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/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 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 models of trophic state (as determined by chlorophyll *a* concentration). The first model estimates trophic state with *in situ* as well as universally available data and the second model 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 the measured and predicted trophic state 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 tropic 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 and 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. 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.

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 is 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 are not without their limitations. Using data mining approaches, such as random forests, avoids many of these 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, we update trophic state modelling efforts with the use of random forests. Second, we assess the accuracy of predicted trophic state in lakes with the full suite of data and then with the universally available data only. Third, we identify important variables for describing lake trophic state and lastly, we 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 Lake 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 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. 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. For this analysis we primarily examined the water quality measurements from the NLA (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 nationally collected 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).

To account for unique aspects of each lake and characterize lake productivity, we also used various 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 inlcuded, Surface Area, Shoreline Length, Shoreline Development, Maximum Depth, Mean Depth, Lake Volume, Maximum Lake Length, Mean Lake Width, Maximum Lake Width, and Fetch. Lastly, to explore associations between trophic state and cyanobacteria, we used total cyanobacteria abundance from the National Lakes Assessment (USEPA 2009).

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 is recursively partitioned according to a given random subset of predictor variables and completely grown without pruning. With each new tree, both the sample data and predictor variable subset is 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 succesfully applied (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, more interpretable set of variables. One issue with the approach in `varSelRF` is that becuase 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 Chlorphyll *a* trophic states (Table 1) as the dependent variables and using all variables or the GIS only variables (i.e. no *in situ* infromation) as the independt variables in the random forest. 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, Jeff Hollister and Kreakie n.d.). 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 calculate confusion matrices, overall accuracy and kappa coeffecients for all 6 models (Liaw and Wiener 2002).

3 Results and Discussion

Our complete dataset includes data on 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 throughought 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 coeffecient had a minimum value of 0.549 and maximum of 0.734. The GIS only models had a total accuracy between 0.489 and 0.489 and kappa coeffecient between 0.302 and 0.302. The importance of variables for the models including the *in situ* data was 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.

3.1.1 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 (Table 2) 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 3).

3.1.2 Model 2: 3 Trophic States ~ All Variables

For Model 2, the reduced model included turbidity, total phosphorus, total nitrogen, total organic carbon, nitrogen:phophorus, longitude, pH, estimated organic anions, elevation, maximum depth, dissolved organic carbon, potassium, latitude, ecoregion, chloride, ammonium and percent cropland (Table 4). 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 5).

3.1.3 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:phophorus, potassium, ecoregion, elevation, total organic carbon, growing degree days, longitude, sodium, maximum depth, estimated organic anions, latitude, and dissolved organic carbon (Table 6). The top three predcitors 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 7).

3.1.4 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 (Table 8). 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 9).

3.1.5 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 decidous forest, percent developed open space, ecoregion, percent woody wetland, and percent shrub/scrub (Table 10). The most important variables for model 5 were ecoregion, perent 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 11).

3.1.6 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 decidous forest, and estimated lake volume (Table 12). 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 13).

3.2 Trophic State Probabilities

One of the powerful features of random forests is that you utilize 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 interepreted as the probability that a lake is in a given trophic state. For instance, for a single lake (COMID = 23491387), the vote probabilities for Model 1 ranged from 0 to 0.81 suggesting little uncertainity in the predicted class (Table 14).

3.2.1 Model Accuracy and Uncertainity

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 probabilites 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 accuarcy.

3.3 Variable Importance

NEED STUFF HERE - Betty you want to take stab?

3.4 Associating Trophic State and Cyanobacteria

Cyanobacteria biomass should be closely related to trophic state as they make up a component of 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 classificaitons (Figures 8, 9, and 10). Furthermore, there is a significant linear relationship (r^2=0.33) between Chl *a* and cyanobacteria abundance (Figure 11). These results suggest that trophic state is indeed an acceptable proxy for cyanobacteria anundance and that in lakes with higher trophic state it is also reasonable to expect higher cyanobacteria.

3.5 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 succesfully 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 accuraccy 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 certainity 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 *in situ* and GIS Models NEED MORE HERE For the GIS only models NEED MORE HERE - summarize from above on varibale importance

Associations between trophic state and cynobacteria show that NEED MORE HERE - need to summarize

