

# Modelling Lake Trophic State: A Random Forest Approach

Jeffrey W. Hollister<sup>\*</sup> <sup>1</sup> W. Bryan Milstead<sup>1</sup> Betty J. Kreakie<sup>1</sup>

<sup>1</sup>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

<sup>\*</sup> corresponding author: [hollister.jeff@epa.gov](mailto:hollister.jeff@epa.gov)

## Abstract

Productivity of lentic ecosystems is well studied and it is widely accepted that as nutrient inputs increase, productivity increases and lakes transition from lower trophic state (e.g. oligotrophic) to higher trophic states (e.g. eutrophic). These broad trophic state classifications are good predictors of ecosystem condition, services, and disservices (e.g. recreation, aesthetics, 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 address this, we take advantage of the availability of a large national lakes water quality database (i.e. the National Lakes Assessment), land use/land cover data, lake morphometry data, other universally available data, and apply modern data mining approaches to predict trophic state. Using this data and random forests, we first model chlorophyll *a*, then classify the resultant predictions into trophic states. The full model estimates chlorophyll *a* with both *in situ* and universally available data. The mean squared error and adjusted  $R^2$  of this model was 0.09 and 0.8, respectively. The second model (i.e. GIS only) uses universally available GIS data only. The mean squared error was 0.22 and the adjusted  $R^2$  was 0.48. The accuracy of the trophic state classifications derived from the chlorophyll *a* predictions were 69% for the full model and 49% for the “GIS only” model. Random forests extend the usefulness of the class predictions by providing prediction probabilities for each lake. This allows us to make trophic state predictions and also indicate the level of uncertainty around those predictions. For the full model, these predicted class probabilities ranged from 0.42 to 1. For the GIS only model, they ranged from 0.33 to 0.96. It is our conclusion that *in situ* data are required for better predictions, yet GIS and universally available data provide trophic state predictions, with estimated uncertainty, that still have the potential for a broad array of applications. The source code and data for this manuscript are available from <https://github.com/USEPA/LakeTrophicModelling>.

## 1 Introduction

Productivity in lentic systems is often categorized across a range of trophic 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 (Rodhe 1969). 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 for rapid assessment of a lakes biological productivity and identification of 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, beach 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 (i.e. from cyanobacteria) or other problems associated with cultural eutrophication. This type of information could be used for setting priorities for management and allow for more efficient use of limited resources.

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, 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 issue 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 regression or linear mixed models (Jones et al. 2001, 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, Fernández-Delgado et al. 2014). 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 two areas in which this research contributes. First, we build, assess, and compare two random forest models of chlorophyll *a* 1) *in situ* and universally available GIS data and then 2) universally available GIS data only. Second, we examine the important predictors for both models. Lastly, this paper, the code, and the data used in the models are made available as an R package from <https://github.com/USEPA/LakeTrophicModelling>.

## 2 Methods

### 2.1 Data and Study Area

We utilized three primary sources of data for this study, the National Lakes Assessment (NLA), the National Land Cover Dataset (NLCD), and lake morphometry modeled from the NHDPlus and National Elevation Data Set (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 extent 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 for detailed description of methods). With consistent methods and metrics collected

at over 1000 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 only use 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 each lake (Homer et al. 2004, Xian et al. 2009). A three kilometer buffer was selected as an intermediate measure of the adjacent neighborhood; the three kilometer buffer size is greater than the immediate parcel but smaller than regional and whole-basin 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). These 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 a predetermined number of decision trees are developed. With each new tree, the sample data subset is randomly selected and with each new split, the subset of predictor variables are randomly selected. A detailed discussion of the benefits of a random forest approach is beyond the scope of this paper. To find out more see Breiman (2001) and Cutler et al. (2007).

Random forests are able to handle numerous correlated variables without a decrease in prediction accuracy; however, one possible shortcoming of this approach is that the resulting model may be difficult to interpret, thus selecting the most important variables is an important first step. Several methods have been proposed to do this with random forest. For instance, 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 the `varSelRF` package (Díaz-Uriarte and De Andres 2006), but this is limited to classification problems. Additionally, others have suggested alternative variable importance measures, but this is only needed with a large number of categorical variables which are selected against with traditional random forest approach (Strobl et al. 2007).

In our case, we are predicting a continuous variable, chlorophyll *a*, directly thus `varSelRF`, does not apply, and all of our variables are continuous so the approach suggested by Strobl (2007) is not necessary. Thus we developed an approach, similar to `varSelRF` but applied to random forest with regression trees. With this approach we fit a full random forest model that includes all variables and a large number of trees. We then rank the variables using the increase in mean square error, which has been shown to be a less biased metric of importance than the mean decrease in the gini coefficient (Strobl et al. 2007). Using this ranking, we then iterate through the variables and create a random forests with the top two variables and record mean square error and adjusted  $R^2$  of the resultant random forest. We then repeat this process by adding the next most important variable in order of importance. With this information we identify the top variables and the point at which adding variables does not improve the fit of the overall model. These variables are selected and used as the “reduced model.” 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.

## 2.3 Model Details

Using `randomForest` R package we ran models to predict chlorophyll *a* with two sets of predictors (Liaw and Wiener 2002); all predictors (*in situ* and universally available GIS predictors) and the GIS only variables (i.e. no *in situ* information). A list of the full suite of variables tested is in Appendix 1. Our separation of predictors was chosen so that we could highlight the additional predictive performance

provided by adding the *in situ* water quality variables on top of the GIS only variables. Lastly, we used only complete cases (i.e. missing data were removed) so the total number of observations varied among models.

Our modelling work flow was as follows:

1. Identify a minimal set of variables that maximize accuracy of the random forest algorithm. This minimal set of variables, the reduced model, is calculated for each of the models.
2. Using R's `randomForest` package, we develop two random forest models (All Variables and GIS only).
3. Assess model performance for both the predicted chlorophyll *a* and for categorical trophic state classifications. Trophic state was defined using the NLA chlorophyll *a* trophic state cut offs (Table 1).

## 2.4 Measures of Model Performance and Variable Importance

We assessed the performance of the random forest two ways. First we compare the root mean square error and the adjusted  $R^2$  of the models. Second, we examine the accuracy of the model predictions when converted to trophic states classes (Table 1). We assess the classifications via a confusion matrix. A confusion matrix shows agreement and disagreement in a tabular form with predicted values forming the columns of the matrix and observed values, the rows. From this tabulated information we calculate the total accuracy (i.e. percent correctly predicted) and the kappa coefficient, which takes into account the error expected by chance alone (i.e. the off diagonal values of the matrix) (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). Negative values are rare and would indicate no agreement between the predicted and observed values. Additionally, random forest builds each tree on bootstrapped, random subsets of the original data, thus, a separate independent validation dataset is not required and random forest error estimates are expected to be unbiased (Breiman 2001).

