A peer-reviewed version of this preprint was published in PeerJ on 12 November 2019.

View the peer-reviewed version (peerj.com/articles/7936), which is the preferred citable publication unless you specifically need to cite this preprint.

Nojavan A. F, Kreakie BJ, Hollister JW, Qian SS. 2019. Rethinking the lake trophic state index. PeerJ 7:e7936 https://doi.org/10.7717/peerj.7936

PeerJ Preprints NOT PEER-REVIEWED

Rethinking the lake Trophic State Index

Farnaz Nojavan A. Corresp., [1] , **Betty J Kreakie** [2] , **Jeffrey W Hollister** [2] , **Song S Qian** [3]

[1] ORISE, Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, Narragansett, RI, United States
[2] Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, Narragansett, RI, United States
[3] Department of Environmental Sciences, The University of Toledo, Toledo, OH, United States

Corresponding Author: Farnaz Nojavan A.
Email address: nojavan-asghari.farnaz@epa.gov

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 based (vs. a complex aggregate of multi-variables), discrete (vs. a continuum), and/or deterministic (vs. an inherent randomness). We present an updated lake trophic classification model using a Bayesian multilevel ordered categorical regression. The model consists of a proportional odds logistic regression (POLR) that models ordered, categorical, lake trophic state using Secchi disk depth, elevation, nitrogen concentration (N), and phosphorus concentration (P). The overall accuracy, when compared to existing classifications of trophic state index (TSI), for the POLR model was 0.68 and the balanced accuracy ranged between 0.72 and 0.93. This work delivers an index that is multi-variable based, 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 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.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

Rethinking the Lake Trophic State Index

Farnaz Nojavan A.[1], **Betty J. Kreakie**[2], **Jeffrey W. Hollister**[2], **and Song S. Qian**[3]

[1]**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.**

[2]**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.**

[3]**Department of Environmental Sciences, The University of Toledo, Toledo, OH 43606 U.S.A.**

Corresponding author:
Farnaz Nojavan A.[1]

Email address: nojavan-asghari.farnaz@epa.gov

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 based (vs. a complex aggregate of multi-variables), discrete (vs. a continuum), and/or deterministic (vs. an inherent randomness). We present an updated lake trophic classification model using a Bayesian multilevel ordered categorical regression. The model consists of a proportional odds logistic regression (POLR) that models ordered, categorical, lake trophic state using Secchi disk depth, elevation, nitrogen concentration (N), and phosphorus concentration (P). The overall accuracy, when compared to existing classifications of trophic state index (TSI), for the POLR model was 0.68 and the balanced accuracy ranged between 0.72 and 0.93. This work delivers an index that is multi-variable based, 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 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.

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). Recreation, habitat and species diversity, property and ecological values are closely related to lake water quality (Keeler et al., 2015; Leggett and Bockstael, 2000). Hence, monitoring water quality is integral to the management of the eutrophication and productivity of lakes. 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.

Given its broad applicability and long history, it is important to periodically review and update the methods used to calculate trophic state. 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 inter-regional variation in lake production. However, current lake trophic state models are one size fits all. Trophic state has been formulated using various indices, the most well known was created by Carlson (1977).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

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)).

Multiparameter index approaches view trophic state as a complex response caused by interaction among various physical, chemical, and biological factors. These approaches use relevant 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). Also, the definition of trophic state should be differentiated from its predictors (Carlson and Simpson, 1996). In other words, the trophic state is based on the biological condition of a lake. The goal of developing a trophic state indicator should be to link a lake's trophic status to the main causes of eutrophication.

Classification procedures 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 the aforementioned critiques by developing a Bayesian ordered categorical regression model to classify lake trophic state. The proposed model builds upon the existing trophic status classification as a starting point and reassesses the trophic state index development and classification methods; hence, "rethinks" the lake trophic state classification and index. The model contributes to literature on trophic state in several ways. First, it generates an index that is multi-variable by using Secchi depth, 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. Fourth, while it is critical to locate a lake across trophic continuum, it is not economically feasible to monitor all lakes by conventional sampling techniques. We extend the developed POLR model by linking easily accessible and universally available GIS variables to nitrogen and phosphorus; hence, allowing prediction of the trophic state of all lakes, even not extensively sampled ones. We present this extended application as one possible use case of the POLR model.

MATERIAL AND METHODS

Data and Study Area

We used data from the United States Environmental Protection Agency's 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; Homer et al., 2004; Xian et al., 2009; Hollister and Milstead, 2010; Hollister et al., 2011; Hollister, 2014; Hollister and Stachelek, 2017). 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, resulting in over 1000 lakes sampled across the continental United States during the summer of 2007 (Figure 1). The source code for data pre-processing and the resultant data are available on GitHub repository `https://github.com/usepa/rethinking_tsi` (Nojavan A. et al., 2017).

Statistical Methods

We developed a proportional odds logistic regression (POLR) to predict lake trophic state using Secchi disk depth, elevation, nitrogen concentration (N), and phosphorus concentration (P). The predictors in the POLR model were selected from in situ and universally available GIS variables using random forest models. We then present an extended application of the developed POLR model using a Bayesian multilevel model. Our modeling work flow was as follows:

1. Variable selection using Random Forest Model: Develop a random forest model, using R's

`randomForest` package (Liaw and Wiener, 2002a), with 5000 trees using all variables (in situ and universally available GIS variables) to identify the best predictor variables for lake trophic state.

2. Develop the POLR model using R function bayespolr from package `arm` (Gelman et al., 2013) and the outputs from previous step.
3. Assess the performance of the POLR model using a hold-out validation method (90% training set, 10% evaluation set).
4. Extended Application
   - Develop a random forest model, using R's `randomForest` package, with 5000 trees using only GIS variables to identify the best predictor variables for nitrogen and phosphorus.
   - Develop the extended application model (the Bayesian multilevel model) using R's rjags package to run Just Another Gibbs Sampler (JAGS) from inside of R.
   - Assess the performance of the extended application model using a hold-out validation method (90% training set, 10% evaluation set).

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, 2002b; R Core Team, 2016). We developed random forest models to select predictor variables to model trophic state. 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). We used percent increase in mean squared error to examine variable importance. We selected the variables that were above 0.1 percent increase in mean squared error.

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 (discussed in the following section) 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.

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). Further, the categories are ordered across the trophic continuum. The proportional odds logistic regression (POLR) model, a generalized linear modeling technique, has been used to account for the ordered categories of the response variable (Gelman and Hill, 2006).

