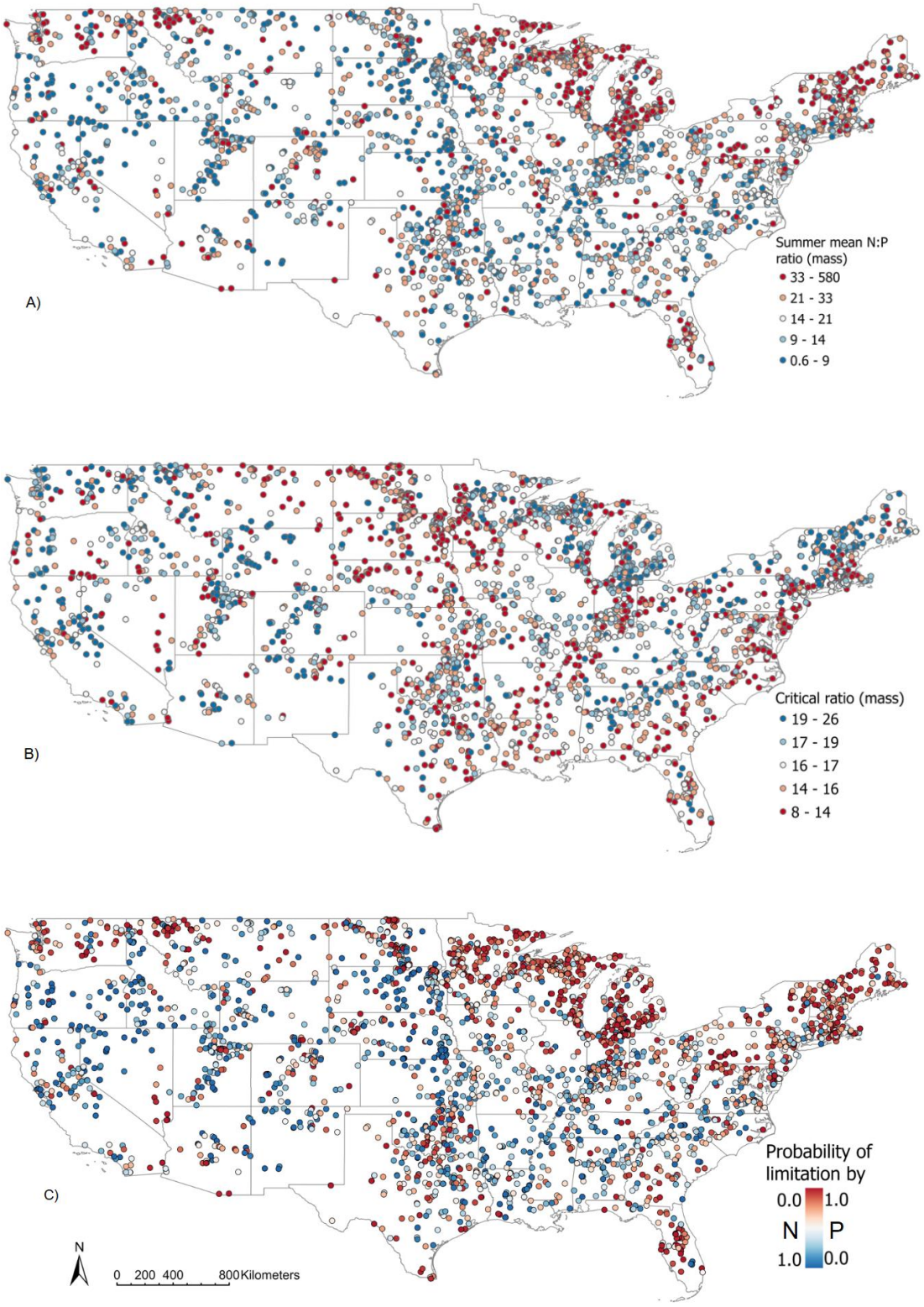


Figure 3.4.1. Estimated (A) summer mean N:P ratio, (B) critical N:P ratio, (C) probability of phosphorus versus nitrogen limitation across CONUS.

4.0 Discussion

4.1 Model structure

The superior performance of the limiting-nutrient IAV models relative to the multiplicative IAV models is expected. While expressing nutrients in a multiplicative fashion may seemingly add flexibility to a model, our results suggest it also inaccurately represents chlorophyll-nutrient relationships mechanistically. This is consistent with the stance of Dolman and Wiedner, who argue that the effects of additional TN in P-limited lakes (as well as the effects of additional TP in N-limited lakes) are overestimated in a multiplicative structure (Dolman & Wiedner, 2015). Dolman and Wiedner observe up to 9% improvement in R^2 values when comparing multiplicative and limiting nutrient structures, which is generally consistent with our range of 5-7% improvement (Section 3.1). Our selection of a limiting nutrient structure also falls in line with the intuition behind Liebig's law of the minimum (de Baar, 1994; Saito et al., 2008). These considerations support the selection of a limiting nutrient structure for the final, MAV model.

The parameterization of auxiliary variable effects as linear expressions within the MAV model is informed by the results of the IAV models. The high correlation of log-depth with the bin-level critical ratios, and temperature with bin-level slopes indicate the sufficiency of a linear formulation in the MAV model. It is also notable that the MAV model yields an R^2 of 0.65 using only 5 parameters, while the IAV models yield R^2 values ranging from 0.59 to 0.61 while utilizing up to 15 parameters (not including variance or hyperparameters). The linear formulations not only greatly reduce MAV model complexity but also allow for enhanced interpretability (below). Increasing model complexity (such as through the addition of squared terms; not shown) has a minimal effect on model performance.

4.2 Interpretation of critical ratio parameters

The estimated critical ratio values for the IAV and MAV models are within our expectations. Across the IAV models and all bin-levels, the critical ratio ranges between 13 and 33 with relatively tight standard deviations (Figure 3.1.1). When expressed as a function of depth in the MAV model, the critical ratio ranges from 11 to 23 (based on the 2.5 and 97.5 percentiles of observed NLA depths). In comparison, Dolman and Wiedner find that optimal critical TN:TP ratios for lakes from different regions vary substantially (25 for a set of German lakes, 16 for a set of CONUS lakes, and 22 for a set of Florida lakes) (Dolman & Wiedner, 2015). The variations in the critical ratio parameter identified in our model align with these ranges and support our hypothesis that the critical ratio has the potential to vary substantially across waterbodies.

The MAV model shows that the critical ratio increases as lake depth increases (Equation 3.1.1). One potential mechanistic reason for this trend could be differences in phosphorus bioavailability between shallow and deep lakes. Phosphorus can be present in many forms in aquatic environments, some of which are sediment-bound and less labile (Søndergaard et al., 2003; Yuan & Jones, 2020). It is more likely for resuspension of sediments to occur in shallow lakes than deep lakes, leading to a larger proportion of the total phosphorus measurement to be in sediment-bound forms that are less conducive to algal growth. Consequently, algae in shallow lakes require a greater amount of TP in order to have the same amount of bioavailable phosphorus (which shifts the critical ratio down).

It is interesting to see the effects that NEI has on the IAV and MAV models. It has been suggested that algal dynamics are dominated by P limitation when lakes are at a low trophic status, and then shift towards joint N and P influences or N limitation as trophic status increases (Liang et al., 2020; Zhao et al., 2023). The results of the IAV model with NEI bins, however, indicate higher critical ratios at lower NEI levels. This suggests a higher probability of N limitation at lower trophic levels, all else being equal. Furthermore, our MAV model suggests that trophic status has no direct impact on nutrient limitation. A likely explanation for these discrepancies can be attributed to the critical ratio varying with depth (Equation 3.1.2), and the correlation between depth and NEI (r = -0.61) (Figure SI 2.4.2). Because these two auxiliary variables are correlated it is natural that one would be removed during the backwards variable selection process when both auxiliary variables are included as candidate variables. For our MAV model, depth outperforms NEI as a predictor variable, and thus NEI is removed. The NEI IAV model relationship is potentially explained by the model capturing the impacts of depth by using NEI as a correlated proxy.

4.3 Interpretation of slope parameters

The slope parameters for the IAV and MAV models calibrate to reasonable values. To understand the implications of the slope parameter, recall that the models predict the natural log of chlorophyll concentrations. When back transformed, the slope and intercept parameters become exponents in a multiplicative formulation (i.e., $a = e^{\beta_0} * LN^{\beta_{LN}}$). With this in mind, a slope coefficient of one indicates a linear relationship between chlorophyll and limiting nutrient concentration. Across the IAV models, the slope coefficients range from 0.87 to 1.04 (Figure 3.1.1). When expressed as a function of temperature in the MAV model, the slope parameter

ranges from 0.77 to 1.06 (based on the 2.5 and 97.5 percentiles of observed NLA temperatures), with lower temperatures yielding lower values. Lower values of the slope parameter indicate a somewhat saturating (concave down) relationship between chlorophyll and temperature, indicating that algae are less able to fully utilize the nutrients present. Lower temperatures are generally associated with reduced rates of algal growth and nutrient cycling. This aligns with the well-established fact that kinetic rates increase with temperature (e.g., Arrhenius equation) as well as studies on algal growth rates, though optimal temperatures for the algal growth rates are species-specific (Singh & Singh, 2015).