170 The random forest algorithm explicitly measures variable importance with two metrics: mean decrease  
 171 in Gini and percent increase in mean squared error. For each of these they measure the impact on  
 172 the overall model when that particular variable is included and thus can be used to assess importance  
 173 (Breiman 2001). The Gini Index has been shown to have a bias that is less apparent than with percent  
 174 increase in mean squared error (Strobl et al. 2007), thus, we use this metric to assess variable importance.  
 175 Lastly, partial dependence plots provide a mechanism to examine the partial relationship between  
 176 individual variables and the response variable (Jones and Linder 2015). We examine these plots for the  
 177 top variables as assigned by percent increase in mean squared error of for each the reduced models.

## 178 2.5 Trophic State Probabilities

179 One of the powerful features of random forests is the ability to aggregate a very large number of  
 180 competing models or trees. Each tree provides an independent prediction or vote for a possible outcome.  
 181 In the context of our chlorophyll *a* models, we have 5,000 estimates of chlorophyll *a* for each lake. We  
 182 convert these values to trophic states (Table 1) then count up total votes for each class and divide by  
 183 total possible votes to get an estimate of the probability that a lake is in a given trophic state. For  
 184 instance, for a single lake (National Lake Assessment ID = NLA06608-0005), the vote probabilities for  
 185 the “All Variables” model were 95% for oligotrophic, 5% for mesotrophic, 0% for eutrophic, and 0% for  
 186 hypereutrophic. The maximum probability provides the predicted class, in this case oligotrophic, and  
 187 suggests little uncertainty in this prediction. We refer to this value as the “prediction probability.”

188 Further, we might expect higher total accuracy for lakes that have more certain predictions. This should  
 189 be evident by looking at the total classification accuracy of lakes given their prediction probability is  
 190 above a certain point. To test this we use an approach similar to one outlined by Paul and MacDonald  
 191 (2005) and implemented by Hollister et al. (2008). We utilize this approach and examine the change in  
 192 total accuracy as a function of the prediction probability for both models.

### 3 Results

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: All Variables

The model built with all predictors had a mean squared error of 0.09 and and  $R^2$  of 0.8. The accuracy of the four trophic states was 68.7% and the kappa coefficient was 0.57 (Table 2). The reduced, all variables model was built using 1080 total observations. The variable selection process identified 20 variables (Figure 2). The five most important variables were turbidity, total phosphorus, total nitrogen, elevation, and total organic carbon (Figures 3 & 4).

#### 3.2 Models: GIS Only Variables

The GIS only model had a mean squared error of 0.22 and and  $R^2$  0.48. Four trophic states were predicted with a total accuracy of 49% and had a kappa coefficient of 0.29 (Table 3). The gis only variables model was built using 1138 total observations. The variable selection process for this model produced a reduced model with 15 variables (Figure 5). The five most important variables were ecoregion, percent cropland, elevation, latitude, and percent evergreen forest (Figures 6 & 4).

#### 3.3 Trophic State Probabilities

The “All Variables” model provides more certain model prediction than the “GIS Only” model with a median prediction probability of 0.81 versus 0.72 (Figure 8). Additionally, total accuracy of the predictions is a function of this uncertainty. Lakes with more certain predictions are more accurately classified (Figure 9). For both models, when prediction probabilities are approximate 0.7 or higher, the



models have an accuracy of 100%. This represents 81% of the lakes for the “All Variables” model and 60% of the lakes for the “GIS Only” model.

The trophic state predictions may be mapped discretely, yet this neglects to provide details on the uncertainty of those predictions. It may be preferable to instead map the prediction probabilities for each of the four classes and given the association between low uncertainty and high accuracy a map of this sort shows the broad spatial patterns of lake trophic state across the United States. The spatial patterns show little variability between the two models, thus we only show the results from the more broadly applicable “GIS only” model (Figure 10). Hypereutrophic lakes are much more commonly predicted in the midwest and southeastern United States. Clear, oligotrophic lakes are in the northwestern United States, throughout the western mountains and in the northeastern United States. The middle trophic states are more evenly distributed across the country.

## 4 Discussion

[JEFF START HERE]

[MOVE SOME OF THIS TO METHODS, CHECK FOR OVERLAP IN RESULTS, DETERMINE WHAT SHOULD GO IN DISCUSS]

### 4.1 Trophic State Probabilities

As expected, lakes with higher maximum vote probabilities were more accurately predicted (Figure ??). The increasing trend suggests that even for models with lower overall accuracy there can also be a large number of individual cases that are predicted with high accuracy.

### 4.2 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 modelling of cyanobacteria abundance and not surprisingly, 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 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.

Our models with the GIS-only variables captured the large scale spatial pattern of the trophic status gradient of lakes across the United States. We reliably saw 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 were also important in describing trophic state patterns. Like elevation and growing degree days, broad scale spatial patterns are 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 likely provide 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 likely shows the importance of in-lake nutrient processing and residence time.

### 4.3 Associating Trophic State and Cyanobacteria

Cyanobacteria biomass should be closely associated with trophic state as cyanobacteria contribute to the chlorophyll concentration in a lake. 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 can be used to differentiate cyanobacteria abundance, we examine distribution of cyanobacteria abundance for each trophic state and also explored linear associations between chlorophyll *a* and cyanobacteria abundance.

The distribution of cyanobacteria abundance showed separation between all of the trophic state classifications (Figures ??, ??, and ??) and there was a significant linear relationship ( $r^2=0.33$ ) between chlorophyll *a* and cyanobacteria abundance (Figure ??). Furthermore, 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 cyanobacteria) 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.

## 5 Conclusions

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

While some of the models (i.e. Model 4) showed 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 indicated 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

288 predicted class. Furthermore, as the certainty of prediction increases, so does overall trophic state  
 289 classification accuracy (Figure ??). These results suggest that our models will provide reasonable  
 290 estimates of trophic state across the United States.

