Modeling Lake Trophic State: A Data Mining Approach

Jeffrey W. Hollister * [1] W. Bryan Milstead [1] Betty J. Kreakie [1]

[1] US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, 27 Tarzwell Drive Narragansett, RI, 02882, USA

* *corresponding author: hollister.jeff@epa.gov*

Abstract

Productivity of lentic ecosystems has been well studied and it is widely accepted that as nutrient inputs increase, productivity increases and lakes transition from low trophic state (e.g. oligotrophic) to higher trophic states (e.g. eutrophic). These broad trophic state classifications are good predictors of ecosystem health and ecosystem services and disservices (e.g. recreation, aesthetics, fisheries, and harmful algal blooms). While the relationship between nutrients and trophic state provides reliable predictions, it requires *in situ* water quality data in order to parameterize the model. This limits the application of these models to lakes with existing and, more importantly, available water quality data. To expand our ability to predict trophic state in lakes without water quality data, we take advantage of the availability of a large national lakes water quality database, land use/land cover data, lake morphometry data, other universally available data, and modern data mining approaches to build and assess models of lake tropic state that may be more universally applied. We use random forests and random forest variable selection to identify variables to be used for predicting trophic state and we compare the performance of two sets of models of trophic state (as determined by chlorophyll *a* concentration). The first set of models estimates trophic state with *in situ* as well as universally available data and the second set of models uses universally available data only. For each of these models we used three separate trophic state categories, for a total of six models. Overall accuracy for models built from *in situ* and universal data ranged from 0.669% to 0.867%. For the universal data only models, overall accuracy ranged from 0.489% to 0.757%. Lastly, it is believed that the presence and abundance of cyanobacteria is strongly associated with trophic state. To test this we examine the association between estimates of cyanobacteria abundance and measured chrlorophyll *a* and find a positive relationship. Expanding these preliminary results to include cyanobacteria taxa indicates that cyanobacteria are significantly more likely to be found in highly eutrophic lakes. These results suggest that predictive models of lake trophic state may be improved with additional information on the landscape surrounding lakes and that those models provide additional information on the presence of potentially harmful cyanobacteria taxa.

1 Introduction

Productivity in lentic systems is often categorized across a range of tropic states (e.g. the trophic continuum) from early successional (i.e. oligotrophic) to late successional lakes (i.e. hypereutrophic) with lakes naturally occurring across this range (Carlson 1977). Oligotrophic lakes occur in nutrient poor areas or have a more recent geologic history, are often found in higher elevations, have clear water, and are usually favored for drinking water or direct contact recreation (e.g. swimming). Lakes with higher productivity (e.g. mesotrophic and eutrophic lakes) have greater nutrient loads, tend to be less clear, have greater density of aquatic plants, and often support more diverse and abundant fish communities. Higher primary productivity is not necessarily a predictor of poor ecological condition as it is natural for lakes to shift from lower to higher trophic states but this is a slow process. However, at the highest productivity levels (hypereutrophic lakes) biological integrity is compromised (Hasler 1969, Smith et al. 1999, Schindler and Vallentyne 2008).

Monitoring trophic state allows the identification of rapid shifts in trophic state or locating lakes with unusually high productivity (e.g. hypereutrophic). These cases are indicative of lakes under greater anthropogenic nutrient loads, also known as cultural eutrophication, and are more likely to be at risk of fish kills, fouling, and harmful algal blooms (Smith 1998, Smith et al. 1999, 2006). Given the association between trophic state and many ecosystem services and disservices, being able to accurately model trophic state could provide a first cut at identifying lakes with the potential for harmful algal blooms or other problems associated with cultural eutrophication.

As trophic state and related indices can be best defined by a number of *in situ* water quality parameters (modeled or measured), most models have used this information as predictors (Imboden and Gächter 1978, Salas and Martino 1991, e.g., Carvalho et al. 2011, Milstead et al. 2013). This leads to accurate models, but also requires data that are often sparse and not always available, thus limiting the population of lakes for which we can make predictions. A possible solution for this is to build models that use widely available data that are correlated to many of the *in situ* variables. For instance, landscape metrics of forests, agriculture, wetlands, and urban land in contributing watersheds have all been shown to explain a significant proportion of the variation (ranging from 50-86%, depending on study) in nutrients in receiving waters (Jones et al. 2001, 2004, Seilheimer et al. 2013). Building on these previously identified associations might allow us to use only landscape and other universally available data to build models. Identifying predictors using this type of ubiquitous data would allow for estimating trophic state in both monitored and unmonitored lakes.

Many published models of nutrients and trophic state in freshwater systems are based on linear modelling methods such as standard least squares methods or linear mixed models (Jones et al. 2001, e.g., 2004). While these methods have proven to be reliable, they have limitations (e.g. independence and distribution assumptions, and outlier sensitivity). Using data mining approaches, such as random forests, avoids many of the limitations, may reduce bias and often provides better predictions (Breiman 2001, Cutler et al. 2007, Peters et al. 2007). For instance, random forests are non-parametric and thus the data do not need to come from a specific distribution (e.g. Gaussian) and can contain collinear variables (Cutler et al. 2007). Second, random forests work well with very large numbers of predictors (Cutler et al. 2007). Lastly, random forests can deal with model selection uncertainty as predictions are based upon a consensus of many models and not just a single model selected with some measure of goodness of fit.

To build on past work we have identified four goals for this research. First, update trophic state modelling efforts with the use of random forests. Second, assess the accuracy of predicted trophic state in lakes with the full suite of data and then with the universally available data only. Third, identify important variables for describing lake trophic state and lastly, explore associations between trophic state and cyanobacteria to begin to understand how changes in trophic state may be linked to an important ecosystem disservice.

2 Methods

2.1 Data and Study Area

We utilize four primary sources of data for this study, the National Lakes Assessment (NLA), the National Land Cover Dataset (NLCD), modeled lake morphometery, and cyanobacteria abundance (Homer et al. 2004, USEPA 2009, Xian et al. 2009, Hollister and Milstead 2010, Hollister et al. 2011, Hollister 2014). All datasets are national in scale and provide a unique snapshot view of the condition of lakes in the conterminous United States' during the summer of 2007.

The NLA data were collected during the summer of 2007 and the final data were released in 2009 (USEPA 2009). With consistent methods and metrics collected at 1056 locations across the conterminous United States (Figure 1), the NLA provides a unique opportunity to examine broad scale patterns in lake productivity. The NLA collected data on biophysical measures of lake water quality and habitat as well as an assessment of the phytoplankton community. For this analysis we examined the water quality measurements and total cyanobacteria abundance from the National Lakes Assessment (USEPA 2009).

Adding to the monitoring data collected via the NLA, we use the 2006 NLCD data to examine landscape-level drivers of trophic status in lakes. The NLCD is a national land use/land cover dataset that also provides estimates of impervious surface. We calculated total proportion of each NLCD land use land cover class and total percent impervious surface within a 3 kilometer buffer surrounding the lake (Homer et al. 2004, Xian et al. 2009). A three kilometer buffer was selected as an intermediate measure of adjacent neighborhood; the three kilometer buffer size is greater than the immediate parcel but smaller than regional measures.

To account for unique aspects of each lake and characterize lake productivity, we also used measures of lake morphometry (i.e. depth, volume, fetch, etc.). As these data are difficult to obtain for large numbers of lakes over broad regions, we used modeled estimates of lake morphometry (Hollister and Milstead 2010, Hollister et al. 2011, Hollister 2014). From these prior efforts we included, Surface Area, Shoreline Length, Shoreline Development, Maximum Depth, Mean Depth, Lake Volume, Maximum Lake Length, Mean Lake Width, Maximum Lake Width, and Fetch.

2.2 Predicting Trophic State with Random Forests

Random forest is a machine learning algorithm that aggregates numerous decision trees in order to obtain a consensus prediction of the response categories (Breiman 2001). Bootstrapped sample data are recursively partitioned according to a given random subset of predictor variables and trees are completely grown without pruning. With each new tree, the sample data subset is randomly selected and with each new split, the subset of predictor variables are randomly selected.

While random forests are able to handle numerous correlated variables without a decrease in prediction accuracy, one possible downfall to this approach is that the resulting model may be difficult to interpret. This is a problem often faced in gene selection and in that field, a variable selection method based on random forest has been successfully applied and implemented in the R Language as `varSelRF` (Díaz-Uriarte and De Andres 2006). With this method, a minimum set of variables that maximizes model accuracy is provided. This allows us to start with a full suite of predictor variables from which to select a minimum, easier to interpret set of variables. One issue with the approach in `varSelRF` is that because of the randomization inherent in random forests it is possible to get variation in the minimum selected set of variables. To account for this we repeated `varSelRF` 100 times. In our case, repeating the procedure 100 times quickly converged on a set of all possible important variables.