4.4 Interpretation of intercept parameters

All intercept parameters for the IAV and MAV models calibrate to values within our expectations. With the back transformed structure in mind, an intercept parameter of zero would indicate a chlorophyll to phosphorus ratio of 1:1. Prior studies have suggested an expected range of 0.4:1 to 2:1 for the chlorophyll to phosphorus ratio when measuring within-cell nutrients (Chapra, 2008). Across the IAV models, the bin-level intercept coefficients range from -1.11 to -0.51 ln(µg/L) (Figure 3.1.1). These values correspond with chlorophyll to total phosphorus ratios ranging from 0.3:1 to 0.6:1. In the MAV model, the intercept is expressed as a constant, intermediate value of -0.85 ln(µg/L). Such values align with expectations, especially when considering that total nutrient measurements (i.e., TN and TP) include both within-cell and extracellular nutrients.

4.5 Nutrient limitation mapping

This study's approach to mapping nutrient limitation in lakes offers insights into various aspects of the chlorophyll-nutrient relationships across CONUS. By utilizing an approach that includes a varying critical ratio and the uncertainty associated with actual within-lake N:P ratios, our methods provide a more quantitative approach to defining nutrient limitation than many methods that rely on generalized nutrient limitation cutoffs independent of lake depth or other covariates. For example, multiple studies have specified N limitation when N:P is less than 9 and P limitation when N:P is greater than 22.6, based on typical values observed in freshwater and marine systems worldwide (Guildford & Hecky, 2000; Naderian et al., 2025; Qin et al., 2020). Additionally, this study reveals large-scale regional trends in nutrient limitation that can be used to inform nutrient management strategies.

Our results can be compared to prior research that assesses spatial variability in nutrient limitation and quantifies the proportions of lakes experiencing N or P limitation across CONUS. One study found N to generally be a better predictor of algal productivity in western CONUS lakes and P to be a better predictor in eastern CONSUS lakes, though there are instances where individual lakes within these regions deviate from these trends due to unique, context-dependent characteristics such as lake depth, surrounding landscape, legacy nutrient accumulation, or other factors. (Rock & Collins, 2024). In the context of our study, this implies a higher proportion of N limited lakes in western CONUS and a higher proportion of P limited lakes in eastern CONUS. The results of our study generally agree with that trend, as well as the expectation of intra-regional variability in nutrient limitation. That same study also explicitly determined that out of the 3270 lakes in their dataset, 436 (13%) were P-limited and 350 (11%) were N-limited based

on a methodology that compared high-yield chl-a and observed chl-a levels at various nutrient concentrations using a “tipping point N/P ratio”. Another study estimated that 17% of CONUS lakes experience P limitation and 14% experience N limitation using the same methodology (Moon et al., 2021). While the criteria used by past studies differ from ours, their results are generally consistent with our findings that 23% of CONUS lakes experience strong P limitation and 16% experience strong N limitation.

4.6 Management implications and model limitations

Observing the chlorophyll trends of specific lakes included in our study helps elucidate the management implications of the MAV model (Figure 3.2.3). Our model illustrates how in scenarios where the alternative (non-limiting) nutrient levels are held constant (solid lines in Figure 3.2.3), an increase in TP in an N-limited system or an increase in TN in a P-limited system has no effect on chlorophyll concentrations until the system switches nutrient limitation states. An alternative way to visualize lake-specific chlorophyll trends is to maintain a constant N:P ratio by increasing nutrients proportional to one another (dashed lines in Figure 3.2.3). Note that in the NLA dataset, TN and TP levels are substantially correlated ($r = 0.78$ for our data on the log scale), so it is likely that real-world nutrient reduction scenarios may typically fall in between the two scenarios shown in Figure 3.2.3, and perhaps closer to the case where nutrient levels change proportionally with one another. Our model also highlights how certain lakes can experience a type of dual-nutrient limitation where an increase in both TN and TP would be necessary to cause significant increases in chl-a (solid blue lines in Figure 3.2.3).

While the primary purpose of our model is to identify large-scale regional trends in nutrient limitation, applying the model to individual lakes can highlight certain aspects of lake-specific

chlorophyll-nutrient dynamics. There are notable limitations when doing this, however. The model in this study does not account for fluctuations in algal concentrations across a lake. The NLA water quality data used to develop this model is collected at a single location roughly in the center of the lake. While there are guidelines for establishing this location, there is some ambiguity introduced during data collection. It is also common for nutrient and algal concentrations to vary across a waterbody based on different water depths, inflows from streams/point sources, sediment loading rates, and various other factors. In addition to this, our model only accounts for summer mean chlorophyll levels and does not consider how nutrient limitation and chlorophyll accumulation can change over time. To further improve the ability of our model to capture chlorophyll-nutrient relationships for individual lakes, steps can be taken such as comprehensively monitoring variations in observed N:P ratios, temperature, and water depth at higher temporal and spatial resolutions for the lake of interest.

5.0 Conclusion

In this study, we developed a model that predicts chlorophyll levels in CONUS lakes based on nutrients, lake depth, and temperature using a Bayesian modeling framework. Our model not only has strong predictive capabilities (R^2=0.65), but also explores the underlying interactions between nutrients, lake characteristics, the critical N:P ratio needed for algal growth, and chlorophyll. Our results indicate that the critical N:P ratio increases with increased lake depth, and that chlorophyll increases with increased temperature and critical (limiting) nutrient levels. Furthermore, this study found that roughly equal amounts of US lakes experience strong nitrogen (16%) and strong phosphorus (23%) limitation, indicating a slight inclination towards phosphorus limitation. These nutrient limitation proportions, however, were shown to vary substantially across CONUS. Our model can be used as an effective tool for making large-scale nutrient management decisions and refines the current understanding of chlorophyll-nutrient relationships in CONUS lakes.

REFERENCES

Anderson, D. M., Glibert, P. M., & Burkholder, J. M. (2002). Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. *Estuaries*, *25*(4), 704–726. https://doi.org/10.1007/BF02804901

de Baar, H. J. W. (1994). von Liebig's law of the minimum and plankton ecology (1899–1991). *Progress in Oceanography*, *33*(4), 347–386. https://doi.org/10.1016/0079-6611(94)90022-1

Backer, L. C., Manassaram-Baptiste, D., LePrell, R., & Bolton, B. (2015). Cyanobacteria and Algae Blooms: Review of Health and Environmental Data from the Harmful Algal Bloom-Related Illness Surveillance System (HABISS) 2007–2011. *Toxins*, *7*(4), 1048–1064. https://doi.org/10.3390/toxins7041048

Brehob, M. M., Pennino, M. J., Handler, A. M., Compton, J. E., Lee, S. S., & Sabo, R. D. (2024). Estimates of Lake Nitrogen, Phosphorus, and Chlorophyll-a Concentrations to Characterize Harmful Algal Bloom Risk Across the United States. *Earth's Future*, *12*(8), e2024EF004493. https://doi.org/10.1029/2024EF004493

Carmichael, W. W., & Boyer, G. L. (2016). Health impacts from cyanobacteria harmful algae blooms: Implications for the North American Great Lakes. *Harmful Algae*, *54*, 194–212. https://doi.org/10.1016/j.hal.2016.02.002

Chapra, S. C. (2008). *Surface Water-Quality Modeling* (1st ed.). Long Grove, IL: Waveland Press.

Del Giudice, D., Fang, S., Scavia, D., Davis, T. W., Evans, M. A., & Obenour, D. R. (2021). Elucidating controls on cyanobacteria bloom timing and intensity via Bayesian

mechanistic modeling. *Science of The Total Environment*, *755*, 142487. https://doi.org/10.1016/j.scitotenv.2020.142487

Dolman, A. M., & Wiedner, C. (2015). Predicting phytoplankton biomass and estimating critical N:P ratios with piecewise models that conform to Liebig's law of the minimum. *Freshwater Biology*, *60*(4), 686–697. https://doi.org/10.1111/fwb.12518

Dolman, A. M., Mischke, U., & Wiedner, C. (2016). Lake-type-specific seasonal patterns of nutrient limitation in German lakes, with target nitrogen and phosphorus concentrations for good ecological status. *Freshwater Biology*, *61*(4), 444–456. https://doi.org/10.1111/fwb.12718

Esri. (2024). ArcGIS Pro (Version 3.2). Redlands, CA: Environmental Systems Research Institute. Retrieved from https://www.esri.com

Fang, C., Song, K., Paerl, H. W., Jacinthe, P.-A., Wen, Z., Liu, G., et al. (2022). Global divergent trends of algal blooms detected by satellite during 1982–2018. *Global Change Biology*, *28*(7), 2327–2340. https://doi.org/10.1111/gcb.16077

Faraway, J. J. (2015). *Linear models with R* (Second edition.). Boca Raton: CRC Press, Taylor & Francis Group.

