

Modelling lake trophic state: A random forest approach

Jeffrey W. Hollister^{*} ¹ W. Bryan Milstead ¹ Betty J. Kreakie ¹

¹US 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, 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 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 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,

38 have greater density of aquatic plants, and often support more diverse and abundant fish communities.
39 Higher primary productivity is not necessarily a predictor of poor ecological condition as it is natural
40 for lakes to shift from lower to higher trophic states but this is a slow process (Rodhe 1969). However,
41 at the highest productivity levels (hypereutrophic lakes) biological integrity is compromised (Hasler
42 1969, Smith et al. 1999, Schindler and Vallentyne 2008).

43 Monitoring trophic state allows for rapid assessment of a lakes biological productivity and identification
44 of lakes with unusually high productivity (e.g. hypereutrophic). These cases are indicative of lakes
45 under greater anthropogenic nutrient loads, also known as cultural eutrophication, and are more likely
46 to be at risk of fish kills, beach fouling, and harmful algal blooms (Smith 1998, Smith et al. 1999, 2006).
47 Given the association between trophic state and many ecosystem services and disservices, being able
48 to accurately model trophic state could provide a first cut at identifying lakes with the potential for
49 harmful algal blooms (i.e. from cyanobacteria) or other problems associated with cultural eutrophication.
50 This type of information could be used for setting priorities for management and allow for more efficient
51 use of limited resources.

52 As trophic state and related indices can be best defined by a number of *in situ* water quality parameters
53 (modeled or measured), most models have used this information as predictors (Imboden and Gächter
54 1978, Salas and Martino 1991, Carvalho et al. 2011, Milstead et al. 2013). This leads to accurate
55 models, but this data is often sparse and not always available, thus limiting the population of lakes for
56 which we can make predictions. A possible solution for this issue is to build models that use widely
57 available data that are correlated to many of the *in situ* variables. For instance, landscape metrics of
58 forests, agriculture, wetlands, and urban land in contributing watersheds have all been shown to explain
59 a significant proportion of the variation (ranging from 50-86%, depending on study) in nutrients in
60 receiving waters (Jones et al. 2001, 2004, Seilheimer et al. 2013). Building on these previously identified
61 associations might allow us to use only landscape and other universally available data to build models.
62 Identifying predictors using this type of ubiquitous data would allow for estimating trophic state in
63 both monitored and unmonitored lakes.

64 Many published models of nutrients and trophic state in freshwater systems are based on linear modelling
65 methods such as standard least squares regression or linear mixed models (Jones et al. 2001, 2004).

66 While these methods have proven to be reliable, they have limitations (e.g. independence, distribution
67 assumptions, and outlier sensitivity). Using data mining approaches, such as random forests, avoids
68 many of the limitations, may reduce bias, and often provides better predictions (Breiman 2001, Cutler
69 et al. 2007, Peters et al. 2007, Fernández-Delgado et al. 2014). For instance, random forests are
70 non-parametric and thus the data do not need to come from a specific distribution (e.g. Gaussian)
71 and can contain collinear variables (Cutler et al. 2007). Second, random forests work well with very
72 large numbers of predictors (Cutler et al. 2007). Lastly, random forests can deal with model selection
73 uncertainty as predictions are based upon a consensus of many models and not just a single model
74 selected with some measure of goodness of fit.

75 The research presented here builds on past work in three areas. First, we built, assessed, and compared
76 two random forest models of chlorophyll *a* 1) *in situ* and universally available GIS data and then 2)
77 universally available GIS data only. Second, we converted the chlorophyll *a* estimates, for both models,
78 to trophic state and assessed prediction accuracy and uncertainty. Third, we examined the important
79 predictors for both models. Lastly, this paper, the code, and the data used in the models are available
80 as an R package from <https://github.com/USEPA/LakeTrophicModelling>.

81 2 Methods

82 2.1 Data and Study Area

83 We utilized three primary sources of data for this study, the National Lakes Assessment (NLA), the
84 National Land Cover Dataset (NLCD), and lake morphometry modeled from the NHDPlus and
85 National Elevation Data Set (Homer et al. 2004, USEPA 2009, Xian et al. 2009, Hollister and Milstead
86 2010, Hollister et al. 2011, Hollister 2014). All datasets are national in extent and provide a unique
87 snapshot view of the condition of lakes in the conterminous United States during the summer of 2007.

88 The NLA dataset was collected during the summer of 2007 and the final datasets were released in 2009
89 (USEPA 2009 for detailed description of methods). With consistent methods and metrics collected
90 at over 1000 locations across the conterminous United States (Figure 1), the NLA provides a unique

91 opportunity to examine broad scale patterns in lake productivity. The NLA collected data on biophysical
92 measures of lake water quality and habitat as well as an assessment of the phytoplankton community. For
93 this analysis, we only use the various water quality measurements from the National Lakes Assessment
94 (USEPA 2009).

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

102 To account for unique aspects of each lake and to characterize lake productivity, we used measures
103 of lake morphometry (i.e. depth, volume, fetch, etc.). As these data are difficult to obtain for large
104 numbers of lakes over broad regions, we used modeled estimates of lake morphometry (Hollister and
105 Milstead 2010, Hollister et al. 2011, Hollister 2014). These included: surface area, shoreline length,
106 Shoreline Development, Maximum Depth, Mean Depth, Lake Volume, Maximum Lake Length, Mean
107 Lake Width, Maximum Lake Width, and Fetch.

108 **2.2 Predicting Trophic State with Random Forests**

109 Random forest is a machine learning algorithm that aggregates numerous decision trees in order to
110 obtain a consensus prediction of the response categories (Breiman 2001). Bootstrapped sample data are
111 recursively partitioned according to a given random subset of predictor variables and a predetermined
112 number of decision trees are developed. With each new tree, the sample data subset is randomly
113 selected and with each new split, the subset of predictor variables are randomly selected. A detailed
114 discussion of the benefits of a random forest approach is beyond the scope of this paper. For a more
115 detail description of random forests see Breiman (2001) and Cutler et al. (2007).

116 Random forests are able to handle numerous correlated variables without a decrease in prediction

117 accuracy; however, one possible shortcoming of this approach is that the resulting model may be difficult
118 to interpret, thus selecting the most important variables is an important first step. Several methods
119 have been proposed to do this with random forest. For instance, this is a problem often faced in gene
120 selection and in that field, a variable selection method based on random forest has been successfully
121 applied and implemented in the R Language as the `varSelRF` package (Díaz-Uriarte and De Andres
122 2006), but this is limited to classification problems. Additionally, others have suggested alternative
123 variable importance measures, but this is only needed with a large number of categorical variables which
124 are selected against with traditional random forest approach (Strobl et al. 2007).

125 In our case, we predicted a continuous variable, chlorophyll *a*, directly thus `varSelRF`, does not apply,
126 and all of our variables are continuous so the approach suggested by Strobl (2007) is not necessary.
127 Thus we developed an approach, similar to `varSelRF` but applied to random forest with regression trees.
128 With this approach we fit a full random forest model that includes all variables and a large number of
129 trees. We then rank the variables using the increase in mean square error, which has been shown to be
130 a less biased metric of importance than the mean decrease in the gini coefficient (Strobl et al. 2007).
131 Using this ranking, we then iterate through the variables and create a random forest with the top two
132 variables and record mean square error and adjusted R^2 of the resultant random forest. We then repeat
133 this process by adding the next most important variable in order of importance. With this information
134 we identify both the top variables and the point at which adding variables does not improve the fit of
135 the overall model. These variables are selected and used as the “reduced model.” With this method, a
136 minimum set of variables that maximizes model accuracy is provided. This allows us to start with a
137 full suite of predictor variables from which to select a minimum, easier to interpret set of variables.

138 2.3 Model Details

139 Using `randomForest` R package we ran models to predict chlorophyll *a* with two sets of predictors (Liaw
140 and Wiener 2002). The first included *in situ* and universally available GIS predictors. We refer to
141 this as the “All variables” model. Second, we use just the univereally available data (i.e. no *in situ*
142 information). This is referred to as the “GIS only” model. A list of the full suite of variables tested
143 is in Appendix 1. Our separation of predictors was chosen so that we could highlight the additional

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

147 Our modelling work flow was as follows:

- 148 1. Identify a minimal set of variables that maximize accuracy of the random forest algorithm. This
149 minimal set of variables, the reduced model, is calculated for each of the models.
- 150 2. Using R's `randomForest` package, we develop two random forest models (“All variables” and
151 “GIS only”).
- 152 3. Assess model performance for both the predicted chlorophyll *a* and for categorical trophic state
153 classifications. Trophic state was defined using the NLA chlorophyll *a* trophic state cut offs (Table
154 1).
- 155 4. Examine importance and partial dependence of the most important variables.

156 2.4 Measures of Model Performance and Variable Importance

157 We assessed the performance of the random forest two ways. First we compare the root mean square
158 error and the adjusted R^2 of the models. Second, we examine the accuracy of the model predictions
159 when converted to trophic states classes (Table 1). We assess the classifications via a confusion matrix.
160 A confusion matrix shows agreement and disagreement in a tabular form with predicted values forming
161 the columns of the matrix and observed values, the rows. From this tabulated information we calculate
162 the total accuracy (i.e. percent correctly predicted) and the kappa coefficient, which takes into account
163 the error expected by chance alone (i.e. the off diagonal values of the matrix) (Cohen 1960, Hubert and
164 Arabie 1985). The kappa coefficient can range from -1 to 1 with 0 equalling the agreement expected by
165 chance alone. Values greater than 0 represent agreement greater than would be expected by chance,
166 with values greater than 0.61 considered “substantial” agreement (Landis and Koch 1977). Negative
167 values are rare and would indicate no agreement between the predicted and observed values. Additionally,
168 random forest builds each tree on bootstrapped, random subsets of the original data, thus, a separate
169 independent validation dataset is not required and random forest error estimates are expected to be

170 unbiased (Breiman 2001).