The ordered categorical response variable, lake trophic status, can be described with a series of logistic regressions in its simplest form as follows:

$$\begin{cases} Pr(\text{lake trophic status} > oligotrophic) = logit^{-1}(\text{trophic state index}) \\ Pr(\text{lake trophic status} > mesotrophic) = logit^{-1}(\text{trophic state index} - \text{cutpoint } 1) \\ Pr(\text{lake trophic status} > eutrophic) = logit^{-1}(\text{trophic state index} - \text{cutpoint } 2) \\ Pr(\text{lake trophic status} > hypereutrophic) = logit^{-1}(\text{trophic state index} - \text{cutpoint } 3) \end{cases} \tag{1}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

The probability of a lake's trophic status being, for example eutrophic, is calculated by $Pr(\text{lake trophic status} > \text{mesotrophic}) - Pr(\text{lake trophic status} > \text{eutrophic})$. The trophic status is eutrophic when the $Pr(\text{lake trophic status} > \text{eutrophic})$ is the highest in comparison to the probability of other trophic categories, which happens when $c_{Meso|Eu} <$ trophic state index $< c_{Eu|Hyper}$.

Figure 2 depicts all the elements of the POLR model. Secchi disk depth (SDD), elevation, nitrogen, and phosphorus are the four predictors of the trophic state index. Associated with each predictor is a coefficient α. There are three cutpoints or thresholds for the four categories of the response variable. The uncertainty of the trophic state index is shown in figure 2 by τ^2. Mathematically, the POLR model was set up as follows:

$$y_i = \begin{cases} Oligotrophic & \text{if } z_i < c_{Oligo|Meso} \\ Mesotrophic & \text{if } z_i \in (c_{Oligo|Meso}, c_{Meso|Eu}) \\ Eutrophic & \text{if } z_i \in (c_{Meso|Eu}, c_{Eu|Hyper}) \\ Hypereutrophic & \text{if } z_i > c_{Eu|Hyper} \end{cases} \tag{2}$$

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

where trophic state index (XA) is equal to Secchi Disk Depth$_i$ $\times$ α_{SDD} $+$ Phosphorus$_i$ $\times$ $\alpha_{Phosphorus}$ $+$ Nitrogen$_i$ $\times$ $\alpha_{Nitrogen}$ $+$ Elevation$_i$ $\times$ $\alpha_{Elevation}$; with the coefficients A: $\{\alpha_{SDD}, \alpha_{phospurous}, \alpha_{Nitrogen}, \alpha_{Elevation}, \}$ and $c_{k|k+1}$ (known as cutpoints or thresholds), the design matrix of predictors X, and scale parameter of τ^2 . The two adjacent cutpoints and XA are used to classify the response variable. The cutpoints and coefficients are estimated simultaneously using maximum likelihood.

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 used a hold-out validation method where we divided the data into two subsets: a training set, used to develop the predictive model, and a validation set, used to assess the performance of the developed model. However, we avoided deviations in our evaluation data by only using 10% of the consistently classified lakes across the three trophic state classification methods as our validation set. We developed the model using the rest of the data. This is similar to the concept of "posterior predictive model checking" described by Gelman et al. (2014), where the model predictions are being compared to the observed data looking for any discrepancies. We decided to use validation as opposed to validating the model with a new data set as a comparable dataset was not available during the model development process. 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.

RESULTS

Variable Selection: Random Forest

The random forest models provided estimates of variable importance for trophic state and the results are reported in figure 4. The number of variables for each response variable was decided using the variable selection plots (Figure **??**) 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, as it is cheaper to measure and readily available for most lakes.

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:

$$y_i = \begin{cases} Oligotrophic & \text{if } z_i < -3.36 \\ Mesotrophic & \text{if } z_i \in (-3.36, -0.18) \\ Eutrophic & \text{if } z_i \in (-0.18, 2.62) \\ Hypereutrophic & \text{if } z_i > 2.62 \end{cases} \tag{3}$$

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

The resulting POLR model has three cutpoints and four slope coefficients (Table 1). Figures 7 and 8 summarize the model uncertainty. The POLR model returns four probabilities associated with each trophic state as opposed to one fixed classification (Figure 8). The off-diagonal elements in table 1 are the mis-classified lakes. There are management implications with misclassification of lakes. However, unlike previous classification schemes, the proposed model keeps the continuous index as well as the discretized classes. Further, each class has a probability. For management applications one needs to pay attention to where the lake is along the continuum (continuous index) and how confident we are about the assigned classification (uncertainty quantification). Figures 7 and 8 illustrate that the proposed trophic index and classification method are now a continuum with quantified probability, hence suggesting a modified eutrophication scale that captures the inherit variability of eutrophication.

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.

EXTENDED APPLICATION

The drawback of the POLR model is the cost of monitoring multiple predictor variables (e.g., nutrients). This is addressed in our extended application by linking nitrogen and phosphorus to universally available GIS variables. We link nitrogen and phosphorus in the POLR model to a separate nutrient model built from universally available GIS data. Thereby, avoiding the need for nitrogen and phosphorus data, costly variables to measure for all lakes. Figure 3 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 nutrient model, in the upper block, estimates the means of nitrogen and phosphorus based on ecoregion, % evergreen forest, and latitude. The predictor variables for the second level models were selected with a random forest modeling approach (see subsection on Variable Selection). The two blocks are connected through the estimated means of nitrogen ($\mu_{Nitrogen}$) and phosphorus ($\mu_{Phosphorus}$) to form the combined model which enables trophic state classification for all lakes without the costly sampling requirement. 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 extended application model and its results are described in detail in the supplementary material.

Random forest modeling, see subsection on variable selection, selected latitude, eco-region, and percent evergreen forest 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 **??** & **??**) depicts high concentrations in the northern and southern extremes of the continental US. The lowest predicted concentrations correspond to the mid-latitudes. The ecoregion variable represents an intermediate scale among these three variables and represents the variation between the regions. Finally, 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. 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.

As mentioned, the extension of our POLR model uses eco-region, latitude, and watershed level percent evergreen forest as predictors for nitrogen and phosphorus. This contrasts with prior trophic state classification models that are applied to all lakes, regardless of the differences across scale. Lake trophic index, and hence lake trophic classes, should be calculated differently in different eco-regions to accommodate variation in landform and climate characteristics and our proposed model and extension bares this out by identifying and including and eco-regional approach to quantifying trophic state. Furthermore, the developed multilevel model structure can be further expanded to lake-specific trophic state index, upon availability of multiple measurements for each lake.

DISCUSSION

Modeling Approach

The modeling approach presented here uses a Bayesian ordered categorical regression model (i.e. the POLR model). The benefits of this approach are that it uses multiple variables to predict lake trophic state and creates a continuous trophic index. A multi-variable predictor model accounts for chemical, biological, and physical aspects of trophic state and quantifies lake trophic state across a continuum. This is important because lake trophic state is a variable that changes gradually across a gradient, yet 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. Additionally, the proposed model quantifies the uncertainty of lake trophic response to changes in nutrients, as the response varies from lake to lake. Lastly, the lake trophic index may also be presented as a classification (e.g., oligotrophic, mesotrophic, etc.) which facilitates organization and communication.