Fernandez-Figueroa, E. G., Rogers, S. R., Waters, M. N., & Wilson, A. E. (2024). The green convergence: United States lakes are collectively moving toward a eutrophic state. *Harmful Algae*, *139*, 102721. https://doi.org/10.1016/j.hal.2024.102721

Garnier, Simon, Ross, Noam, Rudis, Robert, et al. (2023). *viridis(Lite) - Colorblind-Friendly Color Maps for R*. https://doi.org/10.5281/zenodo.4679423

Gelman, A., & Hill, J. (2006, December 17). Data Analysis Using Regression and Multilevel/Hierarchical Models. https://doi.org/10.1017/CBO9780511790942

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Grattan, L. M., Holobaugh, S., & Morris, J. G. (2016). Harmful algal blooms and public health. *Harmful Algae*, *57*, 2–8. https://doi.org/10.1016/j.hal.2016.05.003

Guildford, S. J., & Hecky, R. E. (2000). Total nitrogen, total phosphorus, and nutrient limitation in lakes and oceans: Is there a common relationship? *Limnology and Oceanography*, *45*(6), 1213–1223. https://doi.org/10.4319/lo.2000.45.6.1213

Ho, J. C., & Michalak, A. M. (2020). Exploring temperature and precipitation impacts on harmful algal blooms across continental U.S. lakes. *Limnology and Oceanography*, *65*(5), 992–1009. https://doi.org/10.1002/lno.11365

Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M. H., & Visser, P. M. (2018). Cyanobacterial blooms. *Nature Reviews Microbiology*, *16*(8), 471–483. https://doi.org/10.1038/s41579-018-0040-1

Jackson-Blake, L. A., Clayer, F., Haande, S., Sample, J. E., & Moe, S. J. (2022). Seasonal forecasting of lake water quality and algal bloom risk using a continuous Gaussian Bayesian network. *Hydrology and Earth System Sciences*, *26*(12), 3103–3124. https://doi.org/10.5194/hess-26-3103-2022

Karimi, K., & Obenour, D. R. (2024). Characterizing Spatiotemporal Variability in Phosphorus Export across the United States through Bayesian Hierarchical Modeling. *Environmental Science & Technology*, *58*(22), 9782–9791. https://doi.org/10.1021/acs.est.3c07479

Katin, A., Giudice, D. D., Hall, N. S., Paerl, H. W., & Obenour, D. R. (2021). Simulating algal dynamics within a Bayesian framework to evaluate controls on estuary productivity. *Ecological Modelling*, *447*, 109497. https://doi.org/10.1016/j.ecolmodel.2021.109497

Koenker, R. (2023). *quantreg: Quantile Regression*. Retrieved from https://CRAN.R-project.org/package=quantreg

Lad, A., Breidenbach, J. D., Su, R. C., Murray, J., Kuang, R., Mascarenhas, A., et al. (2022). As We Drink and Breathe: Adverse Health Effects of Microcystins and Other Harmful Algal Bloom Toxins in the Liver, Gut, Lungs and Beyond. *Life*, *12*(3), 418. https://doi.org/10.3390/life12030418

Landis, J. (2022). *ggside: Side Grammar Graphics*. Retrieved from https://CRAN.R-project.org/package=ggside

Liang, Z., Soranno, P. A., & Wagner, T. (2020). The role of phosphorus and nitrogen on chlorophyll *a*: Evidence from hundreds of lakes. *Water Research*, *185*, 116236. https://doi.org/10.1016/j.watres.2020.116236

Liang, Z., Xu, Y., Zhao, G., Lu, W., Fu, Z., Wang, S., & Wagner, T. (2023). Approaching the upper boundary of driver-response relationships: identifying factors using a novel framework integrating quantile regression with interpretable machine learning. *Frontiers of Environmental Science & Engineering*, *17*(6), 76. https://doi.org/10.1007/s11783-023-1676-2

Moon, D. L., Scott, J. T., & Johnson, T. R. (2021). Stoichiometric imbalances complicate prediction of phytoplankton biomass in U.S. lakes: Implications for nutrient criteria. *Limnology and Oceanography*, *66*(8), 2967–2978. https://doi.org/10.1002/lno.11851

Naderian, D., Noori, R., Kim, D., Jun, C., Bateni, S. M., Woolway, R. I., et al. (2025). Environmental controls on the conversion of nutrients to chlorophyll in lakes. *Water Research*, *274*, 123094. https://doi.org/10.1016/j.watres.2025.123094

Ott, W. R. (1990). A Physical Explanation of the Lognormality of Pollutant Concentrations. *Journal of the Air & Waste Management Association*, *40*(10), 1378–1383. https://doi.org/10.1080/10473289.1990.10466789

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, *6*(1), 7–11.

Qian, S. S., Cuffney, T. F., Alameddine, I., McMahon, G., & Reckhow, K. H. (2010). On the application of multilevel modeling in environmental and ecological studies. *Ecology*, *91*(2), 355–361.

Qin, B., Zhou, J., Elser, J. J., Gardner, W. S., Deng, J., & Brookes, J. D. (2020). Water Depth Underpins the Relative Roles and Fates of Nitrogen and Phosphorus in Lakes. *Environmental Science & Technology*, *54*(6), 3191–3198. https://doi.org/10.1021/acs.est.9b05858

Quinlan, R., Filazzola, A., Mahdiyan, O., Shuvo, A., Blagrave, K., Ewins, C., et al. (2021). Relationships of total phosphorus and chlorophyll in lakes worldwide. *Limnology and Oceanography*, *66*(2), 392–404. https://doi.org/10.1002/lno.11611

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rock, L. A., & Collins, S. M. (2024). A Broad-Scale Look at Nutrient Limitation and a Shift toward Co-limitation in United States Lakes. *Environmental Science & Technology*, *58*(26), 11482–11491. https://doi.org/10.1021/acs.est.4c03135

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, PBC. Retrieved from http://www.rstudio.com/

Saito, M. A., Goepfert, T. J., & Ritt, J. T. (2008). Some thoughts on the concept of colimitation: Three definitions and the importance of bioavailability. *Limnology and Oceanography*, *53*(1), 276–290. https://doi.org/10.4319/lo.2008.53.1.0276

Sharma, N. K., Rai, A. K., & Stal, L. J. (2014). *Cyanobacteria: An Economic Perspective*. Newark, UNITED KINGDOM: John Wiley & Sons, Incorporated. Retrieved from http://ebookcentral.proquest.com/lib/ncsu/detail.action?docID=1569020

Singh, S. P., & Singh, P. (2015). Effect of temperature and light on the growth of algae species: A review. *Renewable and Sustainable Energy Reviews*, *50*, 431–444. https://doi.org/10.1016/j.rser.2015.05.024

Søndergaard, M., Jensen, J. P., & Jeppesen, E. (2003). Role of sediment and internal loading of phosphorus in shallow lakes. *Hydrobiologia*, *506*(1), 135–145. https://doi.org/10.1023/B:HYDR.0000008611.12704.dd

Taranu, Z. E., Gregory-Eaves, I., Steele, R. J., Beaulieu, M., & Legendre, P. (2017). Predicting microcystin concentrations in lakes and reservoirs at a continental scale: A new framework for modelling an important health risk factor. *Global Ecology and Biogeography*, *26*(6), 625–637. https://doi.org/10.1111/geb.12569

U.S. Environmental Protection Agency. (2010). National Aquatic Resource Surveys: National Lakes Assessment 2007 (data and metadata files). Retrieved from https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys

U.S. Environmental Protection Agency. (2016). National Aquatic Resource Surveys: National Lakes Assessment 2012 (data and metadata files). Retrieved from

https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys

U.S. Environmental Protection Agency. (2022a). National Aquatic Resource Surveys: National Lakes Assessment 2017 (data and metadata files). Retrieved from https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys

U.S. Environmental Protection Agency. (2022b). *National Lakes Assessment 2022: Field Operations Manual*. Washington, D.C.: U.S. Environmental Protection Agency. Retrieved from https://www.epa.gov/national-aquatic-resource-surveys/nla

U.S. Environmental Protection Agency. (2022c). *National Lakes Assessment 2022 Laboratory Operations Manual*. Retrieved from https://www.epa.gov/system/files/documents/2022-07/LOM%20NLA%202022%20Version1.1_220519_0.pdf

U.S. Environmental Protection Agency. (2024). National Aquatic Resource Surveys: National Lakes Assessment 2022 (data and metadata files). Retrieved from https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys

U.S. Geological Survey. (2024). National Hydrography Dataset (NHD). Retrieved from https://www.usgs.gov/national-hydrography

Valpine, P. de, Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., et al. (2024). NIMBLE User Manual (Version 1.2.1). https://doi.org/10.5281/zenodo.1211190

Wagner, T., Soranno, P. A., Webster, K. E., & Cheruvelil, K. S. (2011). Landscape drivers of regional variation in the relationship between total phosphorus and chlorophyll in lakes.

Freshwater Biology, *56*(9), 1811–1824. https://doi.org/10.1111/j.1365-2427.2011.02621.x

Wei, Z., Alam, S., Verma, M., Hilderbran, M., Wu, Y., Anderson, B., et al. (2024). Integrating water quality data with a Bayesian network model to improve spatial and temporal phosphorus attribution: Application to the Maumee River Basin. *Journal of Environmental Management*, *360*, 121120. https://doi.org/10.1016/j.jenvman.2024.121120

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Yuan, L. L., & Jones, J. R. (2020). Rethinking phosphorus–chlorophyll relationships in lakes. *Limnology and Oceanography*, *65*(8), 1847–1857. https://doi.org/10.1002/lno.11422

Yuan, L. L., & Jones, J. R. (2022). Estimating phytoplankton stoichiometry from routinely collected monitoring data. *Biogeochemistry*, *159*(2), 251–264. https://doi.org/10.1007/s10533-022-00926-8

Yuan, L. L., & Pollard, A. I. (2014). Classifying lakes to improve precision of nutrient–chlorophyll relationships. *Freshwater Science*, *33*(4), 1184–1194. https://doi.org/10.1086/678465

Yuan, L. L., & Pollard, A. I. (2017). Using National-Scale Data To Develop Nutrient–Microcystin Relationships That Guide Management Decisions. *Environmental Science & Technology*, *51*(12), 6972–6980. https://doi.org/10.1021/acs.est.7b01410

Zhao, L., Zhu, R., Zhou, Q., Jeppesen, E., & Yang, K. (2023). Trophic status and lake depth play important roles in determining the nutrient-chlorophyll *a* relationship: Evidence from

thousands of lakes globally. *Water Research*, *242*, 120182.

https://doi.org/10.1016/j.watres.2023.120182

APPENDICES

SI 1.0 Methodology Supplemental Information

SI 1.1 Data processing

The primary datasets used in this analysis were the EPA's National Lakes Assessments (NLA) from the years 2007, 2012, 2017, and 2022 (U.S. Environmental Protection Agency, 2010, 2016, 2022a, 2024). Specifically, the data downloaded for each year included the following variables: unique ID (UID), site ID, visit number, date collected, coordinates of index site, temperature profiles, index site depth, chlorophyll-a concentration, total phosphorus concentration, and total nitrogen concentration.

