Modelling lake trophic state: A random forest 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 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 (e.g., recreation and aesthetics), and disservices (e.g., 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 data mining approaches to predict trophic state. Using these 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 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.

Keywords: Harmful Algal Blooms; Cyanobacteria; Open Science; Nutrients; National Lakes Assessment

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

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 these data 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. Furthermore, being able to classify a large number of lakes would have implications for the management of lakes. A broader discussion of ecological classification and resource management is beyond the scope of this paper, but see (Carpenter 1999) for more information on this topic.

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, 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.

The research presented here builds on past work in three areas. First, we built, assessed, and compared two random forest models of chlorophyll *a* with 1) *in situ* and universally available GIS data and then 2) universally available GIS data only. Second, we converted the chlorophyll *a* estimates, for both models, to trophic state and assessed prediction accuracy and uncertainty. Third, we examined the important predictors for both models. Lastly, to promote transparency in our work, the analysis code and data are 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 morphometery 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 dataset was collected during the summer of 2007 and the final datasets were released in 2009 (USEPA 2009). 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 various water quality measurements from the National Lakes Assessment (USEPA 2009). Additionally, the NLA included ecological regions as defined in the Wadeable Streams Assessment (Figure 2) (Omernik 1987, USEPA 2006).

Adding to the monitoring data collected via the NLA, we used 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). We chose this buffer distance for several reasons. First, in some preliminary efforts we tried a variety of scales (300 m, 1.5 km, and 3 km), and they had little impact on prediction accuracy. Second, since we also include local lake specific variables (see below) as well as the broader scale ecoregions, we chose the 3km buffer as it made intuitive sense as representative of land use impacts that would not be accounted for these other variables. While many regional classifications and scales have been shown to be effective (e.g., Cheruvelil et al. 2013), we chose a three kilometer buffer as it represented an intermediate scale that is greater than immediate parcels but smaller than regional and whole-basin measures.

Local, lake specific characteristics have been show to be important (Read et al. 2015). Thus to account for this, we 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. For a more detail description of random forests 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 predicted a continuous variable, chlorophyll *a*, directly thus `varSelRF`, does not apply, and nearly 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 forest 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 both 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

We used the `randomForest` package in R to build predictive models of chlorophyll *a* with two sets of predictors (Liaw and Wiener 2002). The first included *in situ* and universally available GIS predictors. We refer to this as the "All variables" model. For the second model we used just the universally available data (i.e., no *in situ* information). This is referred to as the "GIS only" model. A list of all considered variables 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 from the full suite of variables (Appendix 1) 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 with 5000 trees ("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).
4. Examine importance and partial dependence of the most important variables.

2.4 Measures of Model Performance and Variable Importance

We assessed the performance of the random forest two ways. First we compared the root mean square error and the adjusted R^2 of the models. Second, we examined the accuracy of the model predictions when converted to trophic states classes via a confusion matrix (Table 1). 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 calculated 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 equaling the agreement expected by chance alone. Values greater than 0 represent agreement greater than would be expected by chance. A kappa coefficient greater than approximately 0.6 is considered “substantial” agreement (Landis and Koch 1977). Negative values are rare and would indicate no agreement between the predicted and observed values. We use kappa as a means of comparison across models as well as within subsets of a given model. 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).

Random forests explicitly measure variable importance with two metrics: mean decrease in Gini and percent increase in mean squared error. These measure the impact on the overall model when a particular variable is included and thus can be used to assess importance (Breiman 2001). The Gini Index has been shown to have a bias (Strobl et al. 2007), thus, we used percent increase in mean squared error to assess variable importance. Lastly, partial dependence plots provide a mechanism to examine the partial relationship between individual variables and the response variable (Jones and Linder 2015). We examined these plots for the top variables as assigned by percent increase in mean squared error for each the reduced models.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

2.5 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 chlorophyll *a* models, we have 5,000 estimates of chlorophyll *a* for each lake. We convert these values to trophic states (Table 1) then count up total votes for each class and divide by total possible votes to get an estimate of 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 the "All variables" model were 95% for oligotrophic, 5% for mesotrophic, 0% for eutrophic, and 0% for hypereutrophic. The maximum probability provides the predicted class, in this case oligotrophic, and suggests little uncertainty in this prediction. We refer to this value as the "prediction probability."

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

3 Results

Our complete dataset included 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 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 used 1080 total observations, 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 variable selection process identified a reduced model with 20 variables (Figure 3). The six most important variables were turbidity, total phosphorus, total nitrogen, elevation, total organic carbon, and N:P ratio (Figures 4). The role that each played in predicting chlorophyll *a* varied (Figure 5).

3.2 Models: GIS Only Variables

The GIS only model was built using 1138 total observations, 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 variable selection process for this model produced a reduced model with 15 variables (Figure 6). The six most important variables were ecoregion, percent cropland, elevation, latitude, percent evergreen forest, and mean lake depth (Figures 7 & 5).

3.3 Trophic State Probabilities

The "All variables" model provides more certain model predictions with a median prediction probability of 0.81 versus 0.72 for the "GIS only" model (Figure 9). Additionally, the Kappa coefficient of the predictions is a function of this uncertainty. Lakes with more certain predictions were more accurately classified and had higher Kappa coefficients (Figure 10). For both models, when prediction probabilities are approximately 0.8 or higher, the models had a Kappa coefficient of ~1. This represents 55% of the lakes for the "All variables" model and 22% of the lakes for the "GIS only" model. A Kappa coefficient of 0.6 or higher is considered "substantial" agreement

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

(Landis and Koch 1977). For the “GIS only” model this is seen with 52% of the lakes. Lastly, as prediction probabilities increased, the difference in kappa coefficient between the two models decreased (Figure 10 & Tables 4 & 5).

4 Discussion

4.1 Trophic State Probabilities

Not surprisingly, lakes with more certain predictions (i.e., higher prediction probabilities) were more accurately predicted (Figure 10). The fact that the difference in accuracy (as measured by the Kappa coefficient) between the two models decreased as certainty in the prediction increased suggests that models with lower overall accuracy, such as the “GIS only” model, may have acceptable accuracy for many individual cases (Tables 4 & 5). Additionally, the prediction probabilities may be mapped for each of the four classes (Figure 11). The spatial patterns show little variability between the “All variables” and “GIS only” models, thus we only show the results from the more broadly applicable “GIS only” model (Figure 11).

This map provides several insights. First, since low uncertainty is associated with high accuracy, this map shows the broad spatial patterns of lake trophic state across the United States (i.e darker colors more likely to be correctly predicted). Hypereutrophic lakes are much more commonly predicted in the Midwest and southeastern United States. Clear, oligotrophic lakes are in the northwestern United States, through the western mountains and in the northeastern united states. The middle trophic states are more evenly distributed across the country. Lastly, this particular map is very similar to simply mapping the raw data. However, it highlights what could be done if the “GIS only” model were used to map data without measured chlorophyll *a* values which would provide probabilities of given trophic states for all lakes in the United States.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

4.2 Partial dependencies of explanatory variables

In line with past predictive modelling of chlorophyll *a* concentrations the “All variables” model selected the water quality variables (turbidity, total organic carbon, total nitrogen, total phosphorus, and N:P ratios) as important variables (Downing et al. 2001). While there is variation in the response of chlorophyll *a* to changes in nutrient concentrations, the general pattern suggests that limiting nutrients have predictable impacts. If we examine the partial dependencies of these variables we see a general linear increase in log chlorophyll *a* with nitrogen, phosphorus and organic carbon concentrations (Figure 5). This relationship holds until nutrient concentrations become saturated. The partial dependency plots (Figure 5) for the nitrogen:phosphorus ratio is more complicated, indicating that for ratios less than ~14 chlorophyll *a* increases but after ~14 there is marked decrease. The effect of the nitrogen phosphorus ratio on chlorophyll has been the subject of considerable research and our results are consistent with the majority of the findings suggesting that at low ratio values nitrogen is limiting (Downing and McCauley 1992, Smith and Schindler 2009). Conversely, at higher ratios the phosphorus levels may be limiting. 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.