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

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

## 311 6 Acknowledgements

312 We would like to thank Farnaz Nojavan, Nathan Schmucker, John Kiddon, Joe LiVolsi, Tim Gleason,  
 313 and Wayne Munns for constructive reviews of this paper. This paper has not been subjected to Agency

314 review. Therefore, it does not necessary reflect the views of the Agency. Mention of trade names or  
315 commercial products does not constitute endorsement or recommendation for use. This contribution is  
316 identified by the tracking number ORD-011075 of the Atlantic Ecology Division, Office of Research  
317 and Development, National Health and Environmental Effects Research Laboratory, US Environmental  
318 Protection Agency.

319 **7 Figures**

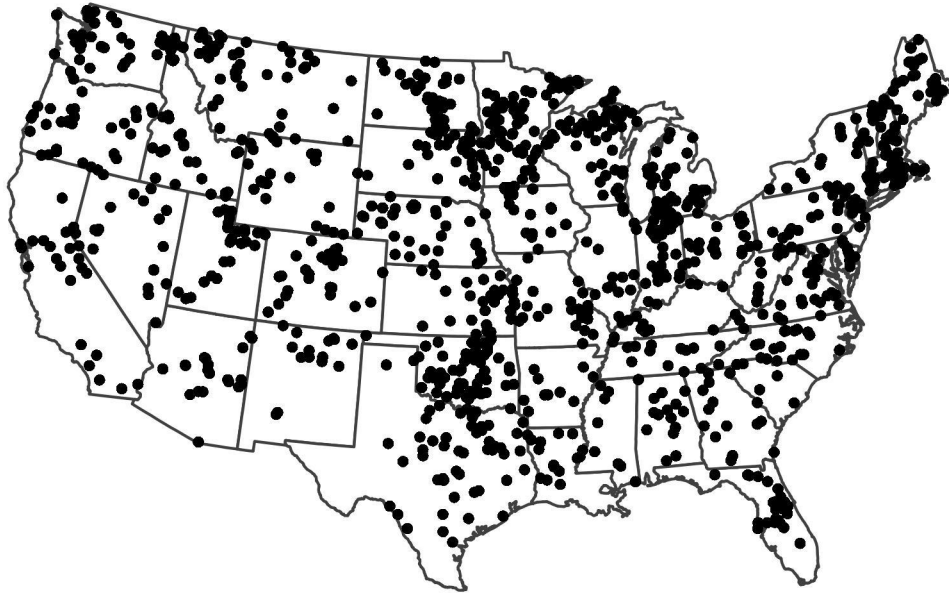


Figure 1: Map of the distribution of National Lakes Assessment Sampling locations

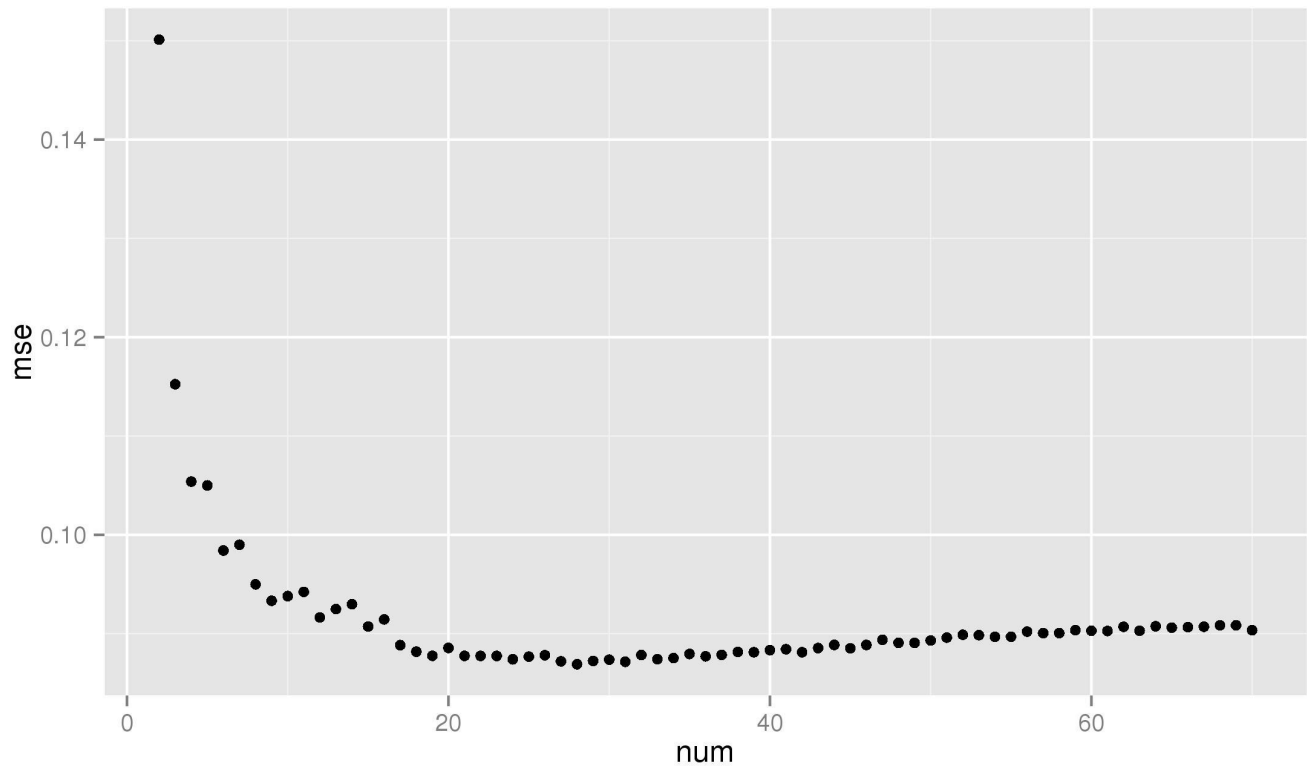


Figure 2: Variable selection plot for all variables. Shows percent increase in mean squared error as a function of the number of variables.