Management Implications

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 undisturbed conditions. It is critical to quantitatively plan and assess the recovery of lakes in restoration projects. Our model has potential as a tool for nutrient management scenario analysis as we can quantify how altering nutrients can move 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 Updating and Model Accuracy

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. The presented model quantifies lake trophic state and the uncertainty around it. The trophic state quantification can help in assessing lake ecological state before and after restoration. Additionally, a key symptom of eutrophication is cyanobacteria dominance in lakes (Conley et al., 2009; Hollister and Kreakie, 2016; Przytulska et al., 2017). 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.

The POLR model has an overall accuracy of 0.68, 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 7 and 8. 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 trophic state classifications. Although we are presenting a novel method, the results are consistent with our intuitive and historical understanding of lake trophic state.

ACKNOWLEDGMENTS

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.

PeerJ Preprints NOT PEER-REVIEWED

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 coordinator's guide to volunteer lake monitoring methods. North American Lake Management Society, 96:305.

Conley, D. J., Paerl, H. W., Howarth, R. W., Boesch, D. F., Seitzinger, S. P., Havens, K. E., Lancelot, C., Likens, G. E., et al. (2009). Controlling eutrophication: nitrogen and phosphorus. Science, 323(5917):1014–1015.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). Bayesian data analysis, volume 2. CRC press Boca Raton, FL.

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.

Gelman, A., Su, Y., Yajima, M., Su, M., and Matrix, I. (2013). Package 'arm': Data analysis using regression and multilevel/hierarchical models. R package version, pages 1–6.

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. and Stachelek, J. (2017). lakemorpho: Calculating lake morphometry metrics in r. F1000Research, 6.

Hollister, J. W. and Kreakie, B. J. (2016). Associations between chlorophyll a and various microcystin health advisory concentrations. F1000Research, 5.

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 landcover 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. (2002a). Classification and regression by randomforest. R News, 2(3):18–22.

Liaw, A. and Wiener, M. (2002b). 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.

Przytulska, A., Bartosiewicz, M., and Vincent, W. F. (2017). Increased risk of cyanobacterial blooms in northern high-latitude lakes through climate warming and phosphorus enrichment. Freshwater Biology, 62(12):1986–1996.

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 nation's 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 nation's 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.

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.

PeerJ Preprints NOT PEER-REVIEWED

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	$C_{Oligo\|Meso}$	-3.36	0.15
	$C_{Meso\|Eu}$	-0.18	0.09
	$C_{Eu\|Hyper}$	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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

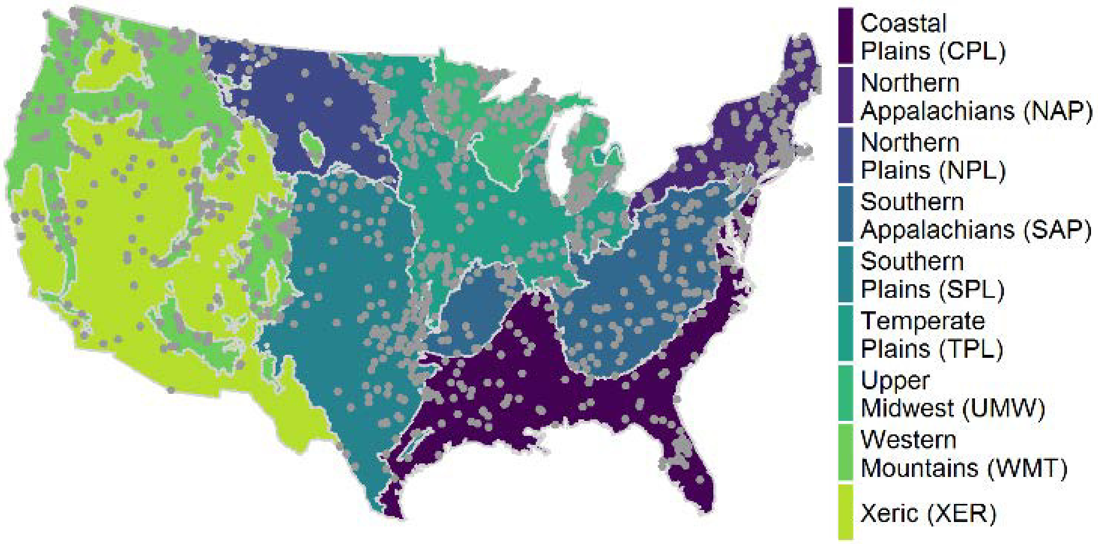

Figure 1. Map of the distribution of National Lakes Assessment sampling locations. Also Wadeable Stream Assessment (WSA) ecoregions are depicted in the map. Areas in an ecoregion have similar landform and climate characteristics.

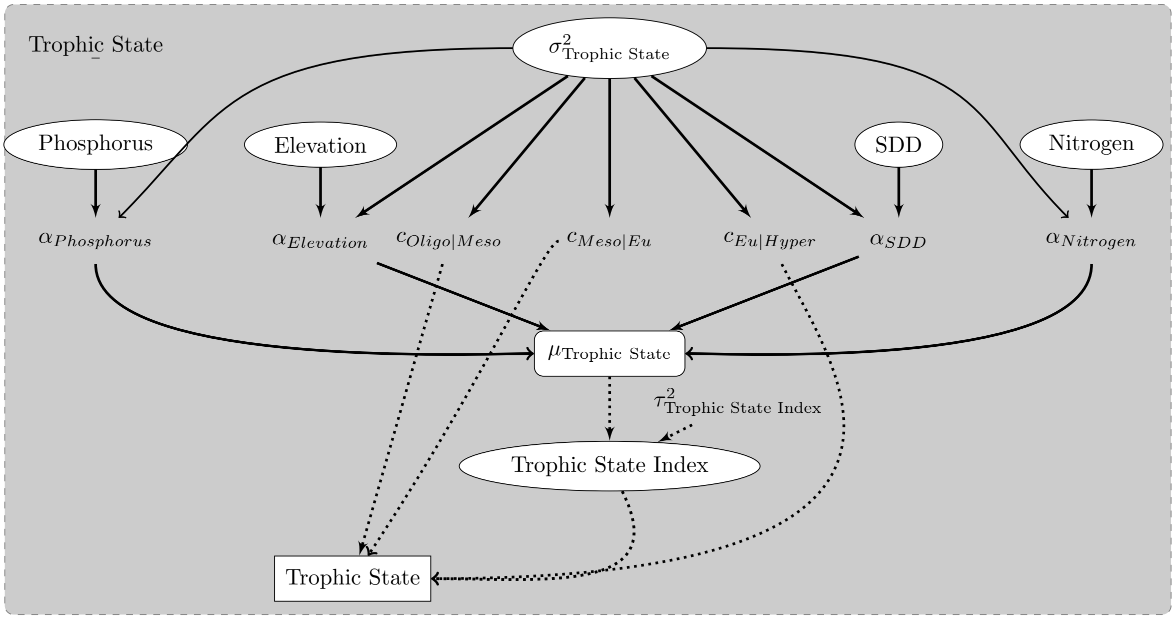

Figure 2. Directed Acyclic Graphical (DAG) model. The figure depicts the developed POLR model with its four predictors of secchi disk depth (SDD), elevation, nitrogen, and phosphorus.

