Rethinking the Lake Trophic State Index

Farnaz Nojavan A.[a,*], Betty J. Kreakie[b], Jeffrey W. Hollister[b], Song S. Qian[c]

[a] *ORISE, Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, 27 Tarzwell Drive Narragansett, RI 02882 U.S.A.*
[b] *Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, 27 Tarzwell Drive Narragansett, RI 02882 U.S.A.*
[c] *Department of Environmental Sciences, The University of Toledo, Toledo, OH 43606 U.S.A.*

Abstract

Lake trophic state classifications provide information about the condition of lentic ecosystems and are indicative of both ecosystem services (e.g., clean water, recreational opportunities, and aesthetics) and disservices (e.g., cyanobacteria blooms). The current classification schemes have been criticized for developing indices that are single variable, discrete, and/or deterministic. We present an updated lake trophic classification model using a Bayesian multilevel ordered categorical regression. The model consists of (1) a proportional odds logistic regression (POLR) that models ordered, categorical, lake trophic state using Secchi disk depth, elevation, nitrogen concentration, and phosphorus concentration and (2) models linking universally available GIS variables to nitrogen (N) and phosphorus (P). The linkage makes it simple and cost effective to predict lake trophic state, especially for lakes without costly monitoring data. We used the National Lakes Assessment 2007 data set. The overall accuracy for POLR model was 0.7 and the balanced accuracy ranged between 0.74 and 0.84. For the POLR model combined with the models of N and P the overall accuracy was 0.6 and balanced accuracy was between 0.68 and 0.78 for each of the classes. This work delivers an index that is multivariate, continuous, and classifies lakes in probabilistic terms. While our model addresses all the limitations of the current approach to lake trophic classification, the addition of uncertainty quantification is exceedingly important, because the trophic state response to predictors varies among lakes. Our model successfully addresses concerns with the current approach and performs well across trophic states in a large spatial extent.

Keywords: Trophic State, Proportional Odds Logistic Regression Model, Bayesian Multilevel Ordered Categorical Regression Model, National Lakes Assessment

1. Introduction

Lake trophic state has become an invaluable tool for lake managers and researchers, and therefore demands due diligence to ensure that the statistical methods and results are robust. Lake trophic state is a proxy for lake productivity, water quality, biological integrity, and fulfillment of designated use criteria (Maloney, 1979; USEPA, 1994). Monitoring water quality is integral to the management of the eutrophication and productivity of lakes. Recreation, habitat and species diversity, property and ecological values are closely related to lake water quality (Keeler et al., 2015; Leggett and Bockstael, 2000). In fact, the Clean Water Act requires that all U.S. lakes be classified according to trophic status in order to provide insight about overall lake quality (USEPA, 1974). Trophic state can be used both as a communication tool with the public and a management tool to provide the scientific accord of eutrophication and character of the lake.

The concept of trophic state, originally proposed by Naumann (1919), is based on lake production and quantified by algal biomass due to their impacts on a lake's biological structure. Naumann (1919) emphasized a regional approach to trophic state due to regional variation in lake production. However, current lake trophic state models are one size fits all. Given its broad applicability and long history, it is important to periodically review, and update the methods used to calculate trophic state. Trophic state has been formulated using various indices, the most well known was created by Carlson (1977). Building on his work, others have developed numerous classification schemes which vary considerably in their approach to classification, variable selection, and category counts. Single parameter indices have been developed using nutrient concentrations, nutrient loading, algal productivity, algal biomass, and hypolimnetic oxygen (for an extensive review see Carlson and Simpson (1996)). A single parameter index simplifies and reduces the expense of monitoring; however, there is no scientific consensus on the best single trophic state predictor (for an extensive review see Carlson and Simpson (1996)).

Multiparameter index approaches view trophic state as a complex response caused by interaction among various physical, chemical and biological factors. Multiparameter indices use any combination of causal factors usually through definition of sub-indices and integrating the sub-indices to calculate a final index (Carlson and Simpson, 1996; Brezonik, 1984). The multivariate approaches are more realistic but costly in terms of data collection. Classification procedures

*Corresponding author
Email address: nojavan.farnaz@epa.gov (Farnaz Nojavan A.)

also differ greatly; some indices are quantitative and continuous whereas others are qualitative and discrete. A continuous index accommodates trophic changes along a production gradient; however, these are often discretized for reasons of convenience and ease of communication. A discrete index classifies lakes into a small number of categories resulting in loss of information on position across the trophic continuum and lack of sensitivity to changes in predictor variables. Lakes have a large degree of variability in their response to a given variable, like nutrient concentrations, and this leads to uncertainty in the trophic response. Hence, trophic state should be formulated in probabilistic terms to quantify this uncertainty.

This paper addresses many of the aforementioned critiques by developing a Bayesian multilevel ordered categorical regression model to classify lake trophic state. The model contributes to literature on trophic state in several ways. First, it generates an index that is multivariate by using Secchi depth (an inexpensive surrogate for algal biomass), elevation, total nitrogen concentration, and total phosphorus concentration. Second, the developed index is continuous and thus captures a given lakes position along the trophic continuum. Third, the index classifies lakes in probabilistic terms. Finally, while it is critical to locate a lake across trophic continuum, it is not economically feasible to monitor all lakes by conventional sampling techniques. The developed multilevel model links easily accessible and universally available GIS variables to nitrogen and phosphorus in the Proportional Odds Logistic Regression (POLR) model allowing it to predict the trophic state of all lakes, even not extensively sampled ones.

2. Material and methods

2.1. Data and Study Area

We used data from the 2007 National Lakes Assessment (NLA), the National Land Cover Dataset (NLCD), and lake morphometry modeled from the NHDPlus and National Elevation Data Set (USEPA, 2009). The national survey was conducted by the U.S. Environmental Protection Agency (USEPA, 2009; Homer et al., 2004; Xian et al., 2009; Hollister and Milstead, 2010; Hollister et al., 2011; Hollister, 2014). Ancillary data, such as Wadeable Streams Assessment ecological regions is also included in the NLA (Omernik, 1987; USEPA, 2006). The sampling population included all permanent non-saline lakes, reservoirs, and ponds within the 48 contiguous United States with a surface area greater than 4 hectares and a depth of greater than 1 meter, omitting the Great Lakes. A Generalized Random Tessellation Stratified (GRTS) survey design

for a finite resource was used with stratification and unequal probability of selection. The design included reverse hierarchical ordering of the selected lakes. The NLA sampled over 1000 lakes across the continental United States (Figure 1). The lakes were sampled at the hypothesized deepest point in the lake (or in the middle of the lake if the lake was deeper than 50 meters), during summer 2007 overlapping with peak summer recreational use of waterbodies, hence, increasing the appeal of classification results to public. The source code and data are available on GitHub repository `https://github.com/usepa/rethinking_tsi` (Nojavan A. et al., 2017).

2.2. Methods

2.2.1. Variable Selection

The goal of variable selection is to identify an optimal reduced subset of predictor variables. Here we used the results from random forest modeling as a means of variable selection. Random forest modeling is a machine learning algorithm that builds numerous statistical decision trees in order to attain a consensus predictor model (Breiman, 2001). Each tree is based on recursively bootstrapped data, and the out-of-bag (OOB) data, cases left out of the sample, provides an unbiased estimation of model error and measure of predictor variable importance. Random forest modeling was conducted in `randomForest` package in R (Liaw and Wiener, 2002; R Core Team, 2016). We developed random forest models to select predictor variables to model trophic state, nitrogen, and phosphorus. The random forest model for trophic state included *in situ* water quality data and universally available GIS data, e.g. landscape data (see Hollister et al. (2016) for detailed methods). The models for nitrogen and phosphorus included only universally available GIS data. We used percent increase in mean square error to examine variable importance.