4 Figures

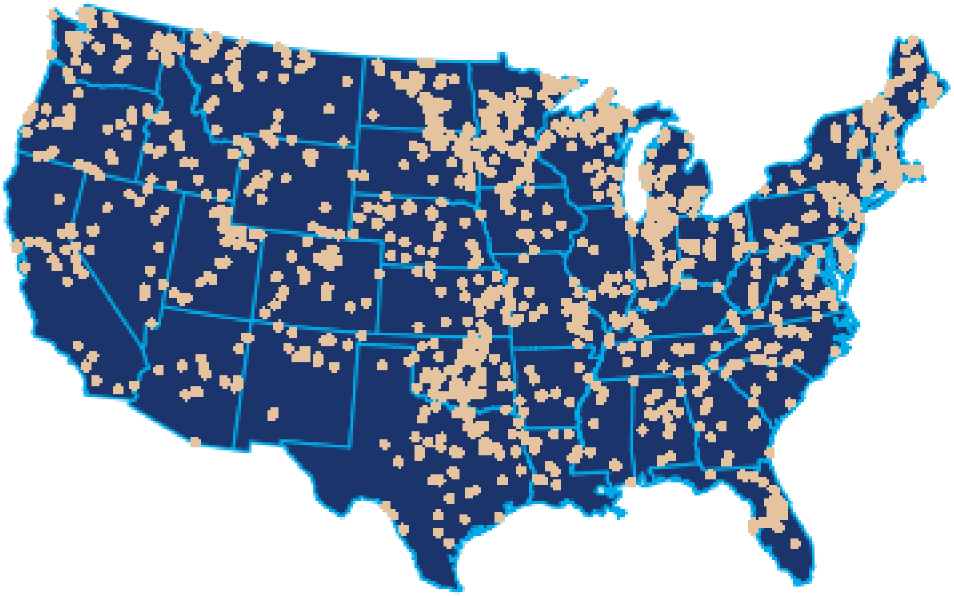

Figure 1: Map of the distribution of National Lakes Assesment Sampling locations

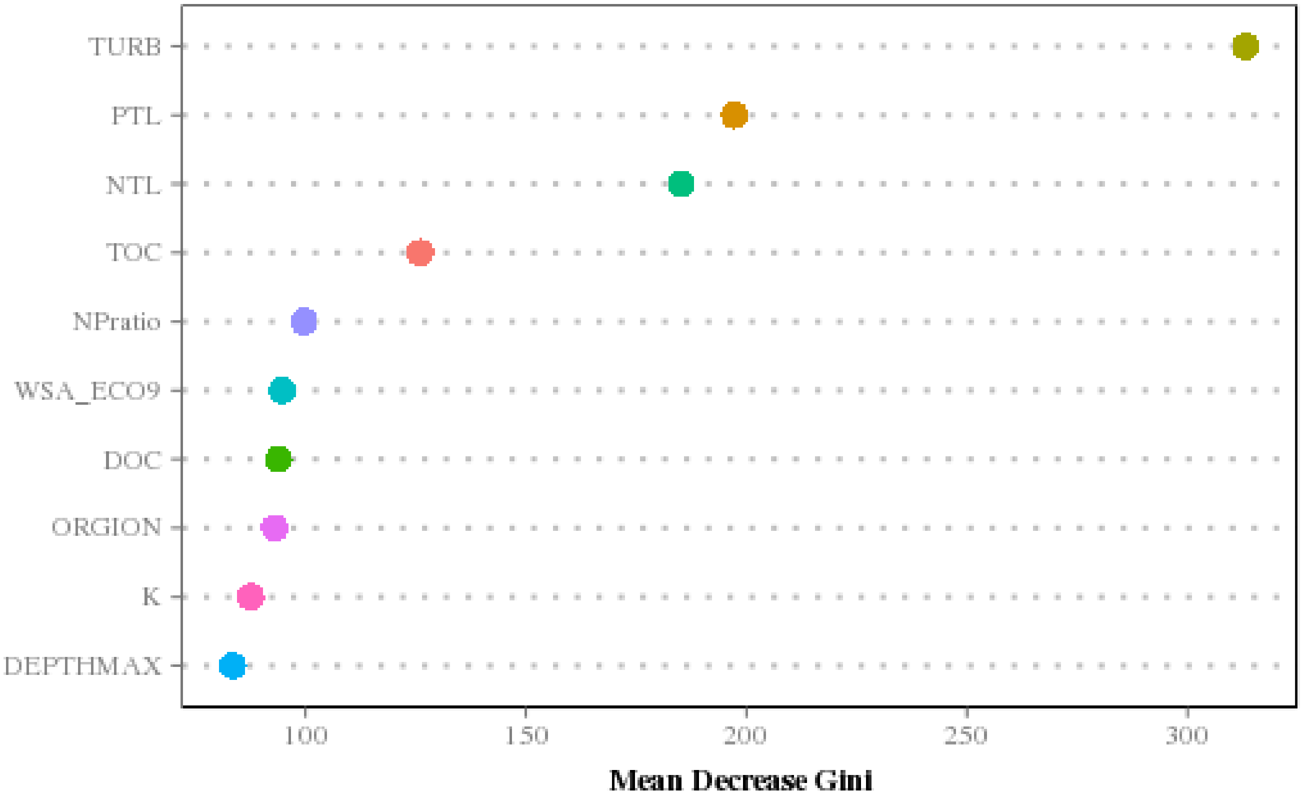