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

179 2.5 Trophic State Probabilities

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

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

194 3 Results

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

199 3.1 Models: All Variables

200 The model built with all predictors used 1080 total observations, had a mean squared error of 0.09 and
201 and R^2 of 0.8. The accuracy of the four trophic states was 68.7% and the kappa coefficient was 0.57
202 (Table 2). The variable selection process identified 20 variables (Figure 2). The six most important
203 variables were turbidity, total phosphorus, total nitrogen, elevation, total organic carbon, and N:P ratio
204 (Figures 3). The role that each played in predicting chlorophyll *a* varied (Figure 4).

205 3.2 Models: GIS Only Variables

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

211 3.3 Trophic State Probabilities

212 The “All variables” model provides more certain model predictions with a median prediction probability
213 of 0.81 versus 0.72 for the “GIS only” model (Figure 8). Additionally, total accuracy of the predictions
214 is a function of this uncertainty. Lakes with more certain predictions were more accurately classified
215 (Figure 9). For both models, when prediction probabilities are approximately 0.8 or higher, the models
216 had an accuracy of ~100%. This represents 55% of the lakes for the “All variables” model and 22% of

217 the lakes for the “GIS only” model. Lastly, as prediction probabilities increased, the difference in total
218 accuracy between the two models decreased (Figure 9 & Table 4).

219 4 Discussion

220 4.1 Trophic State Probabilities

221 Not surprisingly, lakes with more certain predictions (i.e. higher prediction probabilities) were more
222 accurately predicted (Figure 9). The fact that the difference in accuracy between the two models
223 decreased as certainty in the prediction increased suggests that models with lower overall accuracy,
224 such as the “GIS only” model, may have acceptable accuracy for many individual cases (Table 4).
225 Additionally, the prediction probabilities may be mapped for each of the four classes (Figure 10).

226 This map provides several insights. First, since low uncertainty is associated with high accuracy, this
227 map shows the broad spatial patterns of lake trophic state across the United States. The spatial patterns
228 show little variability between the “All variables” and “GIS only” models, thus we only show the results
229 from the more broadly applicable “GIS only” model (Figure 10). Hypereutrophic lakes are much more
230 commonly predicted in the midwest and southeastern United States. Clear, oligotrophic lakes are in
231 the northwestern United States, through the western mountains and in the northeastern United States.
232 The middle trophic states are more evenly distributed across the country. Secondly, instead of mapping
233 the probabilities for each trophic state separately, we can also map the prediction probabilities of the
234 discrete predicted class. (Figure 11). This map shows where the model predicts well and where it is less
235 certain. In general, the map shows most points with higher prediction probabilities than the midpoint
236 of the range and the distribution of prediction probabilities is skewed left (Figure 12). While these
237 patterns are not strong, they suggest that with slight improvements in the “GIS only” model we could
238 skew the prediction probabilities further left and easily improve the overall accuracy of the model. This
239 could be done using modeled, national estimates of nutrient loads (Moore et al. 2011, e.g. Milstead et
240 al. 2013).

241 4.2 Partial dependencies of explanatory variables

242 In line with past predictive modelling of chlorophyll *a* concentrations the “All variables” model selected
243 the water quality variables (turbidity, total organic carbon, total nitrogen, total phosphorus, and N:P
244 ratios) as important variables (Downing et al. 2001). While there is variation in the response of
245 chlorophyll *a* to changes in nutrient concentrations, the general pattern suggests that limiting nutrients
246 have a predictable impacts. If we examine the partial dependencies of these variables we see a general
247 linear increase in log chlorophyll *a* with nitrogen, phosphorus and organic carbon concentrations (Figure
248 4). This relationship holds until nutrient concentrations become saturated. The partial dependency
249 plots (Figure 4) for the nitrogen:phosphorus ratio is more complicated, indicating that for ratios less
250 than ~14 chlorophyll *a* increases but after ~14 there is marked decrease. The effect of the nitrogen
251 phosphorus ratio on chlorophyll has been the subject of considerable research and our results are
252 consistent with the majority of the findings suggesting that at low ratio values nitrogen is limiting
253 (Downing and McCauley 1992, Smith and Schindler 2009). Conversely at higher ratios the phosphorus
254 levels may be limiting. This would be a cause for concern with linear models; however, linearity is not
255 an assumption of tree-based modelling approaches such as random forest.

256 Turbidity was selected as the most important variable in the “All variables” model. The partial
257 dependency analysis shows that, similar to the nutrients discussed above, log chlorophyll *a* increases
258 with increased turbidity. At first this may seem counter intuitive since we might expect productivity to
259 decrease as turbidity increases, and therefore light availability decreases (Tilzer 1988, Bilotta and Brazier
260 2008). However, algal biomass can contribute heavily to measures of turbidity and we expect greater
261 productivity to lead to increased turbidity (Hansson 1992). We interpret this pattern as indicating that
262 as chlorophyll *a* concentrations increase we see a concomitant increase in turbidity.

263 Elevation was selected as an important predictive variable in both the all variables and the GIS only
264 models; the partial dependencies (Figures 4 & 7) indicate a negative relationship between elevation and
265 chlorophyll *a* concentration that is probably due to fact that the location of mountains in the United
266 States is the spatial inverse of the distribution of agricultural and urban lands. As elevation increases
267 we expect decreased loads due to smaller watershed contributing areas. In contrast lower elevation
268 sites will have larger drainage areas and greater potential for increased nutrient loads from urban and

269 agricultural sources.

270 The variables in the “GIS only” model captured the large scale spatial pattern of the trophic status
271 gradient of lakes across the United States. In addition to elevation, mentioned above, the model was
272 most sensitive to latitude and ecoregion. In general, chlorophyll *a* concentrations are highest in the
273 Southern portions of the study area where temperatures can be higher (a known driver of productivity),
274 elevations lower, and agricultural impacts more pronounced. Likewise ecoregion (see Figure 13) has a
275 pronounced affect indicting continental scale effects of land use and geography. Agriculturally dominated
276 landscapes such as the Temperate Plains, Southern Plains, and Coastal Plains show the highest levels
277 of Chlorophyll *a*. Whereas high elevation zones (Western Mountains), arid lands (Xeric), Northern
278 habitats (Upper Midwest) have lower concentrations.

279 Further evidence for the role of land use/land cover variables is shown by the selection of the percent
280 cropland and percent evergreen forest variables. As indicated by the partial dependency plots (Figure
281 7), chlorophyll *a* increases with cropland and decreases with evergreen cover. It is not surprising that
282 croplands were selected given the overwhelming impact of agriculture on the eutrophication process.
283 The negative association of evergreens and chlorophyll *a* concentrations (Figure 7). As the percent of
284 evergreens increases we are likely to see increased elevation and soil differences that limit agriculture.

285 Lastly, morphometry (e.g. depth) also proved to be important in the prediction of lake trophic state
286 (Genkai-Kato and Carpenter 2005). As morphometry shows little to no broad scale spatial pattern and
287 is unique to a given lake, these data are likely illuminating the local, lake scale drivers such as in-lake
288 nutrient processing and residence time.

289 5 Conclusions

290 Our research goals were to explore the utility of a widely used data mining algorithm, random forests,
291 in the modelling of chlorophyll *a* and lake trophic state. Further, we hoped to examine the utility of
292 these models when built with only ubiquitous GIS data, which allows estimation of trophic state for all
293 lakes in the United States. The “All variables” model had an RMSE of 0.09 and an adjusted R^2 of 0.8
294 whereas, the GIS only models had an RMSE of 0.22 and the adjusted R^2 was 0.48. Our total accuracy

295 in predicting chlorophyll *a* based trophic states was 69% for the “All variables” model and 49% for the
296 “GIS only” model.

297 While the “GIS only” model showed lower prediction accuracies than the “All variables” model, the
298 association between the uncertainty of prediction and total accuracy (Figure 9 and Table 4) suggest
299 that the “GIS only” model will provide reasonable estimates of trophic state for many lakes across the
300 United States. Furthermore, we can map the uncertainty of the predictions, thus, we know the spatial
301 patterns and location of the lakes for which we are certain, or not, of their predicted trophic state.
302 Given this and that these models may be applied to any lake in the United States we can recommend
303 using this model. Future iterations of this modelling effort may be able to utilize modeled predictions
304 of nutrients to improve accuracy and also maintain broad applicability (Milstead et al. 2013).

305 For the “All variables” model, the *in situ* water quality variables drove the predictions. This is not
306 surprising. For the “GIS only” model, the results were more nuanced. Three broad categories were
307 routinely being selected as important: broad scale spatial patterns in trophic state, land use/land cover
308 controls of trophic state, and local, lake-scale control driven by lake morphometry.