Turbidity was selected as the most important variable in the “All variables” model. The partial dependency analysis shows that, similar to the nutrients discussed above, log chlorophyll *a* increases with increased turbidity. At first this may seem counter intuitive since we might expect productivity to decrease as turbidity increases, and therefore light availability decreases (Tilzer 1988, Bilotta and Brazier 2008). However, algal biomass can contribute heavily to measures of turbidity and we expect greater productivity to lead to increased turbidity (Hansson 1992). We interpret this pattern as indicating that as chlorophyll *a* concentrations increase we see a concomitant increase in turbidity due to increased algal cell densities.

Elevation was selected as an important predictive variable in both the all variables and the GIS

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

only models; the partial dependencies (Figures 5 & 8) indicate a negative relationship between elevation and chlorophyll *a* concentration that is probably due to fact that the location of mountains in the United States is the spatial inverse of the distribution of agricultural and urban lands. As elevation increases we expect decreased loads due to smaller watershed contributing areas. In contrast lower elevation sites will have larger drainage areas and greater potential for increased nutrient loads from urban and agricultural sources.

The variables in the "GIS only" model captured the large scale spatial pattern of the trophic status gradient of lakes across the United States. In addition to elevation, mentioned above, the model was most sensitive to latitude and ecoregion. In general, chlorophyll *a* concentrations are highest in the Southern portions of the study area where temperatures can be higher (a known driver of productivity), elevations lower, and agricultural impacts more pronounced. Likewise ecoregion (see Figures 2 & 8) has a pronounced affect indicting continental scale effects of land use and geography. Agriculturally dominated landscapes such as the Temperate Plains, Southern Plains, and Coastal Plains show the highest levels of Chlorophyll *a*. Whereas high elevation zones (Western Mountains), arid lands (Xeric), Northern habitats (Upper Midwest) have lower concentrations.

Further evidence for the role of land use/land cover variables, and similar to results from Read et. al. (2015), is shown by the selection of the percent cropland and percent evergreen forest variables. As indicated by the partial dependency plots (Figure 8), chlorophyll *a* increases with cropland and decreases with evergreen cover. It is not surprising that croplands were selected given the overwhelming impact of agriculture on the eutrophication process. Evergreens and chlorophyll *a* concentrations show a negative association (Figure 8). As the percent of evergreens increases we are likely to see increased elevation and soil differences that limit agriculture.

Lastly, morphometry (e.g., depth) also proved to be important in the prediction of lake trophic state (Genkai-Kato and Carpenter 2005). 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 such as in-lake nutrient processing and residence time.

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. The "All variables" model had an RMSE of 0.09 and an adjusted R^2 of 0.8 whereas, the GIS only models had an RMSE of 0.22 and the adjusted R^2 was 0.48. Our total accuracy in predicting chlorophyll *a* based trophic states was 69% for the "All variables" model and 49% for the "GIS only" model.

While the "GIS only" model showed lower prediction accuracies than the "All variables" model, the association between the uncertainty of prediction and total accuracy (Figure 10 and Tables Tables 4 & 5) suggest that the "GIS only" model will provide reasonable estimates of trophic state for many lakes across the United States. Furthermore, we can map the uncertainty of the predictions, thus, we know the spatial patterns and location of the lakes for which we are certain, or not, of their predicted trophic state. Given this and that these models may be applied to any lake in the United States we can recommend using this model.

Future iterations of this modelling effort may be able to utilize modeled predictions of nutrients to improve accuracy and also maintain broad applicability (Milstead et al. 2013). Changes such as these have several advantages. First, this would allow for estimating changes to chlorophyll *a* and trophic state as a function of changing nutrient loads, which are expected due to climate change (Adrian et al. 2009, Jeppesen et al. 2011, Moss et al. 2011, Jones and Brett 2014). Second, with the ability to make predictions for most lakes in the United States, the "GIS only" models could be used as a source of information on national scale phenomena. For example,

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

predictions of chlorophyll *a*, with measures of uncertainty, could be used in efforts to scale up the contributions from lakes to broad scale estimates of gross primary production.

For the “All variables” model, the *in situ* water quality variables drove the predictions. This is not surprising. For the “GIS only”" model, the results were more nuanced. Three broad categories were 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.

Our results raise three important considerations related to managing eutrophication. First, the broad scale patterning, indicated by ecoregion as an important variable, suggests regional trends. This is noteworthy because it suggests that efforts to monitor, model and manage eutrophication and cyanobacteria should be undertaken at both national and regional levels. This corroborates past findings that regional drivers are important for water quality (Cheruvelil et al. 2013). 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, 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.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

6 Acknowledgements

We would like to thank Farnaz Nojavan, Nathan Schmucker, John Kiddon, Joe LiVolsi, Tim Gleason, and Wayne Munns for constructive reviews of this paper. This paper has not been subjected to Agency review. Therefore, it does not necessary reflect the views of the Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. This contribution is identified by the tracking number ORD-011075 of the Atlantic Ecology Division, Office of Research and Development, National Health and Environmental Effects Research Laboratory, US Environmental Protection Agency.