2.3 Model Details

Using both `varSelRF` and `randomForest` we ran models for six sets of variables and trophic state classifications. These included three different combinations of the Chlorophyll *a* trophic states as the dependent variables and using all variables (*in situ* and GIS variables) or the GIS only variables (i.e. no *in situ* information) as the independent variables in the random forest. A listing of all considered variables is in Appendix 1. Trophic state was defined using the NLA chlorophyll *a* trophic state cut offs and the three combinations of trophic state were used to highlight the possible error caused by misclassification of adjacent classes, such as mesotrophic and eutrophic (Table 1). The six model combinations were:

- **Model 1:** Chlorophyll *a* trophic state - 4 class = All variables (*in situ* water quality, lake morphometry, and landscape)
- **Model 2:** Chlorophyll *a* trophic state - 3 class = All variables (*in situ* water quality, lake morphometry, and landscape)
- **Model 3:** Chlorophyll *a* trophic state - 2 class = All variables (*in situ* water quality, lake morphometry, and landscape)
- **Model 4:** Chlorophyll *a* trophic state - 4 class = All variables (lake morphometry, and landscape)
- **Model 5:** Chlorophyll *a* trophic state - 3 class = All variables (lake morphometry, and landscape)
- **Model 6:** Chlorophyll *a* trophic state - 2 class = All variables (lake morphometry, and landscape)

Our modelling work flow was as follows:

1. Use `iterVarSelRF` in the `LakeTrophicModelling` R package to identify a minimal set of variables that maximize accuracy of the random forest algorithm (Diaz-Uriarte 2010, Hollister et al. 2014). This subset of variables, the reduced model, is calculated for each of our 6 models.

2. Using R's `randomForest` package, we pass the reduced models selected with `iterVarSelRF` and assess model performance (Liaw and Wiener 2002).

2.4 Measures of Model Performance and Variable Importance

We assess the performance of the random forest models by comparing the total prediction accuracy and the kappa coefficient of the final confusion matrix. As random forest builds each tree on bootstrapped, random subsets of the original data, a separate independent validation dataset is not required and random forest error estimates expected to be unbiased [breiman2001random]. For each of the models, the final predictions were compared to the original data via a confusion matrix. The total accuracy (i.e. percent correctly predicted) was calculated. Since some agreement can be expected by chance alone, it is also useful to take this type of error into account. For this we calculated the kappa coefficient for each model as well (Cohen 1960, Hubert and Arabie 1985). The kappa coefficient can range from -1 to 1 with 0 equalling the agreement expected by chance alone. Values greater than 0 represent agreement greater than would be expected by chance, with values greater than 0.61 considered "substantial" agreement (Landis and Koch 1977).

Lastly, the random forest algorithm explicitly measures variable importance as mean decrease in Gini. The Gini Index is a measure of how well the data are classified into homogeneous groups. For every node, the splitting variables are permuted and the change in actual Gini and permuted Gini is recorded. The mean decrease Gini is a summed and standardized value for each variable (Breiman 2001). Higher values of mean decrease Gini suggest a higher importance for that variable.

3 Results and Discussion

Our complete dataset includes 1148 lakes; however 5 lakes did not have chlorophyll *a* data. Thus, the base dataset for our modelling was conducted on data for 1143 lakes. The lakes were well distributed both across the four trophic state categories (Table 1) and spatially throughout the United States (Figure 1).

3.1 Models

Accuracy for the models built with all predictors ranged from 0.669 to 0.867 and the kappa coefficient had a minimum value of 0.549 and maximum of 0.734. The GIS only models had a total accuracy between 0.489 and 0.757 and kappa coefficient between 0.302 and 0.515. The importance of variables for the models including the *in situ* data were fairly stable while There was considerably more variation in variable importance for the three different GIS only models. Details for each model are discussed below.

Model 1: 4 Trophic States ~ All Variables The reduced model for Model 1 included potassium, nitrogen:phosphorus, total nitrogen, total phosphorus, total organic carbon, turbidity, ecoregion, organic ions, dissolved organic carbon, and maximum depth and of these, turbidity, total phosphorus, total nitrogen, and total organic carbon were the most four most important predictors of the four classes of trophic state (Figure 2). Total accuracy for Model 1 was 0.669% and the Cohen's Kappa was 0.549 (Table 2).

Model 2: 3 Trophic States ~ All Variables For Model 2, the reduced model included turbidity, total phosphorus, total nitrogen, total organic carbon, nitrogen:phosphorus, longitude, pH, estimated organic anions, elevation, maximum depth, dissolved organic carbon, potassium, latitude, ecoregion, chloride, ammonium and percent cropland (Figure 3). The top predictors for 3 trophic state classes were turbidity, total phosphorus, total nitrogen, and total organic carbon (Figure 3). Model 2 accuracy was 0.795% and the Cohen's Kappa was 0.613 (Table 3).

Model 3: 2 Trophic States ~ All Variables The reduced model for Model 3 was similar to Model 1 and Model 2 and included turbidity, total phosphorus, total nitrogen, nitrogen:phosphorus, potassium, ecoregion, elevation, total organic carbon, growing degree days, longitude, sodium, maximum depth, estimated organic anions, latitude, and dissolved organic carbon (Figure 4). The top three predictors were the same for Model 3; however, elevation and growing degree days had a higher importance than total organic carbon. (Figure 4). Total accuracy for Model 3 was 0.867% and the Cohen's Kappa was 0.734 (Table 4).

Model 4: 4 Trophic States ~ GIS Only Variables The selected variables for the Model 4 were longitude, latitude, elevation, estimated mean lake depth, percent evergreen forest, estimated maximum lake depth, percent cropland, and ecoregion (Figure 5). The most important variables were percent evergreen forest, ecoregion, percent cropland, and longitude (Figure 5). Total accuracy for Model 4 is 0.489% and the Cohen's Kappa is 0.302 (Table 5).

Model 5: 3 Trophic States ~ GIS Only Variables The reduced model for Model 5 included estimated mean lake depth, percent cropland, longitude, latitude, percent evergreen forest, elevation, estimated maximum lake depth, estimated lake volume, percent deciduous forest, percent developed open space, ecoregion, percent woody wetland, and percent shrub/scrub (Figure 6). The most important variables for model 5 were ecoregion, percent evergreen forest, percent cropland, and estimated mean depth. (Figure 6). Total accuracy for Model 5 is 0.676% and the Cohen's Kappa is 0.347 (Table 6).

Model 6: 2 Trophic States ~ GIS Only Variables The variable selection process for Model 6 produced a reduced model with ecoregion, growing degree days, percent evergreen forest, percent cropland, elevation, estimated mean lake depth, longitude, latitude, watershed area, estimated maximum lake depth, percent developed open space, percent deciduous forest, and estimated lake volume (Figure 7). Similar to models 4 and 5, the four most important variables were ecoregion, percent evergreen forest, percent cropland, and elevation (Figure 7). Total accuracy for Model 6 0.757% and the Cohen's Kappa is 0.515 (Table 7).

3.2 Trophic State Probabilities

One of the powerful features of random forests is the ability to aggregate a very large number of competing models or trees. Each tree provides an independent prediction or vote for a possible outcome. In the context of our trophic state models, we have 10,000 votes for each lake. These values may be interpreted as the probability that a lake is in a given trophic state. For instance, for a single lake (National Lake Assessment ID = NLA06608-0005), the vote probabilities for Model 1 were 0.81 for oligotrophic, 0.19 for mesotrophic, 0 for eutrophic, and 0 for hypereutrophic. This suggests little uncertainty in the predicted oligotrophic state.

Further, the maximum probability for each lake can be used as a measure of how certain the random forest model was of the prediction. We would expect higher total accuracy for lakes that had more certain predictions. To test this we can examine the accuracy of trophic state predictions across the full range of trophic state probabilities, similar to an approach outlined by Paul and MacDonald (2005) and implemented by Hollister *et al.* (2008). We utilize this approach and examine the change in total accuracy as a function of the maximum probability for each lake. As expected, lakes with higher maximum vote probabilities are more accurately predicted (Figure 12). This suggest that even for models with low overall accuracy there will also be a large number of cases that are predicted with high accuracy.

3.3 Variable Selection and Importance

There was a great deal of agreement on the important variables for each set of models. In line with past predictive modeling of cyanobacteria abundance, the *in situ* models consistently select the water quality variables (turbidity, total nitrogen, total phosphorus, and N:P ratios) as important variables (Downing et al. 2001). While there is variation in the response of cyanobacteria to changes in relative nutrient concentrations, the general pattern suggests that limiting nutrients have considerable impact once amounts increase beyond expected levels.

The mechanistic role of turbidity on lake trophic state is more complex. Light availability in turbid waters is lower than in clear waters. This would suggest a negative relationship between turbidity and chlorophyll *a*. Second, chlorophyll *a* can also be a component of turbidity and lakes with higher chlorophyll *a* concentrations will also be more turbid. Last, chlorophyll *a* is not the only component of of turbidity and turbid waters can be caused by, for example, increased sediment loads or tannin. This would be a cause for concern with linear models; however, linearity is not an assumption of tree-based modelling approaches such as random forest [need cite].

Our GIS models are capturing the large scale spatial pattern of trophic status gradient of lakes across the United States. We reliably see latitude and longitude and ecoregion selected as important variables. It is also possible that other variables selected as important are also capturing a portion of this trend. For instance, elevation and growing degree days both have obvious spatial components, but may also be accounting for variation in temperature.