NLA data were collected from May to October during all years. Of the points used in our analysis, the following percentages of data were collected during each month: 1.4% in May, 23.2% in June, 29.9% in July, 30.5% in August, 14.7% in September, and 0.2% in October.

An important component of NLA data collection is the lake index site, where the water quality samples are collected. In natural lakes, this site is located at the deepest point in the lake. In reservoirs, this site is located near the midpoint of the lake. For both scenarios, the index site depth must not exceed 50 meters. If the first location exceeds this depth, a new index site is supposed to be established at a point as close to the original location as possible without exceeding a depth of 50 meters (U.S. Environmental Protection Agency, 2022b). Notably, 28 of the 4146 observations (< 1%) used in our analysis included an index site depth equal to or greater than 50 meters.

Water quality data for nutrients and chlorophyll were collected at the index site using an "integrated sampler". The sampler was a two-meter long PVC tube that collects water from the top layer of the water column. Precautions were taken at the time of collection to ensure no contamination of the water samples occurred, and then the samples were transported to a lab for analysis (U.S. Environmental Protection Agency, 2022b).

A multi-parameter water quality sonde was used to obtain temperature, dissolved oxygen, and pH profiles throughout the depth of the water column at the index site. Temperature measurements were taken at regular intervals throughout the entire water column (U.S. Environmental Protection Agency, 2022b). Because water quality samples were taken only over the top two meters, the temperature measurements with a depth less than two meters were averaged together to obtain the temperature used in our analysis for each lake.

For our analysis, it was necessary to determine which lakes were sampled multiple times. While NLA recorded a unique site ID for each lake during each study, these IDs do not align across the different years. Because of this, the National Hydrography Dataset from USGS (U.S. Geological Survey, 2024) was used to identify which lakes were resampled by utilizing geoprocessing tools in ArcGIS Pro (Esri, 2024) and manual inspection.

The raw NLA data specified a minimum reporting limit of 4 µg/L for total phosphorus (U.S. Environmental Protection Agency, 2022c). Consequently, total phosphorus measurements below this threshold (107 of the 4242 samples prior to outlier removal) were set to ½ the minimum reporting limit (2 µg/L)

Based on initial data exploration, it was decided that outlier removal was appropriate, given the presence of extreme values. Outlier removal based on chlorophyll, total nitrogen, and total phosphorus concentrations was conducted using an interquartile rule (IQR) approach on the natural log scale. In this approach, the interquartile range is defined as the difference between the 75^{th} and 25^{th} percentile. Points are identified as outliers and removed from the dataset when the values are below the 25^{th} percentile minus 1.5 times the interquartile range or when the values are above the 75^{th} percentile plus 1.5 times the interquartile range. This process removed nine points based on chlorophyll concentration, 65 additional points based on total nitrogen concentration, and then 13 additional points based on total phosphorus concentration.

In addition to the IQR outlier removal, seven points were removed from the dataset due to having implausible measurements that are most likely errors. These outliers included four points that were removed due to having extreme temperature values (0.8, 91.2, 88.6, and 93.1℃), one point with an unreasonable surface area (2176834 hectares), and two points within the same lake where the depth measurements were incomprehensible. The site ID and NLA years for these points are NLA17_WY-10032 (2017), NLA12_MN-167 (2012), NLA12_RI-102 (2012), NLA12_MN-135 (2012), NES22-10023 (2022), and NLA12_OK-293 (2012), respectively.

The overall outlier removal process included the removal of 94 observations (~2% of the original dataset). Our final dataset consisted of 4148 sampling events occurring in 2755 different lakes across CONUS.

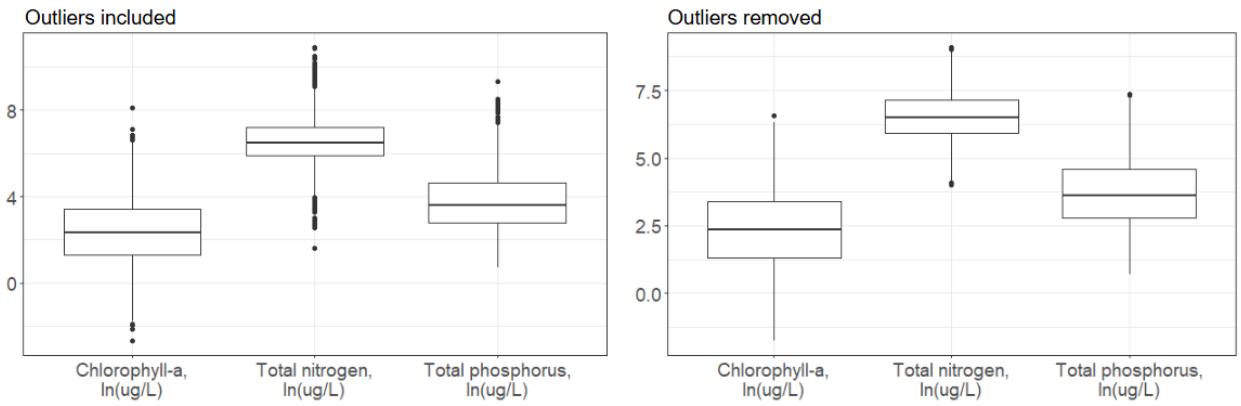


Figure SI 1.1.1. Boxplots of chlorophyll-a, TN, and TP data before and after outlier removal on the natural log scale.

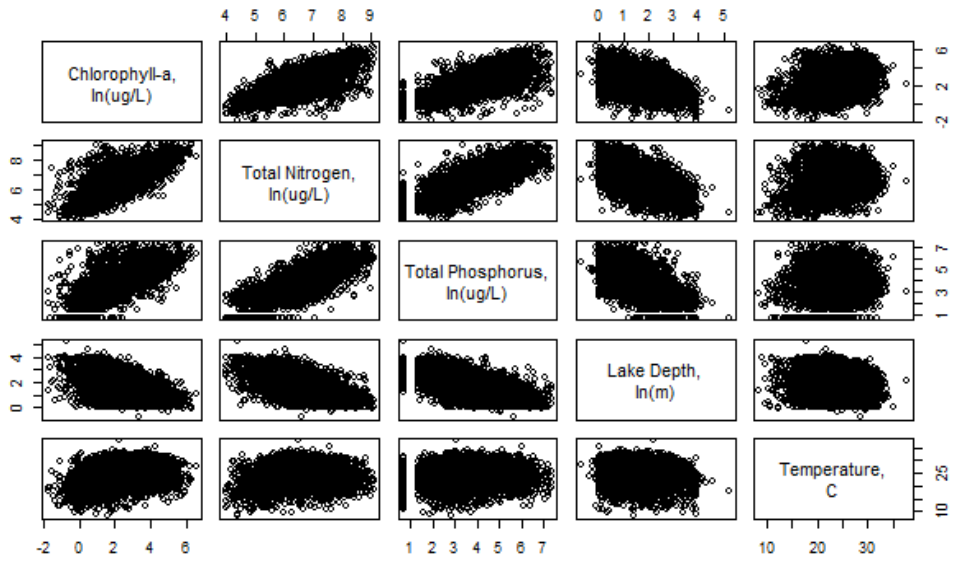
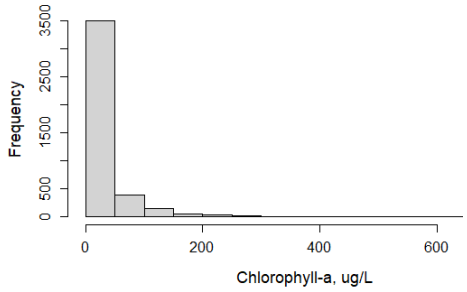
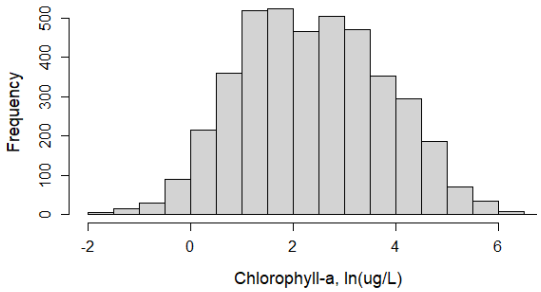
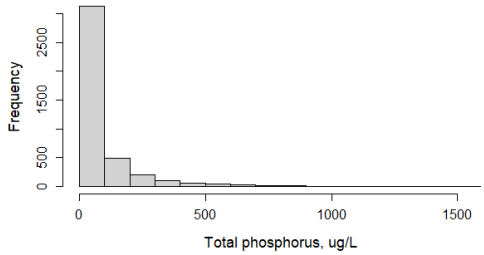


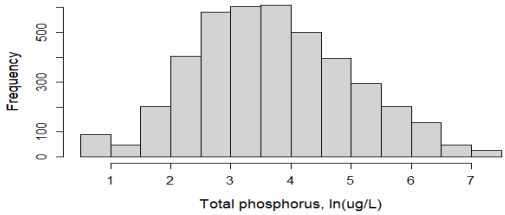
Figure SI 1.1.2. Correlation plots of model variables after outlier removal and data transformations.