2.2.2. Variable Transformation

Using the central limit theorem, Ott (1995) demonstrates that environmental concentration variables are log-normally distributed. As such, we log-transformed total nitrogen concentration, total phosphorus concentration, and secchi disk depth data prior to our statistical analyses. Additionally, we note that the interpretation of regression model coefficients are different when log-transformed (Qian, 2010). Further, all predictors in the POLR model (2.2.3) were standardized based on the discussion of Gelman and Hill (2007) and Gelman (2008) on centering and scaling predictors to simplify the interpretation of the intercept when predictors cannot be

set equal to zero. Scaling also improves the interpretation of coefficients in models with interacting terms, and coefficients can be interpreted on approximately a common scale. Weisberg (2005) also demonstrates that centered predictors would result in uncorrelated regression model coefficients.

2.2.3. Proportional Odds Logistic Regression Model

The response variable, lake trophic status, is a categorical variable that can take on four values, i.e. oligotrophic (1), mesotrophic (2), eutrophic (3), and hypereutrophic (4). The proportional odds logistic regression (POLR) model has been used to account for the ordered categories of the response variable (Gelman and Hill, 2006). The lower block in figure 2 represents the POLR model. The POLR model was set up as follows:

$$y_i = \begin{cases} Oligotrophic & \text{if } z_i < c_{1|2} \\ Mesotrophic & \text{if } z_i \in (c_{1|2}, c_{2|3}) \\ Eutrophic & \text{if } z_i \in (c_{2|3}, c_{3|4}) \\ Hypereutrophic & \text{if } z_i > c_{3|4} \end{cases}$$

$$z_i \sim logistic(XA, \sigma^2)$$

$$\text{where } XA = \text{Secchi Disk Depth}_i \times \alpha_{SDD} + \text{Phosphorus}_i \times \alpha_{Phosphorus} + \text{Nitrogen}_i \times \alpha_{Nitrogen} + \text{Elevation}_i \times \alpha_{Elevation}$$

with the coefficients A and $c_{k|k+1}$ (known as cutpoints or thresholds), and the design matrix of predictors X. The two adjacent cutpoints and XA are used to classify the response variable. Hence, XA can be the equivalent of the lake trophic state index introduced by Carlson (1977). $Pr(y > k)$ represents the probability of a lake's trophic state being higher than k (k= 1, 2, 3, 4). The probability of a lake's trophic state being k, for example $k = 4$ (i.e. hypereutrophic), is calculated $Pr(y > k - 1) - Pr(y > k)$. The $Pr(y = k) > Pr(y =$ all other classes except $k)$ when $c_{k-1|k} < XA < c_{k|k+1}$.

2.2.4. Multilevel Models: Linking Universally Available GIS Data to Nutrients

The multilevel models were developed to extend the developed POLR model from lakes with monitored nitrogen and phosphorus data to all lakes (i.e. unsampled lakes). They link nitrogen

and phosphorus in the POLR model to universally available GIS data. Figure 2 represents the regression models. The model is grouped into two blocks (gray shaded rectangles). The trophic state classification regression, the POLR model in the lower block, includes nitrogen, phosphorus, secchi disk, and elevation as predictors. The nutrients (i.e., nitrogen and phosphorus) model, upper block, estimates the means of nitrogen and phosphorus based on ecoregion, % evergreen forest, and latitude. The two blocks are connected through the estimated means of nitrogen ($\mu_{Nitrogen}$) and phosphorus ($\mu_{Phosphorus}$). The combined model enabled trophic state classification for all lakes without the costly sampling requirement. The predictor variables selection for the second level models used random forest modeling approach (see subsection 2.2.1). The relationship between nitrogen, phosphorus, and their predictors was examined using multilevel linear regression models. The standard deviation of the normal distribution, as well as each parameter in the regression model, were then assigned non-informative prior distributions (uniform, or nearly so, to allow the information from the likelihood to be interpreted probabilistically). The models were set up as follows:

$$\text{Nitrogen}_{\text{ij}} \sim \mathcal{N}(\mu_{Nitrogen_{ij}}, \sigma^2_{Nitrogen})$$
$$\mu_{Nitrogen_{ij}} = X_{Nitrogen}B = \beta_{Ecoregion_j} + \beta_{Latitude} \times Latitude_{ij} + \beta_{\%Evergreen} \times \%Evergreen_{ij}$$

Priors:

$\beta_{Ecoregion_j} \sim \mathcal{N}(0, 10000)$, where $j = 1, ..., 9$

$\beta_{Latitude} \sim \mathcal{N}(0, 10000)$

$\beta_{\%Evergreen} \sim \mathcal{N}(0, 10000)$

Where $X_{Nitrogen}$ is the matrix of predictors and B is the vector of coefficients.

$Nitrogen_{ij}$, $Latitude_{ij}$, $\%Evergreen_{ij}$ are the ith nitrogen, latitude, and %evergreen observations, respectively, in the jth ecoregion.

$\beta_{Ecoregion_j}$ is the varying intercept for $Ecoregion_j$.

$\beta_{Latitude}$ and $\beta_{\%Evergreen}$ are coefficients for latitude and %evergreen, respectively.

$$\text{Phosphorus}_{ij} \sim \mathcal{N}(\mu_{Phosphorus_{ij}}, \sigma^2_{Phosphorus})$$

$$\mu_{Phosphorus_{ij}} = X_{Phosphorus}\Gamma = \gamma_{Ecoregion_j} + \gamma_{Latitude} \times Latitude_{ij} + \gamma_{\%Evergreen} \times \%Evergreen_{ij}$$

Priors:

$\gamma_{Ecoregion_j} \sim \mathcal{N}(0, 10000)$, where $j = 1, ..., 9$

$\gamma_{Latitude} \sim \mathcal{N}(0, 10000)$

$\gamma_{\%Evergreen} \sim \mathcal{N}(0, 10000)$

Where $X_{Phosphorus}$ is the matrix of predictors and Γ is the vector of coefficients.

$Phosphorus_{ij}$, $Latitude_{ij}$, $\%Evergreen_{ij}$ are the *i*th phosphorus, latitude, and %evergreen observations, respectively, in the *j*th ecoregion.

$\gamma_{Ecoregion_j}$is the varying intercept for $Ecoregion_j$.

$\gamma_{Latitude}$ and $\gamma_{\%Evergreen}$ are coefficients for latitude and %evergreen, respectively.

2.2.5. Model Evaluation

The NLA 2007 includes trophic state classification based on chlorophyll *a*, nitrogen, and phosphorus. There is discrepancy in the results of classification based on chlorophyll *a*, nitrogen, and phosphorus. The reasons behind the lack of agreement between the common classification methods is discussed in detail by Carlson and Havens (2005). We avoided deviations in our evaluation data by only using 10% of the consistently classified lakes across three methods as our cross validation data. We developed the model using the rest of the data. We decided to use cross validation as opposed to validating the model with a new data set as a comparable dataset (i.e. 2012 National Lakes Assessment (USEPA, 2016)) was not available during the model development process. It is possible to use NLA 2012 to evaluate and update the model. We evaluated the model using balanced accuracy, the average of the proportion of correct predictions within each class individually, and overall accuracy, the proportion of the total number of correct prediction.

3. Results

3.1. Variable Selection: Random Forest

The random forest models evaluated variable importance for trophic state, nitrogen, and phosphorus and the results are reported in figures 3-5. The number of variables for each response

variable was decided using the variable selection plots (Figures S1-S3) which show percent increase in mean squared error as a function of the number of variables. We used seventy predictor variables in the random forest model for trophic state and it indicated the best representation of trophic state classification could be achieved using four variables, adding more than four variables had incremental (< 0.1) impact on root mean square error. The four most important variables were turbidity, total phosphorus, total nitrogen, and elevation. The NLA uses secchi disk depth as a measure of water clarity and, hence, we used it as a proxy for turbidity. The variable selection narrows down the GIS variables to only three required to predict nitrogen and phosphorus, including %evergreen forest, latitude, and ecoregion.