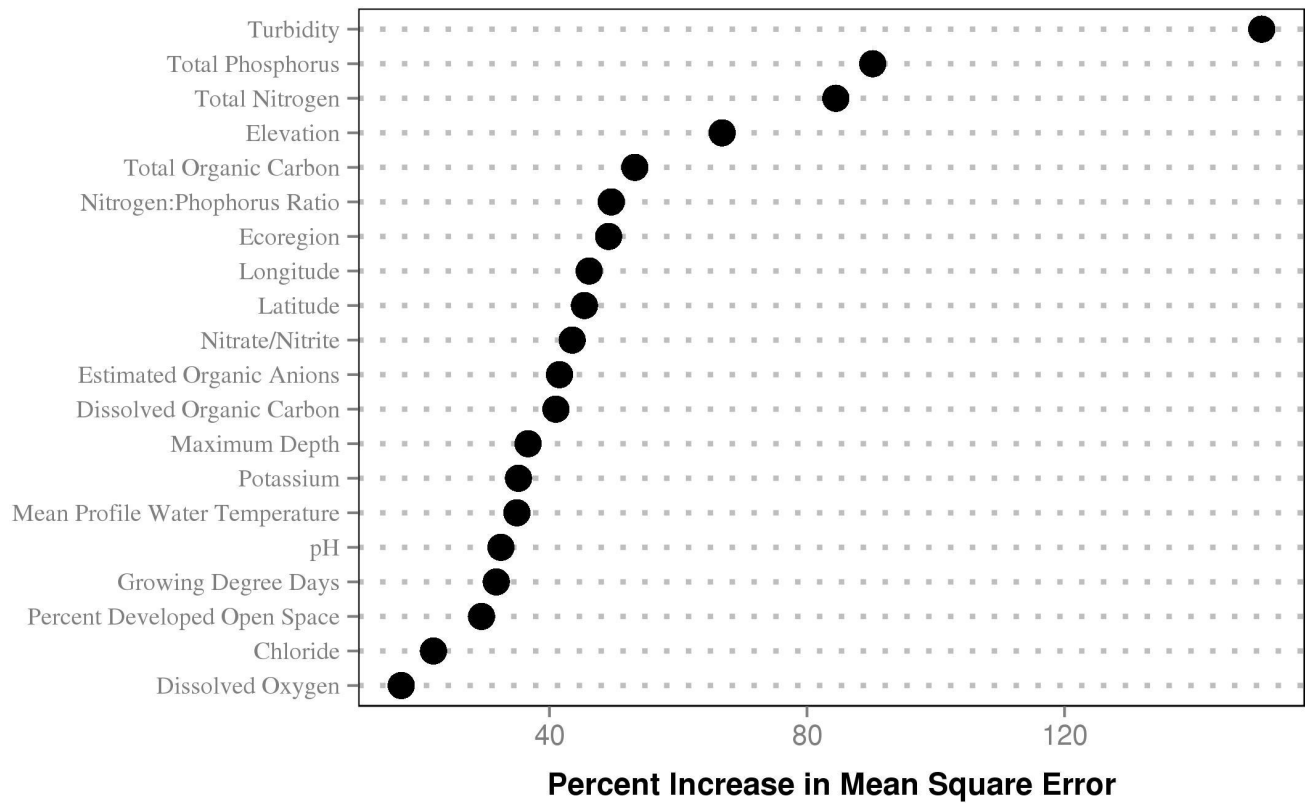


Figure 3: Importance plot for All Variables., shows percent increase in mean square error. Higher values of percent increase in mean squared error indicates higher importance.



320 **8 Tables**

Table 1: Chlorophyll a based trophic state cut-offs.

| Trophic State (4 class) | Trophic State (2 class)  | Concentration Cut-off |
|-------------------------|--------------------------|-----------------------|
| oligotrophic            | oligotrophic/mesotrophic | $\leq 2$              |
| mesotrophic             | oligotrophic/mesotrophic | $>2-7$                |
| eutrophic               | eutrophic/hypereutrophic | $>7-30$               |
| hypereutrophic          | eutrophic/hypereutrophic | $>30$                 |

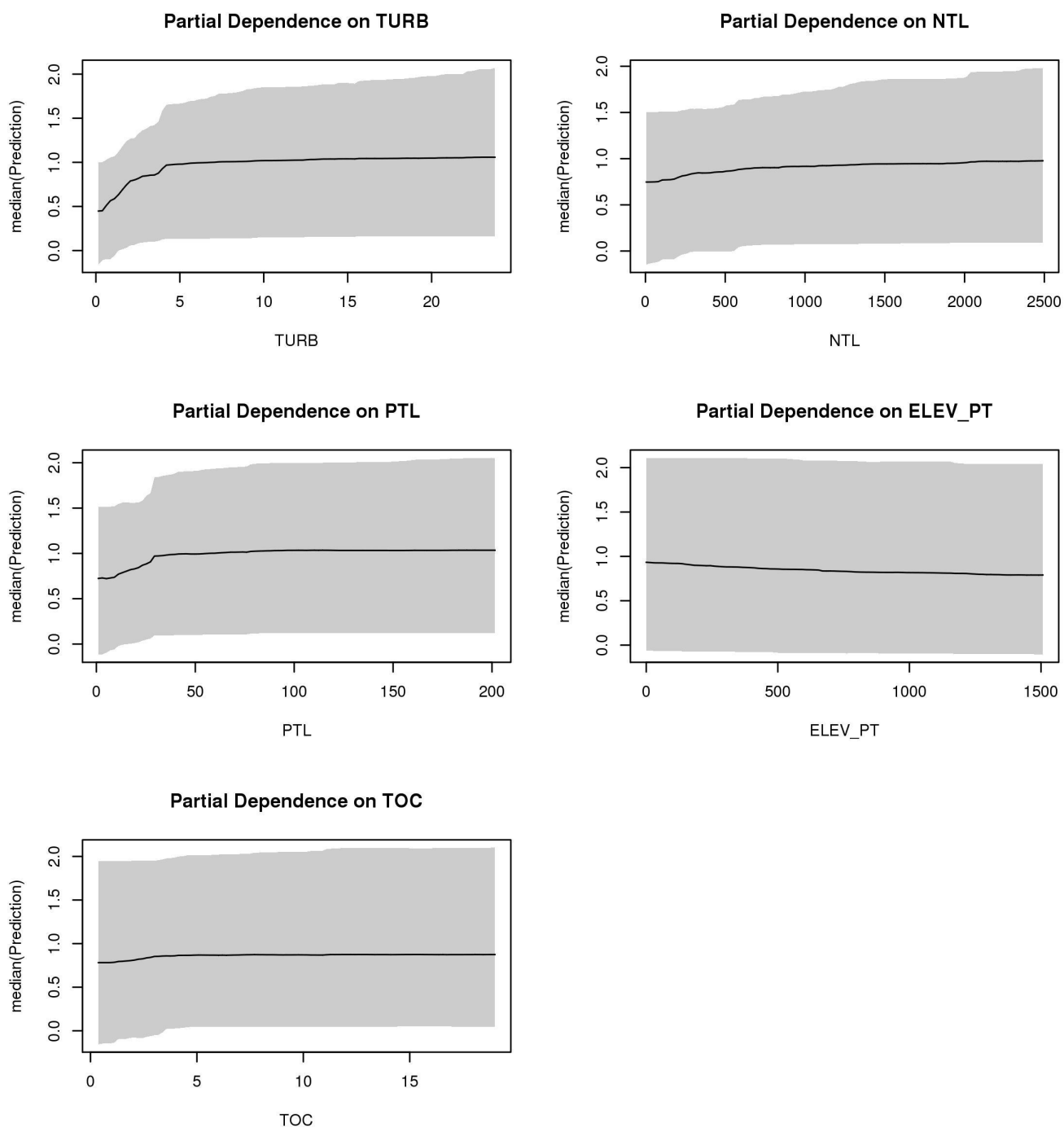


Figure 4: “All Variables” Partial dependence plots for the top 5 most important variables.