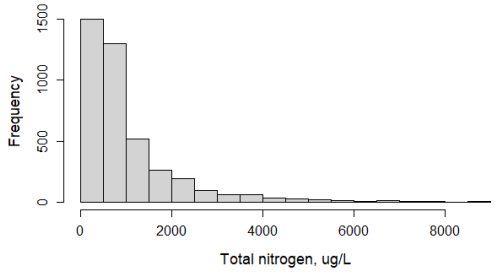
Figure SI 1.1.3. Histograms of model variables after outlier removal.

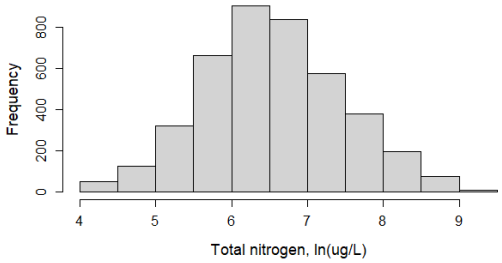

Frequency
0
500
1500
2500
3500
0
200
400
600
Chlorophyll-a, ug/L

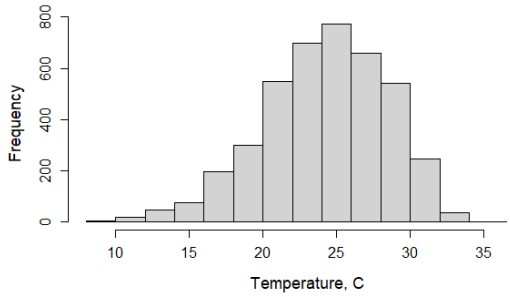

Frequency
0
100
200
300
400
500
-2
0
2
4
6
Chlorophyll-a, ln(ug/L)

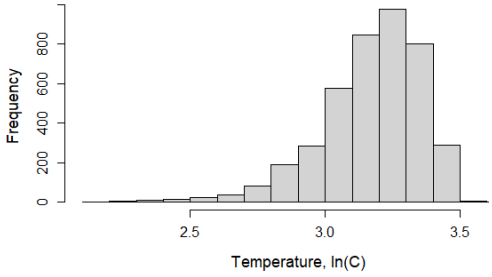

Frequency
0
500
1500
2500
0
500
1000
1500
Total phosphorus, ug/L

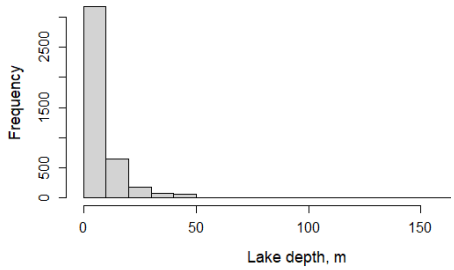

Frequency
0
100
300
500
1
2
3
4
5
6
7
Total phosphorus, ln(ug/L)


Frequency
0
500
1000
1500
0
2000
4000
6000
8000
Total nitrogen, ug/L


Frequency
0
200
400
600
800
4
5
6
7
8
9
Total nitrogen, ln(ug/L)


Frequency
0
200
400
600
800
10
15
20
25
30
35
Temperature, C


Frequency
0
200
400
600
800
2.5
3.0
3.5
Temperature, ln(C)


Frequency
0
500
1500
2500
0
50
100
150
Lake depth, m

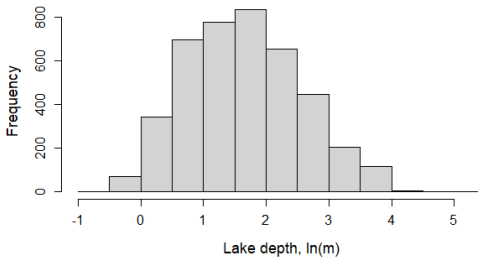
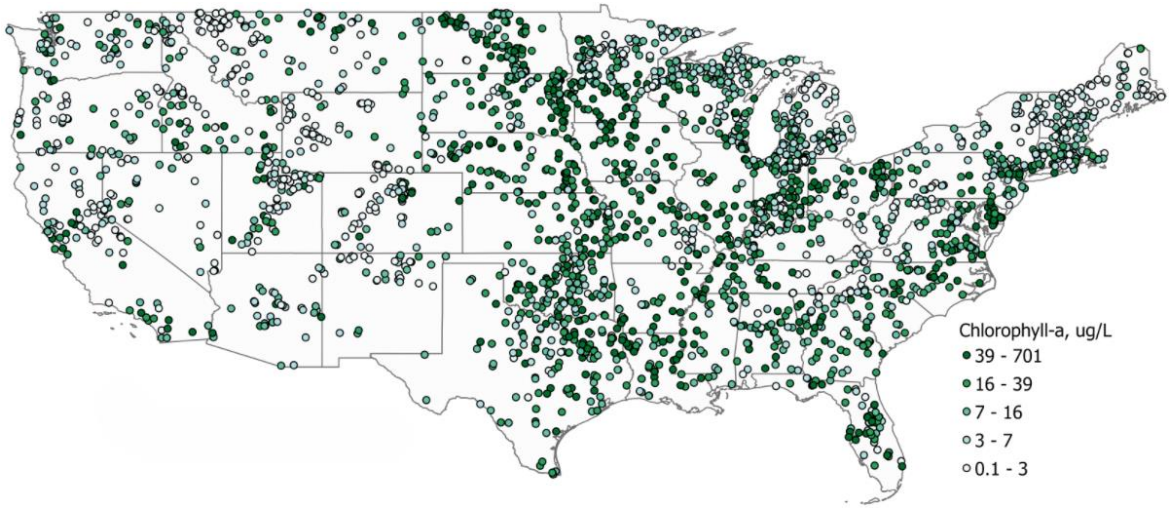
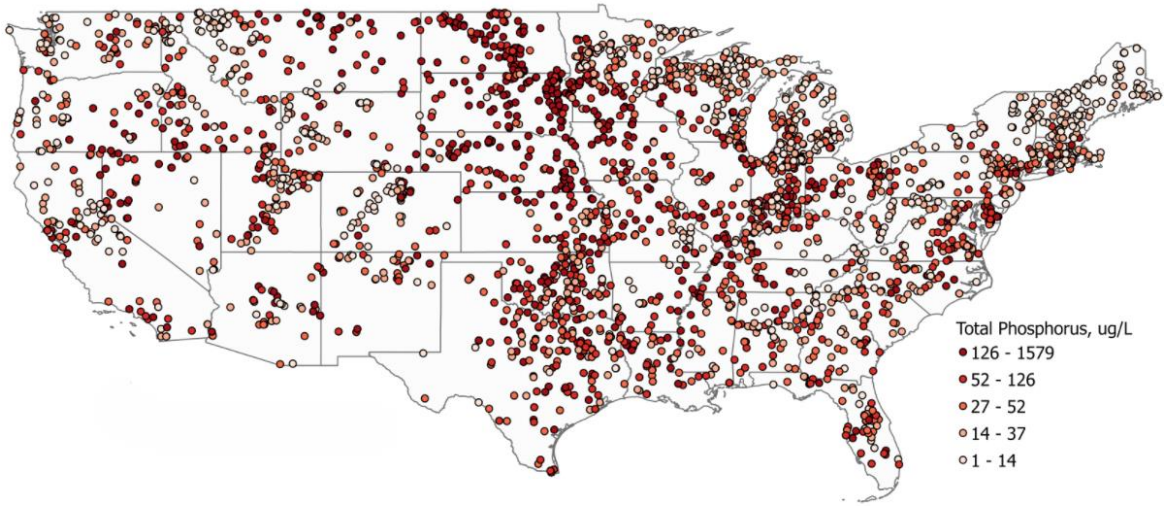
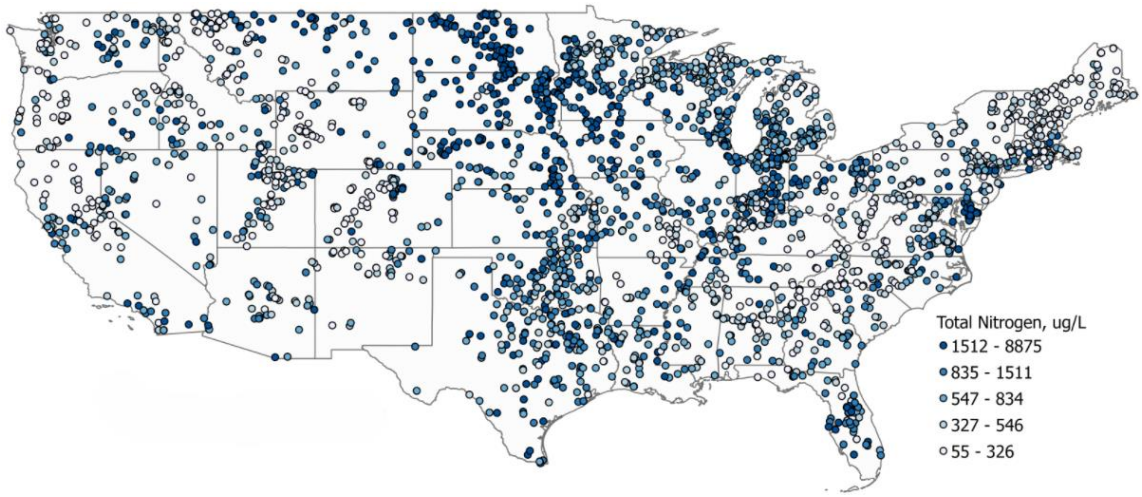
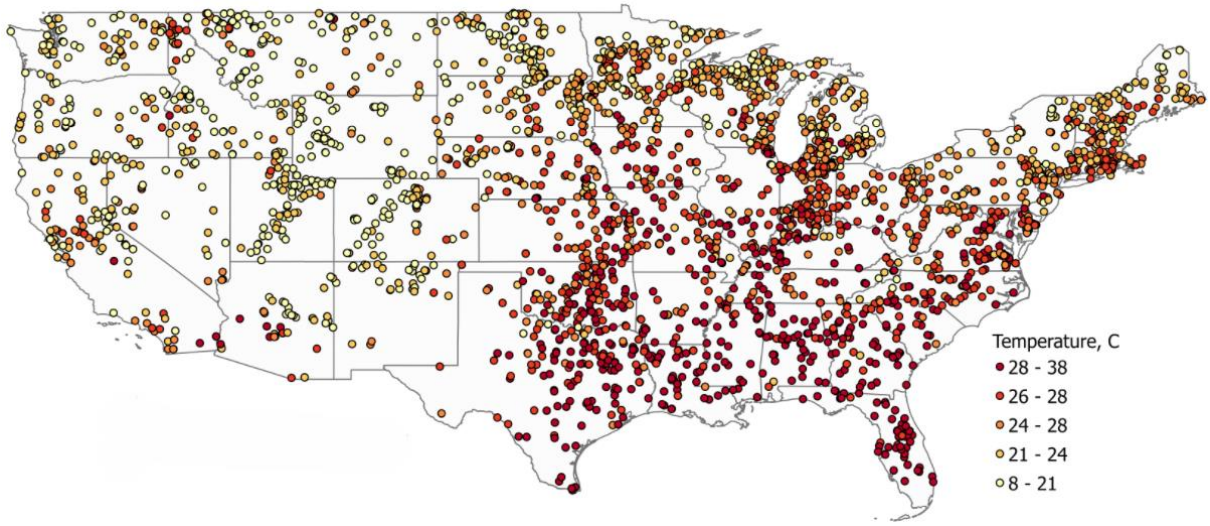