3.2. Proportional Odds Logistic Regression Model

The trophic state index is calculated as: $TSI = -1.69\times\text{Secchi Disk Depth}_i + 0.69\times\text{Nitrogen}_i + 0.55\times\text{Phosphorus}_i - 0.56\times\text{Elevation}_i$. The classification rules, based on cutpoints, are described below:

$$y_i = \begin{cases} Oligotrophic & \text{if } z_i < -3.36 \\ Mesotrophic & \text{if } z_i \in (-3.36, -0.18) \\ Eutrophic & \text{if } z_i \in (-0.18, 2.62) \\ Hypereutrophic & \text{if } z_i > 2.62 \end{cases}$$

$$z_i \sim logistic(TSI, 1)$$

The resulting POLR model has three cutpoints and four slope coefficients (Table 1). Figures 6 and 7 summarize the model uncertainty. The POLR model returns four probabilities associated with each trophic state as opposed to on fixed classification (Figure 7). The overall accuracy is 0.68 and the balanced accuracies are 0.93, 0.83, 0.72, 0,73 for oligotrophic, mesotrophic, eutrophic, and hypereutrophic classes, respectively. Table 2 shows the confusion matrix for the POLR model. Each element of the confusion matrix is the number of cases for which the actual state is the row and the predicted state is the column.

3.3. Multilevel Models: Linking Universally Available GIS Data to Nutrients

The overall accuracy of the multilevel model was 0.6 and the balanced accuracies were 0.78, 0.77, 0.69, 0.68 for oligotrophic, mesotrophic, eutrophic, and hypereutrophic classes, respectively (Table 3). Table 4 shows the confusion matrix for the POLR model.

4. Discussion

The Bayesian multilevel ordered categorical regression model presented uses multiple variables to predict lake trophic state. A multi-variable predictor model accounts for chemical, biological, and physical aspects of trophic state classification. The drawback of such a model is the cost of monitoring multiple predictor variables for trophic state classification. This is addressed in the presented model by linking nitrogen and phosphorus to universally available GIS variables. Hence, only secchi disk depth, an easily measured variable and already available for many lakes, is needed to predict lake trophic state index and classification. The model quantifies lake trophic state across a continuum. The lake trophic classification is done for organization and communication purpose. The lake trophic state is a variable that changes gradually across a gradient and it is important to predict where across the trophic continuum a lake falls, especially for lake restoration and management projects. The continuous trophic index helps us capture lake trophic sensitivity to changes in nitrogen and phosphorus. Finally, the proposed model quantifies the uncertainty of lake trophic response to changes in nutrients, as the response varies from lake to lake.

The multilevel model has an overall accuracy of 0.6; yet this measure of performance fails to capture whether our stated goals were satisfactorily achieved. The accuracy measure requires that we use previous categorization of lakes based on single parameter trophic state. Somewhat circular to our goals, we are relying on discretized classifications to measure the performance of our continuous probabilistic predictions. We partially addressed this problem by using only lakes that were consistently categorized using the three common classification methods (i.e., chlorophyll *a*, nitrogen, and phosphorus) for evaluation data. A continuous scale better summarizes uncertainty, represented in the probability of being in a certain class (i.e. oligotrophic, mesotrophic, and etc.). In an attempt to circumvent this issue, we introduce balanced accuracy to measure performance of each trophic state. Balanced accuracy (as well as the confusion matrix) illustrates that misclassifications are more likely to be in adjacent trophic states. This phenomenon is also

graphically illustrated in Figures 6 and 7. To be clear, the intent of our model is not to accurately predict how lakes are classified currently, rather we show, that our model, while improving upon the statistical foundation for classification, will be comparable to existing notations of trophic state.

The current trophic state classification models are applied to all lakes, regardless of the regional differences. Lake trophic index, and hence lake trophic classes, should be calculated differently in different eco-regions to accommodate variation in landform and climate characteristics. The developed multilevel POLR model uses eco-region, latitude, and watershed level percent evergreen forest as predictors for nitrogen and phosphorus. Hence, the proposed model has an eco-regional approach to trophic state. The developed multilevel model structure can be further expanded to lake-specific trophic state index, upon availability of multiple measurements for each lake.

We built a model to predict total nitrogen and phosphorus concentration in order to provide input nutrient estimates for the trophic state model. Thereby, avoiding the need for nitrogen and phosphorus data, costly variables to measure for all lakes. Random forest modeling selected percent evergreen forest, latitude, and eco-region as the top predictor variables of total nitrogen and phosphorus. These selected variables appear to be capturing patterns of total nutrient concentration at three different spatial scales. The partial dependency plot for latitude (Figure S4 & S5) depicts high concentrations in the northern and southern extremes of the continental US. The lowest predicted concentrations correspond to the mid-latitudes. While the latitude variable is capturing large spatial patterns, the percent evergreen variable is presumably capturing how total nutrients in lakes respond to the land use decisions immediately adjacent to lakes. The percent evergreen forest variable is a measure of forest within a 3 kilometer buffer around each lake. Distinctly, the ecoregion variable represents an intermediate scale among these three variables and represents the variation between the regions. It is striking that total nutrient concentration in lakes across the continental US is most successfully modeled when using predictors summarizing three discrete spatial scales.

Eutrophication has constituted a serious problem for aquatic ecosystems during the past decades, largely due to excess nutrients associated with anthropogenic activities. Lake restoration projects aim to shift water quality of lakes to or closer to their anthropogenically undisturbed conditions. It is critical to quantitatively plan and assess the recovery of lakes in restoration

projects. We developed a model as a tool for nutrient management scenario analysis and to quantify how altering nutrients moves a lake across the trophic continuum. Further, updating the developed model, described in the following, evaluates the efficacy of restoration plans. Ecosystem managers and policy makers need tools that can help them learn from experience and enable them to manage the ecosystem as new knowledge becomes available. Several studies have called for adaptive management of eutrophication (Rabalais et al., 2002; Stow et al., 2003). Bayesian model updating can be implemented using NLA 2012; however, the data was not available during the model development process.

The Bayesian model updating is based on the repeated use of the Bayes theorem, whereby the posterior of the model developed with non-informative priors and the NLA 2007 data can be used as the prior for the Bayesian model updating step. The model can also be used for new sets of lakes not included in the NLA 2007 data and/or without costly sampling data. The spatial model updating steps and procedure are similar to temporal model updating. Finally, eutrophication has led to cyanobacteria dominance in lakes. The presented model quantifies lake trophic state and the uncertainty around it. The trophic state quantification helps in assessing lake ecological state before and after restoration. The trophic state can be used as a gauge to evaluate how prone lakes are to, often toxic, cyanobacteria blooms. The uncertainty quantification helps express the resisting response of cyanobacteria to variation of phosphorus and nitrogen.

5. Acknowledgements

The views expressed in this article are those of the author(s) and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency. Any mention of trade names, products, or services does not imply an endorsement by the U.S. Government or the U.S. Environmental Protection Agency. The EPA does not endorse any commercial products, services, or enterprises. We greatly thank Bryan Misltead, Jason Gear, and Autumn Oczkowski for their helpful and constructive comments that contributed to improving the manuscript.

References

Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

Brezonik, P. L. (1984). Trophic state indices: rationale for multivariate approaches. Lake and Reservoir Management, 1(1):441–445.

Carlson, R. E. (1977). A trophic state index for lakes. Limnology and oceanography, 22(2):361–369.