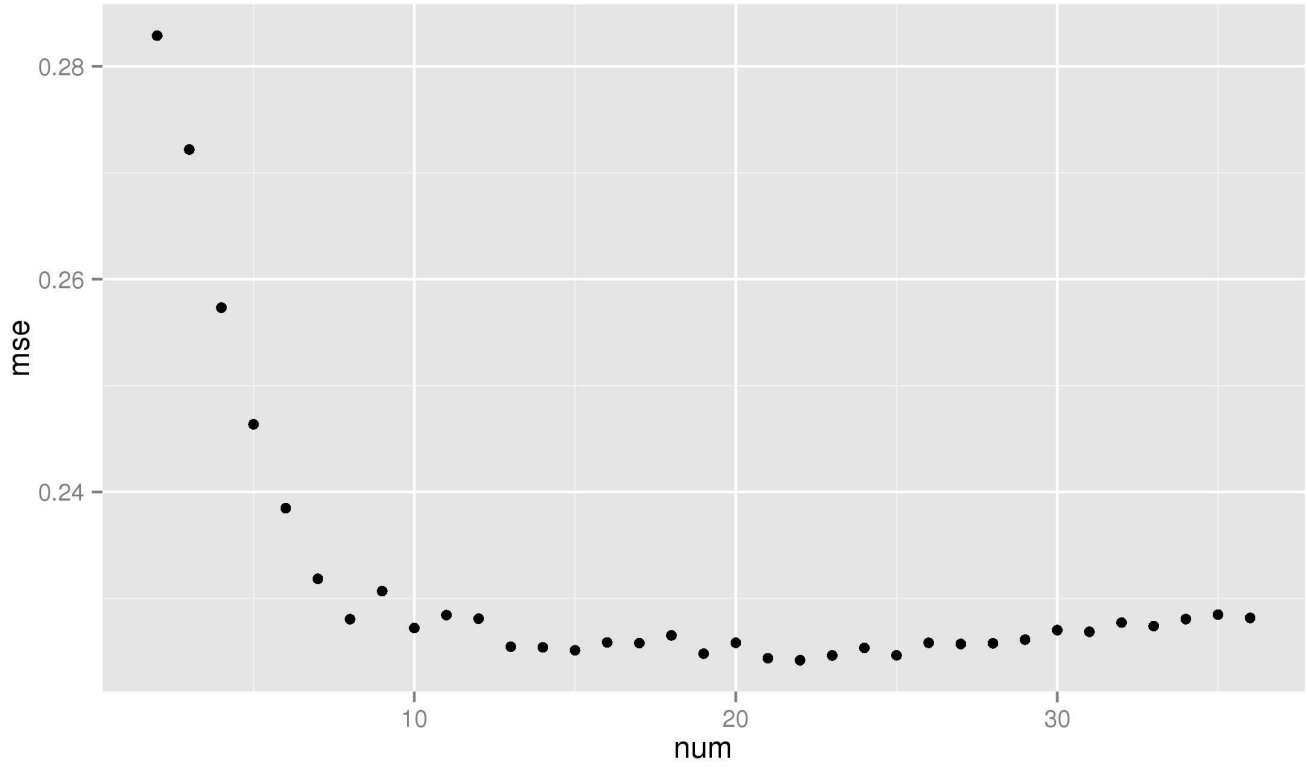


Figure 5: Variable selection plot for GIS only variables. Shows percent increase in mean squared error as a function of the number of variables.

Table 2: Random Forest confusion matrix for All Variables model converted to 4 trophic states. Columns show predicted values and rows show observed values. Agreement indicated on diagonal and accuracy for each trophic state indicated in 'Class Accuracy' column.

|       | oligo | meso | eu  | hyper | Class Accuracy (%) |
|-------|-------|------|-----|-------|--------------------|
| oligo | 115   | 31   | 0   | 0     | 78.77              |
| meso  | 67    | 251  | 63  | 0     | 65.88              |
| eu    | 7     | 61   | 217 | 75    | 60.28              |
| hyper | 0     | 5    | 29  | 159   | 82.38              |