Frequency
0
200
400
600
800
-1
0
1
2
3
4
5
Lake depth, ln(m)

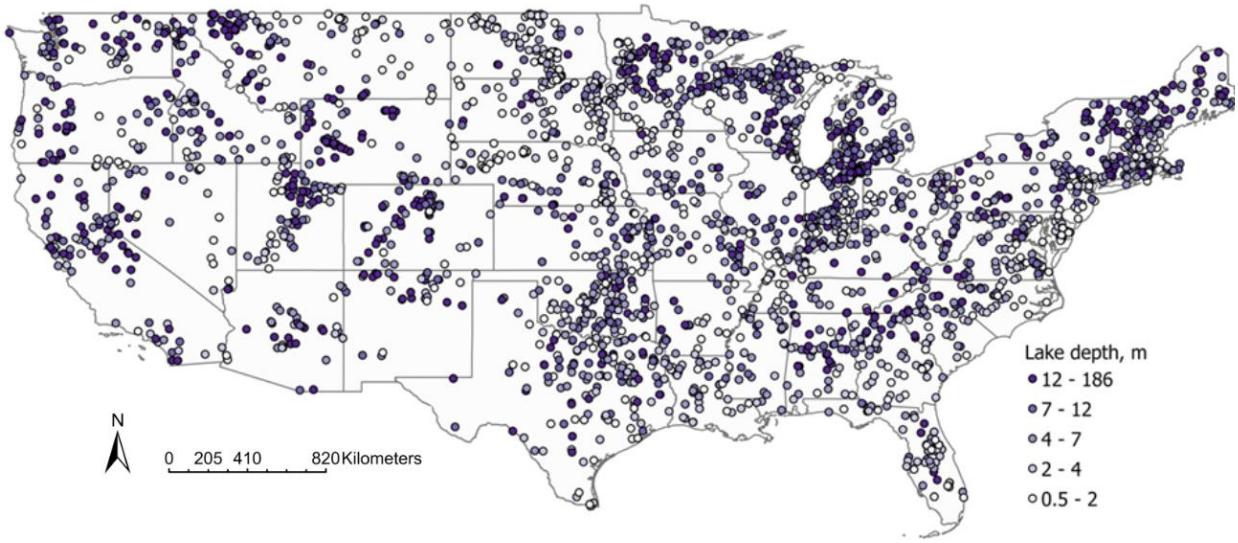
Figure SI 1.1.4. Spatial distribution of model variables after outlier removal.


Chlorophyll-a, ug/L
39 - 701
16 - 39
7 - 16
3 - 7
0.1 - 3


Total Phosphorus, ug/L
126 - 1579
52 - 126
27 - 52
14 - 37
1 - 14


Total Nitrogen, ug/L
1512 - 8875
835 - 1511
547 - 834
327 - 546
55 - 326


Temperature, C
28 - 38
26 - 28
24 - 28
21 - 24
8 - 21


Lake depth, m
12 - 186
7 - 12
4 - 7
2 - 4
0.5 - 2
N
0 205 410 820 Kilometers

SI 2.0 Results Supplemental Information

SI 2.1 Algal models including individual auxiliary variables

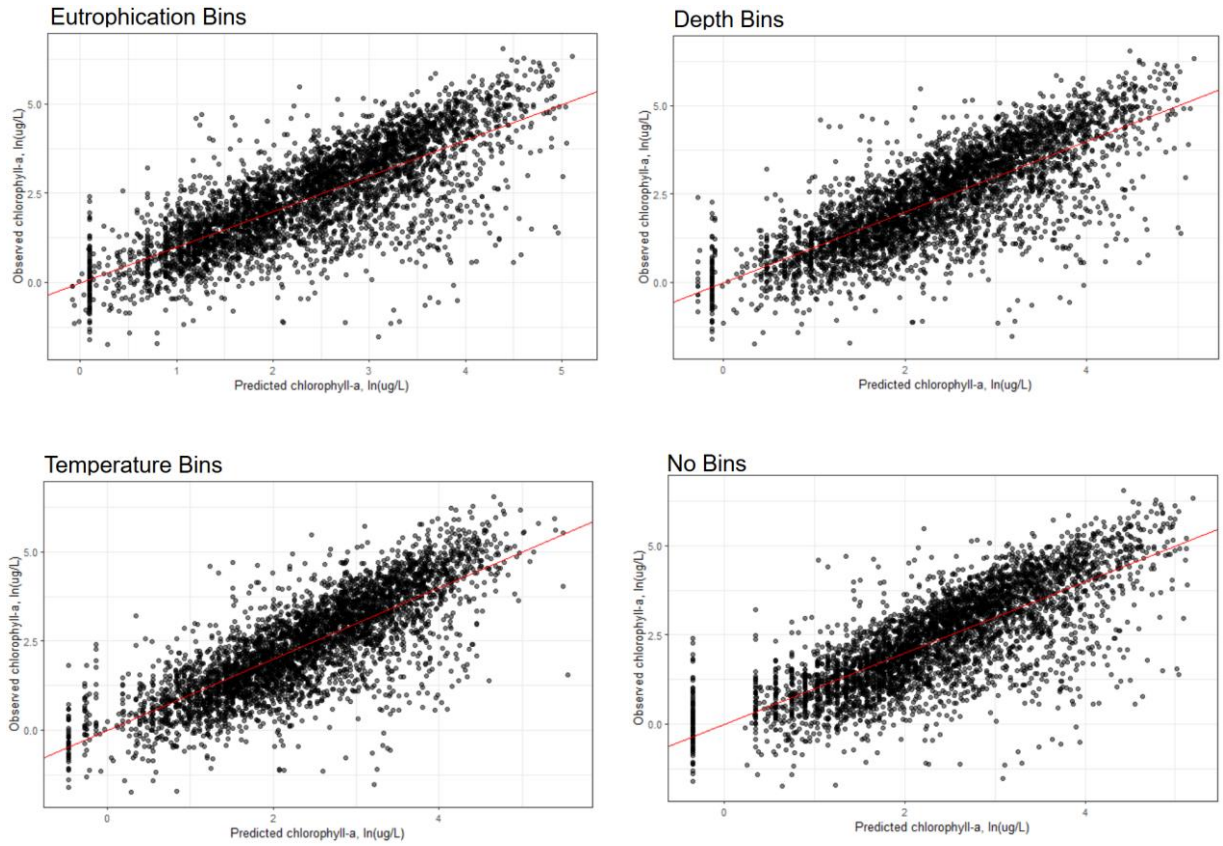


Figure SI 2.1.1. Residual plots for limiting nutrient IAV models.

Table SI 2.1.1. Summary of multiplicative Individual Auxiliary Variable (IAV) models. β_{CR}, β_{LN}, and β_0 metrics provided for the hyperparameter of binned models. Standard deviations of respective parameter means are expressed in parentheticals.