309 A potentially useful benefit of models of trophic state and chlorophyll *a* are their use in assessing risk due
310 to cyanobacteria. Cyanobacteria biomass should be closely associated with chlorophyll *a* and trophic
311 state as cyanobacteria contribute to the chlorophyll concentration in a lake. If these associations are
312 strong enough we may be able to expand models such as those reported here to also predict probability
313 of cyanobacteria blooms and other indices related to cyanobacteria (e.g. toxin presence). Others have
314 seen these associations. For instance, Yuan et al. (2014) used the 2007 NLA to demonstrate that
315 total nitrogen and chlorophyll *a* concentrations were good predictors of World Health Organization
316 microcystin (a toxin produced by some cyanobacteria) criteria exceedences. Using this same data, we
317 see a positive trend between chlorophyll *a* and cyanobacteria abundance (Figure 14). Both of these
318 suggest that trophic state may be an acceptable proxy for cyanobacteria abundance or presence of
319 microcystin.

320 Our results raise three important considerations related to managing eutrophication. First, the broad
321 scale patterning, indicated by ecoregion as an important variable, suggests regional trends. This
322 is noteworthy because it suggests that efforts to monitor, model and manage eutrophication and

323 cyanobacteria should be undertaken at both national and regional levels. Second, while direct control
324 of water quality in lakes would have a large impact, the land use/land cover drivers (i.e. non-point
325 sources) of water quality are also important, and better management of the spatial distribution of
326 important classes such as forest and agriculture can provide some level of control on trophic state and
327 amount of cyanobacteria present. Third, in-lake processes (i.e. residence time, nutrient cycling, etc.)
328 are, as expected, important and need to be part of any management strategy. Building on these efforts
329 through updated models, direct prediction of cyanobacteria, and additional information on the regional
330 differences will help us get a better handle on the broad scale dynamics of productivity in lakes and the
331 potential risk to human health from cyanobacteria blooms.

332 **6 Acknowledgements**

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

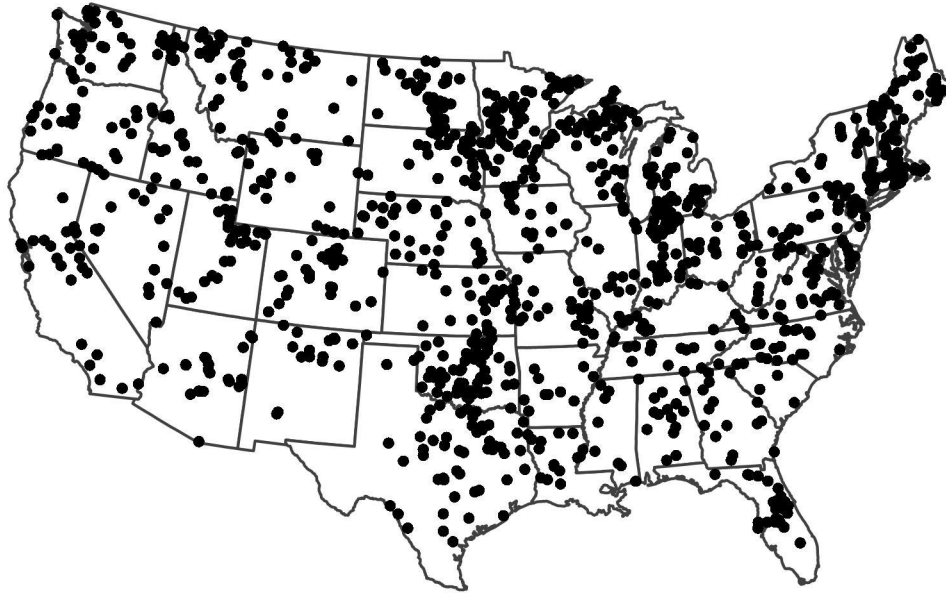


Figure 1: Map of the distribution of National Lakes Assessment Sampling locations

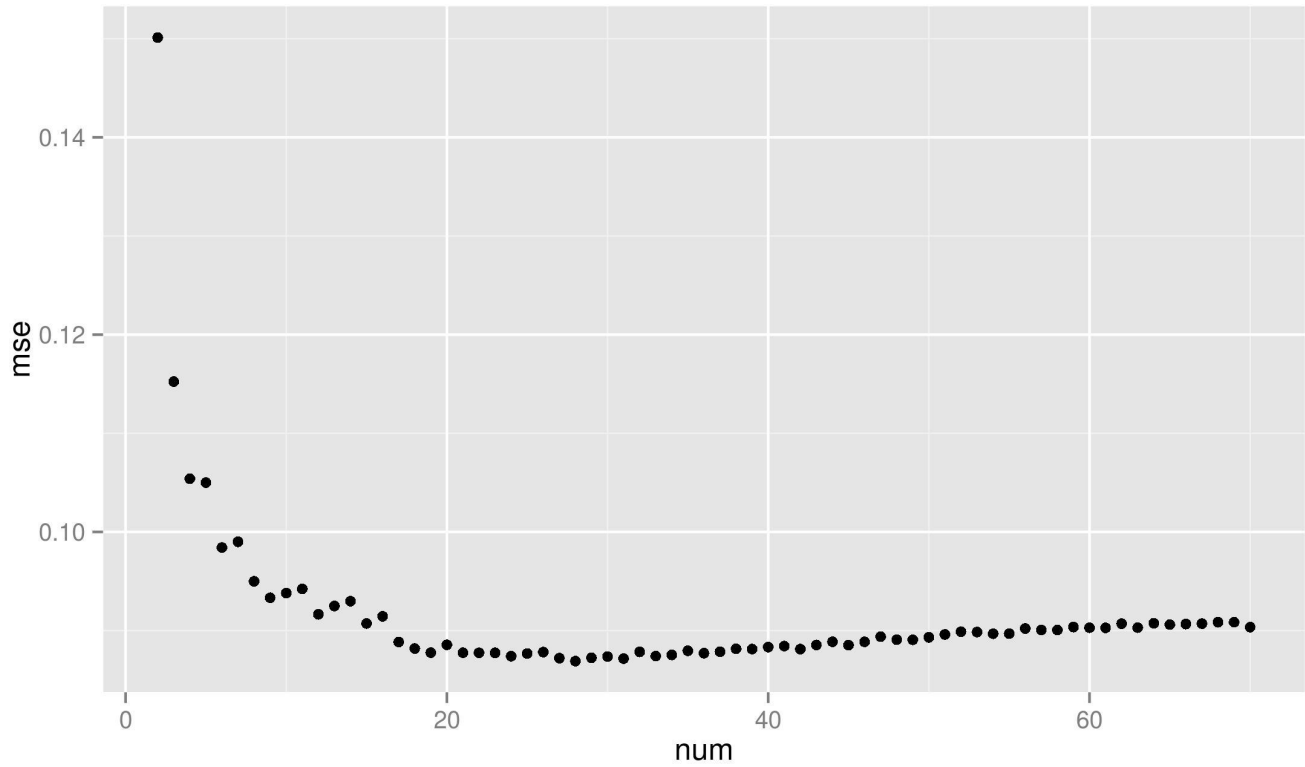


Figure 2: Variable selection plot for all variables. Shows percent increase in mean squared error as a function of the number of variables.

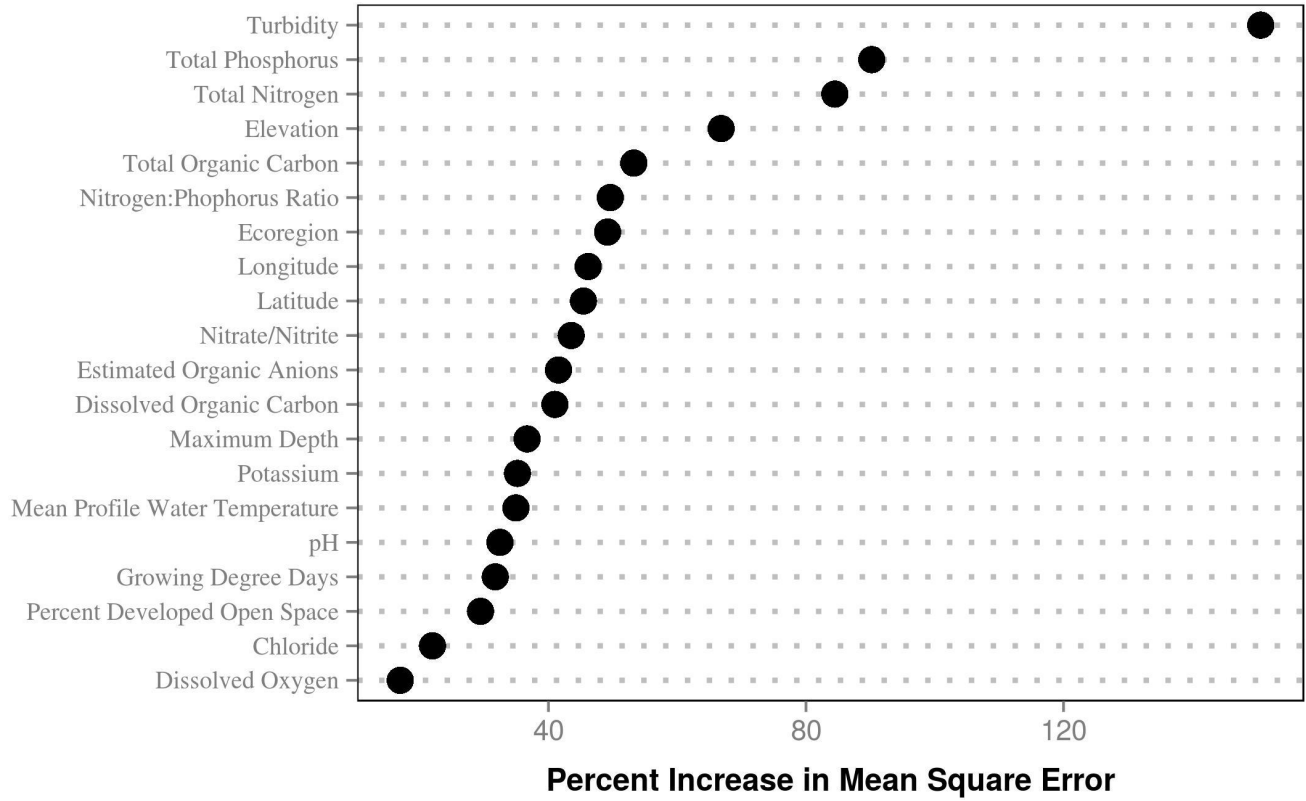


Figure 3: Importance plot for All Variables., shows percent increase in mean square error. Higher values of percent increase in mean squared error indicates higher importance.