The land use/land cover variables are also important variables in describing trophic state patterns. Like elevation and growing degree days, there are broad scale spatial patterns inherent in the data. For instance, the relative continental position of mountains in the United States is the spatial inverse of the distribution of agricultural lands. However, it is known that forests are positively associated with lower nutrient loads where as agricultural land shows a negative association. These more local scale relationships with land use/land cover are likely providing additional predictive power to the information in the broader scale data.

Lastly, morphometry (e.g. depth and volume) also proved to be important in the prediction of lake trophic state. As morphometry shows little to no broad scale spatial pattern and is unique to a given lake, these data are likely illuminating the local, lake scale drivers of trophic state. As only depth and volume were selected this probably shows the importance of in lake nutrient processing and residence time.

3.4 Associating Trophic State and Cyanobacteria

Cyanobacteria biomass should be closely related to trophic state as they contribute to the chlorophyll concentration in a lake. These associations have been seen by others. If these associations are strong enough we may be able to expand models such as those reported here to also predict probability of cyanobacteria blooms. To test if trophic state may be used to differentiate cyanobacteria abundance we examine distribution of cyanobacteria abundance for each trophic state and we also explore linear associations between Chlorohyll *a* and cyanobacteria abundance.

The distribution of cyanobacteria abundance shows separation between all of the trophic state classifications (Figures 8, 9, and 10). Furthermore, there is a significant linear relationship (r^2=0.33) between chlorophyll *a* and cyanobacteria abundance (Figure 11). Further, Yuan et al. (2014) used the 2007 NLA to demonstrate that total nitrogen and chlorophyll *a* concentrations were good predictors of World Health Organization microcystin (a toxin produced by some cyanobacteris) criteria exceedences. These results suggest that trophic state is indeed an acceptable proxy for cyanobacteria abundance and that in lakes with higher trophic state it is also reasonable to expect higher cyanobacteria.

4 Conclusions

Our research goals were to explore the utility of a widely used data mining algorithm, random forests, in the modelling of lake trophic state. Further, we hoped to examine the utility of these models when built with only ubiquitous GIS data, which would allow for making trophic state estimates for all lakes in the United States. We were able to successfully predict a variety of trophic state classes. With the GIS only data models our total accuracy ranged from 0.4894552 to 0.7574692 and with the full suite of data our model accuracy had a minimum accuracy of 0.6690018 and maximum accuracy of 0.8669002.

While some of the models (i.e. Model 4) show relatively low prediction accuracies, another feature of the random forest, votes, can provide additional information. In addition to providing a single estimate of trophic state for each lake, our models also indicate the probability that a lake was classified in any of the categories. These probabilities may be mapped directly to show the uncertainty of a given predicted class. Furthermore as the certainty of prediction increases so to does the overall trophic state classification accuracy (Figure 12). These results suggest that our models will provide reasonable estimates of trophic state across the United States.

There was great deal of agreement on the important variables for each set of models. For the combined *in situ* and GIS Models, the *in situ* water quality variables drove the predictions. This is expected. For the GIS only models, the results were more nuanced with three broad categories routinely being selected as important: broad scale spatial patterns in trophic state, land use/land cover controls of trophic state, and local, lake-scale control driven by lake morphometry. Lastly, associations between trophic state and cyanobacteria show that at the broad scale of the 2007 NLA there is a linear relationship between chlorophyll *a* and cyanobacteria abundance and that using trophic state as a proxy for cyanobacteria has potential.

These broad categories and the association between trophic state and total cyanobacteria abundance raise three important considerations related to managing eutrophication. First, the broad scale patterning suggests regional trends. This is important because it suggests that efforts to monitor, model and manage eutrophication and cyanobacteria should be undertaken at both national and regional levels. Second, while direct control of water quality in lakes would have a large impact, the land use/land cover drivers (i.e. non-point sources) of water quality are also important and better management of the spatial distribution of important classes such as forest and agriculture can provide some level of control on trophic state and amount of cyanobacteria present. Third, in-lake processes (i.e. residence time, nutrient cycling, etc.) are, as expected, very important and need to be part of any management strategy. Building on these efforts through updated models, direct prediction of cyanobacteria, and additional information on the regional differences will help us get a better handle on the broad scale dynamics of productivity in lakes and the potential risk to human health from cyanobacteria blooms.