Figure 2: Importance plot for Model 1

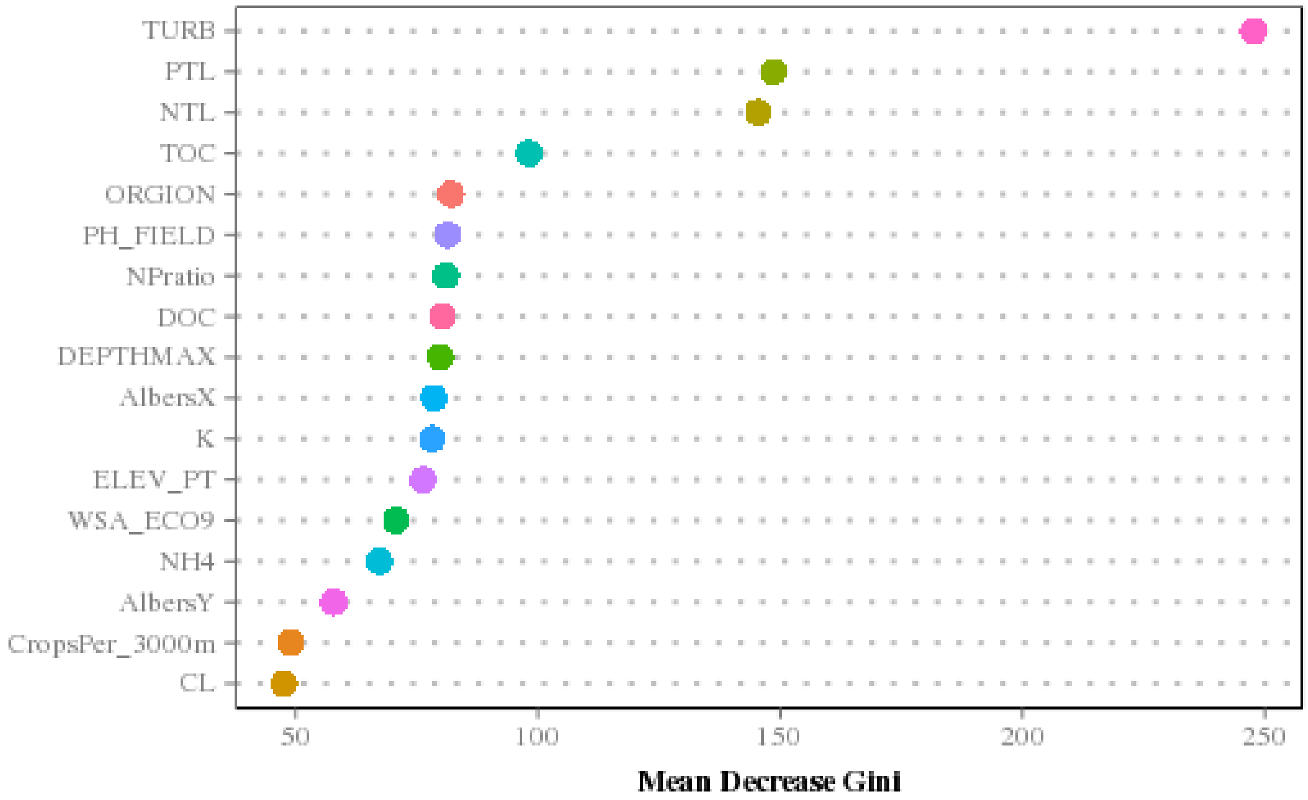

Figure 3: Importance plot for Model 2

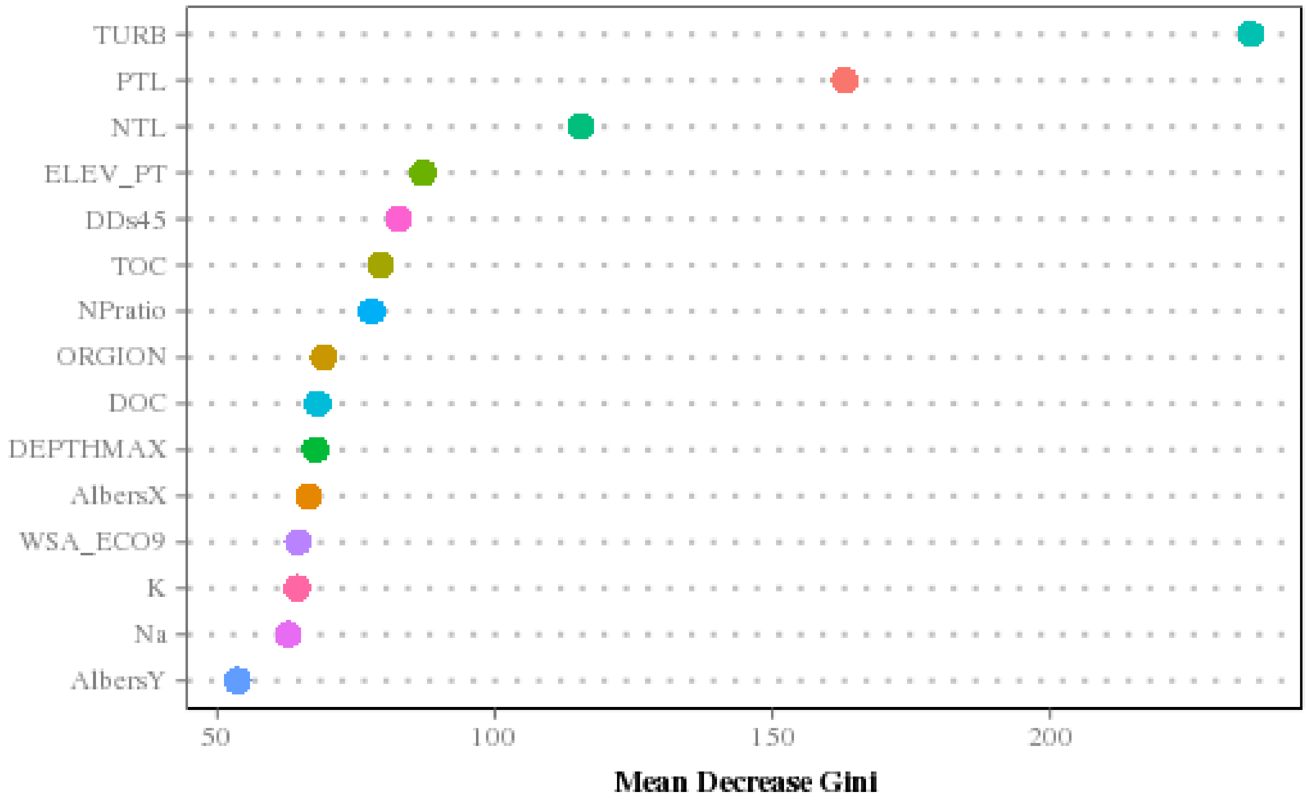

Figure 4: Importance plot for Model 3