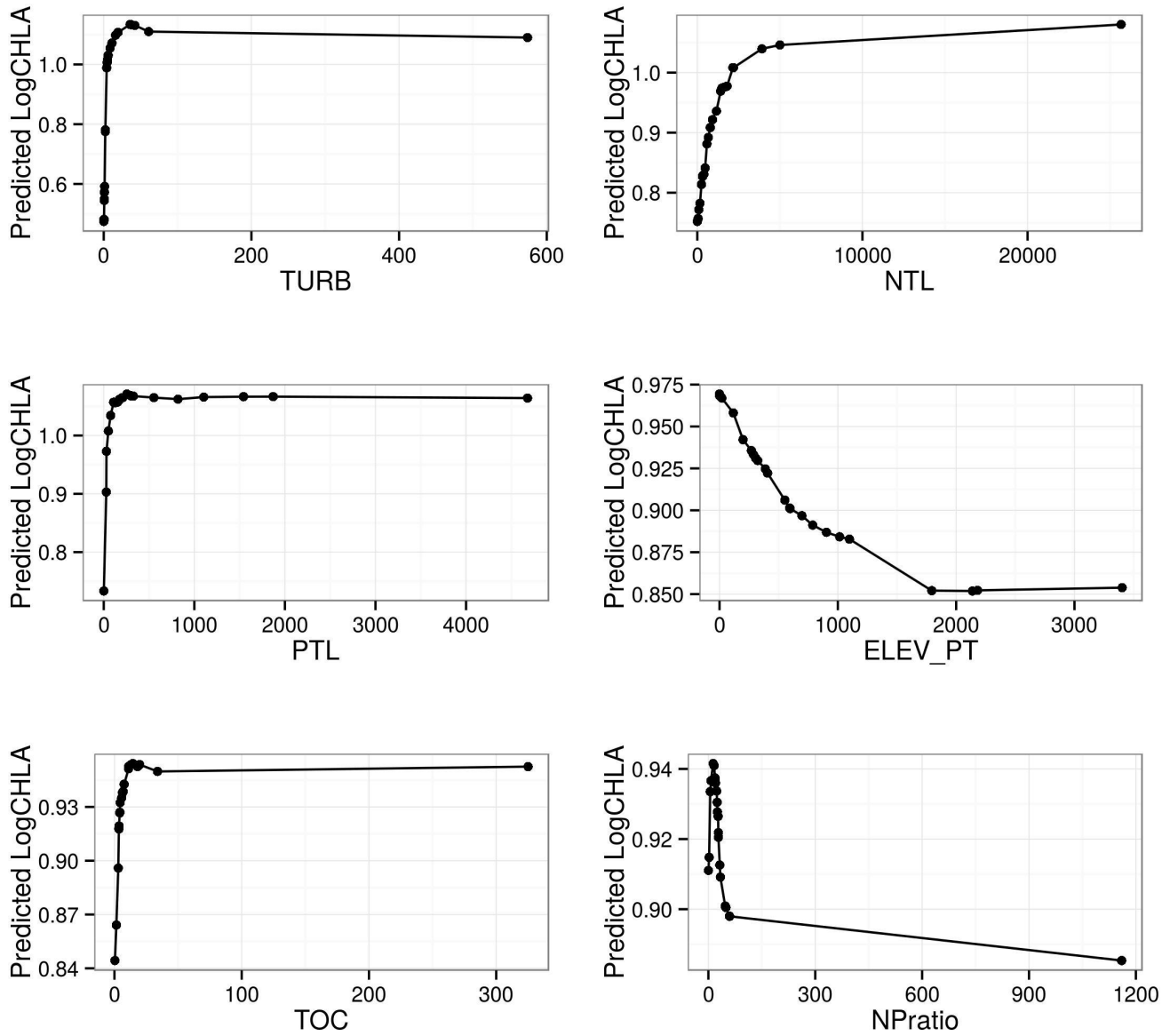


Figure 4: All Variables partial dependence plots for the top 5 most important variables.

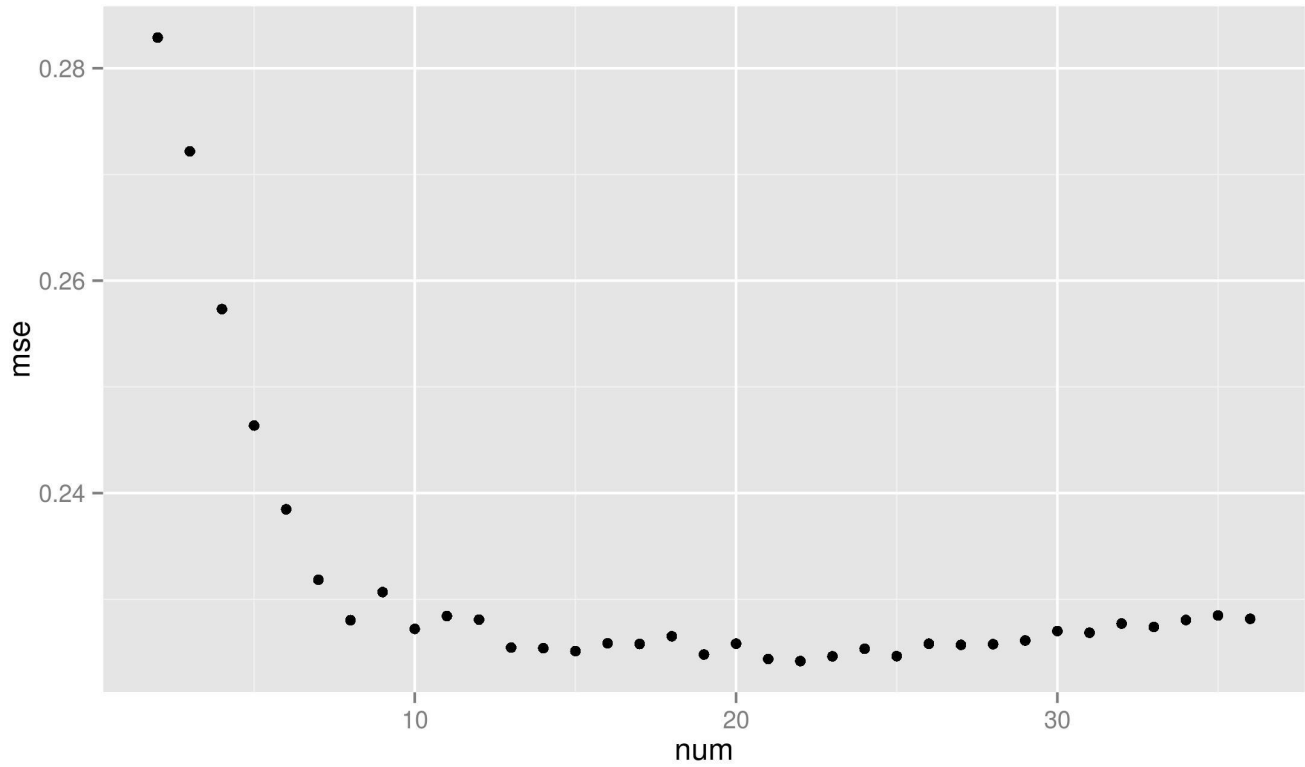


Figure 5: Variable selection plot for GIS only variables. Shows percent increase in mean squared error as a function of the number of variables.