Carlson, R. E. and Havens, K. E. (2005). Simple graphical methods for the interpretation of relationships between trophic state variables. Lake and Reservoir Management, 21(1):107–118.

Carlson, R. E. and Simpson, J. (1996). A coordinators guide to volunteer lake monitoring methods. North American Lake Management Society, 96:305.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine, 27(15):2865–2873.

Gelman, A. and Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

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

Hollister, J. (2014). Lakemorpho: Lake morphometry in r. r package version 1.0.

Hollister, J. and Milstead, W. B. (2010). Using gis to estimate lake volume from limited data. Lake and Reservoir Management, 26(3):194–199.

Hollister, J. W., Milstead, W. B., and Kreakie, B. J. (2016). Modeling lake trophic state: a random forest approach. Ecosphere, 7(3).

Hollister, J. W., Milstead, W. B., and Urrutia, M. A. (2011). Predicting maximum lake depth from surrounding topography. PLoS One, 6(9):e25764.

Homer, C., Huang, C., Yang, L., Wylie, B., and Coan, M. (2004). Development of a 2001 national land-cover database for the united states. Photogrammetric Engineering & Remote Sensing, 70(7):829–840.

Keeler, B. L., Wood, S. A., Polasky, S., Kling, C., Filstrup, C. T., and Downing, J. A. (2015). Recreational demand for clean water: evidence from geotagged photographs by visitors to lakes. Frontiers in Ecology and the Environment, 13(2):76–81.

Leggett, C. G. and Bockstael, N. E. (2000). Evidence of the effects of water quality on residential land prices. Journal of Environmental Economics and Management, 39(2):121–144.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. r news 2 (3): 18–22. URL: http://CRAN. R-project. org/doc/Rnews.

Maloney, T. E. (1979). Lake and reservoir classification systems. Environmental Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Corvallis, OR.

Naumann, E. (1919). Några synpunkter angående limnoplanktons ökologi med särskild hänsyn till fytoplankton. Svensk Botanisk Tidskrift, 13:129–163.

Nojavan A., F., Kreakie, B. J., Hollister, J. W., and Qian, S. S. (2017). Rethinking the lake trophic state index. GitHub Repository. doi:10.5281/zenodo.556175.

Omernik, J. M. (1987). Ecoregions of the conterminous united states. Annals of the Association of American geographers, 77(1):118–125.

Ott, W. (1995). Environmental Statistics and Data Analysis. Lewis Publishers, Boca Raton.

Qian, S. (2010). Environmental and Ecological Statistics with R. Chapman and Hall/CRC Press.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabalais, N. N., Turner, R. E., and Scavia, D. (2002). Beyond science into policy: Gulf of mexico hypoxia and the mississippi river. BioScience, 52(2):129–142.

Stow, C. A., Roessler, C., Borsuk, M. E., Bowen, J. D., and Reckhow, K. H. (2003). Comparison of estuarine water quality models for total maximum daily load development in neuse river estuary. Journal of Water Resources Planning and Management, 129(4):307–314.

USEPA (1974). An approach to a relative trophic index system for classifying lakes and reservoirs. Technical Report 24, U.S. Environmental Protection Agency (USEPA), U.S. Environmental Protection Agency, Office of Water and Office of Research and Development, Corvallis, OR.

USEPA (1994). Water quality standards handbook. Technical Report EPA-823-B-94-005a, U.S. Environmental Protection Agency (USEPA), U.S. Environmental Protection Agency, Office of Water and Office of Research and Development, Washington, D.C.

USEPA (2006). Wadeable streams assessment: A collaborative survey of the nations streams. Technical Report EPA 841-b-06-002, U.S. Environmental Protection Agency (USEPA), U.S. Environmental Protection Agency, Office of Water, Office of Research and Development, Washington, D.C.

USEPA (2009). National lakes assessment: A collaborative survey of the nations lakes. Technical Report EPA 841-R-09-001, U.S. Environmental Protection Agency (USEPA), U.S. Environmental Protection Agency, Office of Water and Office of Research and Development, Washington, D.C.

USEPA (2016). National lakes assessment: A collaborative survey of the nations lakes. Technical Report EPA 841-R-16-113, U.S. Environmental Protection Agency (USEPA), U.S. Environmental Protection Agency, Office of Water and Office of Research and Development, Washington, D.C.

Weisberg, S. (2005). Applied Linear Regression. Wiley.

Xian, G., Homer, C., and Fry, J. (2009). Updating the 2001 national land cover database land cover classification to 2006 by using landsat imagery change detection methods. Remote Sensing of Environment, 113(6):1133–1147.

Table 1: Estimated POLR model coefficients and standard errors

		Mean	Std. Error
Slope Coefficients	$\alpha_{\text{Secchi Disk Depth}}$	-1.69	0.13
	$\alpha_{Nitrogen}$	0.69	0.13
	$\alpha_{Phosphorus}$	0.56	0.14
	$\alpha_{Elevation}$	-0.56	0.08
Cutpoints	$C_{1\|2}$	-3.36	0.15
	$C_{2\|3}$	-0.18	0.09
	$C_{3\|4}$	2.62	0.13

Table 2: Confusion matrix for POLR model. Each element of the matrix is the number of cases for which the actual state is the row and the predicted state is the column.

	Oligo	Meso	Eu	Hyper
Oligo	7	1	0	0
Meso	1	14	9	2
Eu	0	0	16	8
Hyper	0	1	4	10

Table 3: Coefficients for the multilevel model.

		Mean	Standard Deviation
Cutoff points/Thresholds	$C_{1\|2}$	-156.60	44.04
	$C_{2\|3}$	-6.18	8.29
	$C_{3\|4}$	121.32	35.04
POLR model coefficients	$\alpha_{Elevation}$	-40.20	12.86
	$\alpha_{Nitrogen}$	-44.33	29.29
	$\alpha_{Phosphorus}$	165.90	46.96
	$\alpha_{\text{Secchi Disk}}$	0.18	5.23
Multilevel model coefficients for nitrogen	$\beta_{\%Evergreen}$	0.00	0.01
	$\beta_{Ecoregion_1}$	0.34	0.13
	$\beta_{Ecoregion_2}$	-0.78	0.12
	$\beta_{Ecoregion_3}$	0.96	0.15
	$\beta_{Ecoregion_4}$	-0.37	0.10
	$\beta_{Ecoregion_5}$	0.59	0.10
	$\beta_{Ecoregion_6}$	0.68	0.09
	$\beta_{Ecoregion_7}$	-0.01	0.10
	$\beta_{Ecoregion_8}$	-1.00	0.10
	$\beta_{Ecoregion_9}$	0.11	0.12
	$\beta_{Latitude}$	0.11	0.05
Multilevel model coefficients for phosphorus	$\gamma_{\%Evergreen}$	-0.00	0.01
	$\gamma_{Ecoregion_1}$	0.40	0.09
	$\gamma_{Ecoregion_2}$	-0.90	0.09
	$\gamma_{Ecoregion_3}$	0.73	0.11
	$\gamma_{Ecoregion_4}$	-0.38	0.08
	$\gamma_{Ecoregion_5}$	0.53	0.08
	$\gamma_{Ecoregion_6}$	0.71	0.07
	$\gamma_{Ecoregion_7}$	-0.32	0.08
	$\gamma_{Ecoregion_8}$	-0.69	0.08
	$\gamma_{Ecoregion_9}$	0.07	0.09
	$\gamma_{Latitude}$	-0.03	0.03
Logistic distribution's scale parameter	σ	75.64	21.27

Table 4: Confusion matrix for multilevel POLR model. Each element of the matrix is the number of cases for which the actual state is the row and the predicted state is the column.