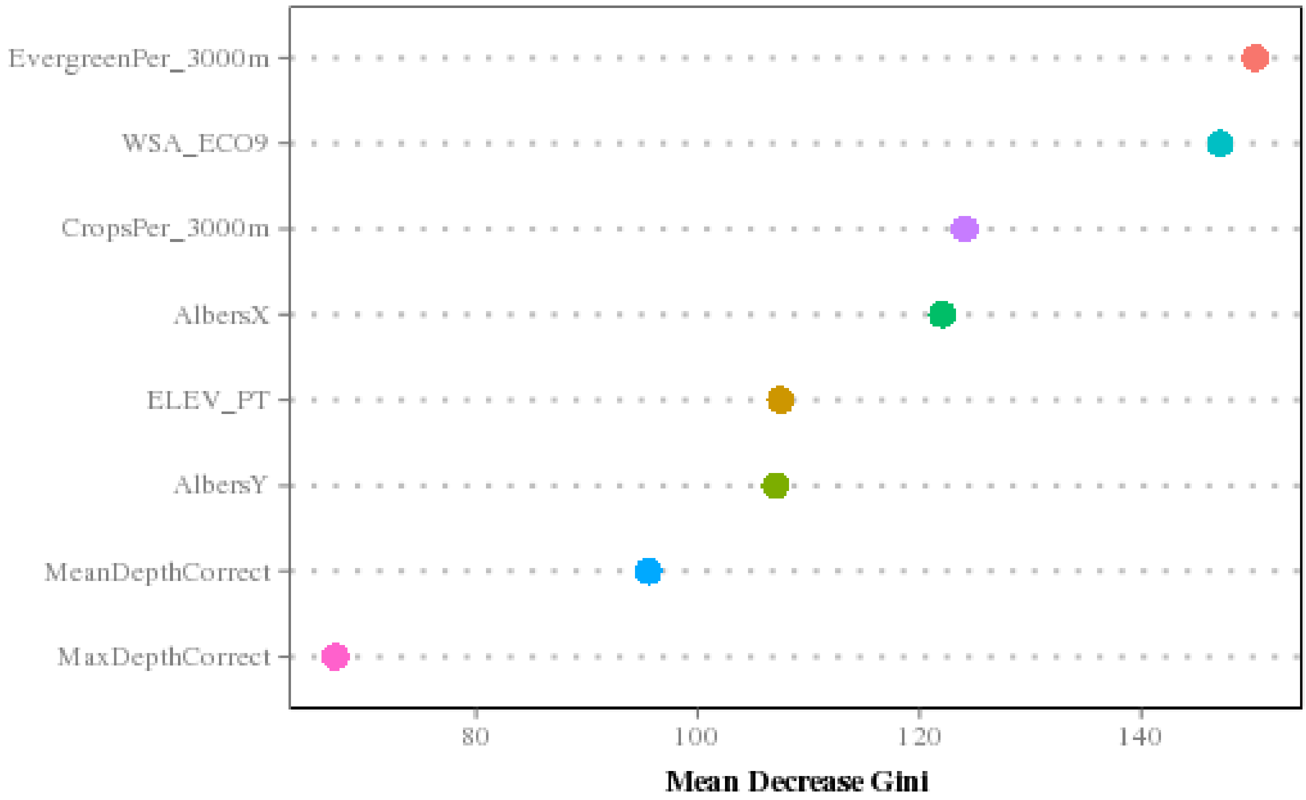

Figure 5: Importance plot for Model 4

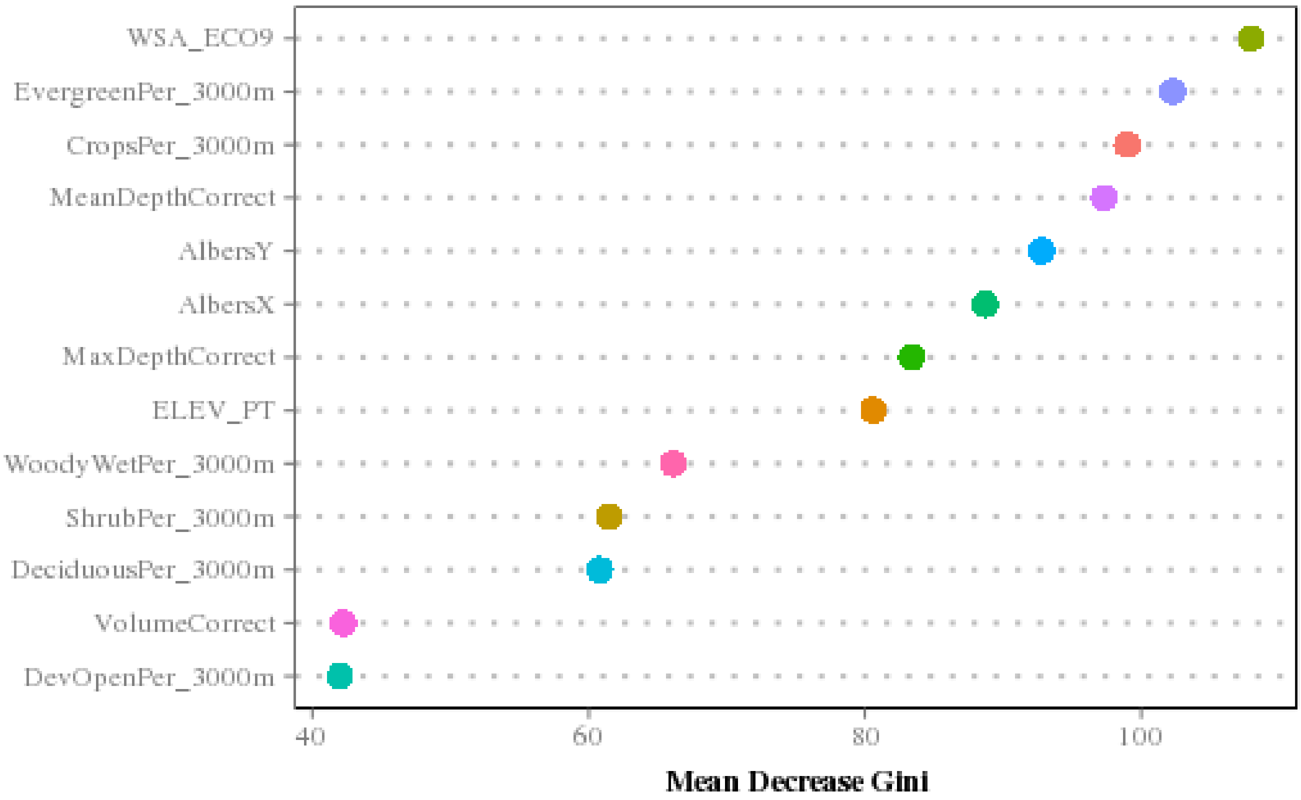

Figure 6: Importance plot for Model 5

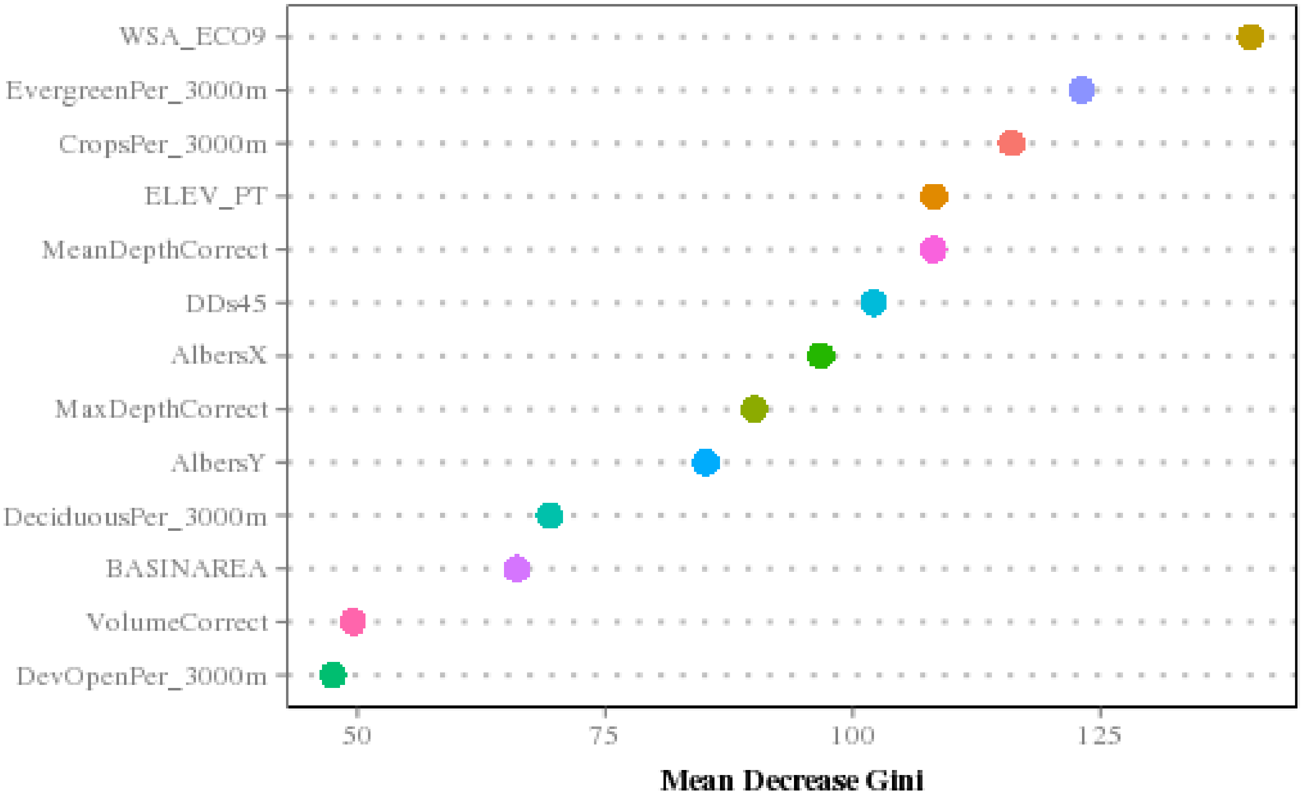