5 Figures

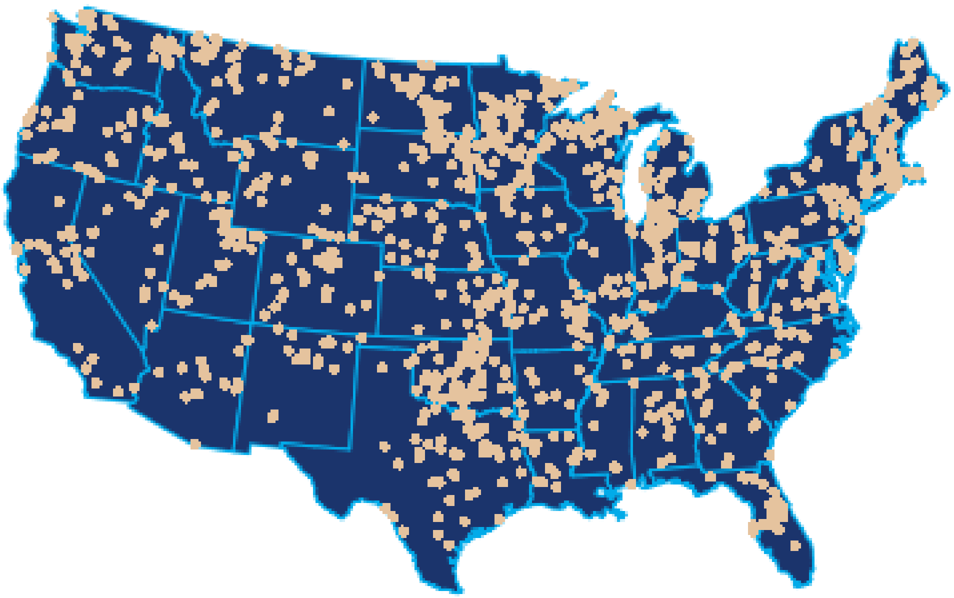

Figure 1: Map of the distribution of National Lakes Assesment Sampling locations

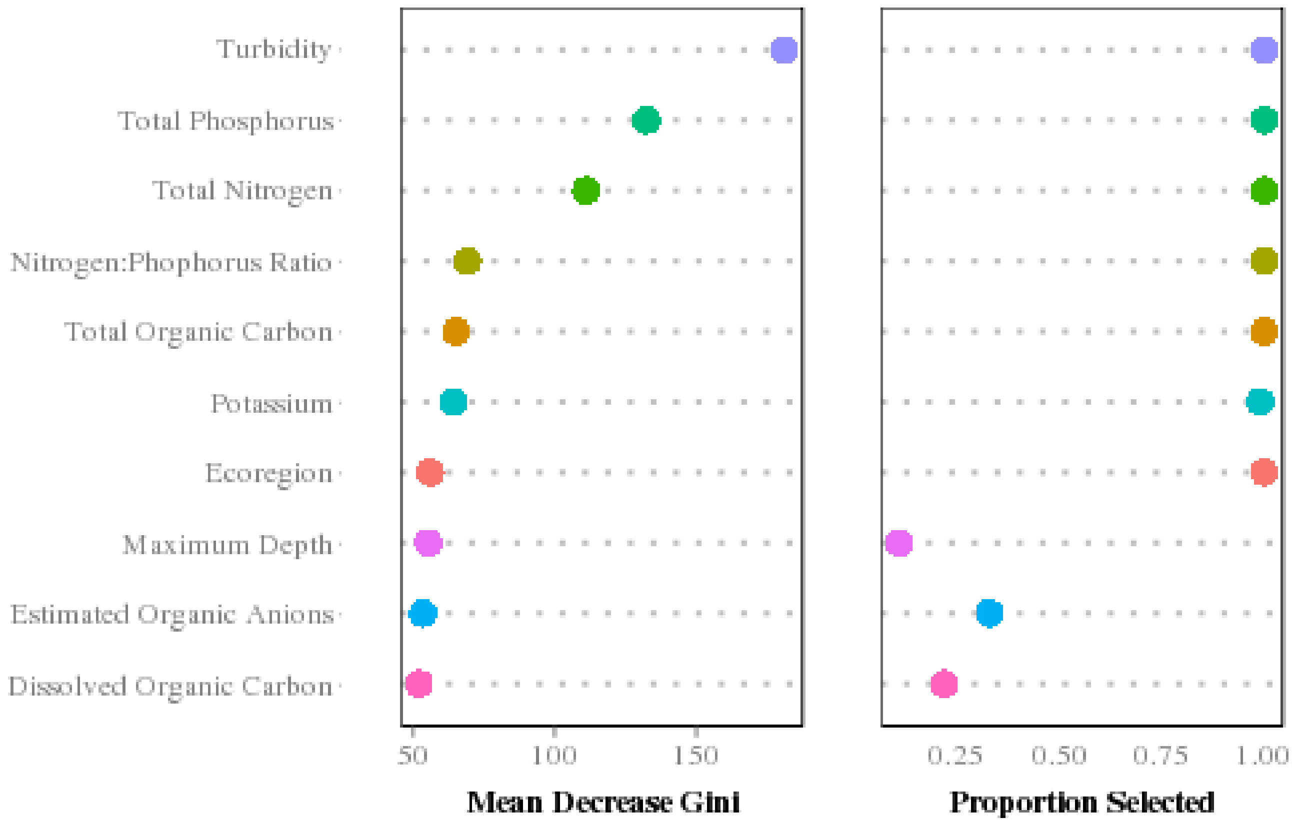

Figure 2: Importance plot for Model 1

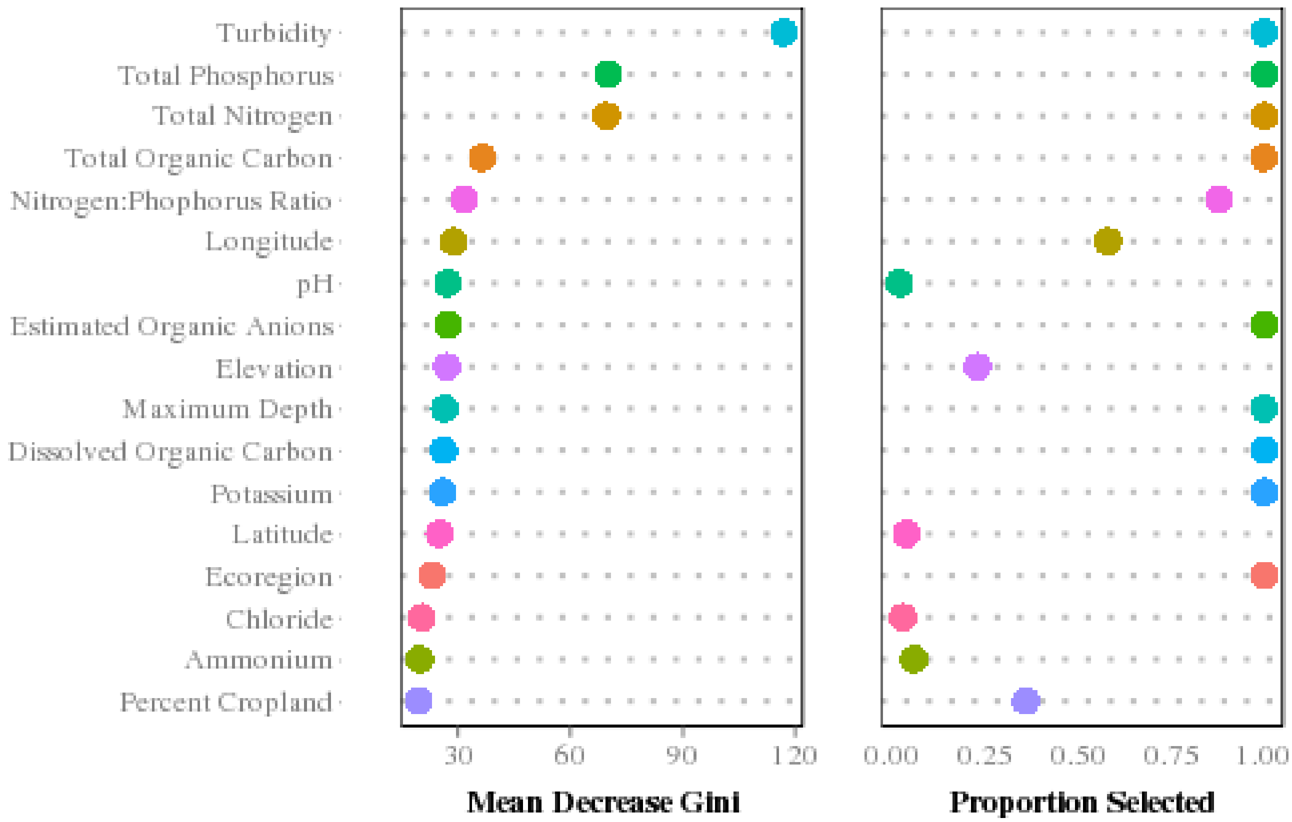

Figure 3: Importance plot for Model 2