Metric/Parameter	**Units**	**Auxiliary Variable**			
		Eutrophication	**Depth**	**Temperature**	**None**
R^2	unitless	0.58	0.57	0.61	0.56
RMSE	ln(µg/L)	0.91	0.92	0.88	0.93
β_{TN}	unitless	0.58 (0.07)	0.64 (0.13)	0.58 (0.04)	0.59 (0.03)
β_{TP}	unitless	0.43 (0.10)	0.38 (0.03)	0.44 (0.07)	0.44 (0.02)
β_0	ln(µg/L)	-3.47 (1.12)	-3.52 (1.08)	-3.10 (0.20)	-3.10 (0.12)

Table SI 2.1.2. Summary of parameters for limiting nutrient IAV models. Standard deviations of respective parameter means are expressed in parentheticals.

Parameter	Units	Eutrophication	Depth	Temperature
$\beta_{CR,1}$	unitless	33.4 (4.3)	12.2 (0.8)	18.5 (1.6)
$\beta_{CR,2}$	unitless	15.8 (1.6)	13.5 (0.8)	16.0 (1.5)
$\beta_{CR,3}$	unitless	16.3 (1.1)	18.6 (1.2)	17.4 (1.3)
$\beta_{CR,4}$	unitless	13.8 (0.9)	20.0 (1.4)	13.4 (1.0)
$\beta_{CR,5}$	unitless	13.7 (0.8)	24.3 (2.1)	13.2 (1.1)
$\beta_{0,1}$	ln(µg/L)	-0.51 (0.11)	-0.88 (0.11)	-1.11 (0.09)
$\beta_{0,2}$	ln(µg/L)	-0.83 (0.16)	-0.92 (0.12)	-0.91 (0.08)
$\beta_{0,3}$	ln(µg/L)	-0.70 (0.17)	-0.75 (0.10)	-0.93 (0.08)
$\beta_{0,4}$	ln(µg/L)	-0.76 (0.21)	-0.97 (0.12)	-0.80 (0.10)
$\beta_{0,5}$	ln(µg/L)	-0.90 (0.25)	-0.73 (0.10)	-0.93 (0.11)
$\beta_{LN,1}$	unitless	0.87 (0.06)	0.93 (0.03)	0.93 (0.03)
$\beta_{LN,2}$	unitless	0.89 (0.06)	0.96 (0.03)	0.91 (0.03)
$\beta_{LN,3}$	unitless	0.91 (0.05)	0.95 (0.03)	0.98 (0.03)
$\beta_{LN,4}$	unitless	0.93 (0.05)	1.00 (0.04)	0.97 (0.03)
$\beta_{LN,5}$	unitless	0.94 (0.05)	0.87 (0.04)	1.04 (0.03)
σ_l	ln(µg/L)	0.61 (0.02)	0.62 (0.02)	0.53 (0.02)
σ_s	ln(µg/L)	0.64 (0.01)	0.65 (0.01)	0.67 (0.01)
τ_{CR}	unitless	7.6 (1.5)	6.0 (1.8)	3.6 (1.8)
τ_{LN}	unitless	0.07 (0.07)	0.12 (0.08)	0.09 (0.06)
τ_0	ln(µg/L)	0.29 (0.22)	0.33 (0.22)	0.21 (0.15)

SI 2.2 Algal model with multiple auxiliary variables

Equations SI 2.2.1 and 2.2.2 display the parameter calculation process where $\beta_{i,j,k}$ represents the the *i*-th parameter (slopes or intercept) for the *j*-th model parameter (critical ratio, limiting nutrient) dependent upon the *k*-th auxiliary variable (lake depth, temperature). *D* and *T* represent lake depth (m) and temperature (C), respectively.

$$\ln(a) = \beta_0 + \ln(LN) * (\beta_{1,LN,T} * T + \beta_{0,LN}) + \epsilon_l + \epsilon_s \quad \text{Eq. SI 2.2.1}$$

$$LN = \min\left[TP, \frac{TN}{(\beta_{1,CR,D} * \ln(D) + \beta_{0,CR})}\right] \quad \text{Eq. SI 2.2.2}$$

Table SI 2.2.1. Summary of parameters for algal model with multiple auxiliary variables.

Parameter	Units	Mean	St. Deviation
$\beta_{1,CR,D}$	$\ln(m^{-1})$	3.32	0.45
$\beta_{0,CR}$	unitless	16.51	0.57
$\beta_{1,LN,T}$	C^{-1}	0.018	0.00
$\beta_{0,LN}$	unitless	0.94	0.02
$\beta_{0,0}$	ln(µg/L)	-0.85	0.05
σ_l	ln(µg/L)	0.52	0.02
σ_s	ln(µg/L)	0.66	0.01

The specific lakes used to develop Figure 3.2.3 are as follows:

- A P-limited Michigan lake (Clear Lake). This lake has two samples with the site ID NLA17_MI-10012 from the 2017 NLA dataset and NLA12_MI-165 from the 2012 NLA dataset. The average observed TP, TN, and chl-a values are 15.1 µg/L, 577 µg/L, and 13.8 µg/L, respectively. The average index site depth is 4.1 meters, and the average lake temperature is 25.4℃. The percentage of samples indicating P-limitation for this lake is 99%.
- An N-limited Nevada lake (Bilk Creek Reservoir). This lake has the site ID NLA12_NV-119 and is from the 2012 NLA dataset. The observed TP, TN, and chl-a values are 108 µg/L, 510 µg/L, and 5.5 µg/L, respectively. The index site depth is 3.0 meters, and the lake temperature is 21.6℃. The percentage of samples indicating P-limitation for this lake is 15%.
- An Ohio lake with uncertain nutrient limitation (Summit Lake) This lake has one sample with the site ID NLA12_OH-137 and is from the 2012 NLA dataset. The observed TP, TN, and chl-a values are 41 µg/L, 845 µg/L, and 36 µg/L, respectively. The index site depth is 12.5 meters, and the lake temperature is 27℃. The percentage of samples indicating P-limitation for this lake is 50%.

SI 2.3 Model diagnostics and cross validation

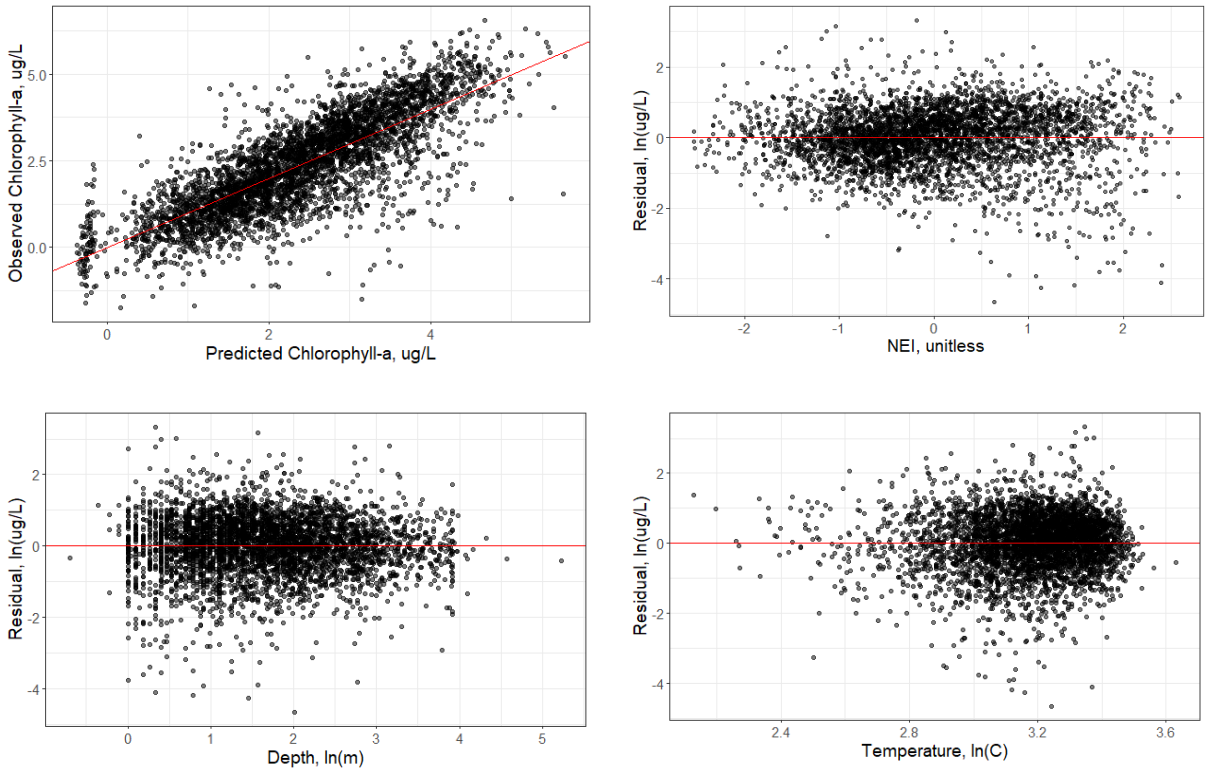


Figure SI 2.3.1. Chlorophyll-a residuals (observations minus predictions) of multiple auxiliary variable (MAV) model. Note that the NEI was not included in the final MAV model, but is shown here for reference.

Table SI 2.3.1. Cross validation results by fold.