Figure 7: Importance plot for Model 6

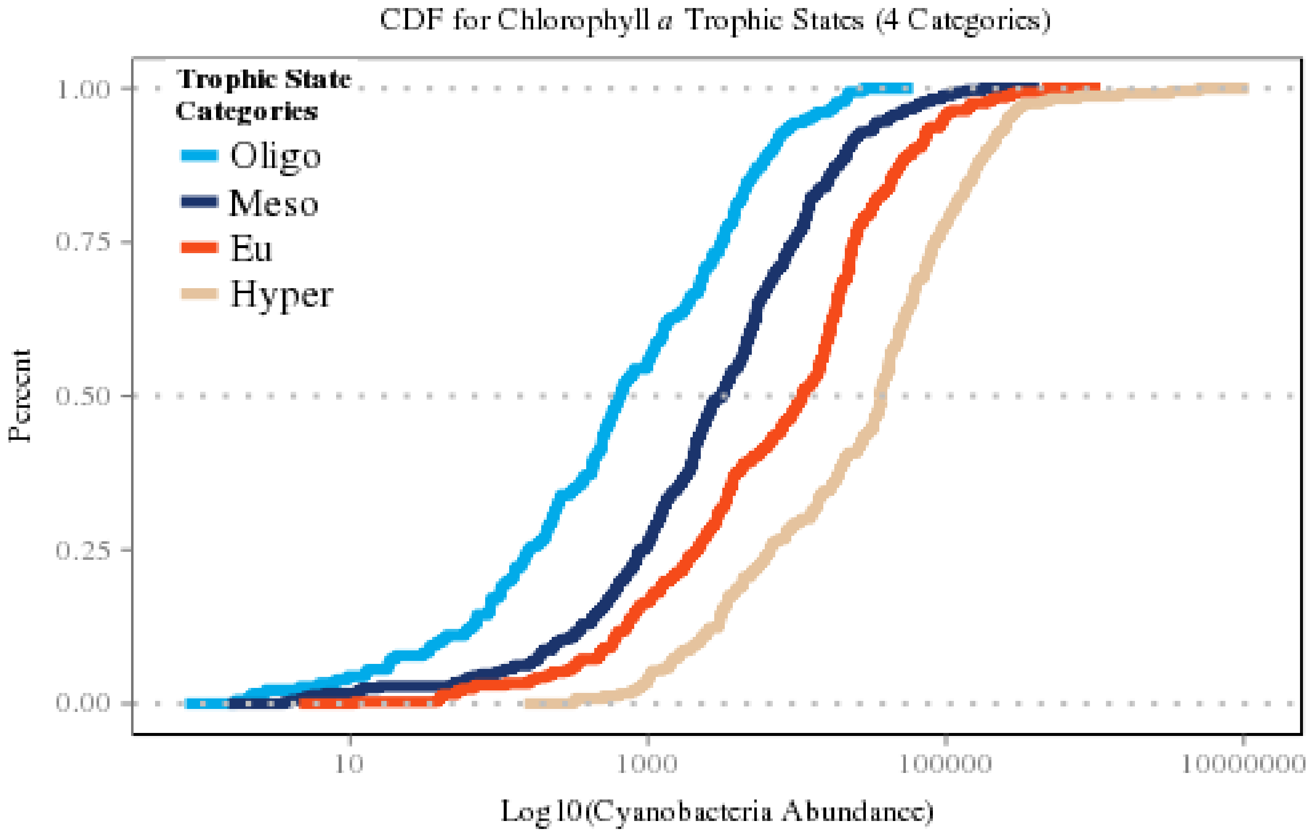

Figure 8: Cumulative distribution function of cyanobacetria abundance for 4 trophic state classes