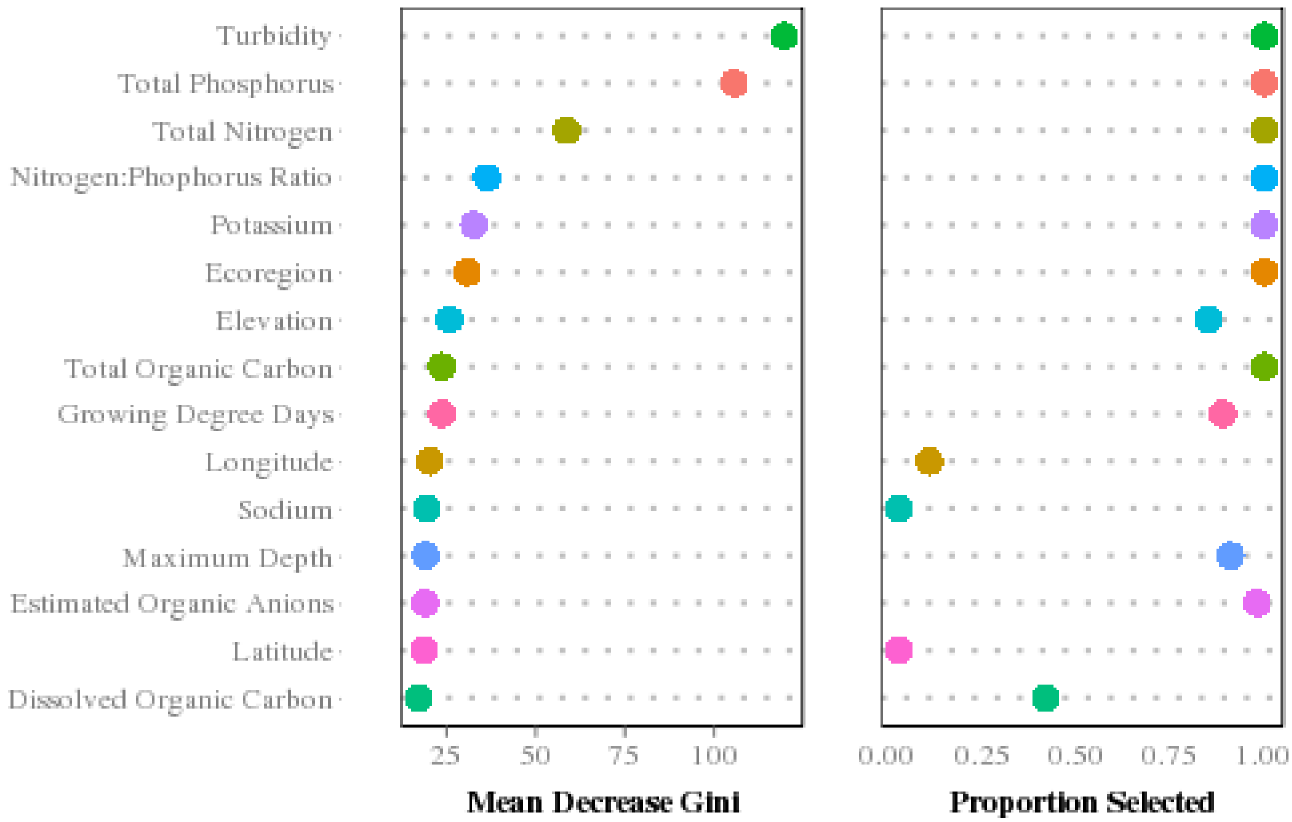

Figure 4: Importance plot for Model 3

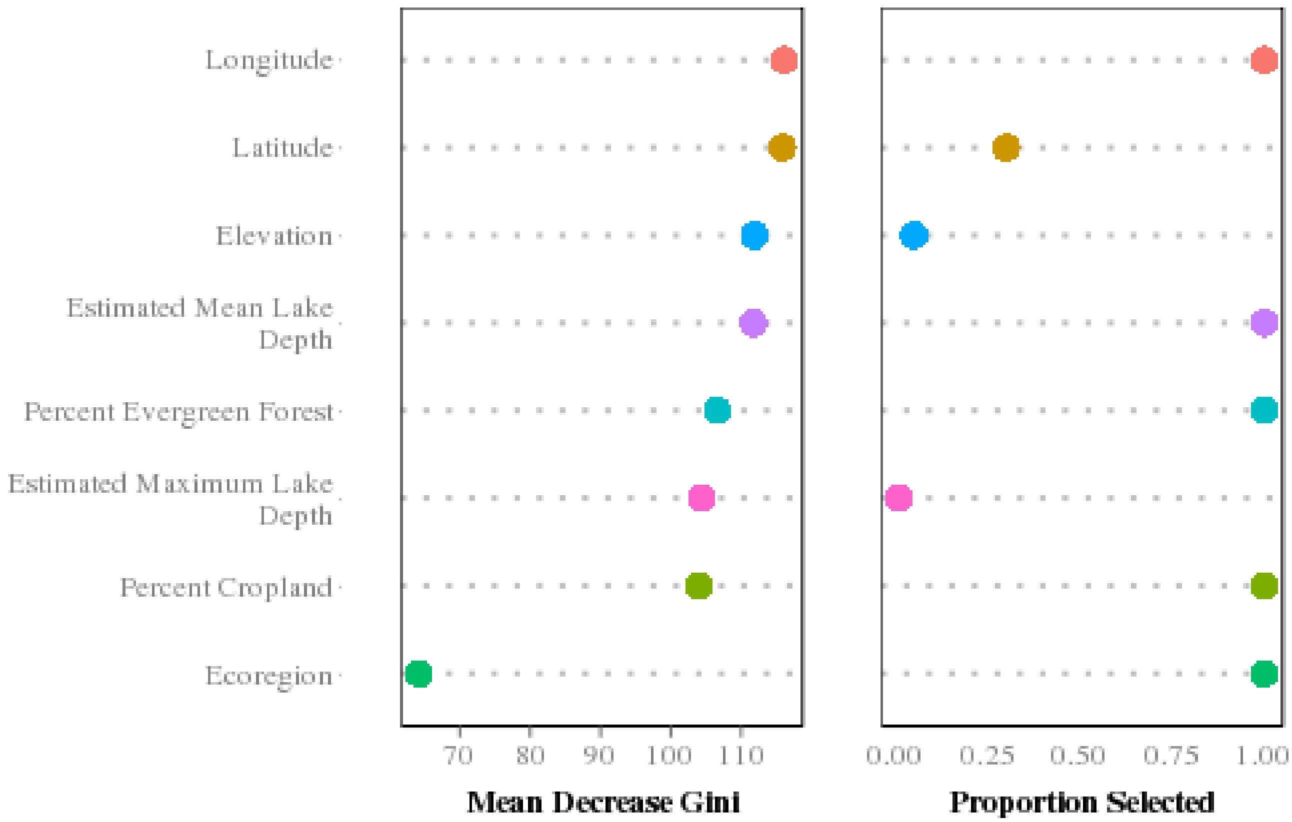

Figure 5: Importance plot for Model 4

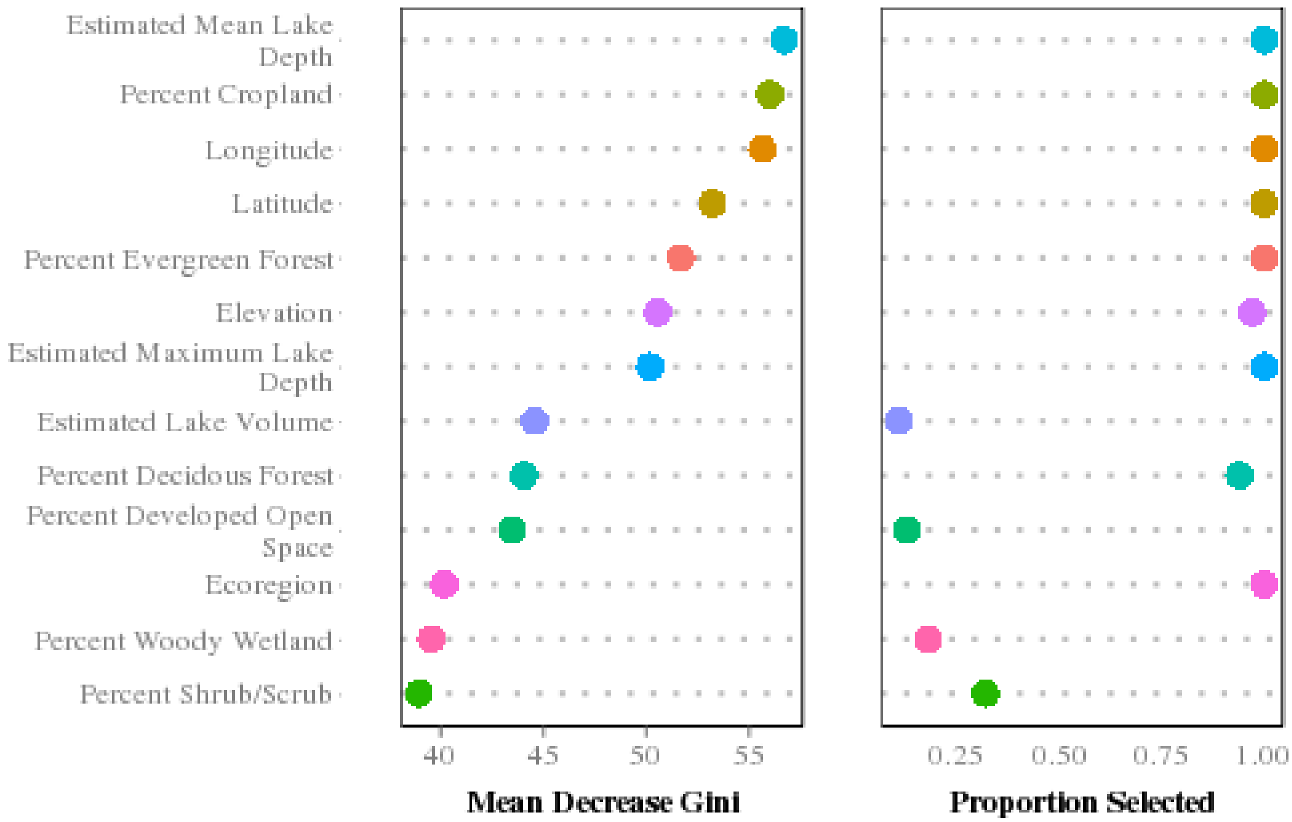

Figure 6: Importance plot for Model 5

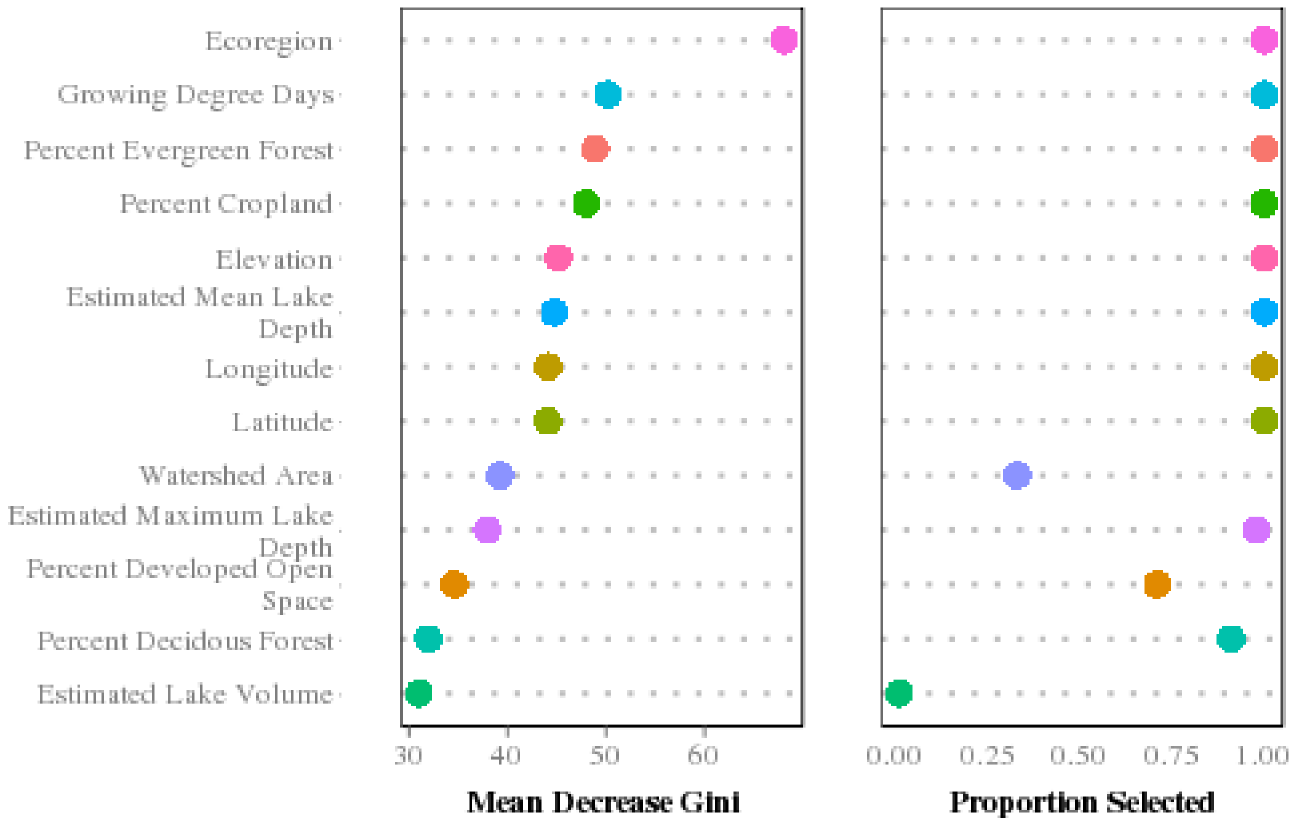

Figure 7: Importance plot for Model 6