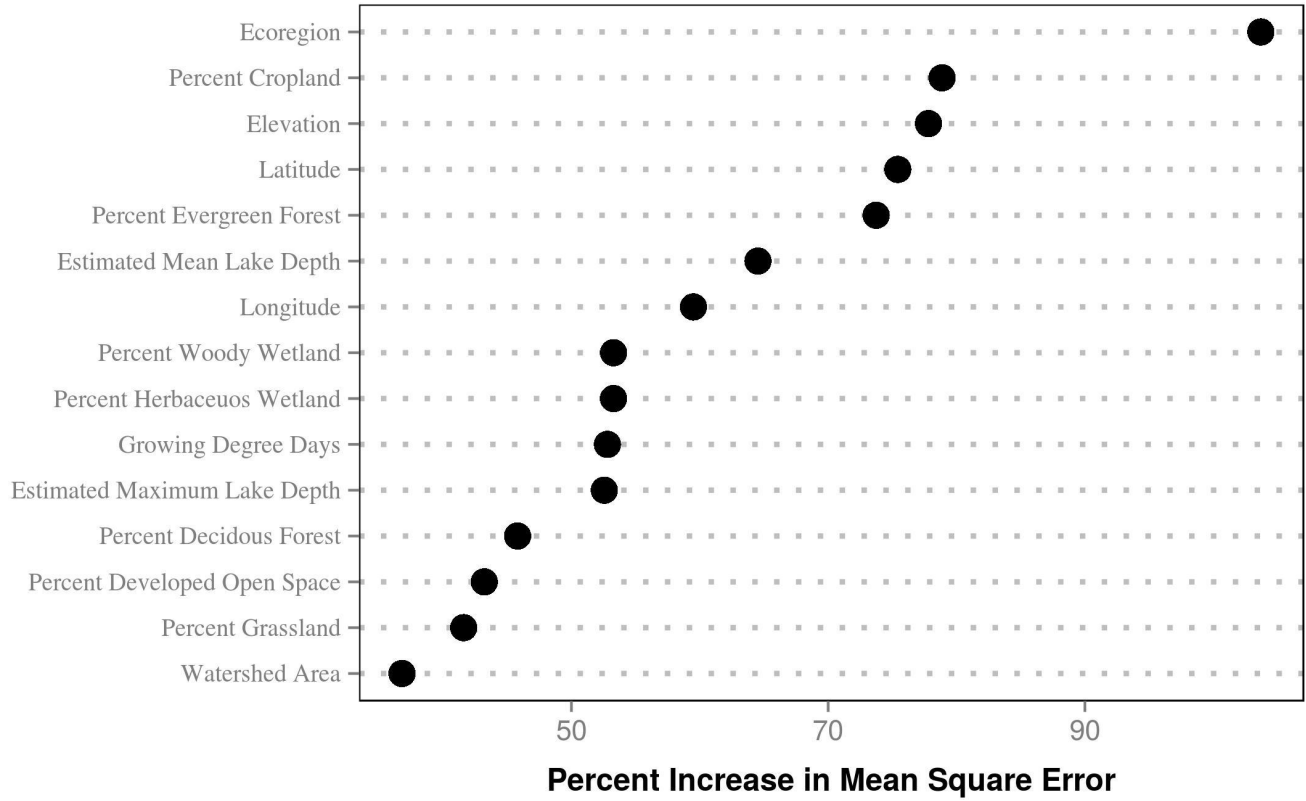


Figure 6: Importance plot for GIS Only Variables., shows percent increase in mean square error. Higher values of percent increase in mean squared error indicates higher importance.

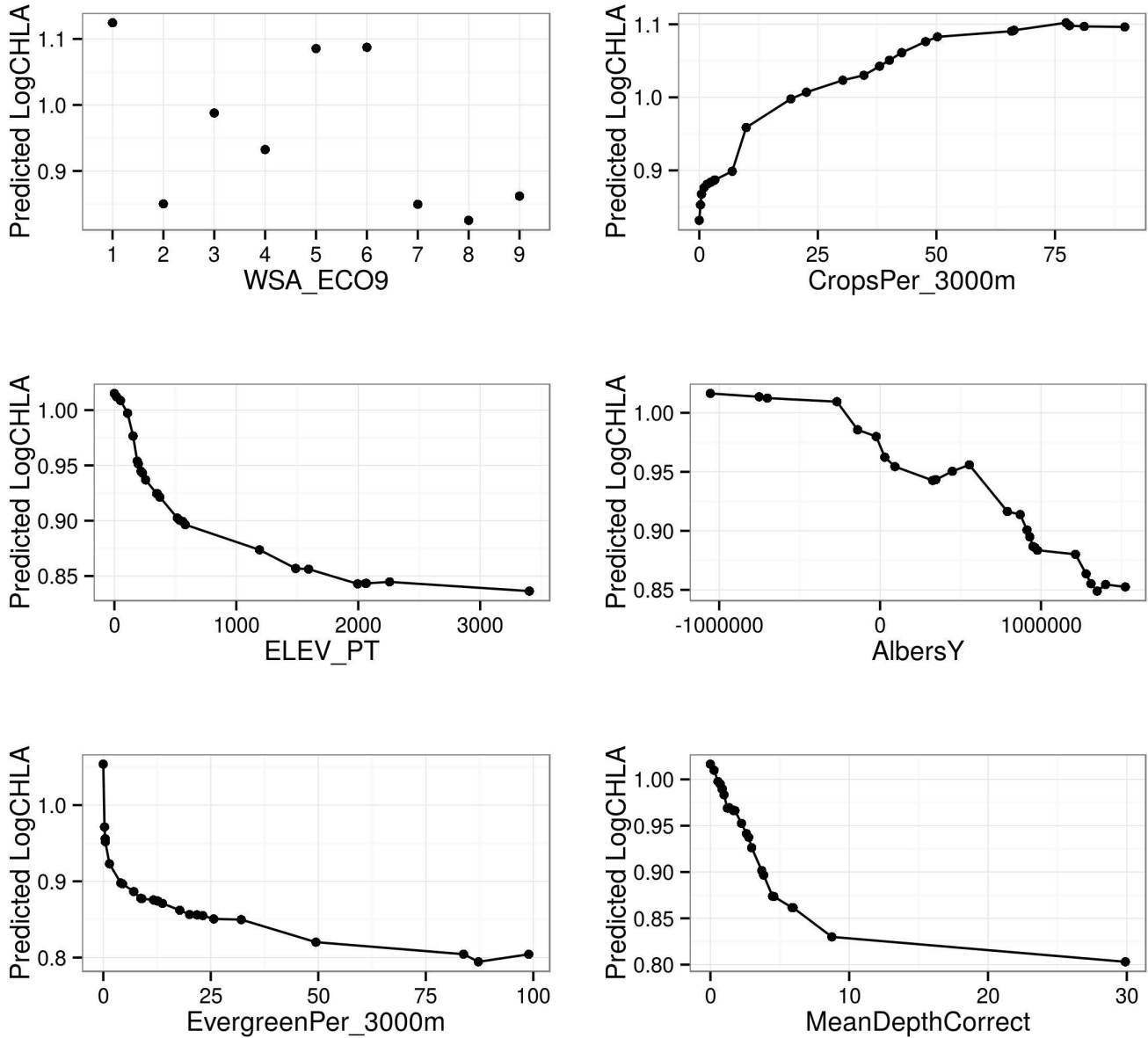


Figure 7: GIS Only Variables partial dependence plots for the top 5 most important variables.

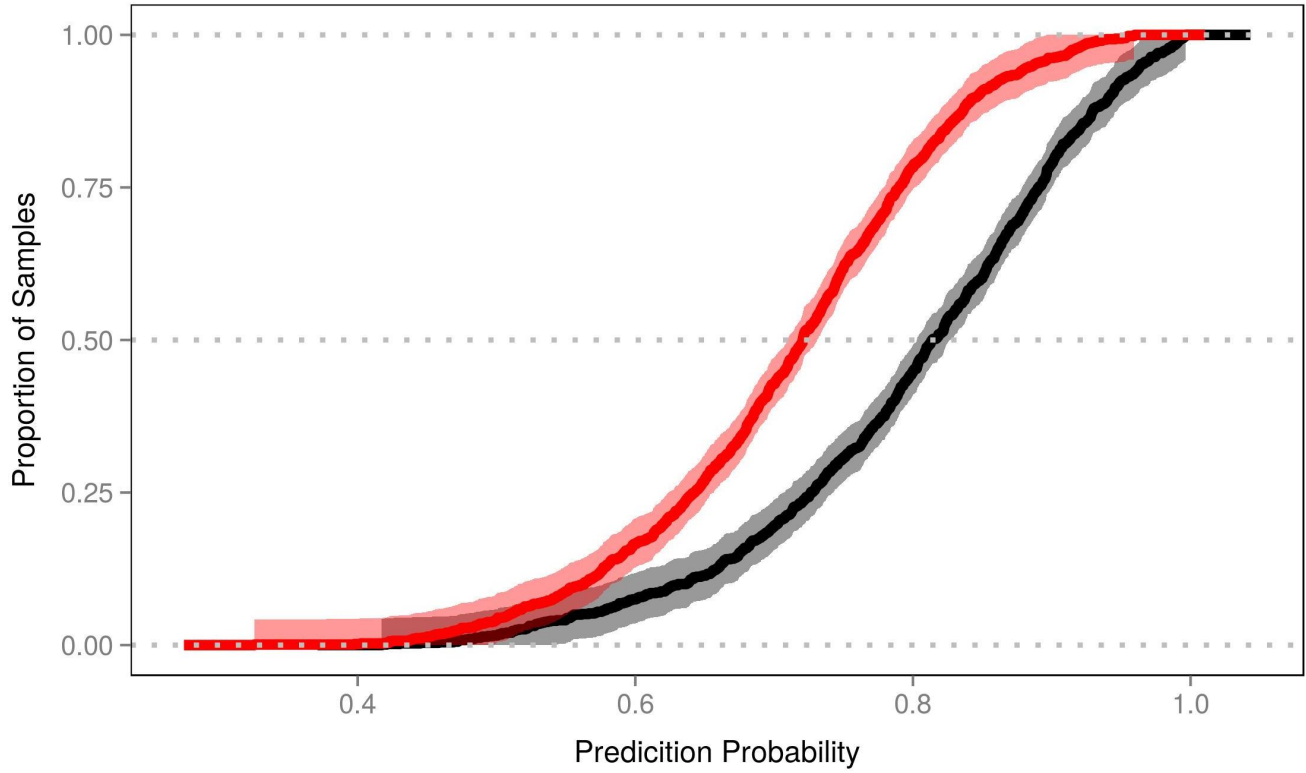


Figure 8: Prediction probabilities for the All Variables and GIS Only models.