7 Figures

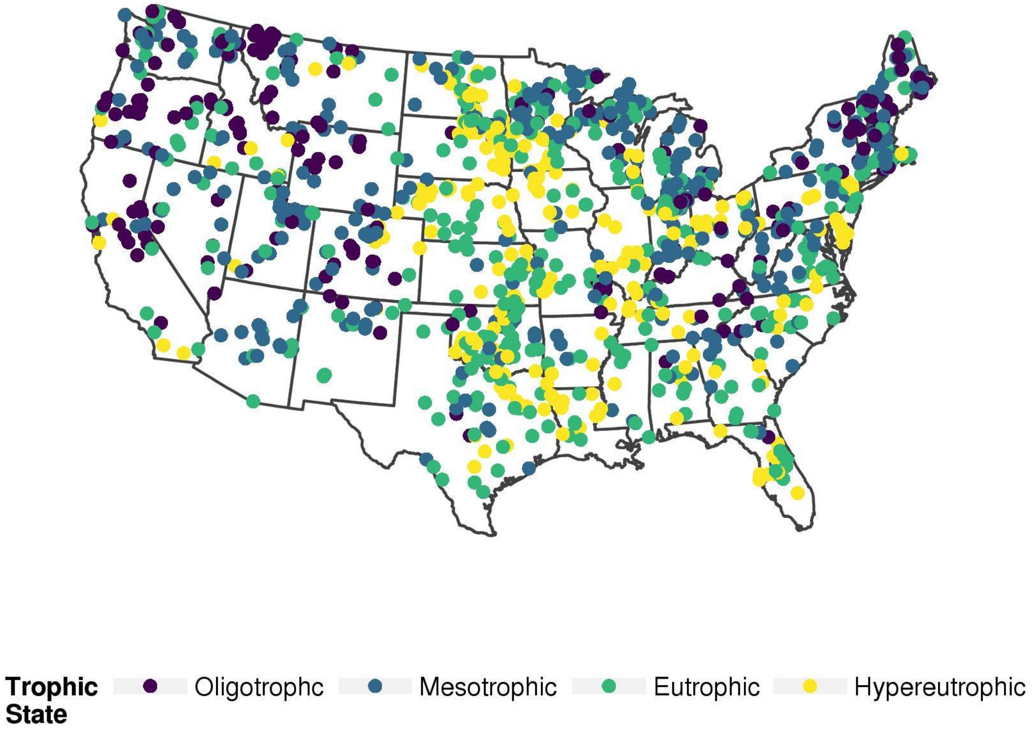

Figure 1: Map of the distribution of National Lakes Assesment Sampling locations

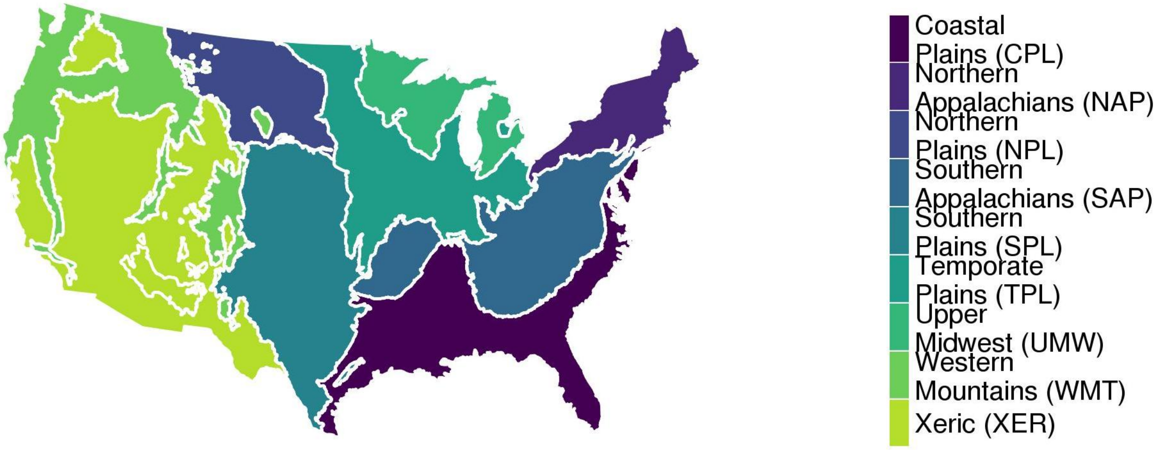

Figure 2: Wadeable Streams Assesment ecoregions

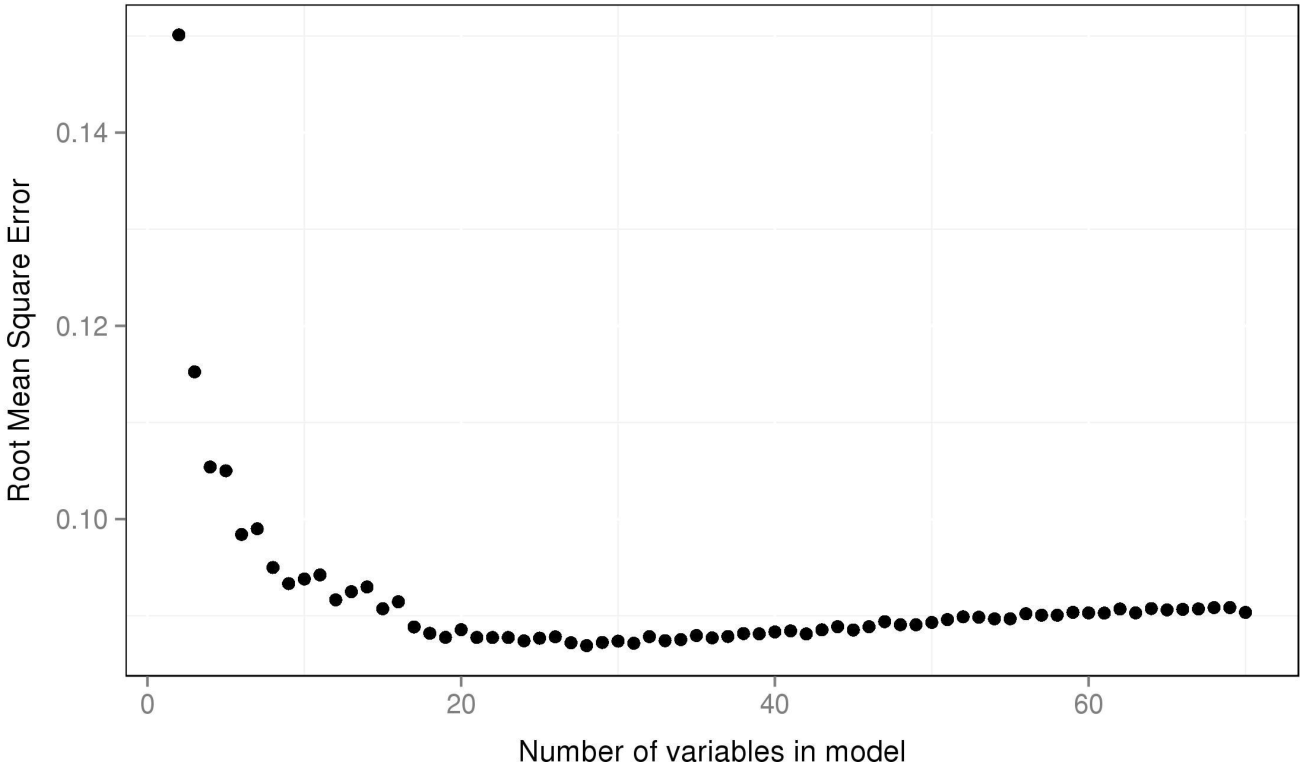

Figure 3: Variable selection plot for all variables. Shows percent increase in mean squared error as a function of the number of variables.