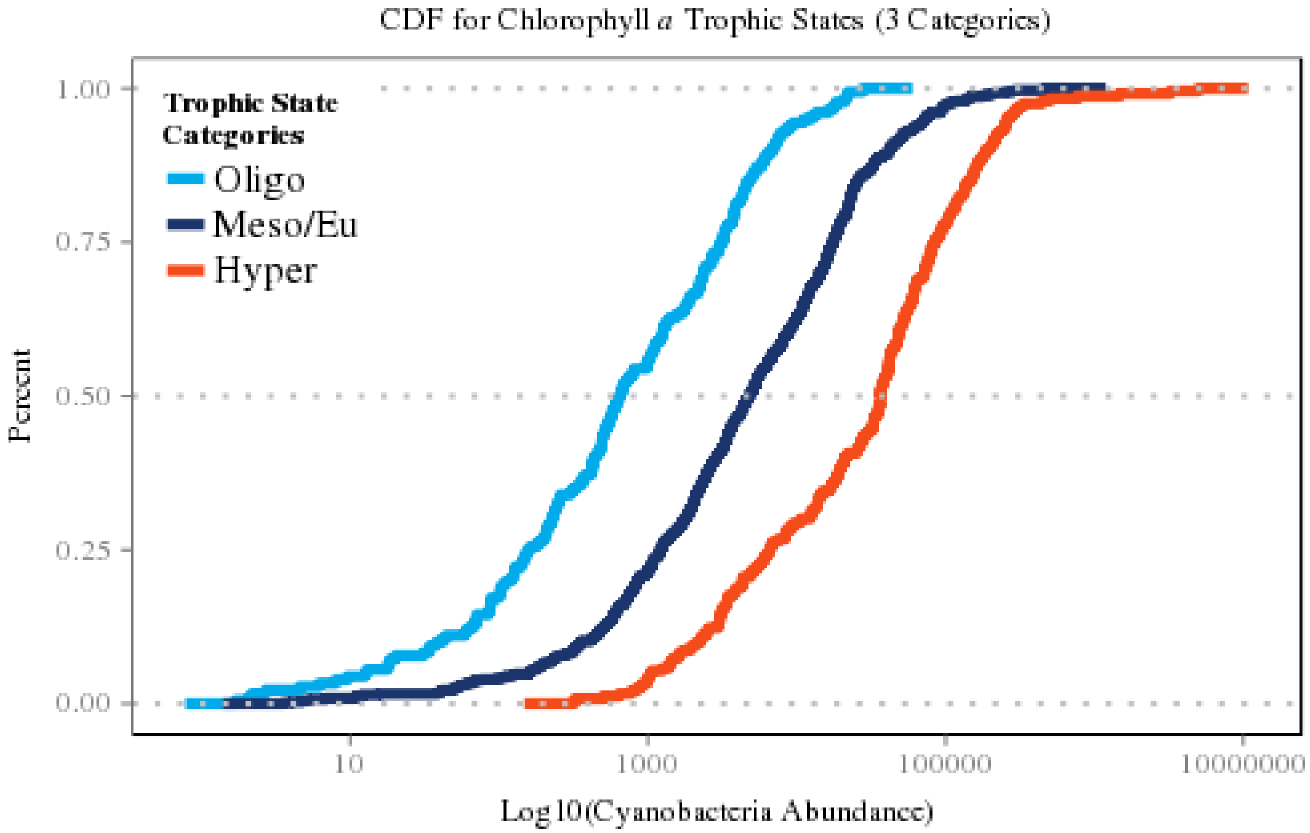

Figure 9: Cumulative distribution function of cyanobacetria abundance for 3 trophic state classes