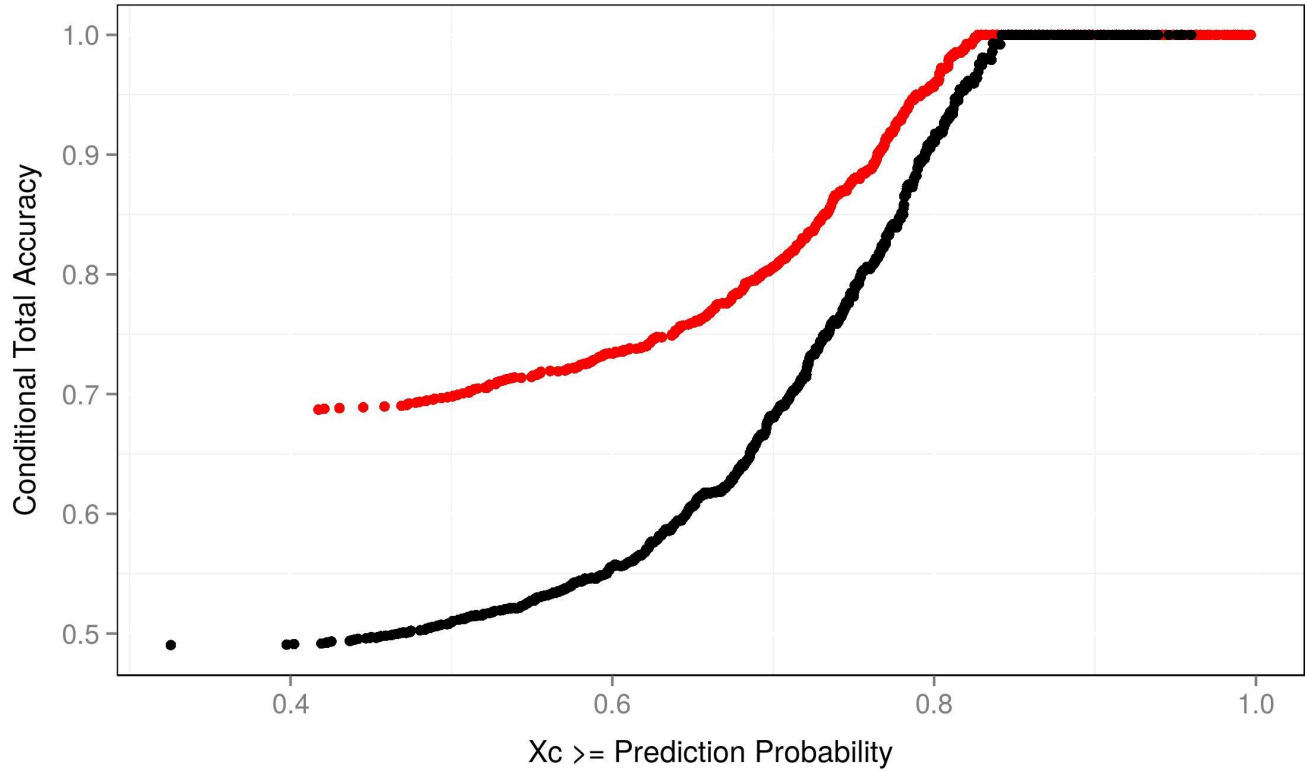


Figure 9: Accuracy of predictions as a function of lake prediction probability. The x-axis represents lakes with a prediction probability at a given level or higher.

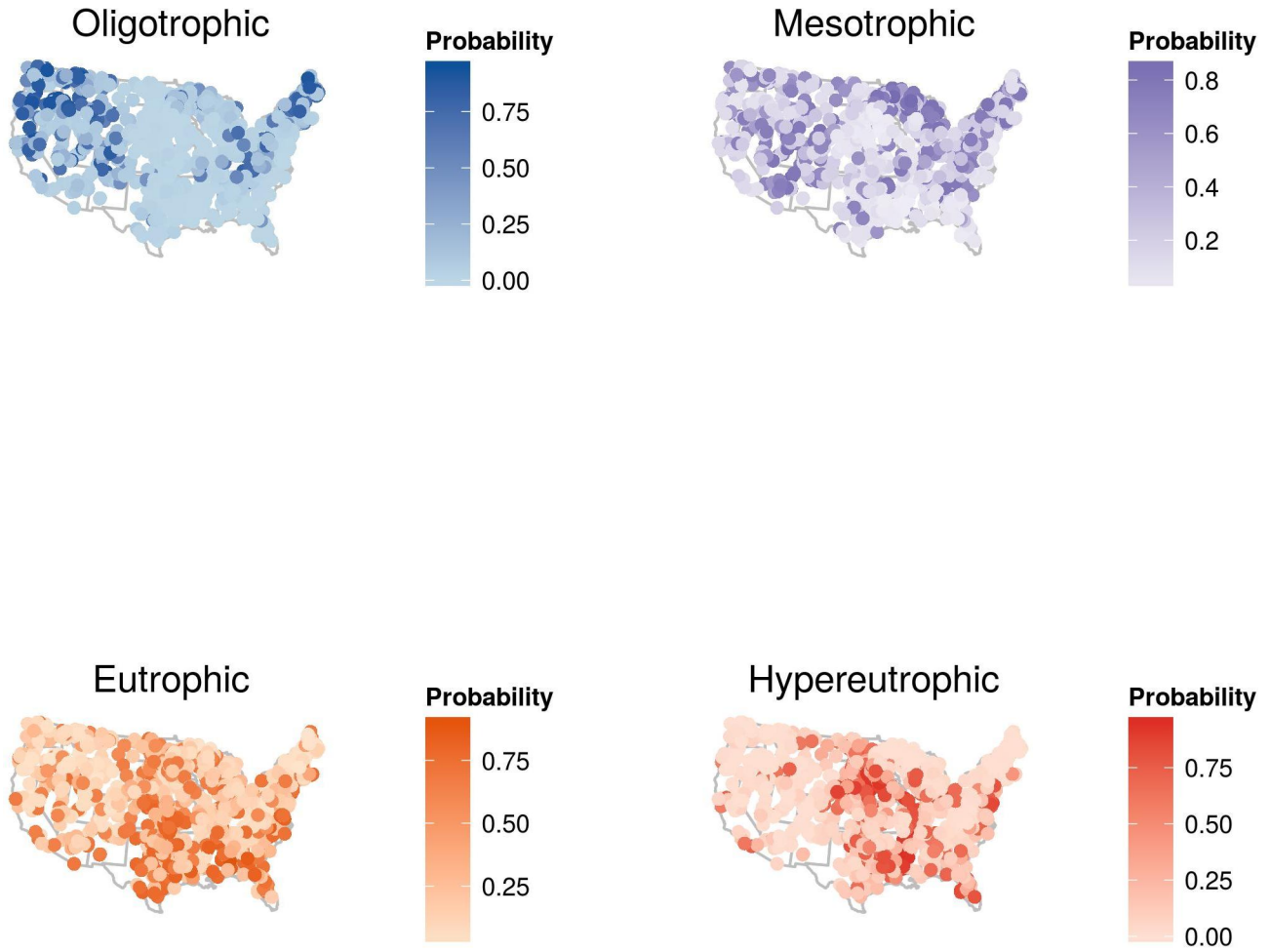


Figure 10: Maps of prediction probabilities for each of the four chlorophyll *a* trophic states

Predicted Probability

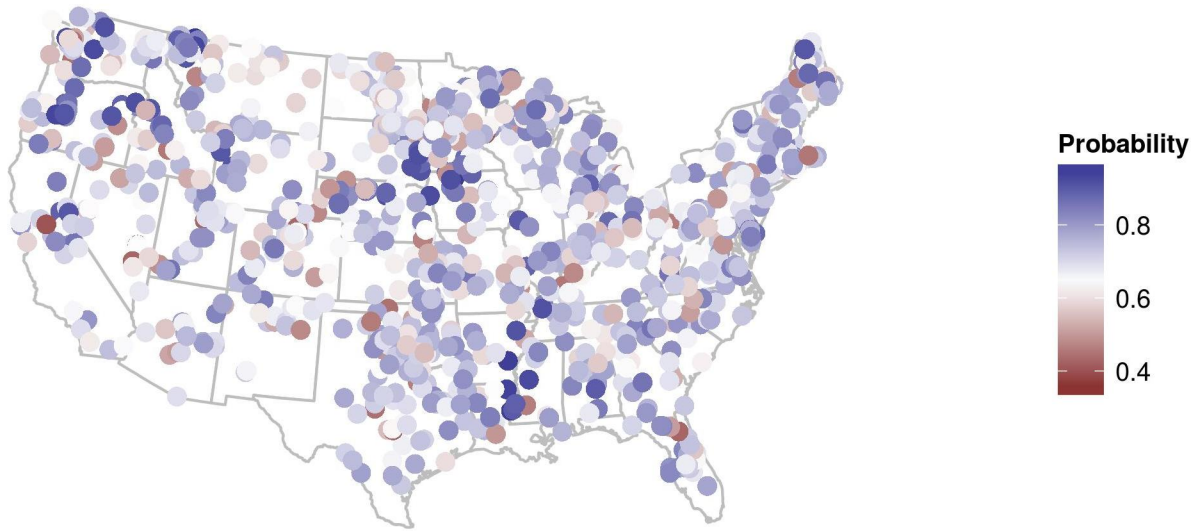


Figure 11: Maps of prediction probabilities for the discrete, predicted chlorophyll *a* trophic state. Shows spatial patterns of prediction uncertainty.