Validation Data	Calibration		Validation	
	R^2	RMSE	R^2	RMSE
NLA 2022	0.65	0.83	0.62	0.86
NLA 2017	0.64	0.86	0.68	0.76
NLA 2012	0.66	0.82	0.59	0.90
NLA 2007	0.64	0.83	0.65	0.86
Combined Validation Results	-	-	0.64	0.85
Calibration Using All NLA Years	0.65	0.84	-	-

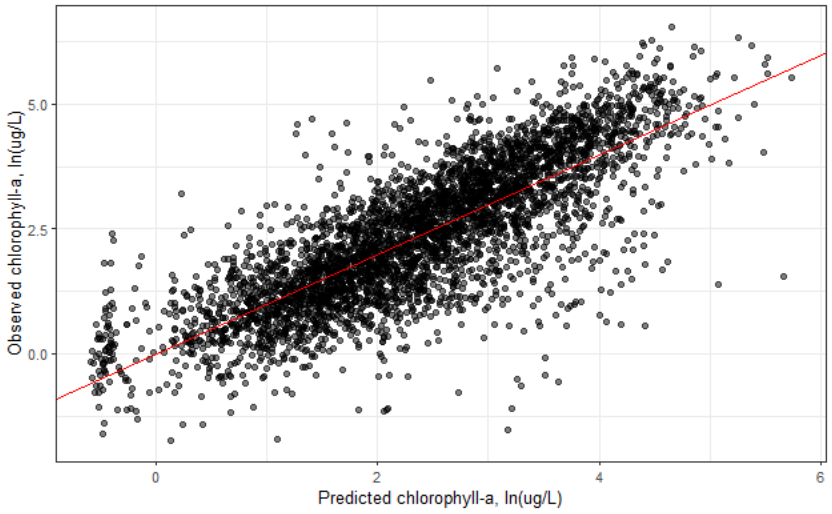


Figure SI 2.3.2. Cross validation observations versus predictions plots for validation dataset.

Table SI 2.3.2. Parameter estimates of cross validation folds. Percent deviations from the base MAV model calibrated to all data are shown in parentheses.

Parameter	Units	Base	Fold 1	Fold 2	Fold 3	Fold 4
$\beta_{1,CR,D}$	$\ln(m^{-1})$	3.32	4.12 (24%)	3.17 (-5%)	2.93 (-12%)	2.71 (-18%)
$\beta_{0,CR}$	unitless	16.51	17.06 (3%)	17.35 (3%)	16.07 (-3%)	15.49 (-6%)
$\beta_{1,LN,T}$	C^{-1}	0.018	0.017 (-4%)	0.019 (7%)	0.020 (12%)	0.018 (-2%)
$\beta_{0,LN}$	unitless	0.94	0.93 (-1%)	0.94 (0%)	0.93 (-1%)	0.99 (6%)
$\beta_{0,0}$	ln(µg/L)	-0.85	-0.81 (-5%)	-0.85 (-1%)	-0.79 (-7%)	-1.10 (29%)
σ_l	ln(µg/L)	0.52	0.51 (-2%)	0.54 (5%)	0.48 (-8%)	0.53 (2%)
σ_s	ln(µg/L)	0.66	0.66 (-1%)	0.67 (0%)	0.67 (1%)	0.65 (-2%)

SI 2.4 Model application and mapping

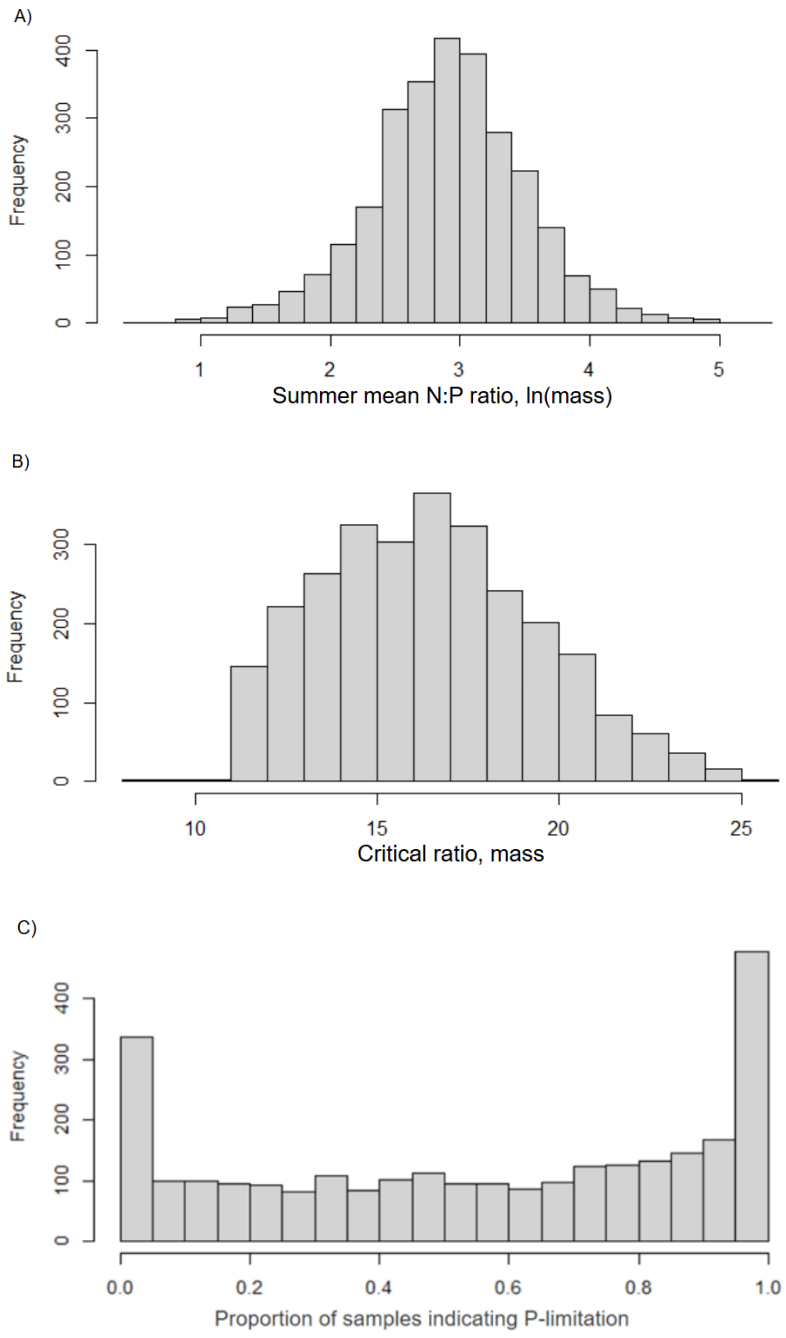


Figure SI 2.4.1. Histograms of estimated (A) summer mean N:P ratio, (B) critical N:P ratio, (C) probability of phosphorus versus nitrogen limitation.

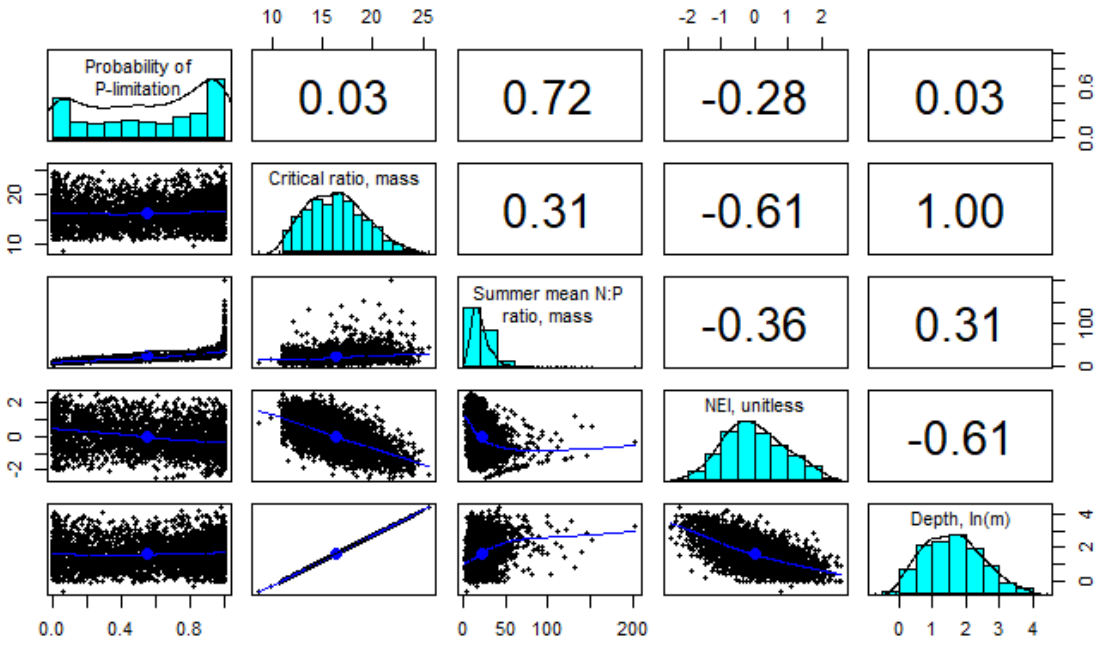


Figure SI 2.4.2. Correlation plots, histograms, and correlation coefficients of estimated summer mean N:P ratio, critical N:P ratio, probability of phosphorus limitation, nutrient enrichment index (NEI), and water depth.