PeerJ Preprints NOT PEER-REVIEWED

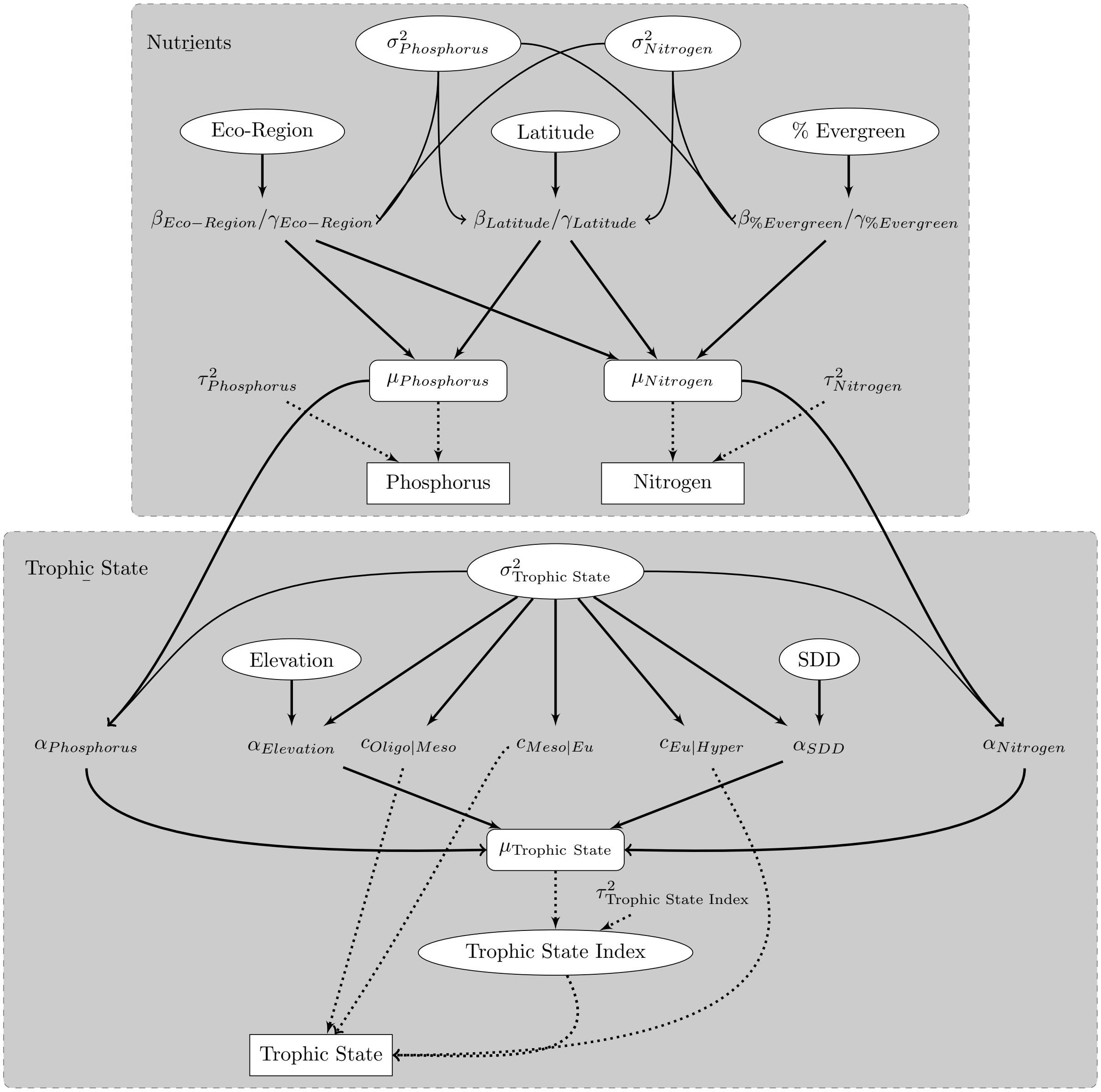

Figure 3. Directed Acyclic Graphical (DAG) model. The lower box depicts the POLR model with its four predictors of secchi disk depth (SDD), elevation, nitrogen, and phosphorus. The upper box is the extension to the POLR model to predict nitrogen and phosphorus using universally available GIS variables.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