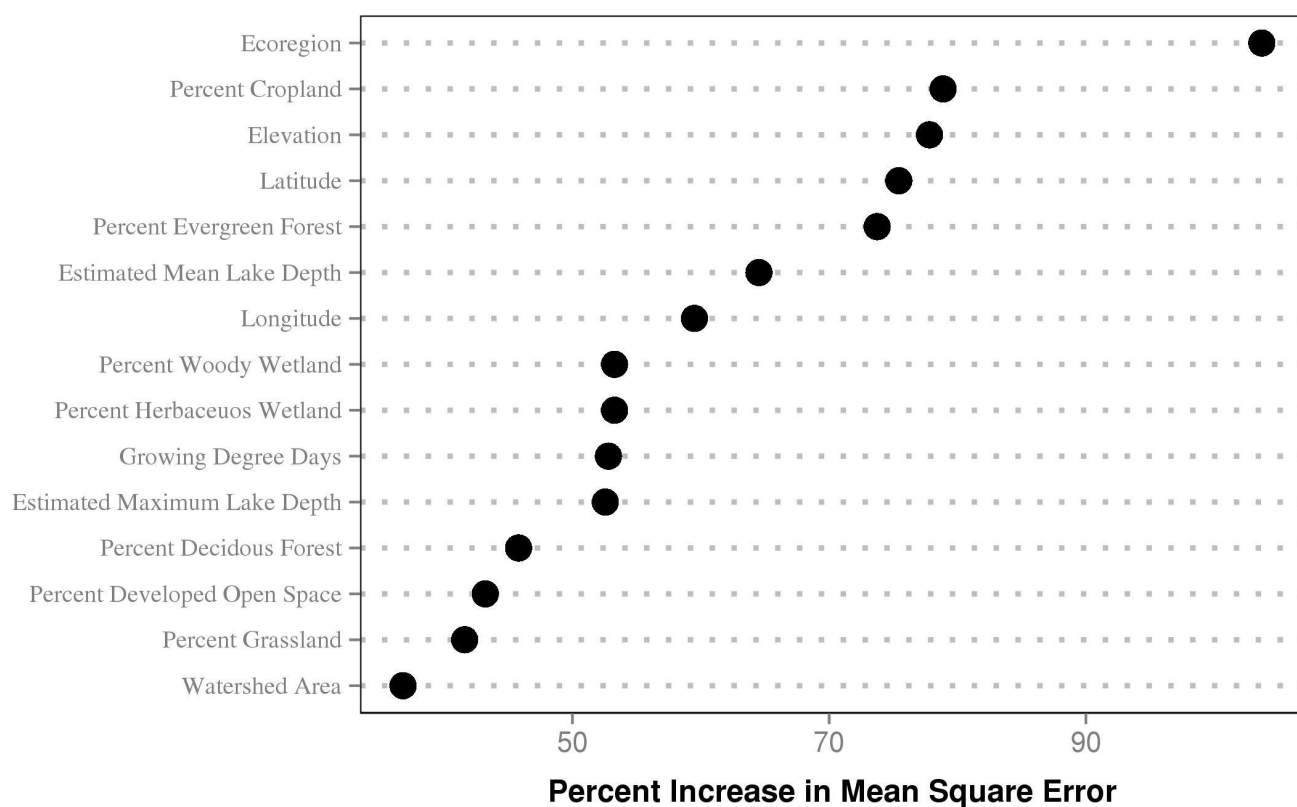


Figure 6: Importance plot for GIS Only Variables., shows percent increase in mean square error. Higher values of percent increase in mean squared error indicates higher importance.

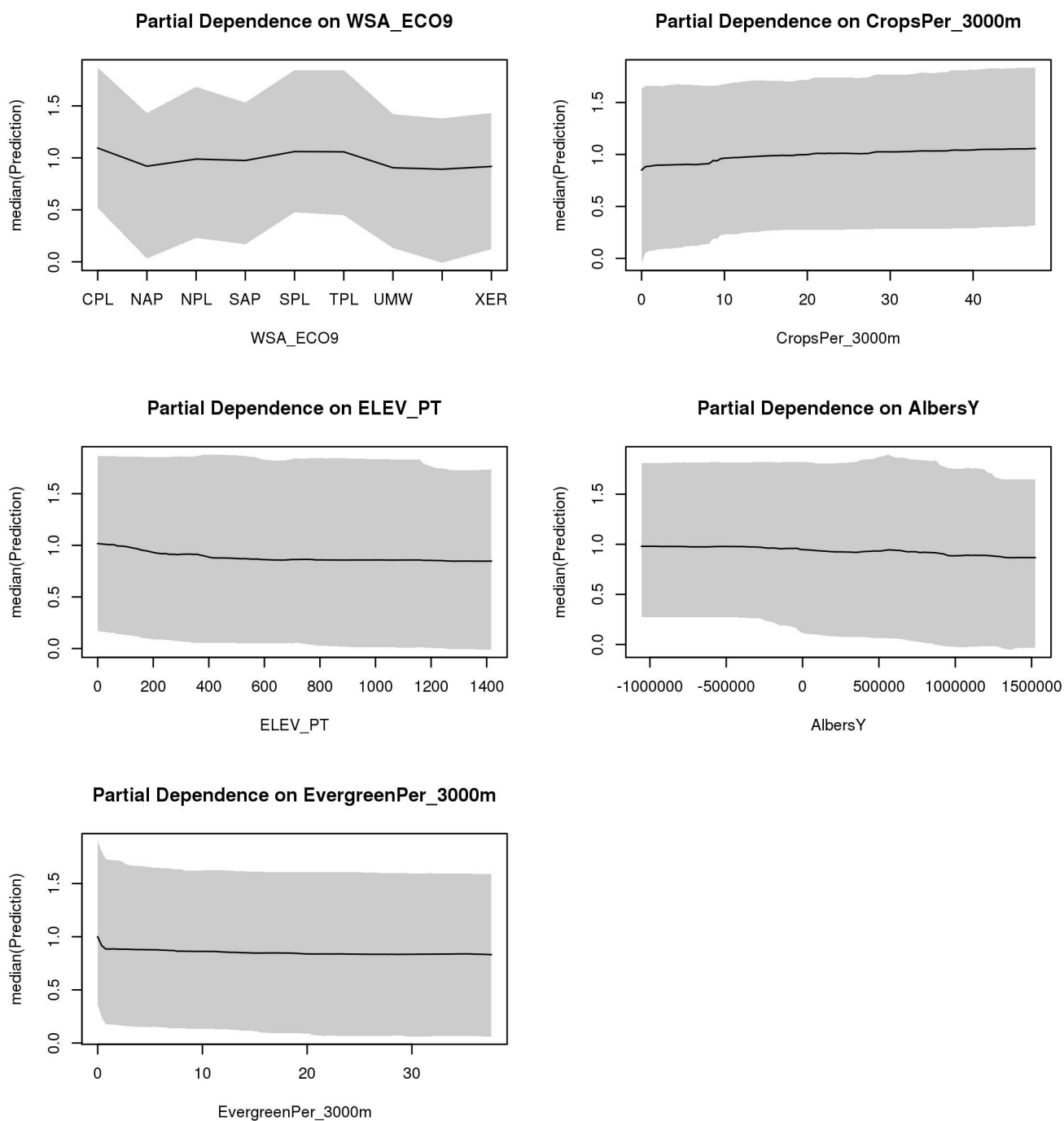


Figure 7: “GIS Only Variables” Partial dependence plots for the top 5 most important variables.