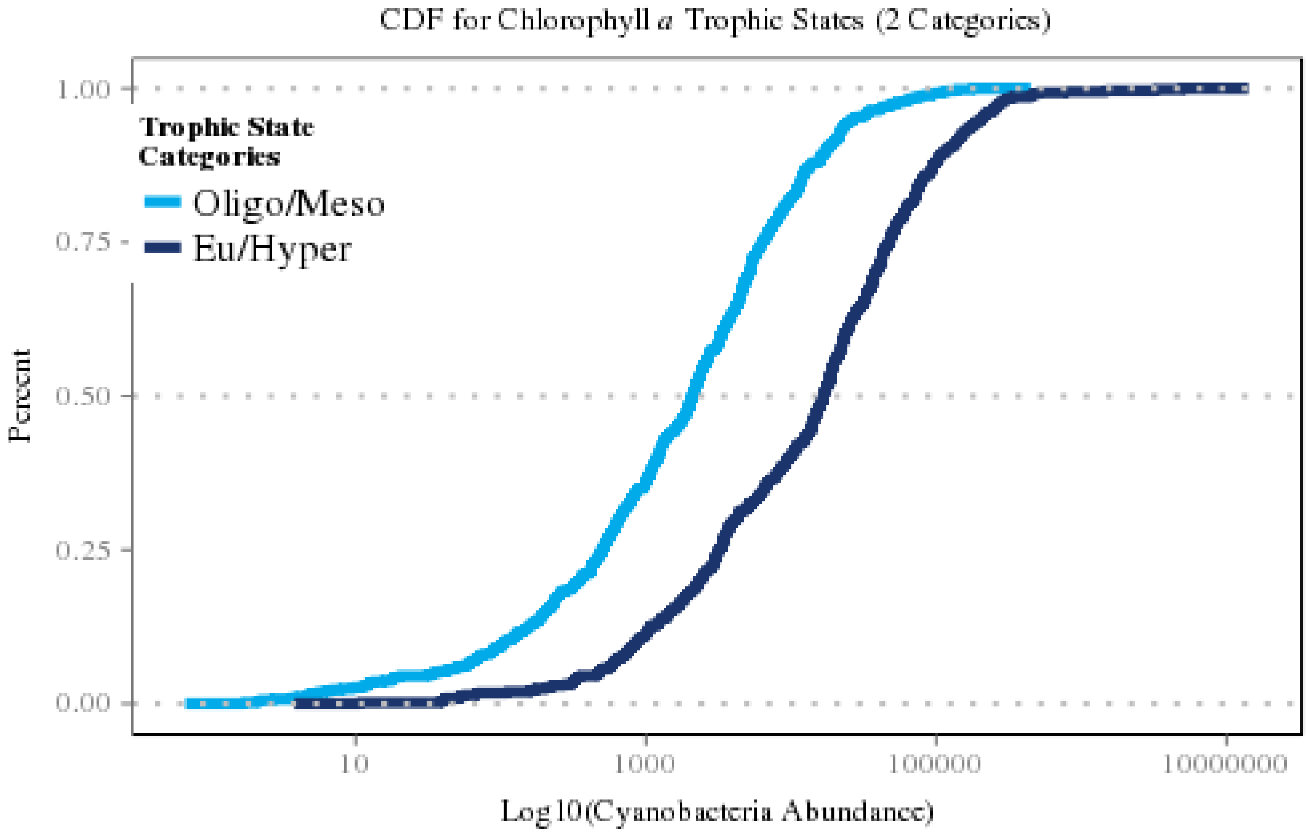

Figure 10: Cumulative distribution function of cyanobacetria abundance for 2 trophic state classes