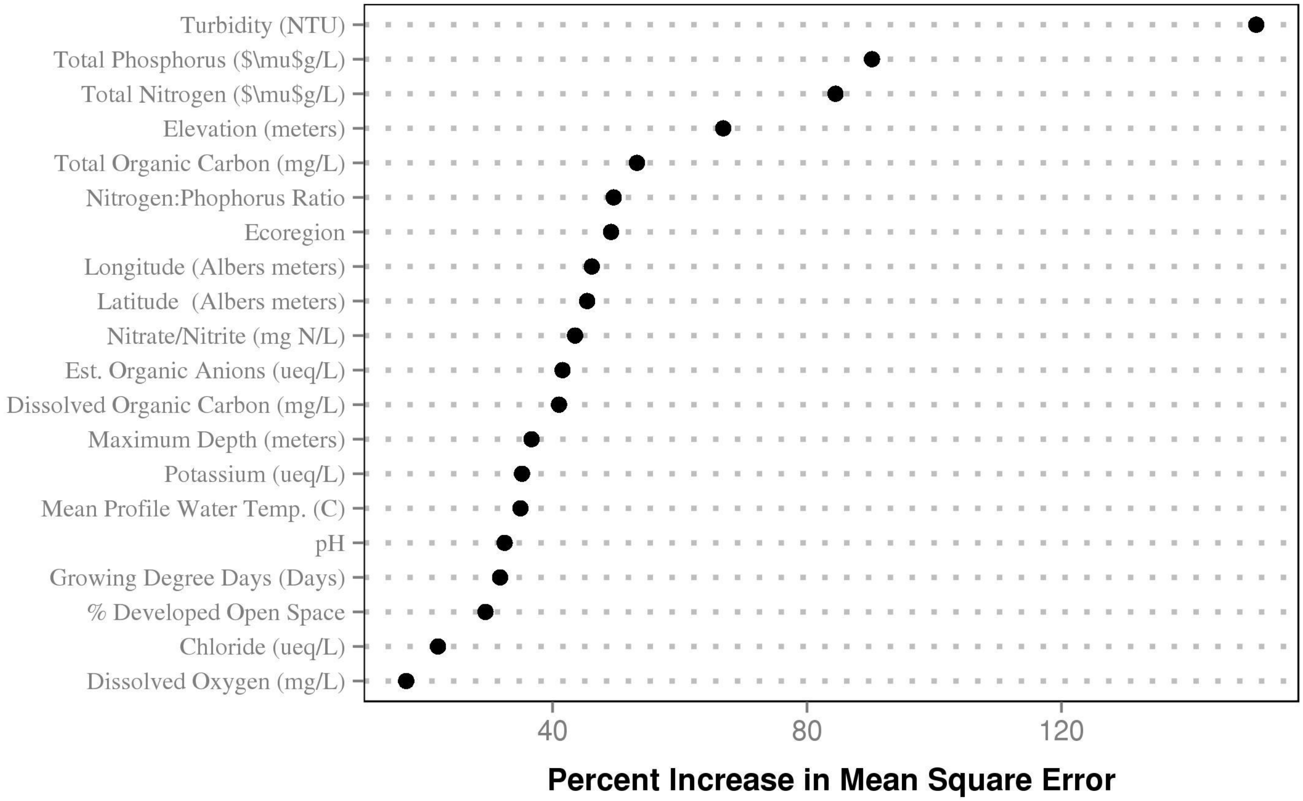

Figure 4: Importance plot for All Variables., shows percent increase in mean square error. Higher values of percent increase in mean squared error indicates higher importance.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

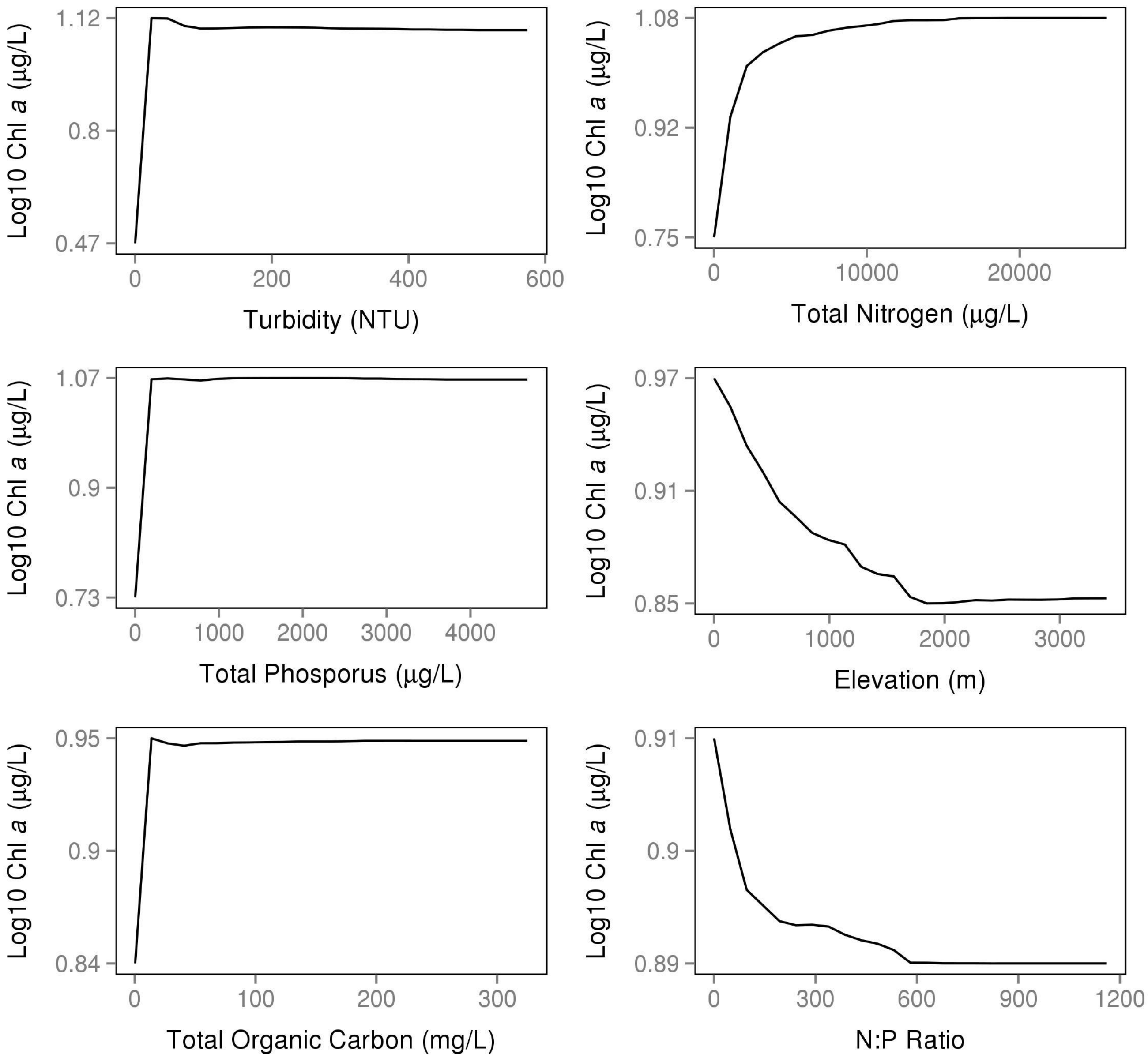

Figure 5: All Variables partial dependence plots for the top 5 most important variables.