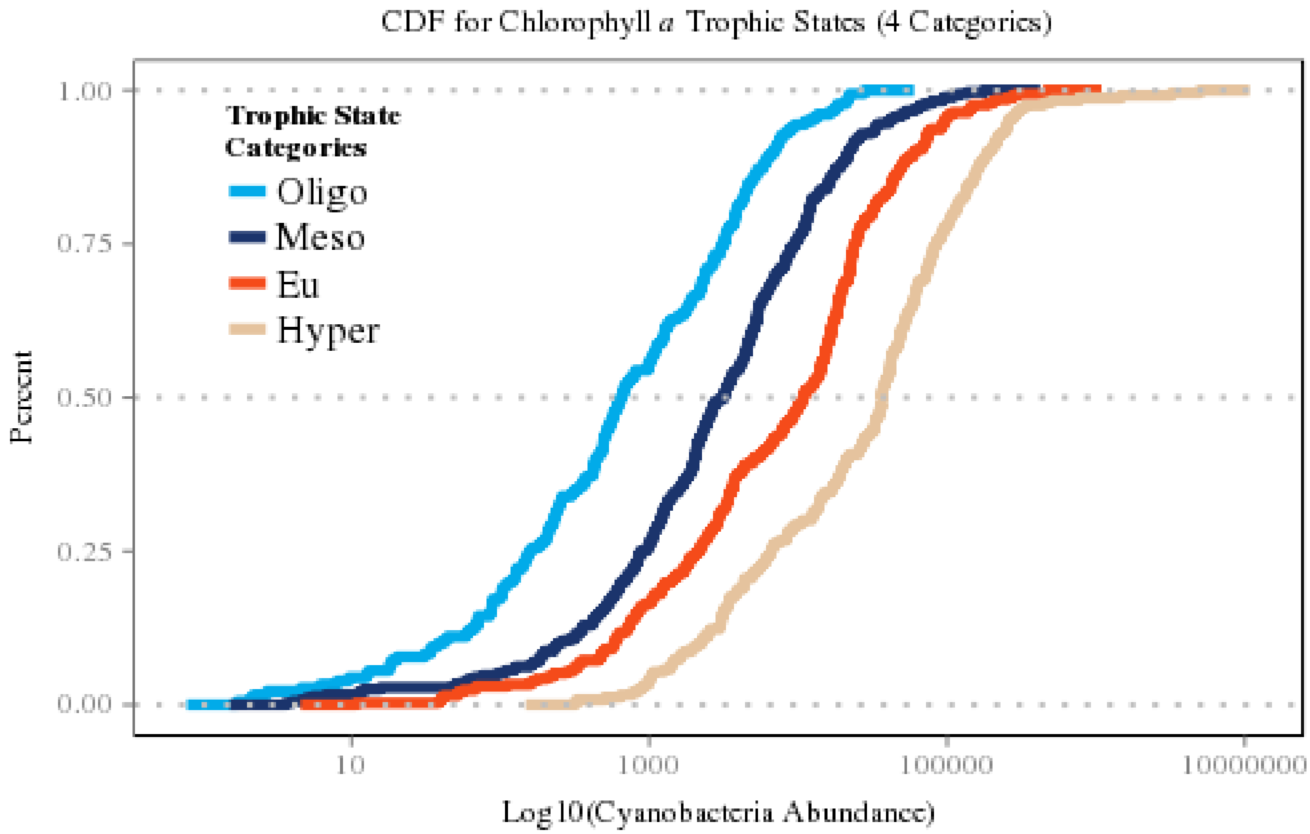

Figure 8: Cumulative distribution function of cyanobacetria abundance for 4 trophic state classes

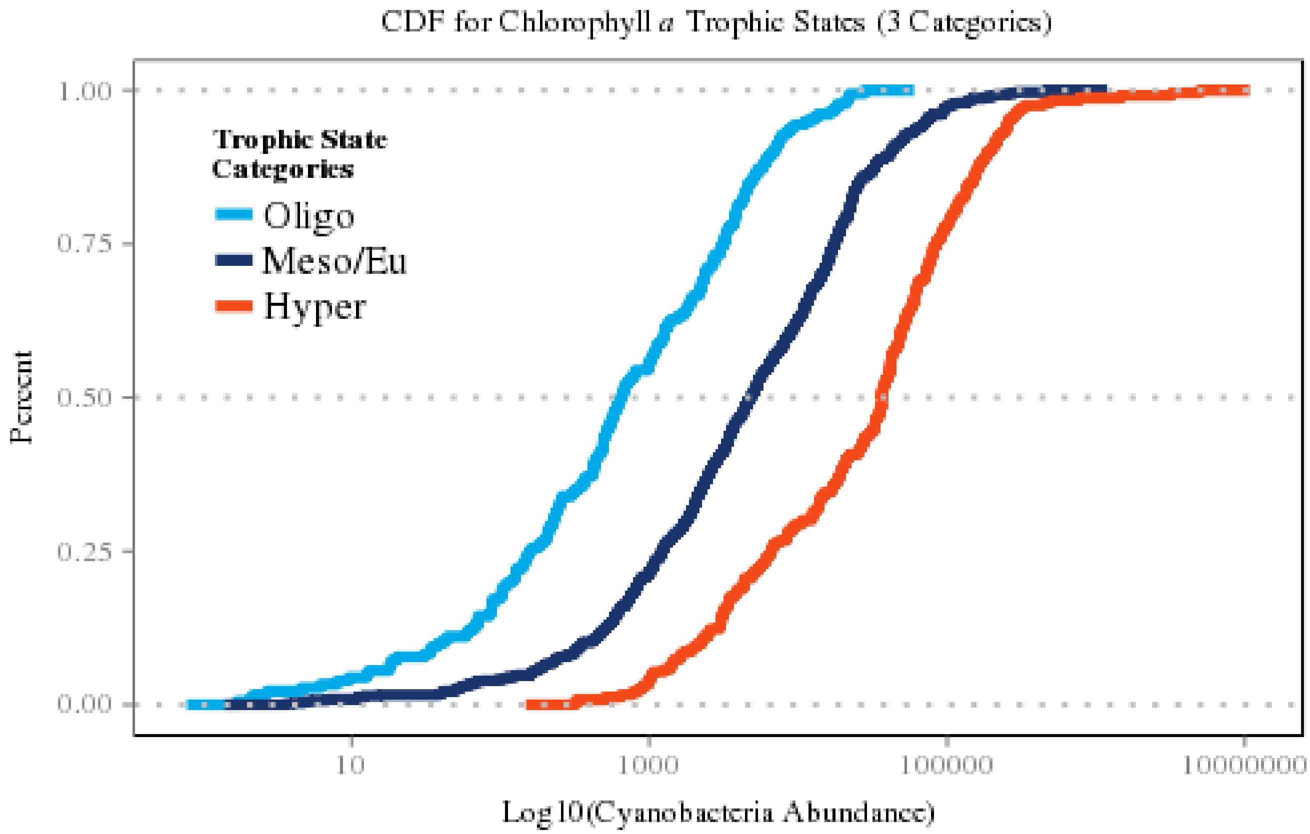

Figure 9: Cumulative distribution function of cyanobacetria abundance for 3 trophic state classes