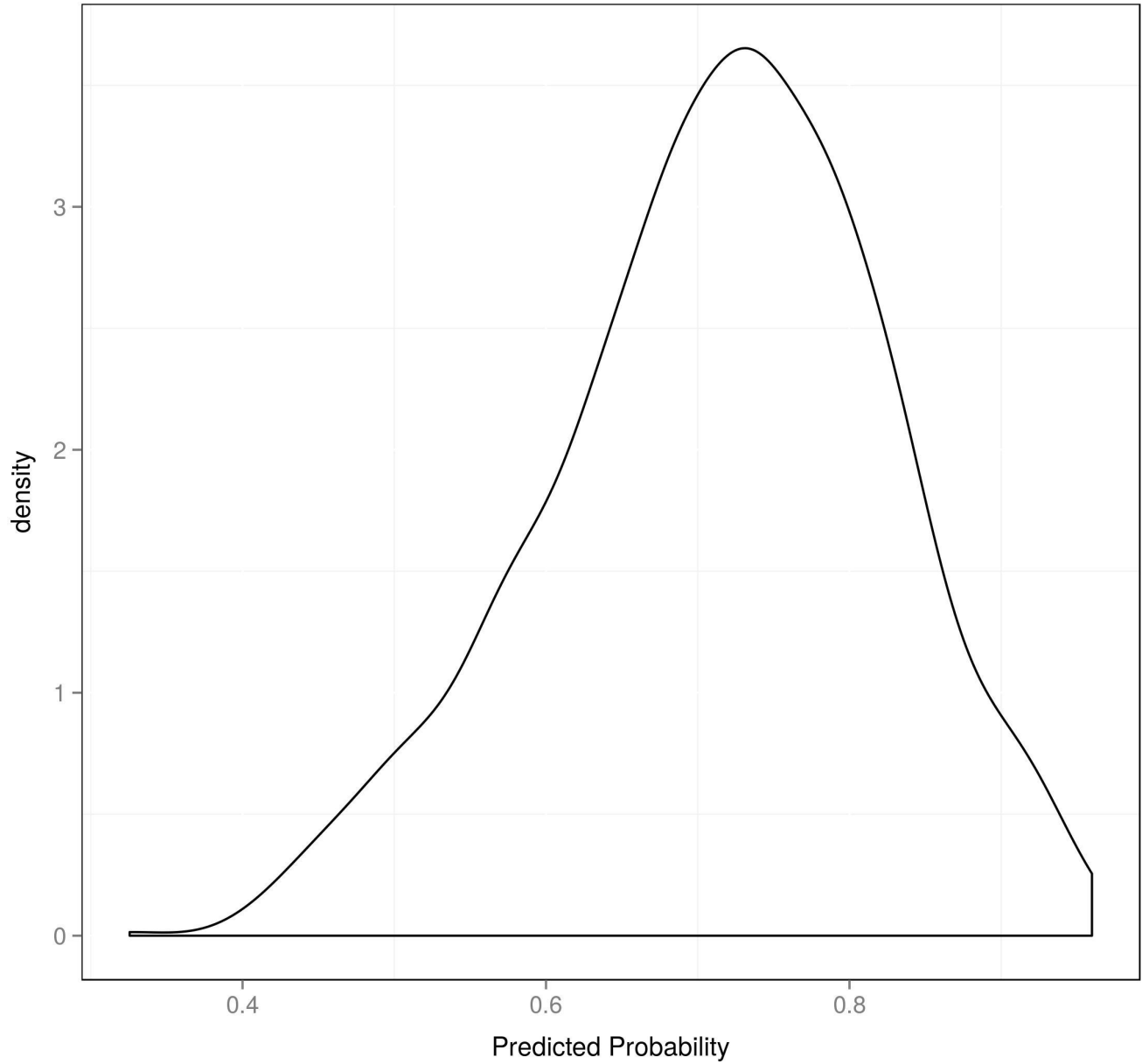


Figure 12: Distribution of predicted probabilities.

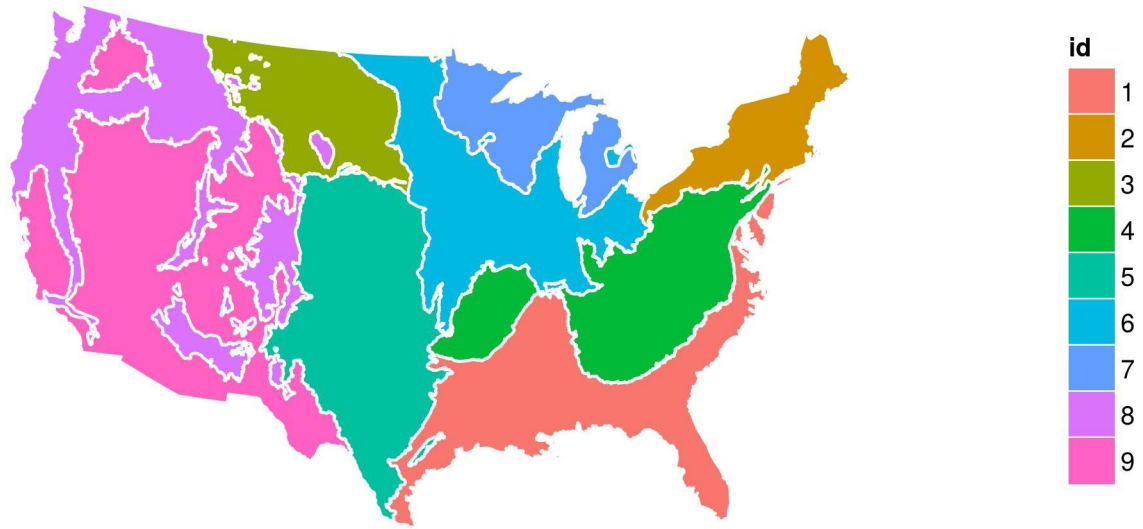


Figure 13: Wadeable Streams assesment ecoregions

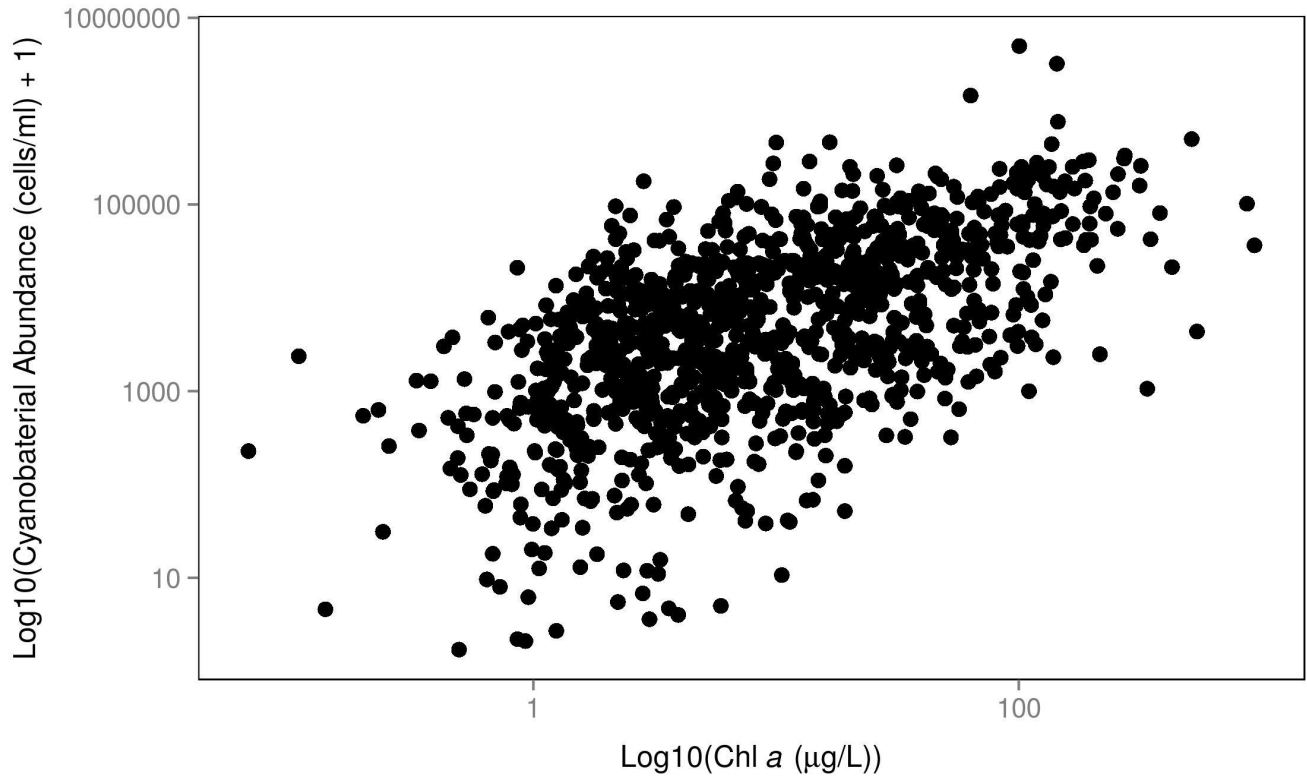


Figure 14: Cholorphyll *a* and cyanobacteria abundance scatterplot

341 **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	≤ 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

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

Table 4: Summary of relationship between prediction probabilities, total accuracy, and number of lakes.