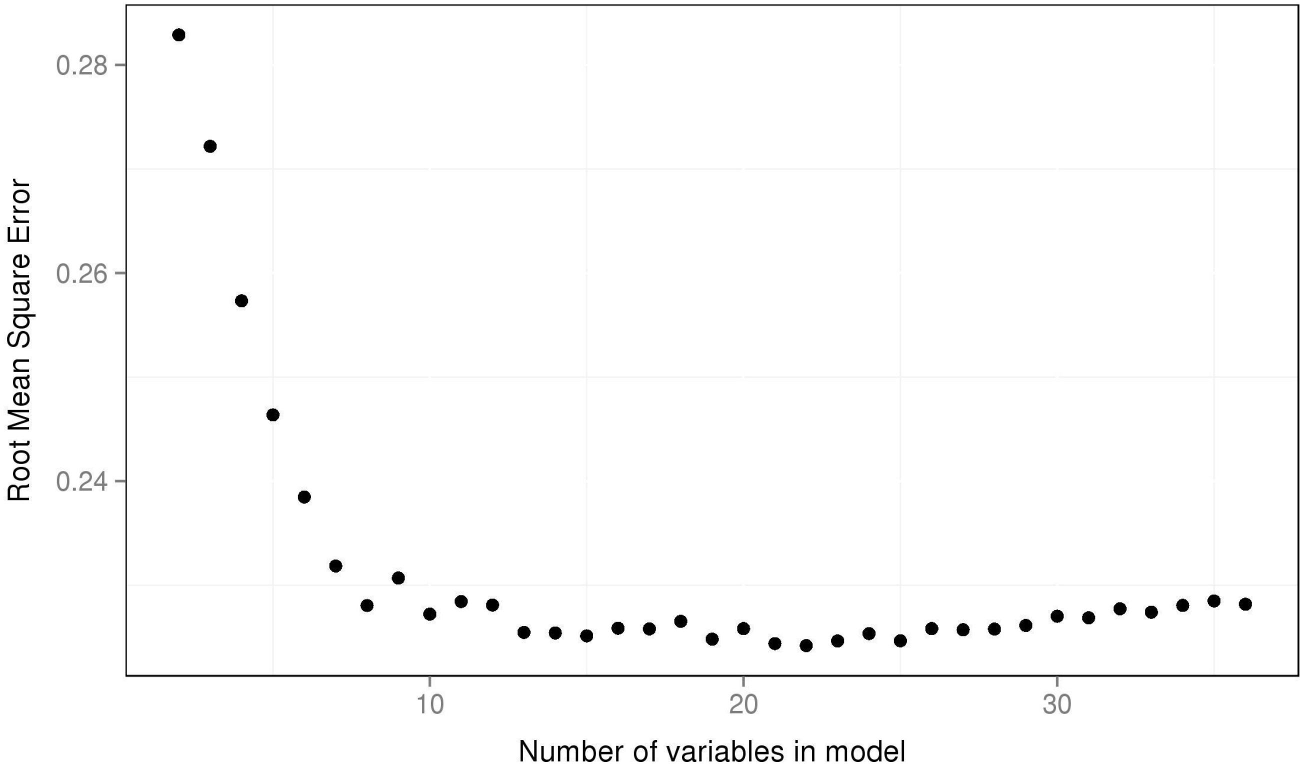

Figure 6: Variable selection plot for GIS only variables. Shows percent increase in mean squared error as a function of the number of variables.

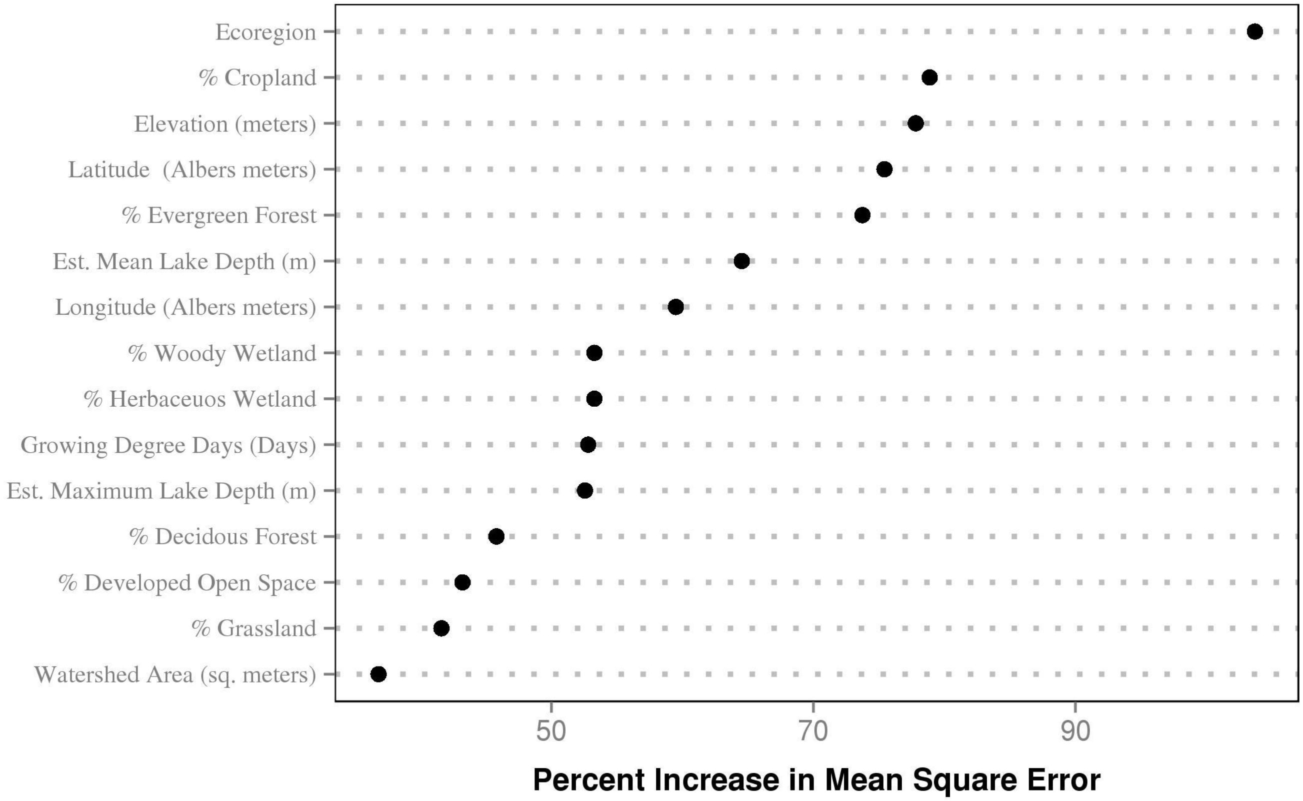

Figure 7: Importance plot for GIS Only Variables., shows percent increase in mean square error. Higher values of percent increase in mean squared error indicates higher importance.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