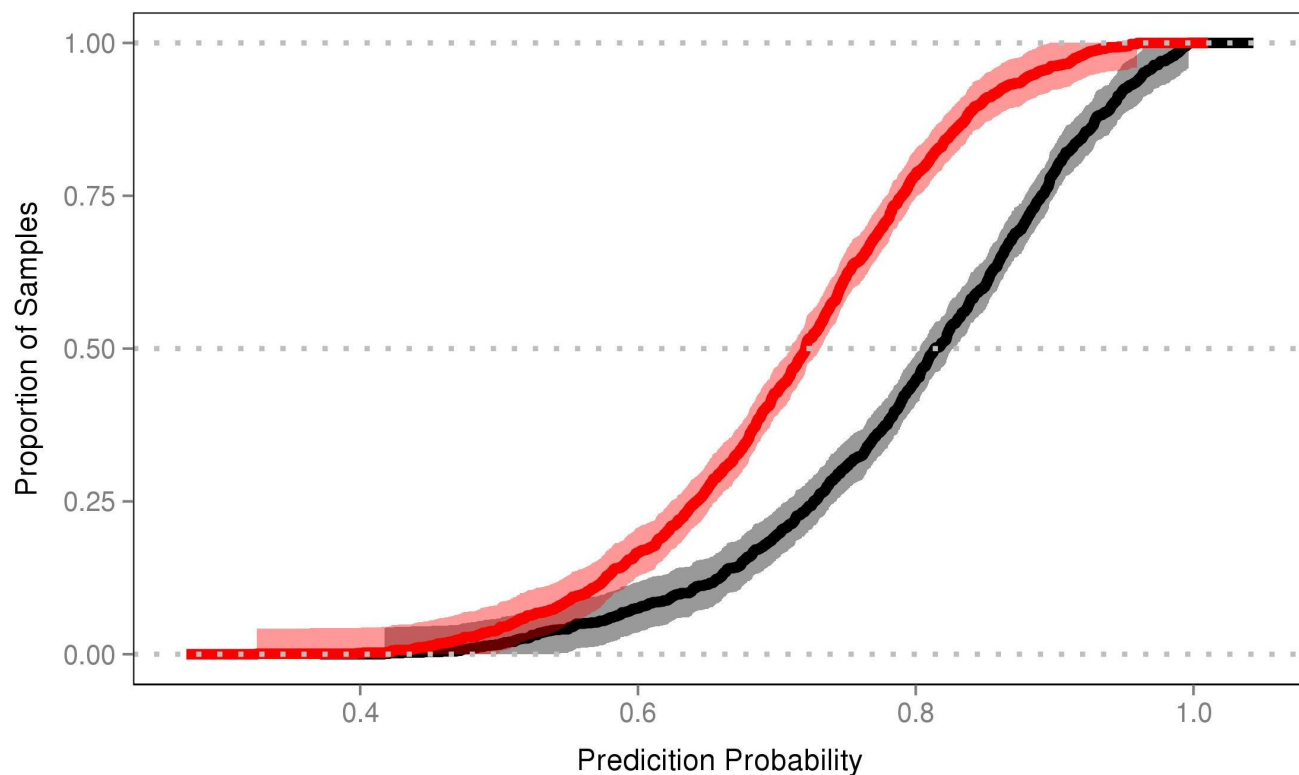


Figure 8: Prediction probabilities for the All Variables and GIS Only models.

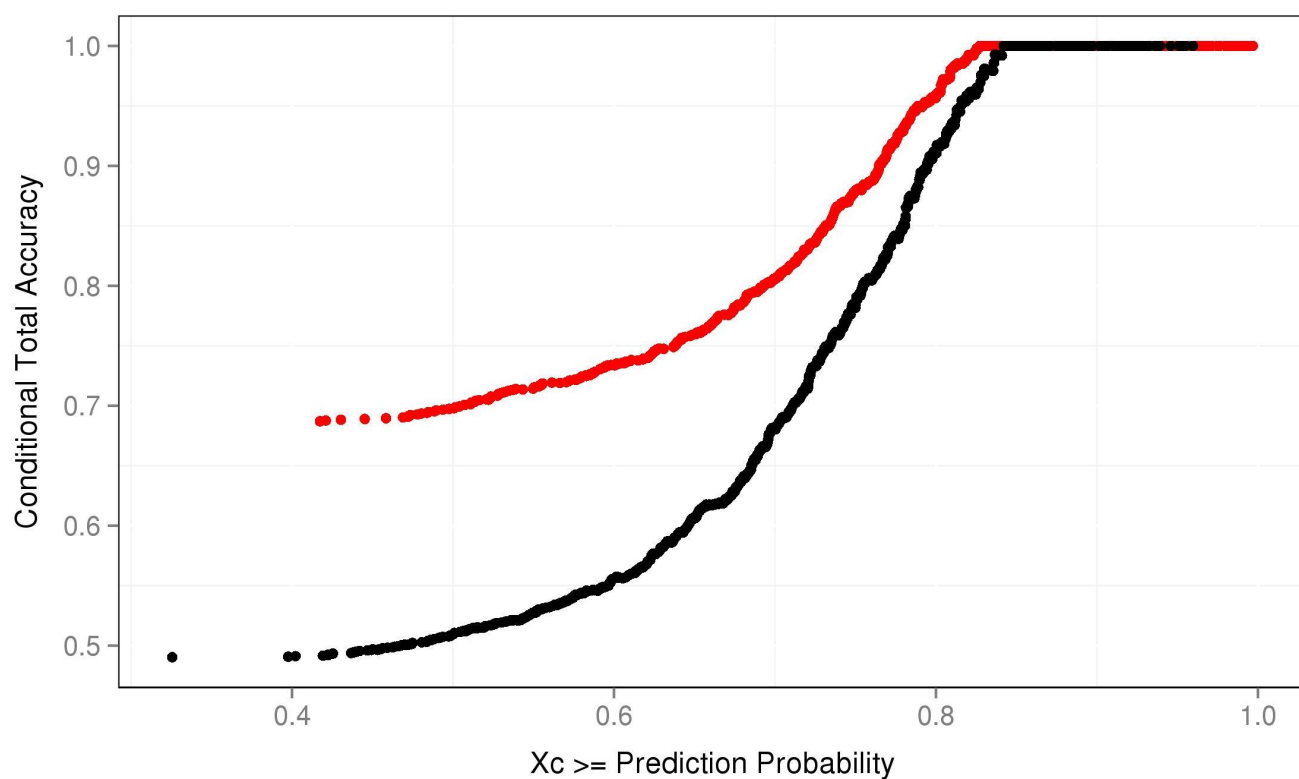


Figure 9: Accuracy of predictions as a function of lake prediction probability. The x-axis represents lakes with a prediction probability at a given level or higher.