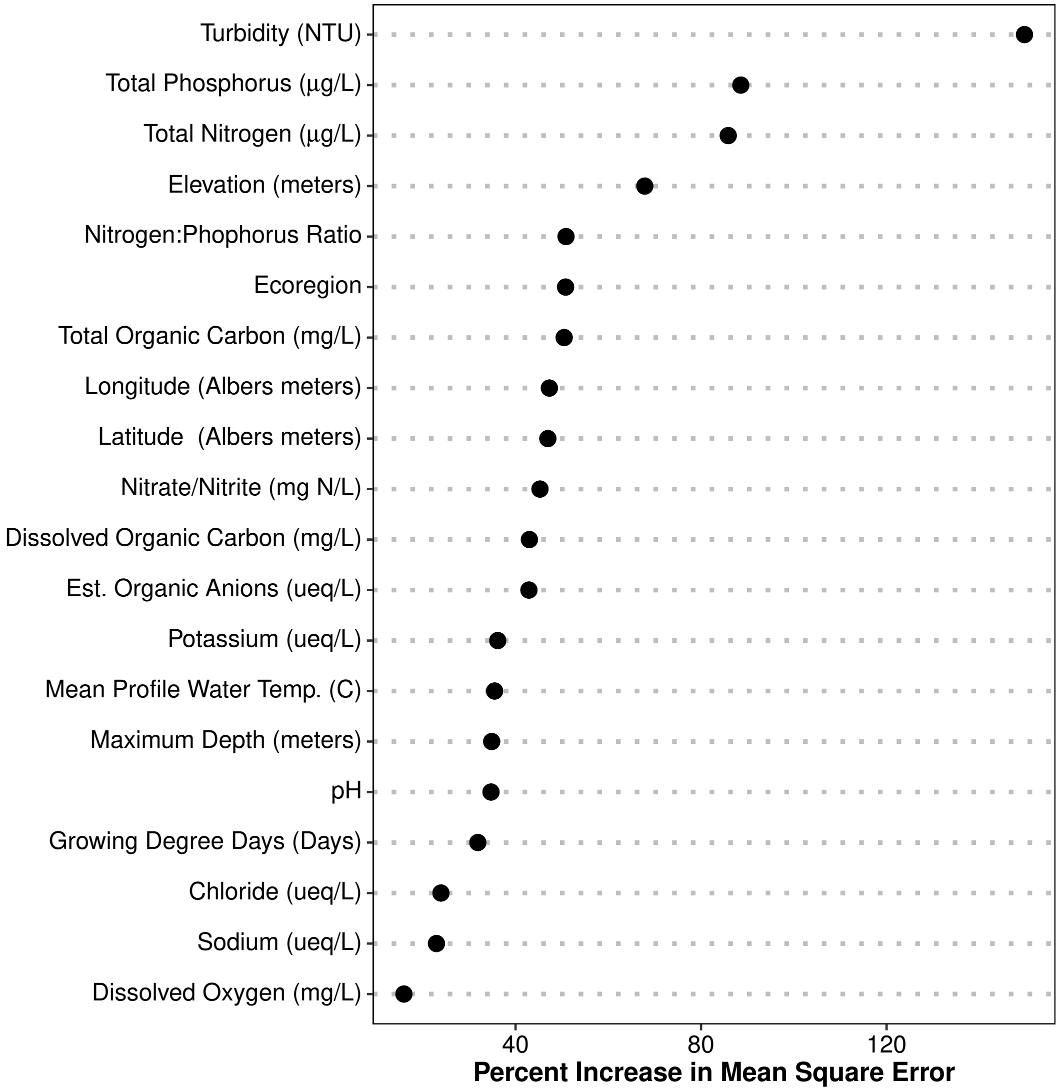

Figure 4. Random Forest model's output for POLR model predictors. Importance plot for all variables. Shows percent increase in mean squared error. Higher values of percent increase in mean squared error indicates higher importance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

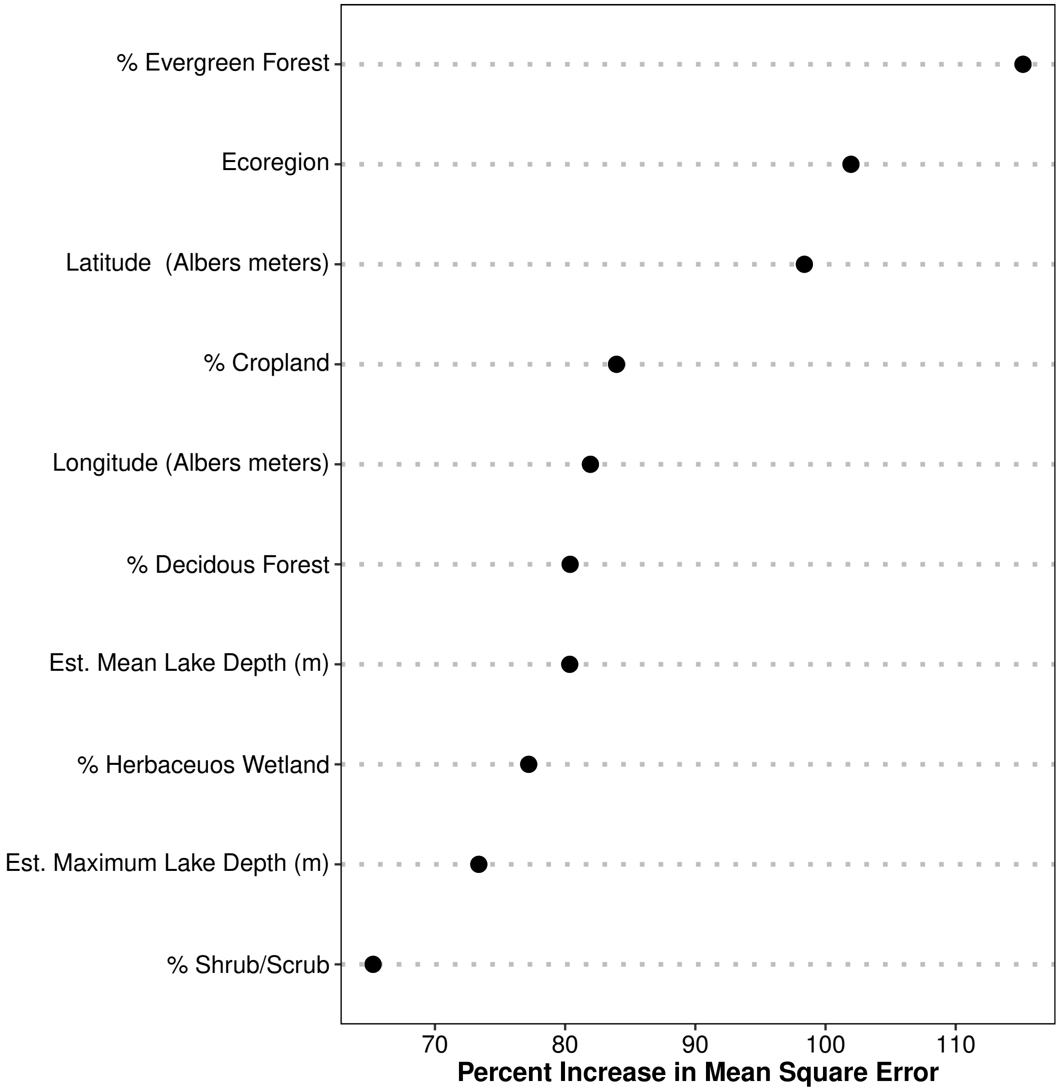

Figure 5. Random Forest model's output for nitrogen predictors. Importance plot for GIS variables. Shows percent increase in mean squared error. Higher values of percent increase in mean squared error indicates higher importance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