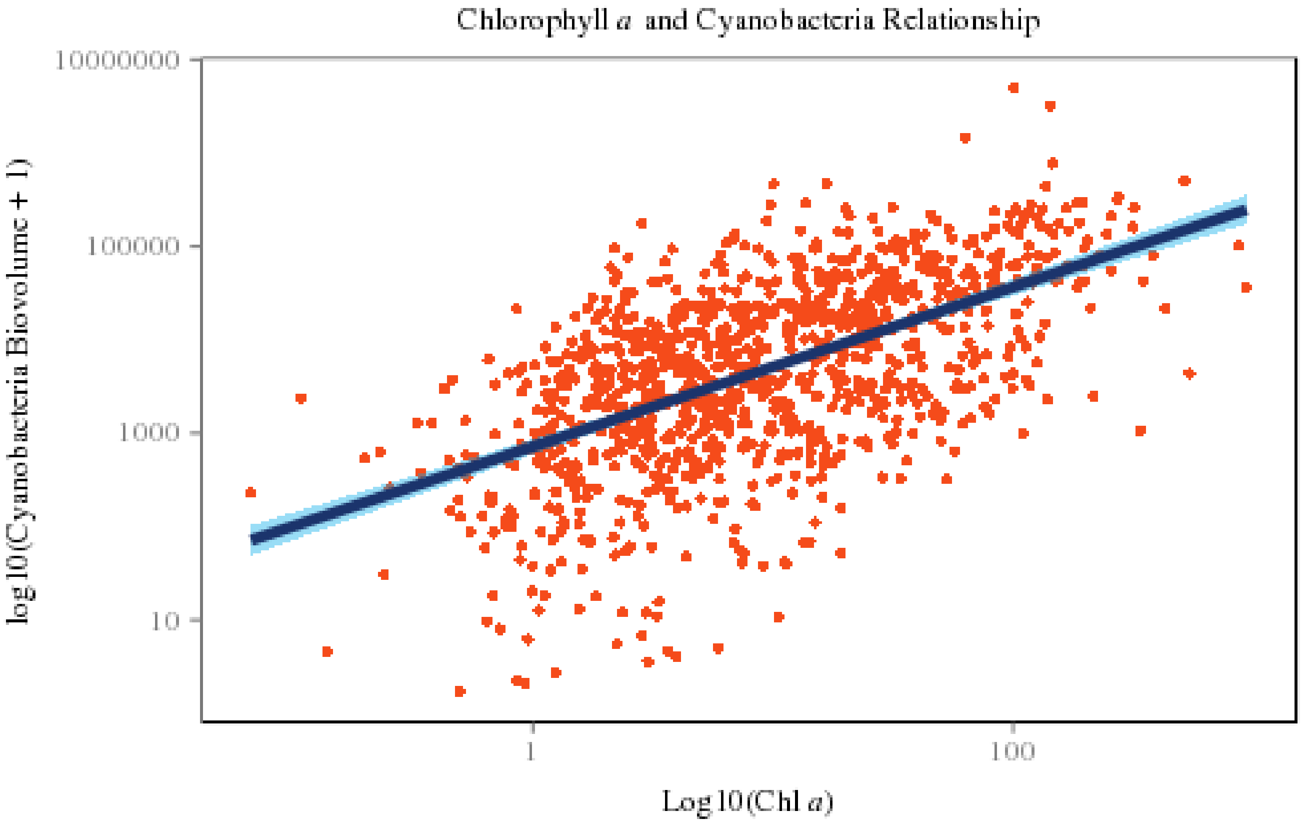

Figure 11: Cholorphyll *a* and cyanobacteria abundance scatterplot

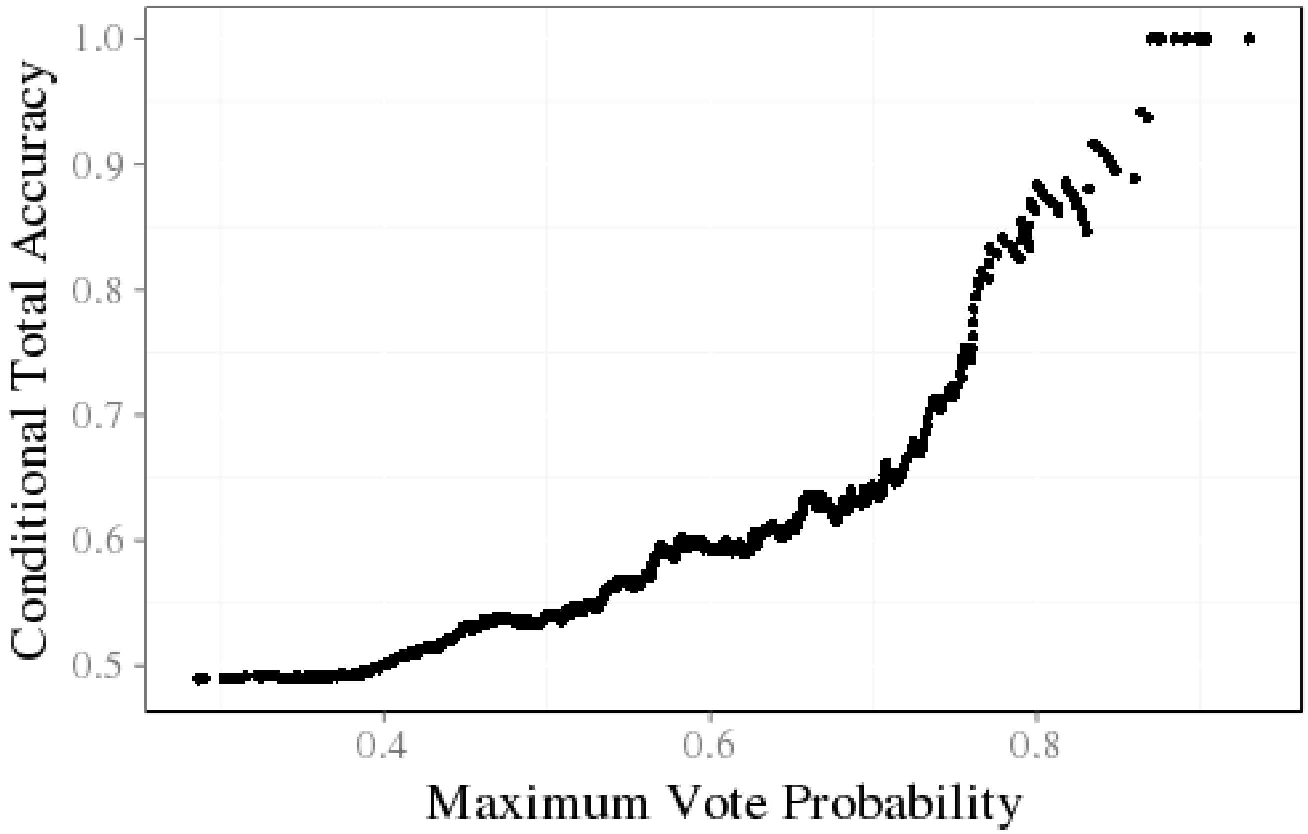

Figure 12: Comparison of certainity of trophic state prediction and total accuracy

5 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

Variable	Percent
NPratio	1.00
NTL	1.00
PTL	1.00
TOC	1.00
TURB	1.00
WSA_ECO9	1.00
K	0.99
ORGION	0.33
DOC	0.22
DEPTHMAX	0.11

Table 2: Variable selection results for Model 1

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 3: Random Forest confusion matrix for Model 1

Variable	Percent
DEPTHMAX	1.00
DOC	1.00
K	1.00
NTL	1.00
ORGION	1.00
PTL	1.00
TOC	1.00
TURB	1.00
WSA_ECO9	1.00
NPratio	0.88
AlbersX	0.58
CropsPer_3000m	0.36
ELEV_PT	0.23
NH4	0.06
AlbersY	0.04
CL	0.03
PH_FIELD	0.02

Table 4: Variable selection results for Model 2

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

Table 5: Random Forest confusion matrix for Model 2

Variable	Percent
K	1.00
NPratio	1.00
NTL	1.00
PTL	1.00
TOC	1.00
TURB	1.00
WSA_ECO9	1.00
ORGION	0.98
DEPTHMAX	0.91
DDs45	0.89
ELEV_PT	0.85
DOC	0.42
AlbersX	0.11
AlbersY	0.03
Na	0.03

Table 6: Variable selection results for Model 3

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

Table 7: Random Forest confusion matrix for Model 3

Variable	Percent
AlbersX	1.00
CropsPer_3000m	1.00
EvergreenPer_3000m	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
AlbersY	0.30
ELEV_PT	0.05
MaxDepthCorrect	0.01

Table 8: Variable selection results for Model 4

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 9: Random Forest confusion matrix for Model 4

Variable	Percent
AlbersX	1.00
AlbersY	1.00
CropsPer_3000m	1.00
EvergreenPer_3000m	1.00
MaxDepthCorrect	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
ELEV_PT	0.97
DeciduousPer_3000m	0.94
ShrubPer_3000m	0.32
WoodyWetPer_3000m	0.18
DevOpenPer_3000m	0.13
VolumeCorrect	0.11

Table 10: Variable selection results for Model 5

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

Table 11: Random Forest confusion matrix for Model 5

Variable	Percent
AlbersX	1.00
AlbersY	1.00
CropsPer_3000m	1.00
DDs45	1.00
ELEV_PT	1.00
EvergreenPer_3000m	1.00
MeanDepthCorrect	1.00
WSA_ECO9	1.00
MaxDepthCorrect	0.98
DeciduousPer_3000m	0.91
DevOpenPer_3000m	0.71
BASINAREA	0.33
VolumeCorrect	0.01

Table 12: Variable selection results for Model 6

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

Table 13: Random forest confusion matrix for Model 6

Oligo	Meso	Eu	Hyper
0.8127741	0.1858553	0.0013706	0

Table 14: Example Lake Random Forest Vote Results

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.

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.

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

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.

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

Jeff Hollister, B. M., and B. Kreakie. (n.d.). LakeTrophicModelling: Package to reproduce hollister et al. (2014) modeling lake trophic state: A data mining approach.

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.

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 aPPROACH1. 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.

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.