	Oligo	Meso	Eu	Hyper
Oligo	5	3	0	0
Meso	3	12	7	1
Eu	0	0	16	10
Hyper	0	1	3	9

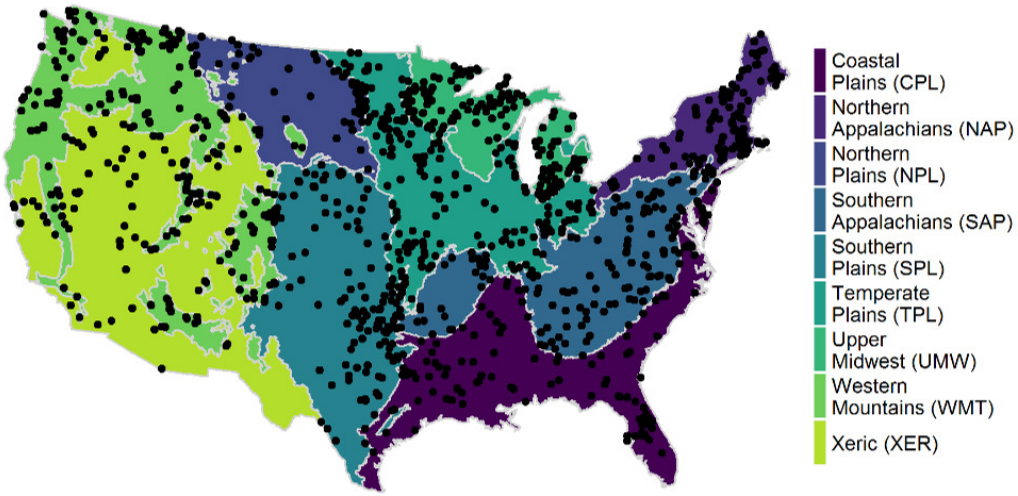

Figure 1: Map of the distribution of National Lakes Assessment sampling locations. Also Wadeable Stream Assessment (WSA) ecoregions are depicted in the map. Areas in an ecoregion have similar landform and climate characteristics.

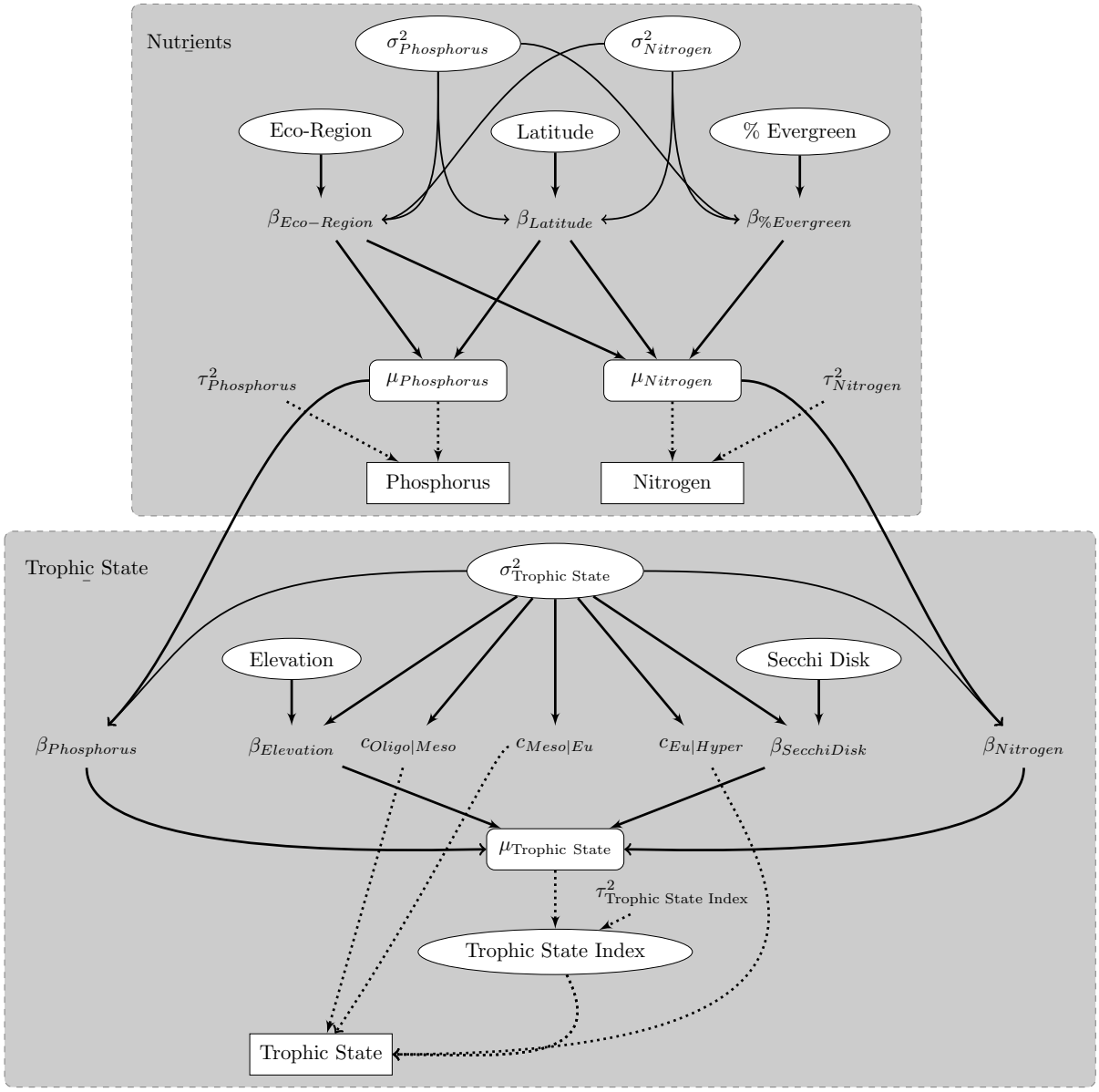

Figure 2: Directed Acyclic Graphical (DAG) model. The lower box depicts the POLR model with its four predictors of secchi disk depth, elevation, nitrogen, and phosphorus. The upper box adds a second level to the POLR model to predict nitrogen and phosphorus using universally available GIS variables.