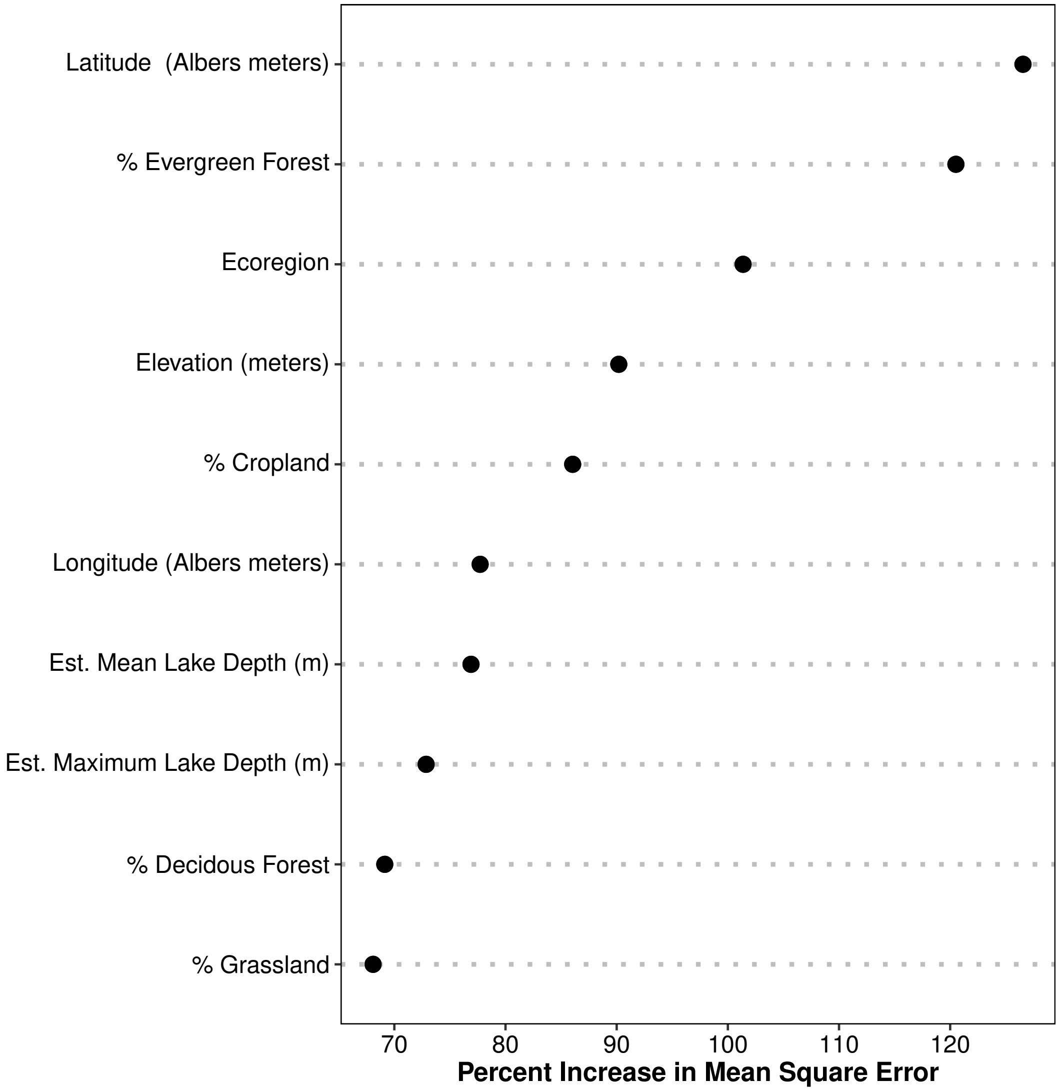

Figure 6. Random Forest model's output for phosphorus predictors. Importance plot for GIS variables. Shows percent increase in mean squared error. Higher values of percent increase in mean squared error indicates higher importance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

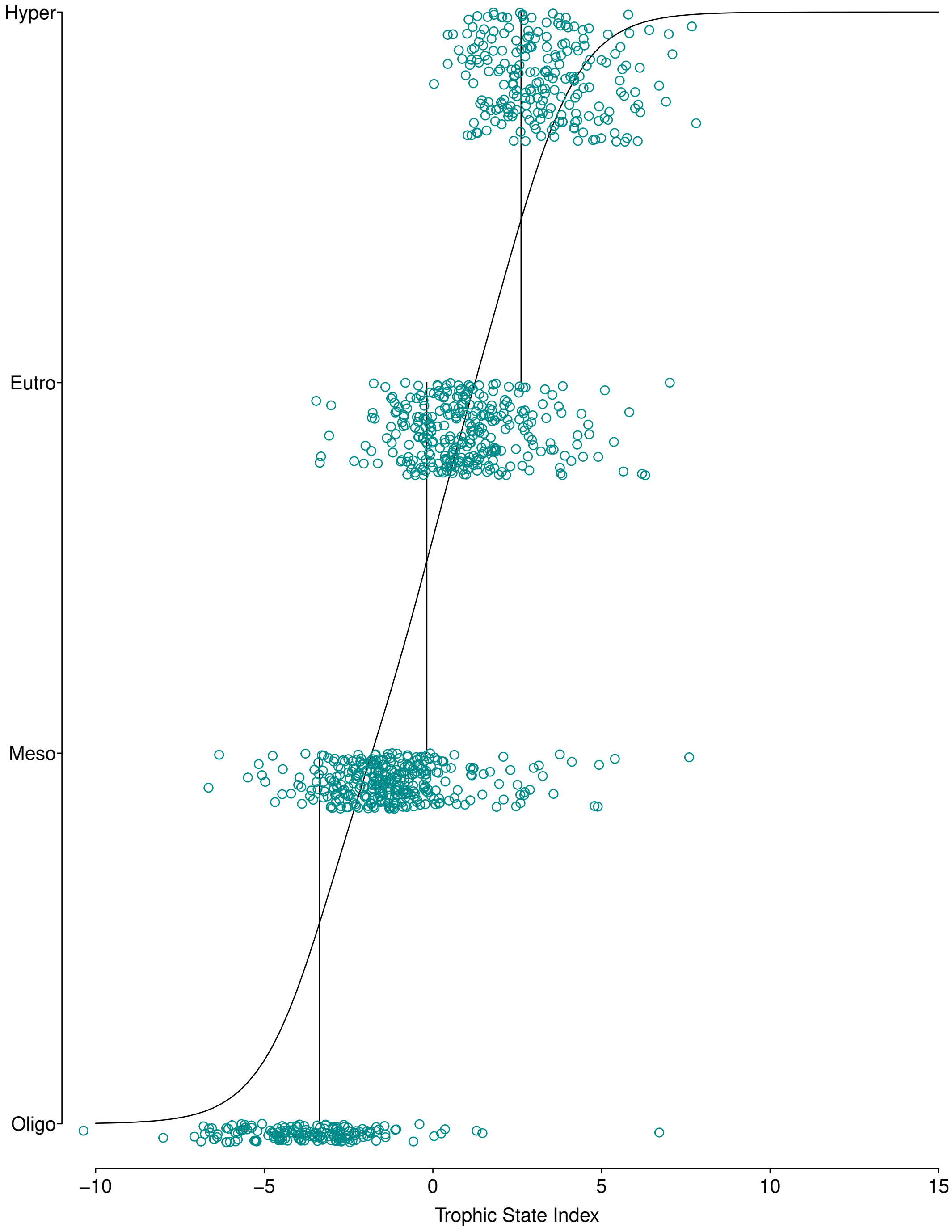

Figure 7. Graphical presentation of the POLR model. The x-axis is the trophic state index, the y-axis is each lake's trophic state, vertical lines show estimated cutpoints, and curve shows expected trophic state as estimated using ordered logistic regression.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