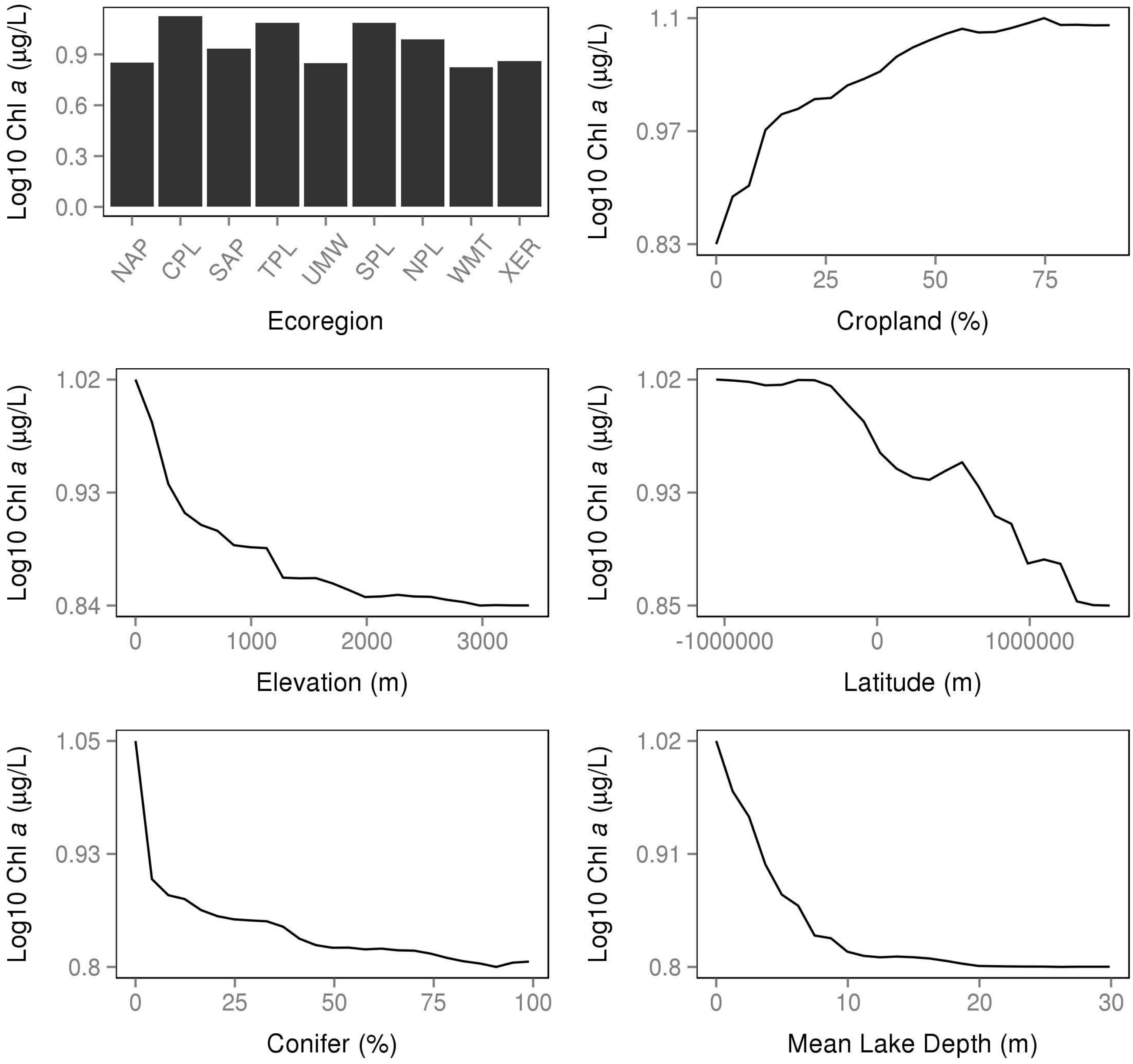

Figure 8: GIS Only Variables partial dependence plots for the top 5 most important variables.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

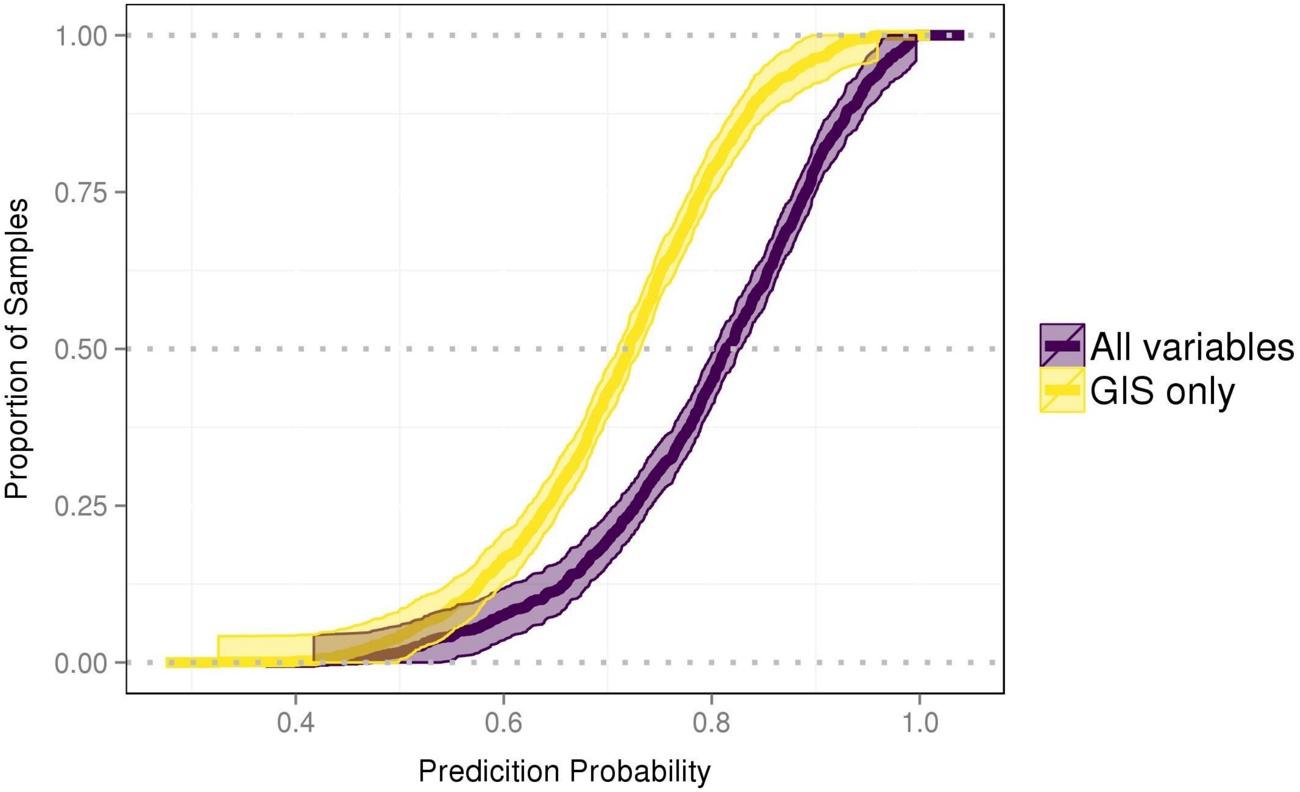

Figure 9: Prediction probabilities for the All Variables and GIS Only models.

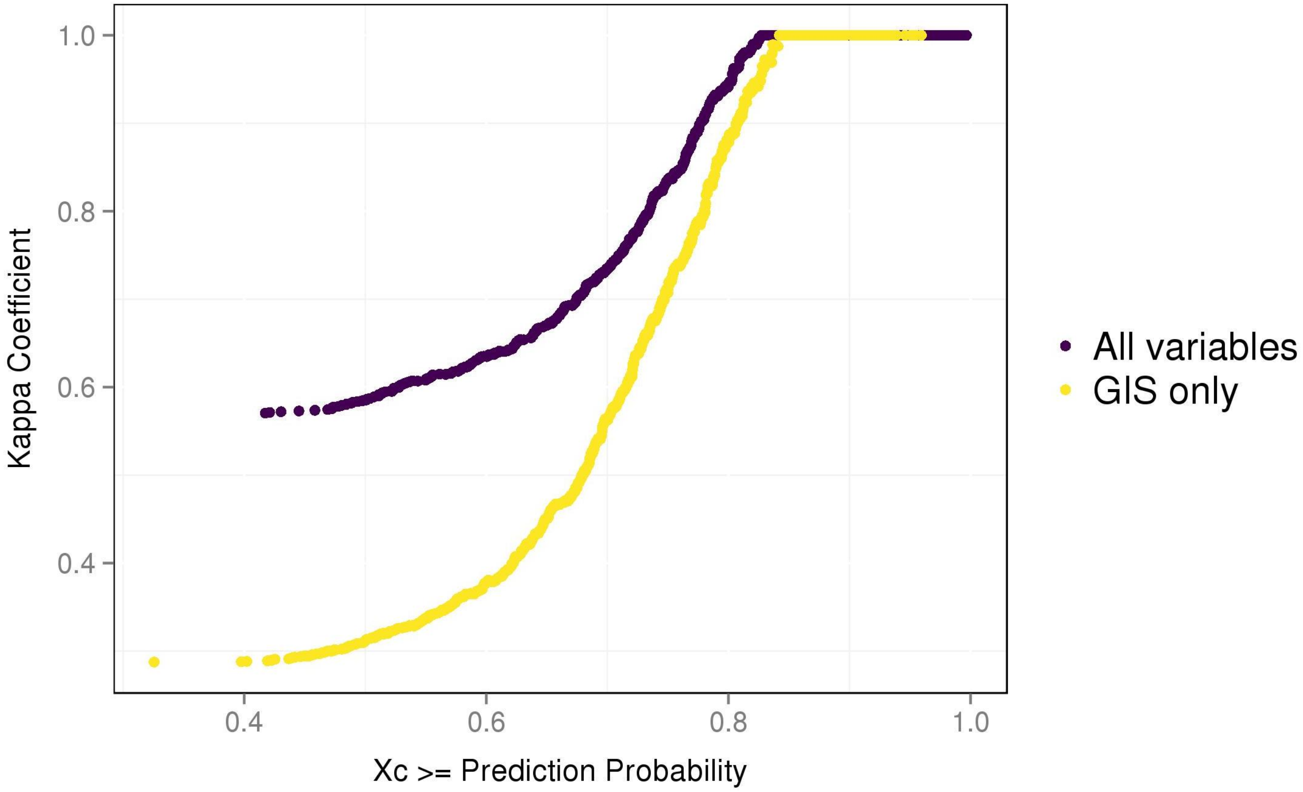

Figure 10: Accuracy of predictions as a function of lake prediction probability. The x-axis represents lakes with a prediction probability at a given level or higher.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