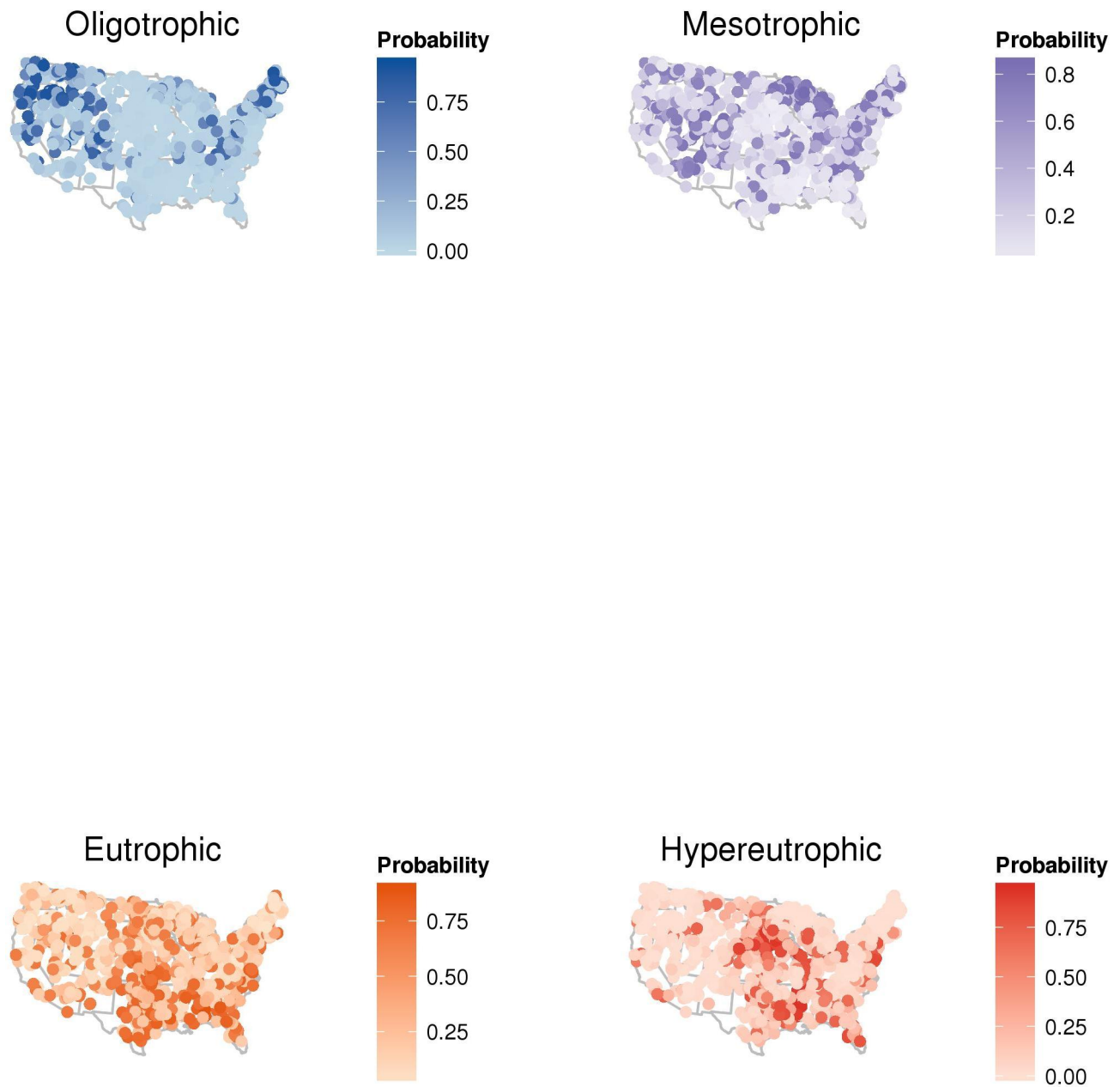


Figure 10: Maps of prediction probabilities for each of the four chlorophyll *a* trophic states

Table 3: Random Forest confusion matrix for GIS Only model converted to 4 trophic states. Columns show predicted values and rows show observed values. Agreement indicated on diagonal and accuracy for each trophic state indicated in ‘Class Accuracy’ column.

|       | oligo | meso | eu  | hyper | Class Accuracy (%) |
|-------|-------|------|-----|-------|--------------------|
| oligo | 65    | 14   | 6   | 0     | 76.47              |
| meso  | 101   | 213  | 98  | 18    | 49.53              |
| eu    | 29    | 126  | 193 | 141   | 39.47              |
| hyper | 1     | 8    | 38  | 87    | 64.93              |



## 321 9 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 Deciduous 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 Herbaceous Wetland           | GIS           |
| AlbersX             | Longitude                            | GIS           |
| AlbersY             | Latitude                             | GIS           |
| LakeArea            | Lake Surface Area                    | GIS           |
| LakePerim           | Lake Perimeter                       | GIS           |
| 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 |

| variable_names | description                | type          |
|----------------|----------------------------|---------------|
| 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 |
| 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 |

| variable_names   | description                    | type          |
|------------------|--------------------------------|---------------|
| 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           |
| 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, 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.
- 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.
- FERNÁNDEZ-DELGADO, M., E. CERNADAS, S. BARRO, AND D. AMORIM. 2014. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* 15:3133–3181.
- HASLER, A. D. 1969. Cultural eutrophication is reversible. *BioScience* 19:425–431.
- HOLLISTER, J. W. 2014. Lakemorpho: Lake morphometry in R. R package version 1.0. <http://CRAN.R-project.org/package=lakemorpho>.
- 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.
- JONES, Z., AND F. LINDER. 2015. Exploratory data analysis using random forests. *in* The 73rd annual mPSA conference. MPSA.
- LANDIS, J. R., AND G. G. KOCH. 1977. The measurement of observer agreement for categorical data. *biometrics* 33:159–174.
- 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 41:1211–1223.
- 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.
- RODHE, W. 1969. Crystallization of eutrophication concepts in northern europe.
- 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. W., AND J. R. VALLENTYNE. 2008. The algal bowl: Overfertilization of the world’s freshwaters and estuaries. Page 334. University of Alberta Press Edmonton.
- 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.
- STROBL, C., A.-L. BOULESTEIX, A. ZEILEIS, AND T. HOTHORN. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics* 8:25.
- 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

393 cover classification to 2006 by using landsat imagery change detection methods. Remote Sensing of  
394 Environment 113:1133–1147.

395 YUAN, L. L., A. I. POLLARD, S. PATHER, J. L. OLIVER, AND L. D'ANGLADA. 2014. Managing  
396 microcystin: Identifying national-scale thresholds for total nitrogen and chlorophyll a. Freshwater  
397 Biology 59:1970–1981.