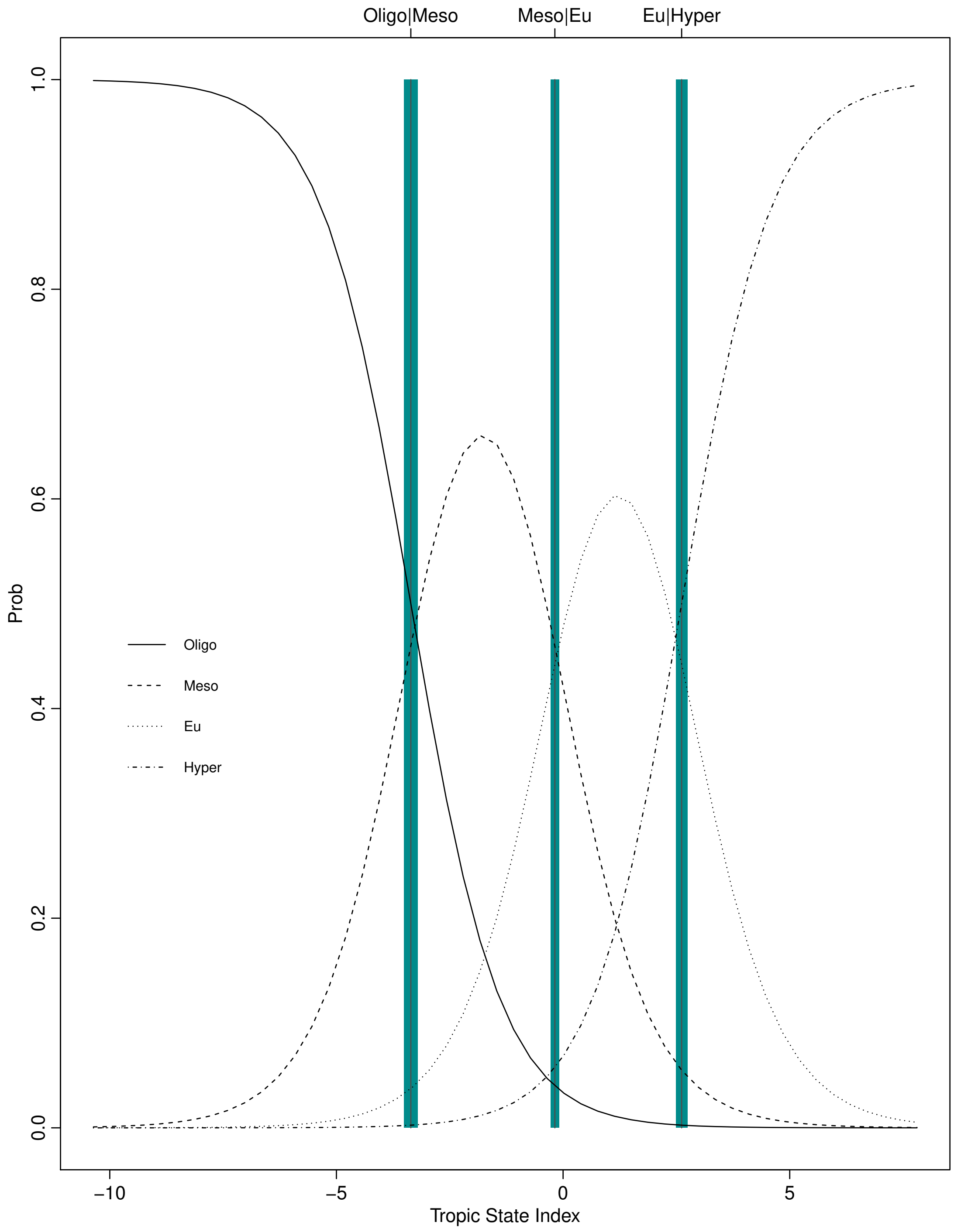

Figure 8. Graphical presentation of the POLR model. The x-axis is the trophic state index, the y-axis is the probability of being classified into one of the 4 trophic state classes, and the vertical lines and blue bars are the cutpoints ± one standard error.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

Supplementary Material for "Rethinking the Lake Trophic State Index"

The models were set up as follows:

$$\text{Nitrogen}_{\text{ij}} \sim \mathcal{N}(\mu_{Nitrogen_{ij}}, \sigma^2_{Nitrogen}) \tag{S1}$$

where $\mu_{Nitrogen_{ij}} = X_{Nitrogen}B$, $X_{Nitrogen}$ is the matrix of predictors, and B is the vector of coefficients. $Nitrogen_{ij}$ is the ith nitrogen observation in the jth ecoregion.

$$\text{Phosphorus}_{\text{ij}} \sim \mathcal{N}(\mu_{Phosphorus_{ij}}, \sigma^2_{Phosphorus}) \tag{S2}$$

where $\mu_{Phosphorus_{ij}} = X_{Phosphorus}\Gamma$, $X_{Phosphorus}$ is the matrix of predictors, and Γ is the vector of coefficients. $Phosphorus_{ij}$ is the ith phosphorus observation in the jth ecoregion.

The overall accuracy of the extension 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 S1). Table S2 shows the confusion matrix for the POLR model.

The extension model calculates lake trophic index and classes differently for different eco-regions. Please refer to Table S1 for varying coefficients in different eco-regions. For example, eco-regions 3, 6, and 5, corresponding to Northern Plains, Temperate Plains, and Southern Plains, have the highest positive coefficients for nitrogen. Hence, nitrogen plays a significant role in moving the trophic state index and class toward the eutrophic/hypereutrophic side of the trophic continuum. As another example, in eco-regions 3, 6, and 5, corresponding to Northern Plains, Temperate Plains, and Southern Plains phosphorus plays a significant role in moving the trophic state index and class toward the eutrophic/hypereutrophic side of the trophic continuum. Further Table S1 shows the coefficients for latitude and percent evergreen. We included these predictors even though they were not statistically significant. Based on the discussions in gelman2007data it is generally fine to keep a statistically insignificant predictor with the correct sign in. It may not help predictions dramatically but is also probably not hurting them.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

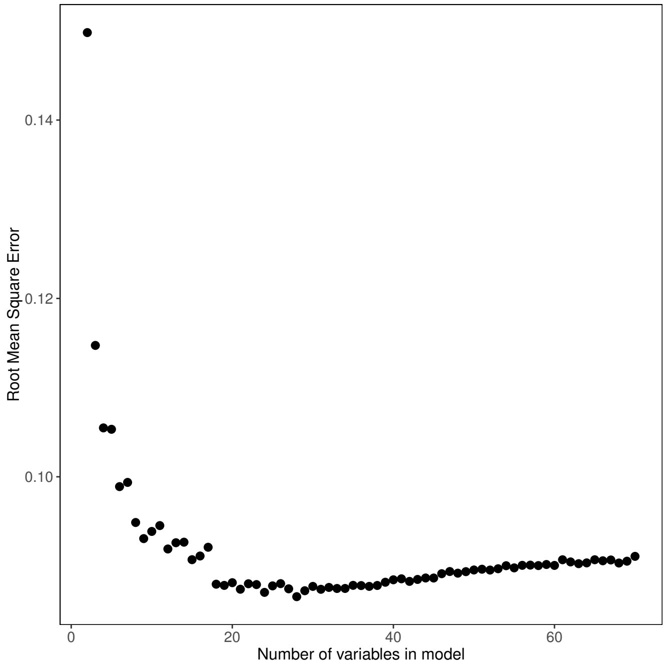

Figure S1: Random Forest model's output for POLR model. Shows percent increase in mean squared error as a function of the number of variables.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