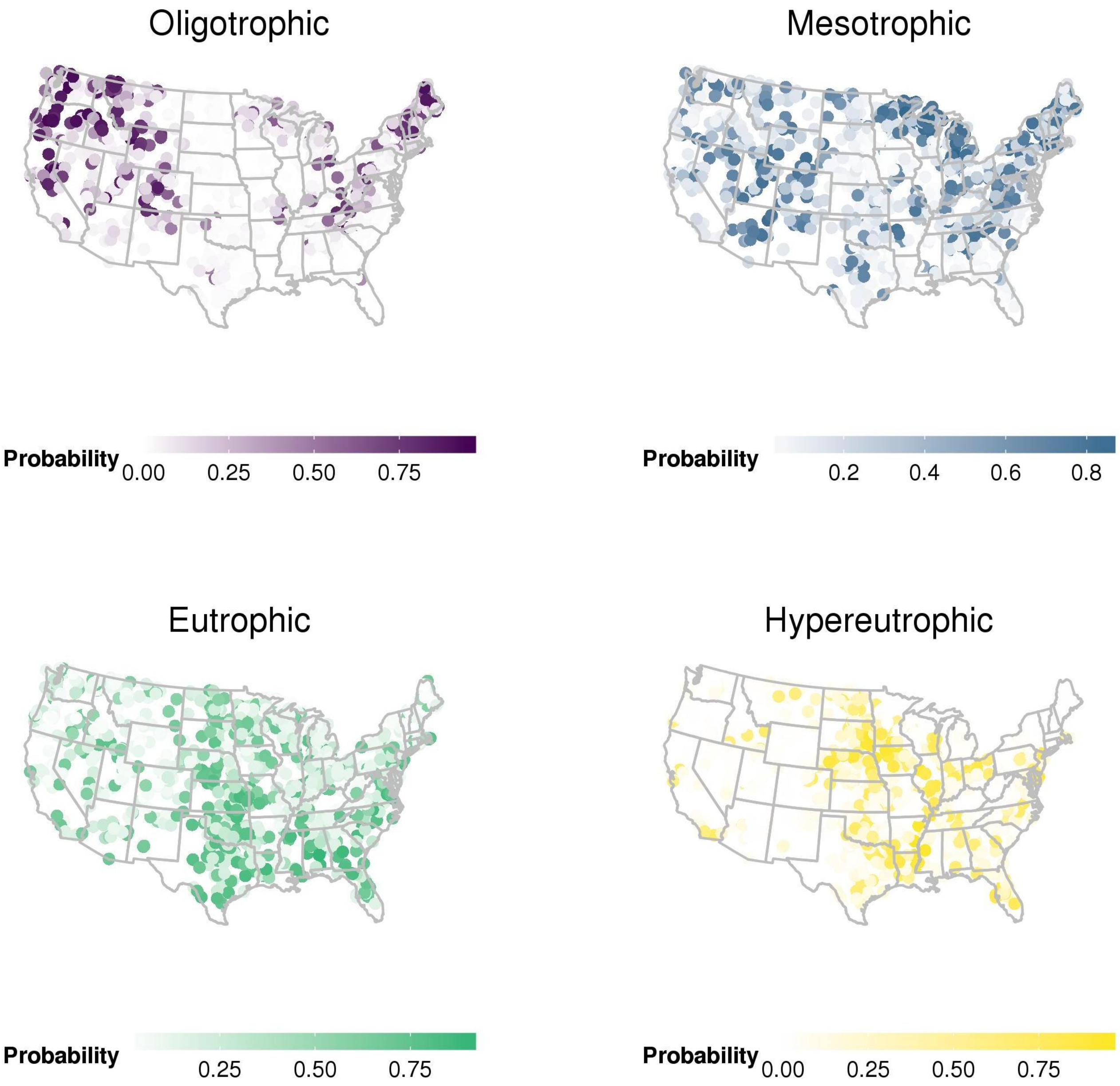

Figure 11: Maps of prediction probabilities for each of the four chlorophyll a trophic states

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

374
8 Tables

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

Trophic State (4 class)	Trophic State (2 class)	μg/L Cut-off
oligotrophic	oligotrophic/mesotrophic	<= 2
mesotrophic	oligotrophic/mesotrophic	>2-7
eutrophic	eutrophic/hypereutrophic	>7-30
hypereutrophic	eutrophic/hypereutrophic	>30

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

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

Table 3: Random Forest confusion matrix for GIS Only model converted to 4 tropic 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

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

Table 4: Summary of relationship between prediction probabilities, total accuracy, and number of lakes for the All variables model.

Prediction Prob.	Kappa Coefficient	Percent of Sample	Number of Samples
All	57	100	1080
0.50	59	98	1063
0.60	63	92	999
0.70	73	81	870
0.80	95	55	596
0.90	100	21	227

Table 5: Summary of relationship between prediction probabilities, total accuracy, and number of lakes for the GIS only model.

Prediction Prob.	Kappa Coefficient	Percent of Sample	Number of Samples
All	29	100	1138
0.50	31	96	1091
0.60	38	83	949
0.70	56	57	651
0.80	88	22	247
0.90	100	4	43

375 9 Appendix 1. Variable Definitions

Variable Names	Description	Source	Mean	Std. Error
AlbersX	Longitude (Albers meters)	GIS	126757.1	34305.5
AlbersY	Latitude (Albers meters)	GIS	436908.1	17367.2
BarrenPer_3000m	% Barren	GIS	0.7	0.1
BASINAREA	Watershed Area (sq. meters)	GIS	3208.5	788.1
CropsPer_3000m	% Cropland	GIS	13.3	0.6
DDs45	Growing Degree Days (Days)	GIS	2750.0	41.0
DeciduousPer_3000m	% Decidous Forest	GIS	17.1	0.6
DevHighPer_3000m	% High Intensity Development	GIS	0.4	0.0
DevLowPer_3000m	% Low Intensity Development	GIS	3.0	0.2
DevMedPer_3000m	% Medium Intensity Development	GIS	1.4	0.1
DevOpenPer_3000m	% Developed Open Space	GIS	5.4	0.2
ELEV_PT	Elevation (meters)	GIS	607.6	20.1
EvergreenPer_3000m	% Evergreen Forest	GIS	12.2	0.6
FetchE	Fetch from East (m)	GIS	1652.8	80.3
FetchN	Fetch from North (m)	GIS	2009.6	106.9
FetchNE	Fetch form Northeast (m)	GIS	1645.0	80.9
FetchSE	Fetch from Southeast (m)	GIS	1642.0	80.5
GrassPer_3000m	% Grassland	GIS	13.8	0.7
HerbWetPer_3000m	% Herbaceuos Wetland	GIS	1.7	0.1
IceSnowPer_3000m	% Ice/Snow	GIS	0.0	0.0
LakeArea	Lake Surface Area (sq. meters)	GIS	12.2	2.3
LakePerim	Lake Perimeter (meters)	GIS	33.6	4.5
MaxDepthCorrect	Est. Maximum Lake Depth (m)	GIS	8.4	0.3
MaxLength	Maximum Lake Length (m)	GIS	2972.1	137.2