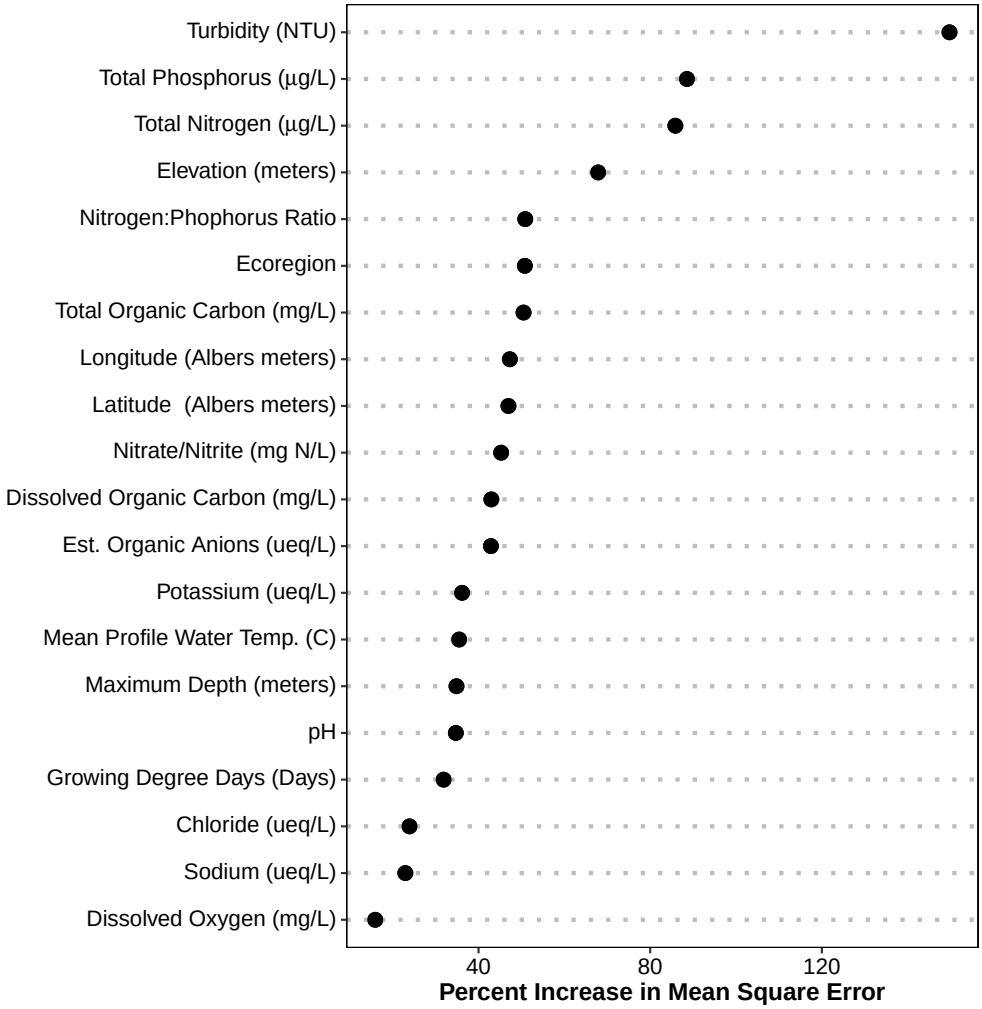

Figure 3: Random Forest model's output for POLR model predictors. Importance plot for all variables. Shows percent increase in mean squared error. Higher values of percent increase in mean squared error indicates higher importance.

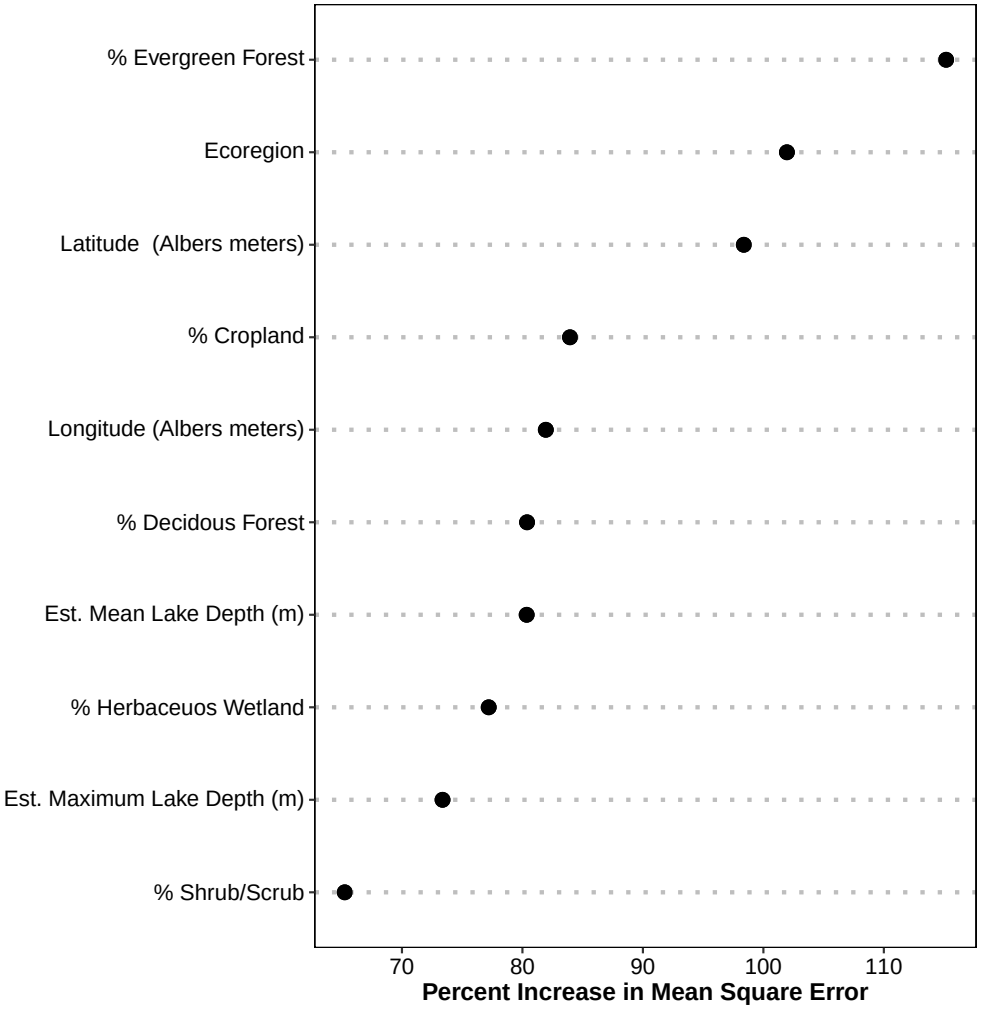

Figure 4: Random Forest model's output for nitrogen predictors. Importance plot for GIS variables. Shows percent increase in mean squared error. Higher values of percent increase in mean squared error indicates higher importance.

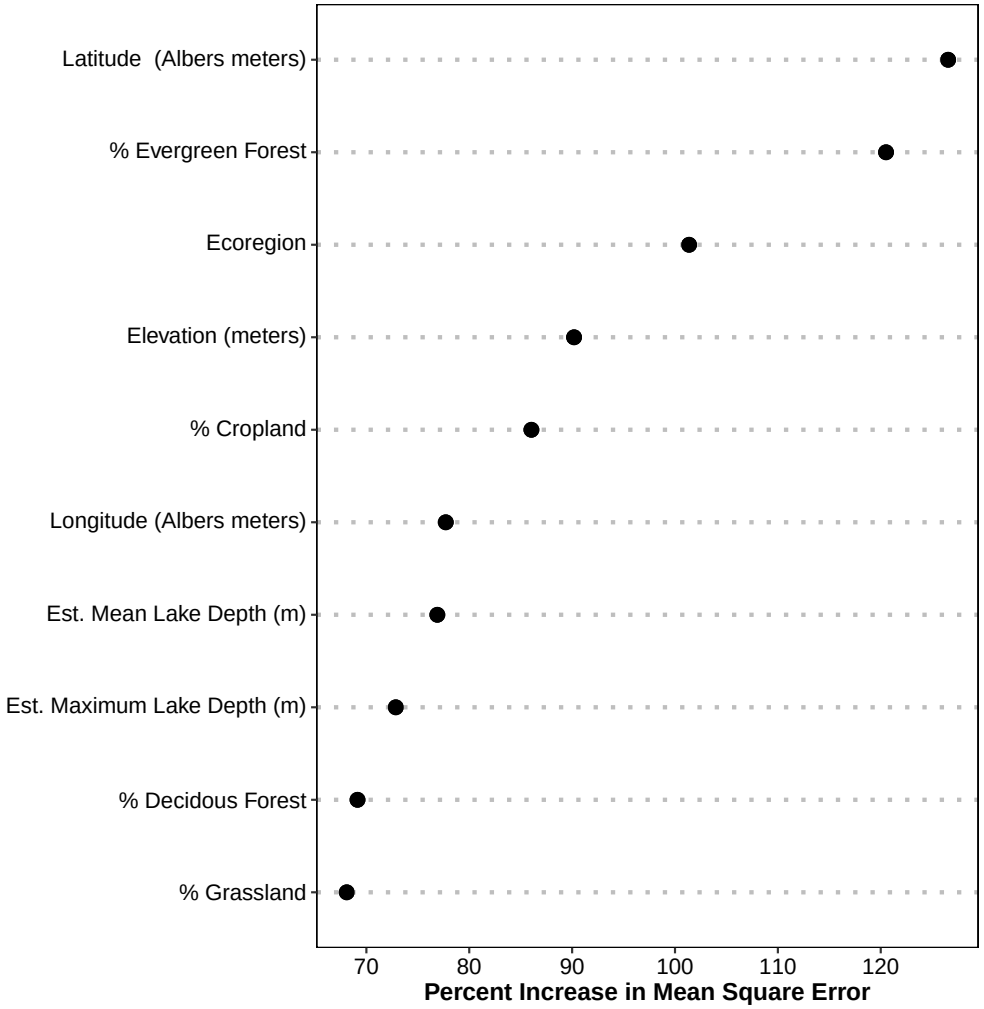

Figure 5: Random Forest model's output for phosphorus predictors. Importance plot for GIS variables. Shows percent increase in mean squared error. Higher values of percent increase in mean squared error indicates higher importance.