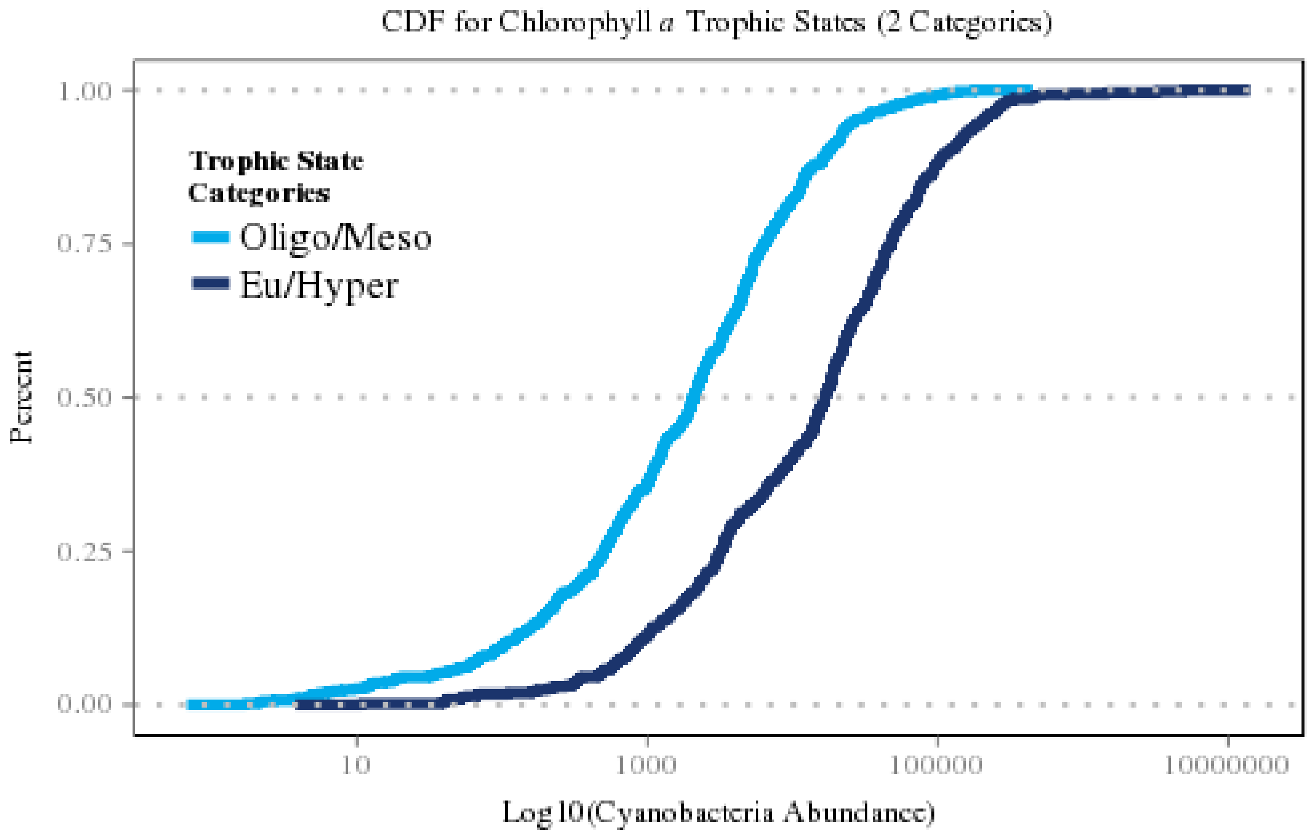

Figure 10: Cumulative distribution function of cyanobacetria abundance for 2 trophic state classes

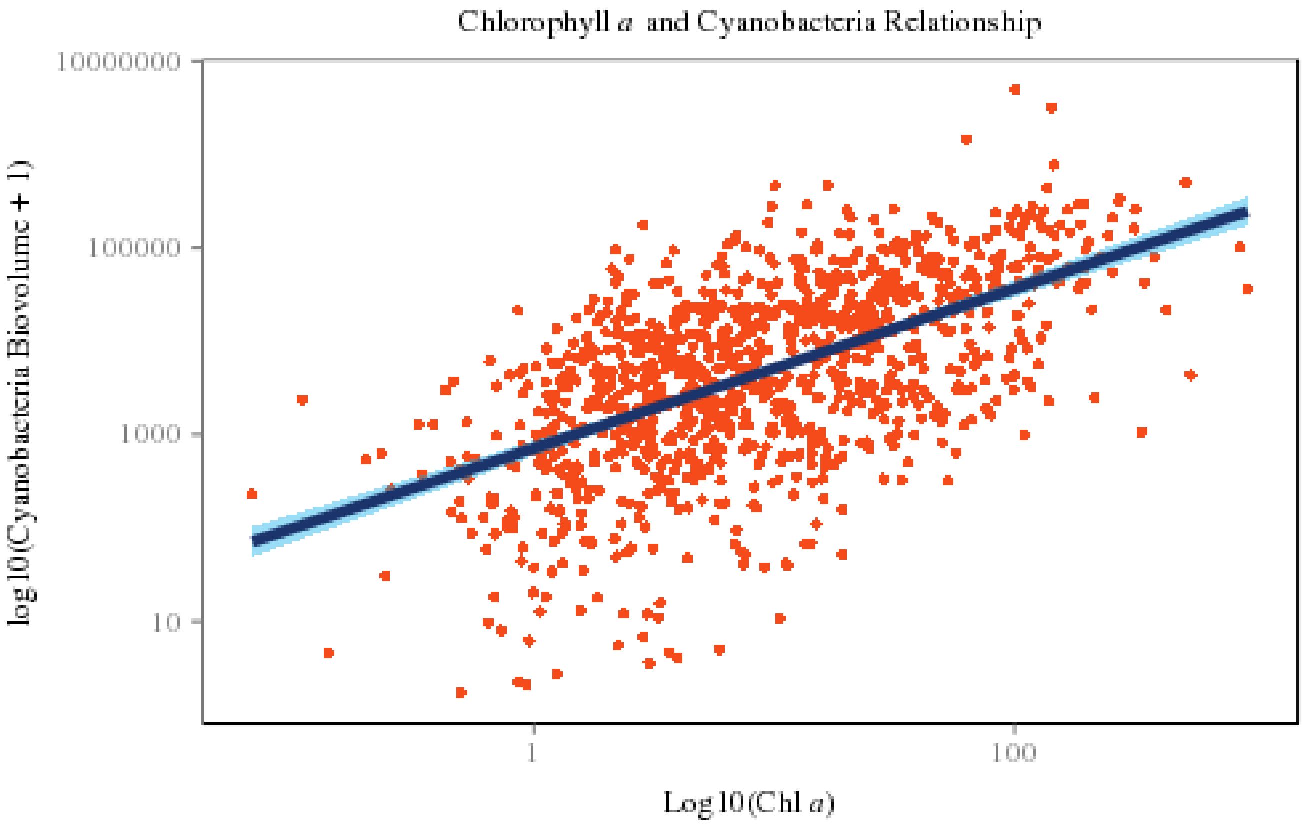

Figure 11: Cholorphyll a and cyanobacteria abundance scatterplot

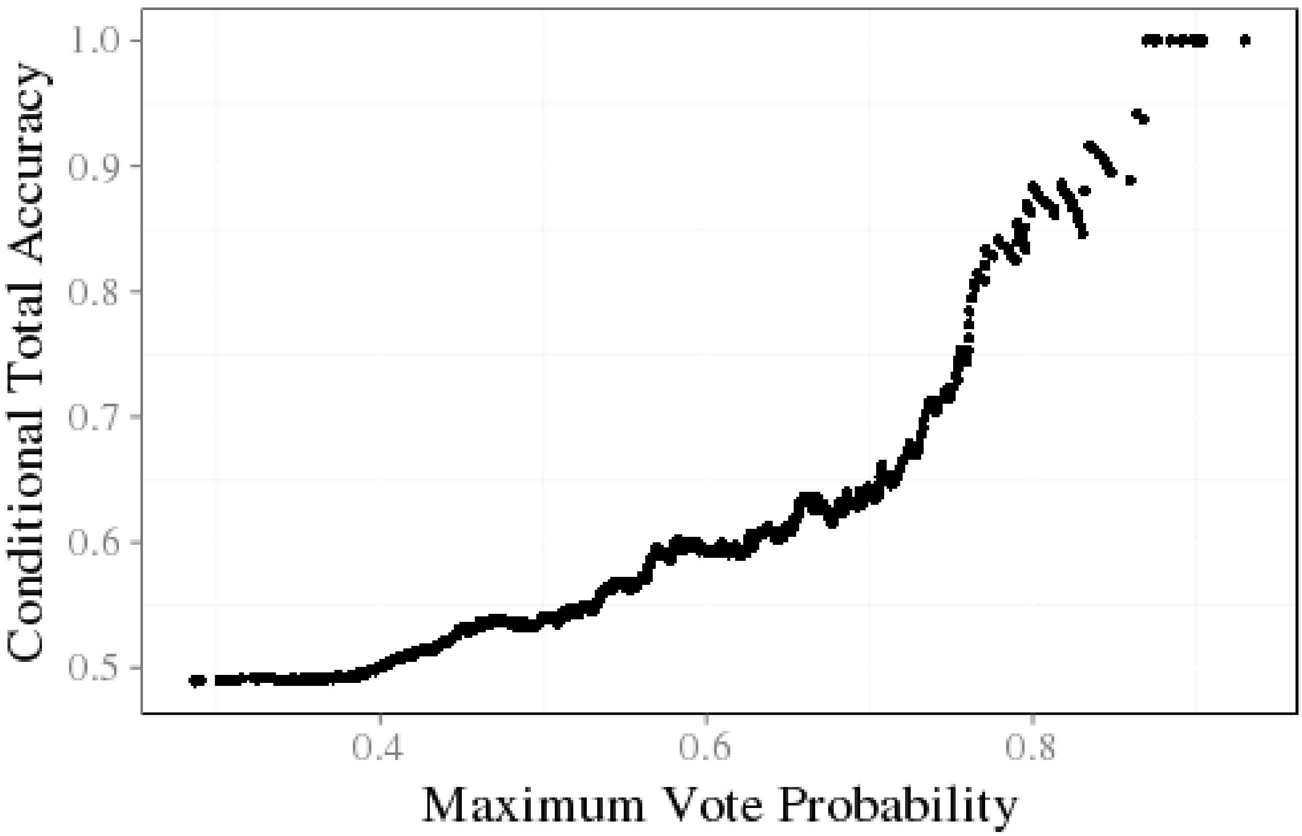

Figure 12: Comparison of certainity of trophic state prediction and total accuracy

6 Tables

Trophic State (4)	Trophic State (3)	Trophic State (2)	Cut-off	n
oligo	oligo	oligo/meso	<= 0.2	198
meso	meso/eu	oligo/meso	>2-7	362
eu	meso/eu	eu/hyper	>7-30	337
hyper	hyper	eu/hyper	>30	246