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

Variable Names	Description	Source	Mean	Std. Error
MaxWidth	Maximum Lake Width (m)	GIS	1567.5	76.0
MeanDepthCorrect	Est. Mean Lake Depth (m)	GIS	2.9	0.1
MeanWidth	Mean Lake Width (m)	GIS	1370.1	122.6
MixedForPer_3000m	% Mixed Forest	GIS	3.8	0.3
PasturePer_3000m	% Pasture	GIS	7.7	0.3
PercentImperv_3000m	% Impervious	GIS	2.6	0.2
ShoreDevel	Shoreline Development Index	GIS	2.7	0.1
ShrubPer_3000m	% Shrub/Scrub	GIS	10.4	0.6
VolumeCorrect	Est. Lake Volume (cubic meters)	GIS	101211909.9	27438696.4
WaterPer_3000m	% Water	GIS	4.1	0.2
WoodyWetPer_3000m	% Woody Wetland	GIS	5.2	0.3
WSA_ECO9	Ecoregion	GIS	NA	NA
ANC	Acid Neutralizing Capacity (ueq/L)	NLA	2584.2	171.7
ANDEF2	Anion Deficit (ueq/L)	NLA	-506.4	143.2
ANSUM2	Sum of Anions using ANC (ueq/L)	NLA	8043.1	1197.9
BALANCE2	Ion Balance (%)	NLA	-0.7	0.1
CA	Calcium (ueq/L)	NLA	1388.3	54.0
CATSUM	Sum of Cations (ueq/L)	NLA	7536.7	1105.0
CL	Chloride (ueq/L)	NLA	1600.3	438.2
COLOR	Color (PCU)	NLA	16.1	0.5
CONCAL2	Calculated Conductivity (uS/cm)	NLA	949.0	148.1
COND	Conductivity (uS/cm)	NLA	656.0	72.6
CONDHO2	D-H-O Calculated Conductivity (uS/cm)	NLA	618.6	55.1
DATE_COL	Date Samples Collected	NLA	NA	NA
DEPTHMAX	Maximum Depth (meters)	NLA	9.6	0.3

Variable Names	Description	Source	Mean	Std. Error
DO2_2M	Dissolved Oxygen (mg/L)	NLA	7.9	0.1
DOC	Dissolved Organic Carbon (mg/L)	NLA	8.6	0.5
H	Hydrogen Ions (ueq/L)	NLA	0.2	0.1
K	Potassium (ueq/L)	NLA	245.6	40.6
MG	Magnesium (ueq/L)	NLA	2190.4	282.2
Na	Sodium (ueq/L)	NLA	3709.7	816.3
NH4	Ammonium (mg/L)	NLA	2.9	0.2
NH4ION	Calculated Ammonium (ueq/L)	NLA	2.5	0.2
NO3	Nitrate (ueq/L)	NLA	5.4	0.7
NO3_NO2	Nitrate/Nitrite (mg N/L)	NLA	0.1	0.0
NPratio	Nitrogen:Phophorus Ratio	NLA	34.5	1.8
NTL	Total Nitrogen (μg/L)	NLA	1109.9	56.4
OH	Hydroxide (ueq/L)	NLA	3.1	0.2
ORGION	Est. Organic Anions (ueq/L)	NLA	85.9	4.8
PH_FIELD	pH	NLA	8.1	0.0
PTL	Total Phosphorus (μg/L)	NLA	103.1	7.8
SIO2	Silica (mg/L)	NLA	8.6	0.3
SO4	Sulfate (ueq/L)	NLA	3853.4	935.7
SOBC	Sum of Base Cation (ueq/L)	NLA	7534.1	1105.0
TmeanW	Mean Profile Water Temp. (C)	NLA	24.1	0.1
TOC	Total Organic Carbon (mg/L)	NLA	9.6	0.6
TURB	Turbidity (NTU)	NLA	12.3	1.0

References

Adrian, R., C. M. O'Reilly, H. Zagarese, S. B. Baines, D. O. Hessen, W. Keller, D. M. Livingstone, R. Sommaruga, D. Straile, E. Van Donk, and others. 2009. Lakes as sentinels of climate change. Limnology and Oceanography 54:2283–2297.

Bilotta, G., and R. Brazier. 2008. Understanding the influence of suspended solids on water quality and aquatic biota. Water research 42:2849–2861.

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.

Carpenter, W.-D. C., Constance A.; Busch. 1999. The use of ecological classification in management. Pages 395–430 *in* R. Szaro, N. Johnson, W. Sexton, and A. Malk, editors. Ecological stewardship: A common reference for ecosystem management.

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.

Cheruvelil, K., P. Soranno, K. Webster, and M. Bremigan. 2013. Multi-scaled drivers of ecosystem state: Quantifying the importance of the regional spatial scale. Ecological Applications 23:1603–1618.

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

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

data using random forest. BMC bioinformatics 7:3.

Downing, J. A., and E. McCauley. 1992. The nitrogen:phosphorus relationship in lakes. Limnology and Oceanography 37:936–945.

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.

Genkai-Kato, M., and S. R. Carpenter. 2005. Eutrophication due to phosphorus recycling in relation to lake morphometry, temperature, and macrophytes. Ecology 86:210–219.

Hansson, L.-A. 1992. Factors regulating periphytic algal biomass. Limnology and Oceanography 37:322–328.

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.

Jeppesen, E., B. Kronvang, J. E. Olesen, J. Audet, M. Søndergaard, C. C. Hoffmann, H. E. Andersen, T. L. Lauridsen, L. Liboriussen, S. E. Larsen, and others. 2011. Climate change effects on nitrogen loading from cultivated catchments in europe: Implications for nitrogen retention, ecological state of lakes and adaptation. Hydrobiologia 663:1–21.

Jones, J., and M. T. Brett. 2014. Lake nutrients, eutrophication, and climate change. Pages 273–279 *in* Global environmental change. Springer.

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.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1319v3 | CC-BY 4.0 Open Access | rec: 3 Dec 2015, publ: 3 Dec 2015

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.

Moss, B., S. Kosten, M. Meerhof, R. Battarbee, E. Jeppesen, N. Mazzeo, K. Havens, G. Lacerot, Z. Liu, L. De Meester, and others. 2011. Allied attack: Climate change and eutrophication. Inland waters 1:101–105.

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

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.

Read, E. K., V. P. Patil, S. K. Oliver, A. L. Hetherington, J. A. Brentrup, J. A. Zwart, K. M. Winters, J. R. Corman, E. R. Nodine, R. I. Woolway, and others. 2015. The importance of lake-specific characteristics for water quality across the continental united states. Ecological Applications 25:943–955.

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., and D. W. Schindler. 2009. Eutrophication science: Where do we go from here? Trends in Ecology & Evolution 24:201–207.

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.

Tilzer, M. M. 1988. Secchi disk—chlorophyll relationships in a lake with highly variable phytoplankton biomass. Hydrobiologia 162:163–171.

USEPA. 2006. Wadeable streams assessment: A collaborative survey of the nation's streams. ePA 841-b-06-002. Office of Water; Office of Research; Development, US Environmental Protection Agency Washington, DC.

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.