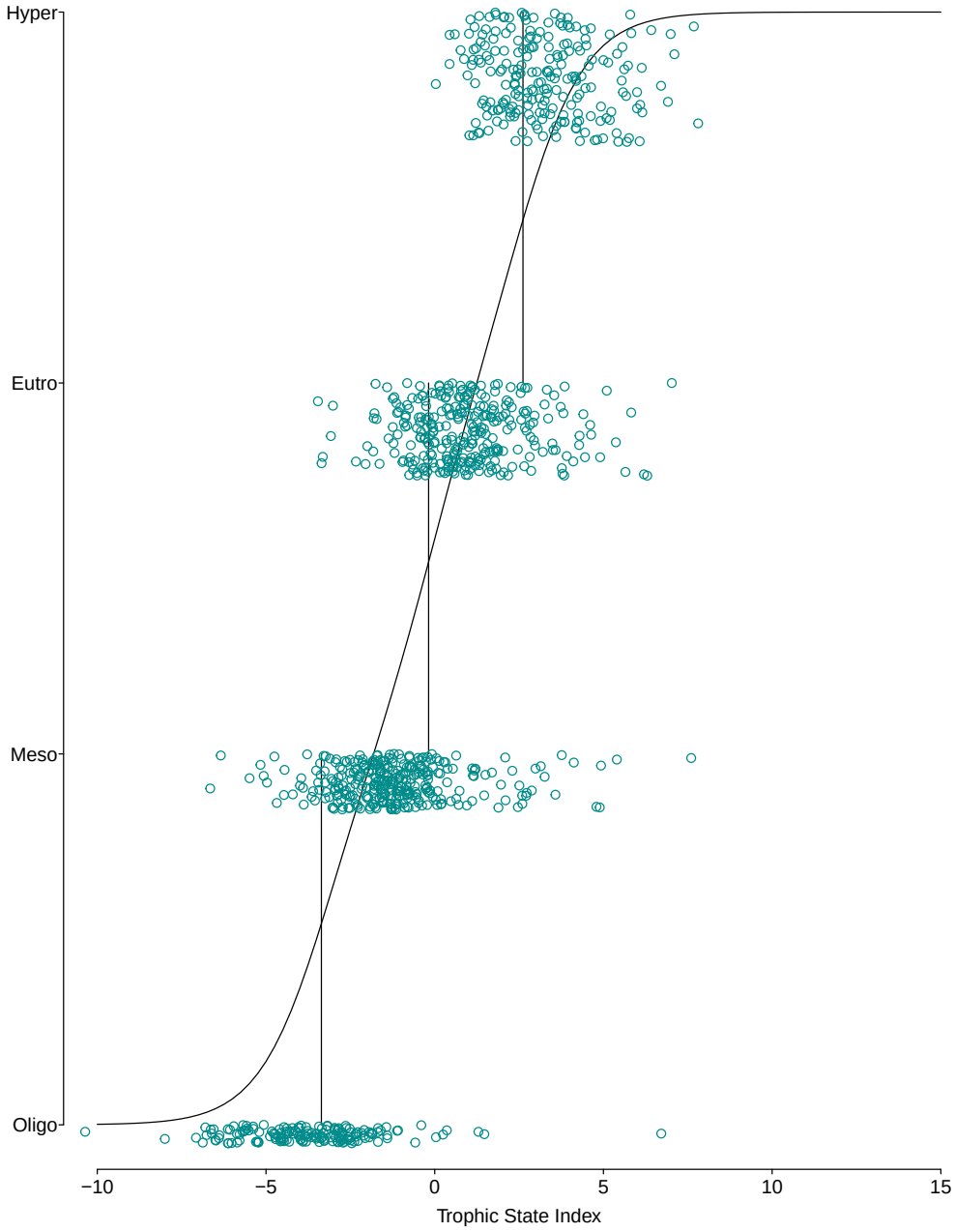

Figure 6: Graphical presentation of the POLR model. The x-axis is the trophic state index, the y-axis is each lake's trophic state, vertical lines show estimated cutpoints, and curve shows expected trophic state as estimated using ordered logistic regression.

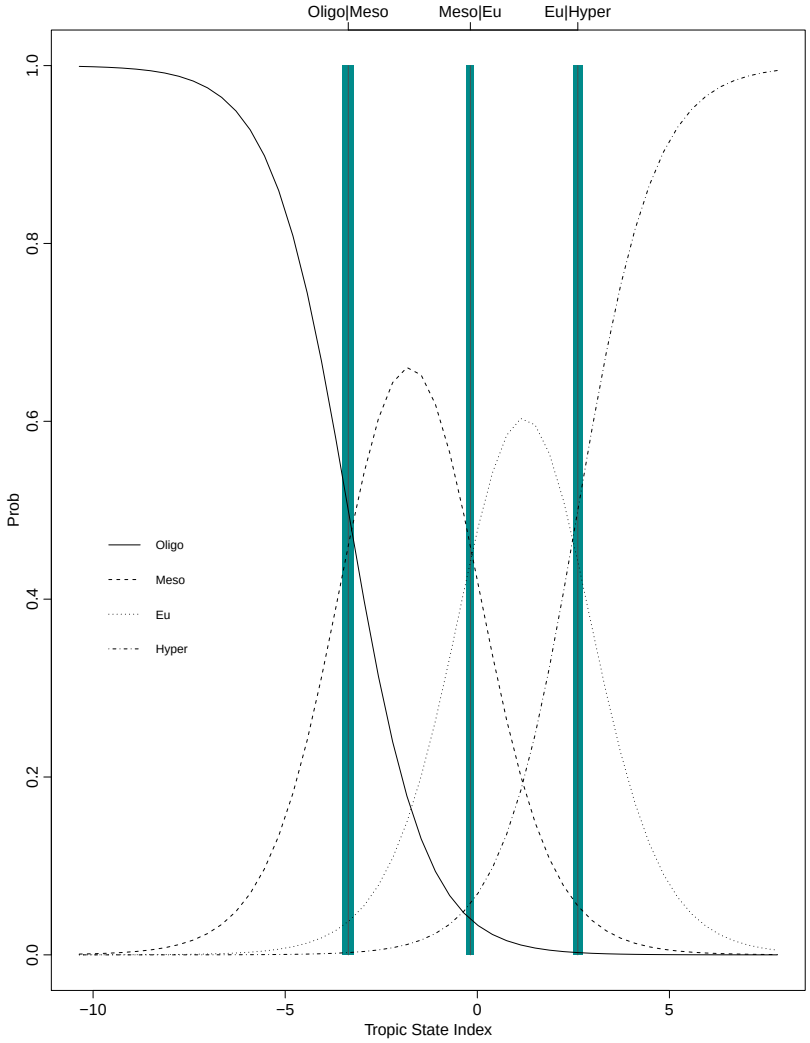

Figure 7: Graphical presentation of the POLR model. The x-axis is the trophic state index, the y-axis is the probability of being classified into one of the 4 trophic state classes, and the vertical lines and blue bars are the cutpoints ± one standard error.