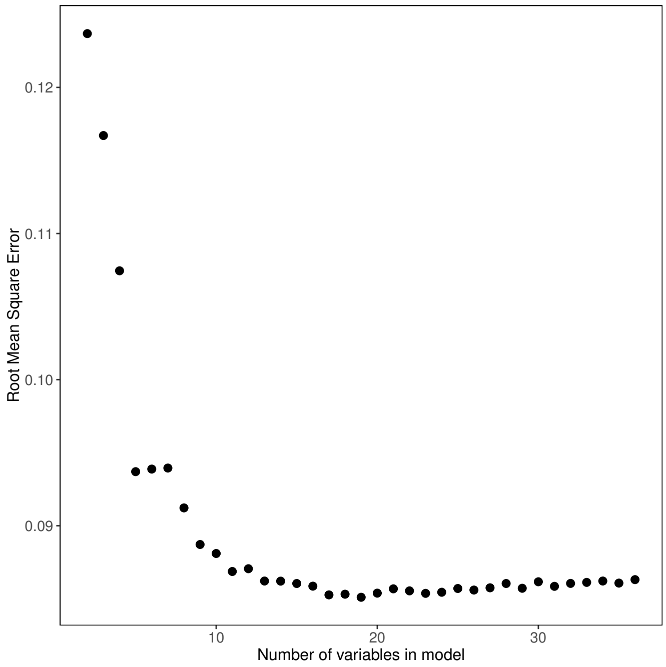

Figure S2: Random Forest model's output for nitrogen with GIS only variables as predictors. Shows percent increase in mean squared error as a function of the number of variables.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

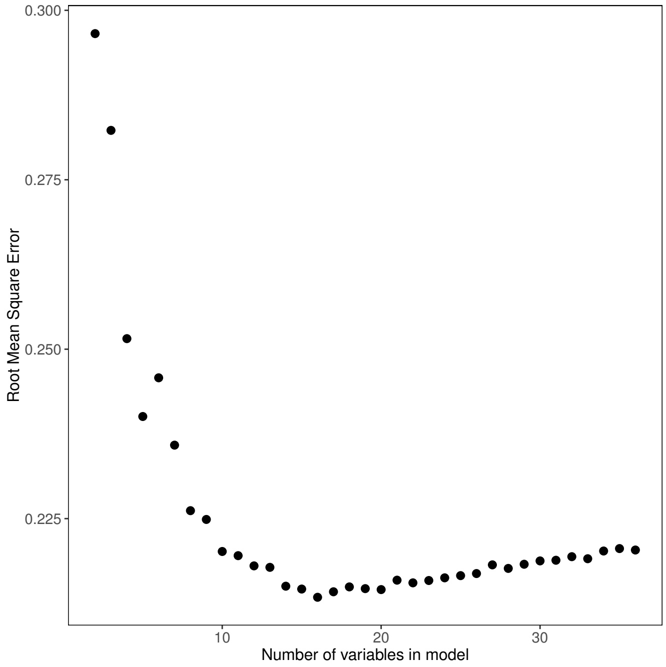

Figure S3: Random Forest model's output for phosphorus with GIS only variables as predictors. Shows percent increase in mean squared error as a function of the number of variables.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

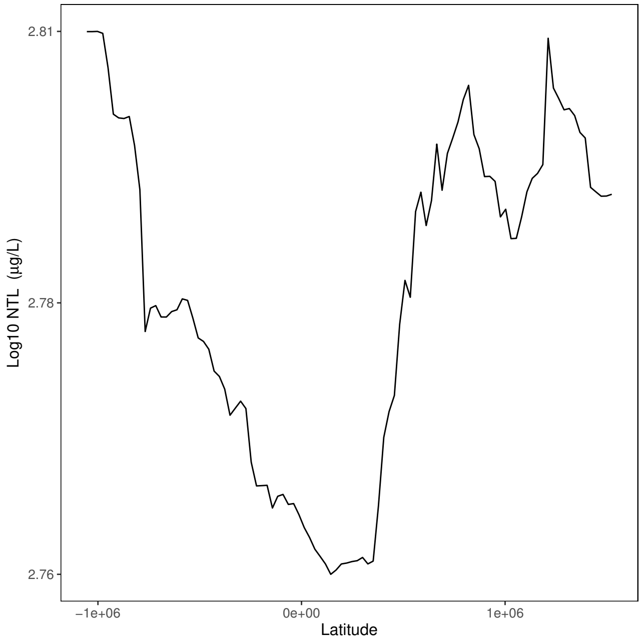

Figure S4: Partial dependency plot for total nitrogen versus latitude: the effect of latitude on total nitrogen when the rest of the predictors are held constant.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

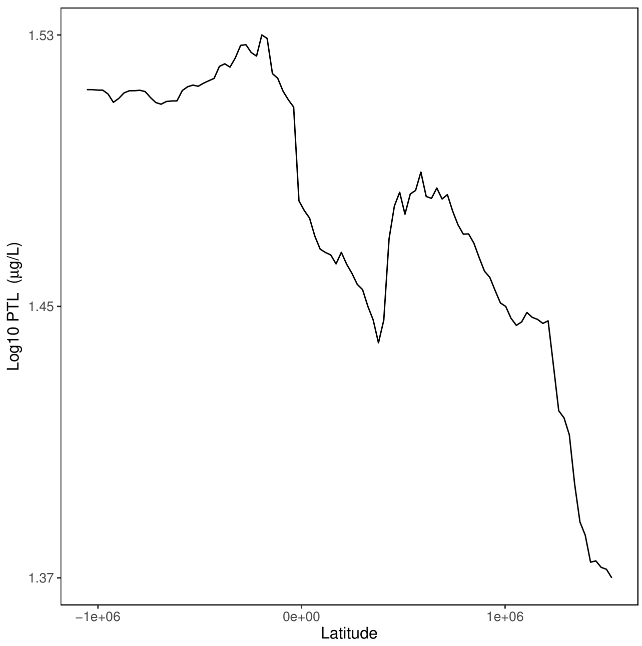

Figure S5: Partial dependency plot for total phosphorus versus latitude: the effect of latitude on total phosphorus when the rest of the predictors are held constant.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019

PeerJ Preprints NOT PEER-REVIEWED

Table S1: Coefficients for the multilevel model.

		Mean	Standard Deviation
Cutoff points/Thresholds	$C_{Oligo\|Meso}$	-156.60	44.04
	$C_{Meso\|Eu}$	-6.18	8.29
	$C_{Eu\|Hyper}$	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
	$\alpha_{SecchiDiskDepth}$	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 S2: 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27536v1 | CC BY 4.0 Open Access | rec: 14 Feb 2019, publ: 14 Feb 2019