Table 1: Chlorophyll a based trophic state cut-offs

Oligo	Meso	Eu	Hyper	class.error
135	58	4	1	0.32
42	233	77	10	0.36
2	66	222	46	0.34
0	3	69	174	0.29

Table 2: Random Forest confusion matrix for Model 1

Oligo	Meso/Eu	Hyper	class.error
122	74	0	0.38
43	604	42	0.12
0	72	173	0.29

Table 3: Random Forest confusion matrix for Model 2

Oligo/Meso	Eu/Hyper	class.error
485	75	0.13
77	505	0.13

Table 4: Random Forest confusion matrix for Model 3

Oligo	Meso	Eu	Hyper	class.error
94	72	28	2	0.52
50	201	80	30	0.44
21	110	131	73	0.61
1	34	80	131	0.47

Table 5: Random Forest confusion matrix for Model 4

Oligo	Meso/Eu	Hyper	class.error
80	115	1	0.59
50	585	61	0.16
0	142	104	0.58

Table 6: Random Forest confusion matrix for Model 5

Oligo/Meso	Eu/Hyper	class.error
428	129	0.23
147	434	0.25

Table 7: Random forest confusion matrix for Model 6

7 Appendix 1. Variable Definitions

variable_names	description	type
PercentImperv_3000m	Percent Impervious	GIS
WaterPer_3000m	Percent Water	GIS
IceSnowPer_3000m	Percent Ice/Snow	GIS
DevOpenPer_3000m	Percent Developed Open Space	GIS
DevLowPer_3000m	Percent Low Intensity Development	GIS
DevMedPer_3000m	Percent Medium Intensity Development	GIS
DevHighPer_3000m	Percent High Intensity Development	GIS
BarrenPer_3000m	Percent Barren	GIS
DeciduousPer_3000m	Percent Decidous Forest	GIS
EvergreenPer_3000m	Percent Evergreen Forest	GIS
MixedForPer_3000m	Percent Mixed Forest	GIS
ShrubPer_3000m	Percent Shrub/Scrub	GIS
GrassPer_3000m	Percent Grassland	GIS
PasturePer_3000m	Percent Pasture	GIS
CropsPer_3000m	Percent Cropland	GIS
WoodyWetPer_3000m	Percent Woody Wetland	GIS
HerbWetPer_3000m	Percent Herbaceuos Wetland	GIS
AlbersX	Longitude	GIS
AlbersY	Latitude	GIS
LakeArea	Lake Surface Area	GIS
LakePerim	Lake Perimeter	GIS

variable_names	description	type
ShoreDevel	Shoreline Development Index	GIS
DATE_COL	Date Samples Collected	Water Quality
WSA_ECO9	Ecoregion	GIS
BASINAREA	Watershed Area	GIS
DEPTHMAX	Maximum Depth	Water Quality
ELEV_PT	Elevation	GIS
DO2_2M	Dissolved Oxygen	Water Quality
PH_FIELD	pH	Water Quality
COND	Conductivity	Water Quality
ANC	Acid Neutralizing Capacity	Water Quality
TURB	Turbidity	Water Quality
TOC	Total Organic Carbon	Water Quality
DOC	Dissolved Organic Carbon	Water Quality
NH4	Ammonium	Water Quality
NO3_NO2	Nitrate/Nitrite	Water Quality
NTL	Total Nitrogen	Water Quality
PTL	Total Phosphorus	Water Quality
CL	Chloride	Water Quality
NO3	Nitrate	Water Quality
SO4	Sulfate	Water Quality
CA	Calcium	Water Quality
MG	Magnesium	Water Quality

variable_names	description	type
Na	Sodium	Water Quality
K	Potassium	Water Quality
COLOR	Color	Water Quality
SIO2	Silica	Water Quality
H	Hydrogen Ions	Water Quality
OH	Hydroxide	Water Quality
NH4ION	Calculate Ammonium	Water Quality
CATSUM	Cation Sum	Water Quality
ANSUM2	Anion Sum	Water Quality
ANDEF2	Anion Deficit	Water Quality
SOBC	Base Cation Sum	Water Quality
BALANCE2	Ion Balance	Water Quality
ORGION	Estimated Organic Anions	Water Quality
CONCAL2	Calculated Conductivity	Water Quality
CONDHO2	D-H-O Calculated Conductivity	Water Quality
TmeanW	Mean Profile Water Temperature	Water Quality
DDs45	Growing Degree Days	GIS
MaxLength	Maximum Lake Length	GIS
MaxWidth	Maximum Lake Width	GIS
MeanWidth	Mean Lake Width	GIS
FetchN	Fetch from North	GIS
FetchNE	Fetch form Northeast	GIS

variable_names	description	type
FetchE	Fetch from East	GIS
FetchSE	Fetch from Southeast	GIS
MaxDepthCorrect	Estimated Maximum Lake Depth	GIS
VolumeCorrect	Estimated Lake Volume	GIS
MeanDepthCorrect	Estimated Mean Lake Depth	GIS
NPratio	Nitrogen:Phophorus Ratio	Water Quality

References

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

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

Carvalho, L., C. A. Miller (nee Ferguson), E. M. Scott, G. A. Codd, P. S. Davies, and A. N. Tyler. 2011. Cyanobacterial blooms: Statistical models describing risk factors for national-scale lake assessment and lake management. Science of The Total Environment 409:5353–5358.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:37–46.

Cutler, D. R., T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

Diaz-Uriarte, R. 2010. varSelRF: Variable selection using random forests.

Díaz-Uriarte, R., and S. A. De Andres. 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics 7:3.

Downing, J. A., S. B. Watson, and E. McCauley. 2001. Predicting cyanobacteria dominance in lakes. Canadian journal of fisheries and aquatic sciences 58:1905–1908.

Hasler, A. D. 1969. Cultural eutrophication is reversible. BioScience 19:425–431.

Hollister, J. W. 2014. lakemorpho: Lake morphometry in r.

Hollister, J. W., W. B. Milstead, and B. J. Kreakie. 2014. LakeTrophicModelling: Package to reproduce hollister et al. (2014) modeling lake trophic state: A data mining approach.

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

Hollister, J. W., H. A. Walker, and J. F. Paul. 2008. CProb: a computational tool for conducting conditional probability analysis. Journal of environmental quality 37:2392–2396.

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

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

Hubert, L., and P. Arabie. 1985. Comparing partitions. Journal of classification 2:193–218.

Imboden, D., and R. Gächter. 1978. A dynamic lake model for trophic state prediction. Ecological modelling 4:77–98.

Jones, J., M. Knowlton, D. Obrecht, and E. Cook. 2004. Importance of landscape variables and morphology on nutrients in missouri reservoirs. Canadian Journal of Fisheries and Aquatic Sciences 61:1503–1512.

Jones, K. B., A. C. Neale, M. S. Nash, R. D. Van Remortel, J. D. Wickham, K. H. Riitters, and R. V. O'Neill. 2001. Predicting nutrient and sediment loadings to streams from landscape metrics: a multiple watershed study from the united states mid-atlantic region. Landscape Ecology 16:301–312.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. biometrics:159–174.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.

Milstead, W. B., J. W. Hollister, R. B. Moore, and H. A. Walker. 2013. Estimating summer nutrient concentrations in northeastern lakes from sPARROW load predictions and modeled lake depth and volume. PloS one 8:e81457.

Paul, J. F., and M. E. McDonald. 2005. Development of empirical, geographically specific water quality criteria: a conditional probability analysis approach. Wiley Online Library.

Peters, J., B. D. Baets, N. E. Verhoest, R. Samson, S. Degroeve, P. D. Becker, and W. Huybrechts. 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling 207:304–318.

Salas, H. J., and P. Martino. 1991. A simplified phosphorus trophic state model for warm-water tropical lakes. Water research 25:341–350.

Schindler, D., and J. Vallentyne. 2008. Algal bowl: overfertilization of the world's freshwaters and estuaries.

Seilheimer, T. S., P. L. Zimmerman, K. M. Stueve, and C. H. Perry. 2013. Landscape-scale modeling of water quality in lake superior and lake michigan watersheds: How useful are forest-based indicators? Journal of Great Lakes Research 39:211–223.

Smith, V. H. 1998. Cultural eutrophication of inland, estuarine, and coastal waters. Pages 7–49 *in* Successes, limitations, and frontiers in ecosystem science. Springer.

Smith, V. H., S. B. Joye, R. W. Howarth, and others. 2006. Eutrophication of freshwater and marine ecosystems. Limnology and Oceanography 51:351–355.

Smith, V. H., G. D. Tilman, and J. C. Nekola. 1999. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environmental pollution 100:179–196.

USEPA. 2009. National lakes assessment: a collaborative survey of the nation's lakes. ePA 841-r-09-001. Office of Water; Office of Research; Development, US Environmental Protection Agency Washington, DC.

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

Yuan, L. L., A. I. Pollard, S. Pather, J. L. Oliver, and L. D'Anglada. 2014. Managing microcystin: identifying national-scale thresholds for total nitrogen and chlorophyll a. Freshwater Biology 59:1970–1981.