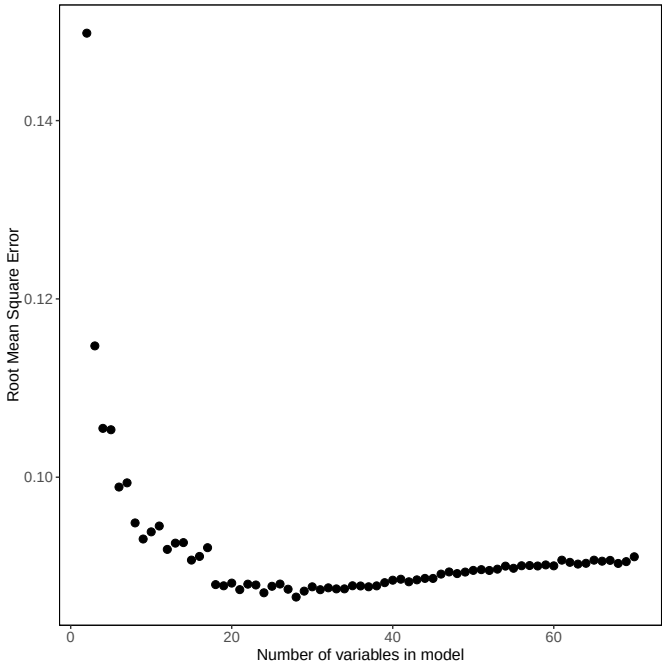

Figure S1: Random Forest model's output for POLR model. Shows percent increase in mean squared error as a function of the number of variables.

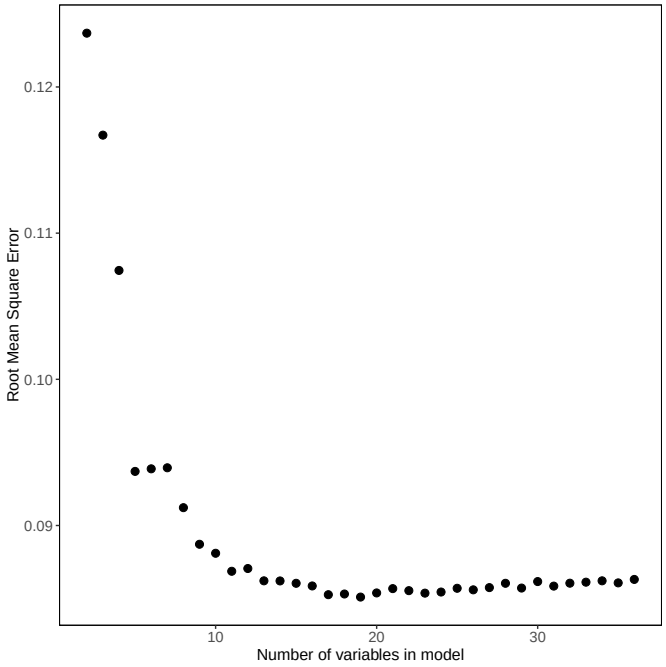

Figure S2: Random Forest model's output for nitrogen with GIS only variables as predictors. Shows percent increase in mean squared error as a function of the number of variables.

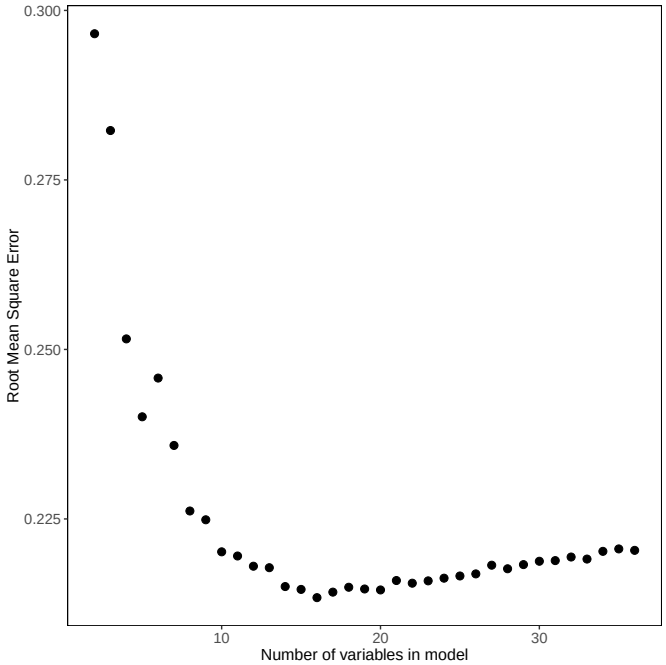

Figure S3: Random Forest model's output for phosphorus with GIS only variables as predictors. Shows percent increase in mean squared error as a function of the number of variables.

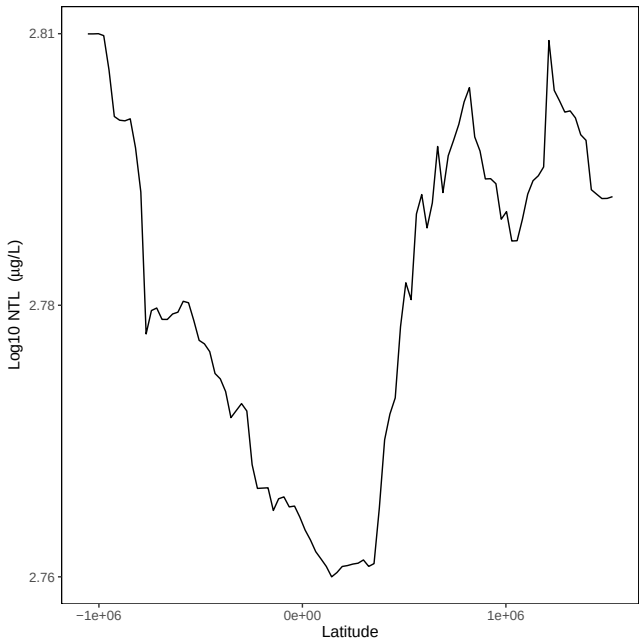

Figure S4: Partial dependency

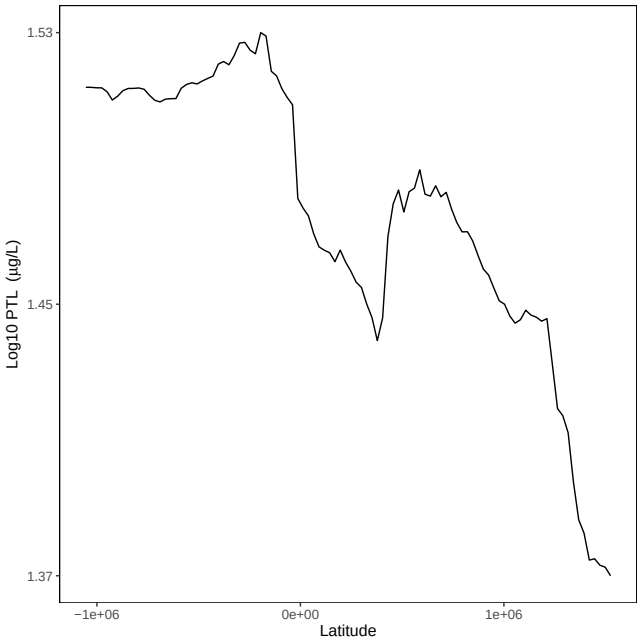

Figure S5: Partial dependency