Prediction Prob.	“All Var.”			“GIS Only”		
	Total Accuracy	Percent of Sample	Number of Samples	Total Accuracy	Percent of Sample	Number of Samples
All	69	100	846	49	100	878
0.50	70	98	829	51	95	834
0.60	73	91	770	56	81	711
0.70	81	77	654	68	56	490
0.80	96	51	434	91	24	212
0.90	100	20	173	100	5	41

342 **9 Appendix 1. Variable Definitions**

Variable Names	Description	Source
AlbersX	Longitude	GIS
AlbersY	Latitude	GIS
BASINAREA	Watershed Area	GIS
BarrenPer_3000m	Percent Barren	GIS
CropsPer_3000m	Percent Cropland	GIS
DDs45	Growing Degree Days	GIS
DeciduousPer_3000m	Percent Deciduous Forest	GIS
DevHighPer_3000m	Percent High Intensity Development	GIS
DevLowPer_3000m	Percent Low Intensity Development	GIS
DevMedPer_3000m	Percent Medium Intensity Development	GIS
DevOpenPer_3000m	Percent Developed Open Space	GIS
ELEV_PT	Elevation	GIS
EvergreenPer_3000m	Percent Evergreen Forest	GIS
FetchE	Fetch from East	GIS
FetchN	Fetch from North	GIS
FetchNE	Fetch from Northeast	GIS
FetchSE	Fetch from Southeast	GIS
GrassPer_3000m	Percent Grassland	GIS
HerbWetPer_3000m	Percent Herbaceous Wetland	GIS
IceSnowPer_3000m	Percent Ice/Snow	GIS
LakeArea	Lake Surface Area	GIS
LakePerim	Lake Perimeter	GIS
MaxDepthCorrect	Estimated Maximum Lake Depth	GIS
MaxLength	Maximum Lake Length	GIS
MaxWidth	Maximum Lake Width	GIS
MeanDepthCorrect	Estimated Mean Lake Depth	GIS

Variable Names	Description	Source
MeanWidth	Mean Lake Width	GIS
MixedForPer_3000m	Percent Mixed Forest	GIS
PasturePer_3000m	Percent Pasture	GIS
PercentImperv_3000m	Percent Impervious	GIS
ShoreDevel	Shoreline Development Index	GIS
ShrubPer_3000m	Percent Shrub/Scrub	GIS
VolumeCorrect	Estimated Lake Volume	GIS
WSA_ECO9	Ecoregion	GIS
WaterPer_3000m	Percent Water	GIS
WoodyWetPer_3000m	Percent Woody Wetland	GIS
ANC	Acid Neutralizing Capacity	Water Quality
ANDEF2	Anion Deficit	Water Quality
ANSUM2	Anion Sum	Water Quality
BALANCE2	Ion Balance	Water Quality
CA	Calcium	Water Quality
CATSUM	Cation Sum	Water Quality
CL	Chloride	Water Quality
COLOR	Color	Water Quality
CONCAL2	Calculated Conductivity	Water Quality
COND	Conductivity	Water Quality
CONDHO2	D-H-O Calculated Conductivity	Water Quality
DATE_COL	Date Samples Collected	Water Quality
DEPTHMAX	Maximum Depth	Water Quality
DO2_2M	Dissolved Oxygen	Water Quality
DOC	Dissolved Organic Carbon	Water Quality
H	Hydrogen Ions	Water Quality
K	Potassium	Water Quality
MG	Magnesium	Water Quality

Variable Names	Description	Source
NH4	Ammonium	Water Quality
NH4ION	Calculate Ammonium	Water Quality
NO3	Nitrate	Water Quality
NO3_NO2	Nitrate/Nitrite	Water Quality
NPratio	Nitrogen:Phophorus Ratio	Water Quality
NTL	Total Nitrogen	Water Quality
Na	Sodium	Water Quality
OH	Hydroxide	Water Quality
ORGION	Estimated Organic Anions	Water Quality
PH_FIELD	pH	Water Quality
PTL	Total Phosphorus	Water Quality
SIO2	Silica	Water Quality
SO4	Sulfate	Water Quality
SOBC	Base Cation Sum	Water Quality
TOC	Total Organic Carbon	Water Quality
TURB	Turbidity	Water Quality
TmeanW	Mean Profile Water Temperature	Water Quality

343 **References**

- 344 BILOTTA, G., AND R. BRAZIER. 2008. Understanding the influence of suspended solids on water
345 quality and aquatic biota. *Water research* 42:2849–2861.
- 346 BREIMAN, L. 2001. Random forests. *Machine learning* 45:5–32.
- 347 CARLSON, R. E. 1977. A trophic state index for lakes. *Limnology and oceanography* 22:361–369.
- 348 CARVALHO, L., C. A. MILLER, E. M. SCOTT, G. A. CODD, P. S. DAVIES, AND A. N. TYLER. 2011.
349 Cyanobacterial blooms: Statistical models describing risk factors for national-scale lake assessment and
350 lake management. *Science of The Total Environment* 409:5353–5358.
- 351 COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological*
352 *Measurement* 20:37–46.
- 353 CUTLER, D. R., T. C. EDWARDS JR, K. H. BEARD, A. CUTLER, K. T. HESS, J. GIBSON, AND J. J.
354 LAWLER. 2007. Random forests for classification in ecology. *Ecology* 88:2783–2792.
- 355 DÍAZ-URIARTE, R., AND S. A. DE ANDRES. 2006. Gene selection and classification of microarray
356 data using random forest. *BMC bioinformatics* 7:3.
- 357 DOWNING, J. A., AND E. McCAULEY. 1992. The nitrogen: Phosphorus relationship in lakes.
358 *Limnology and Oceanography* 37:936–945.
- 359 DOWNING, J. A., S. B. WATSON, AND E. McCAULEY. 2001. Predicting cyanobacteria dominance in
360 lakes. *Canadian journal of fisheries and aquatic sciences* 58:1905–1908.
- 361 FERNÁNDEZ-DELGADO, M., E. CERNADAS, S. BARRO, AND D. AMORIM. 2014. Do we need
362 hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*
363 15:3133–3181.
- 364 GENKAI-KATO, M., AND S. R. CARPENTER. 2005. Eutrophication due to phosphorus recycling in
365 relation to lake morphometry, temperature, and macrophytes. *Ecology* 86:210–219.

- 366 HANSSON, L.-A. 1992. Factors regulating periphytic algal biomass. *Limnology and Oceanography*
367 37:322–328.
- 368 HASLER, A. D. 1969. Cultural eutrophication is reversible. *BioScience* 19:425–431.
- 369 HOLLISTER, J. W. 2014. Lakemorpho: Lake morphometry in R. R package version 1.0. [http://CRAN.R-](http://CRAN.R-project.org/package=lakemorpho)
370 [project.org/package=lakemorpho](http://CRAN.R-project.org/package=lakemorpho).
- 371 HOLLISTER, J. W., W. B. MILSTEAD, AND M. A. URRUTIA. 2011. Predicting maximum lake depth
372 from surrounding topography. *PLoS ONE* 6:e25764.
- 373 HOLLISTER, J. W., H. A. WALKER, AND J. F. PAUL. 2008. CProb: A computational tool for
374 conducting conditional probability analysis. *Journal of environmental quality* 37:2392–2396.
- 375 HOLLISTER, J., AND W. B. MILSTEAD. 2010. Using gIS to estimate lake volume from limited data.
376 *Lake and Reservoir Management* 26:194–199.
- 377 HOMER, C., C. HUANG, L. YANG, B. WYLIE, AND M. COAN. 2004. Development of a 2001 national
378 land-cover database for the united states. *Photogrammetric Engineering & Remote Sensing* 70:829–840.
- 379 HUBERT, L., AND P. ARABIE. 1985. Comparing partitions. *Journal of classification* 2:193–218.
- 380 IMBODEN, D., AND R. GÄCHTER. 1978. A dynamic lake model for trophic state prediction. *Ecological*
381 *modelling* 4:77–98.
- 382 JONES, J., M. KNOWLTON, D. OBRECHT, AND E. COOK. 2004. Importance of landscape variables
383 and morphology on nutrients in missouri reservoirs. *Canadian Journal of Fisheries and Aquatic Sciences*
384 61:1503–1512.
- 385 JONES, K. B., A. C. NEALE, M. S. NASH, R. D. VAN REMORTEL, J. D. WICKHAM, K. H. RIITERS,
386 AND R. V. O'NEILL. 2001. Predicting nutrient and sediment loadings to streams from landscape
387 metrics: A multiple watershed study from the united states mid-atlantic region. *Landscape Ecology*
388 16:301–312.
- 389 JONES, Z., AND F. LINDER. 2015. Exploratory data analysis using random forests. *in* *The 73rd annual*

390 mPSA conference. MPSA.

391 LANDIS, J. R., AND G. G. KOCH. 1977. The measurement of observer agreement for categorical data.
392 *biometrics* 33:159–174.

393 LIAW, A., AND M. WIENER. 2002. Classification and regression by randomForest. *R News* 2:18–22.

394 MILSTEAD, W. B., J. W. HOLLISTER, R. B. MOORE, AND H. A. WALKER. 2013. Estimating summer
395 nutrient concentrations in northeastern lakes from sPARROW load predictions and modeled lake depth
396 and volume. *PLoS one* 8:e81457.

397 MOORE, R. B., C. M. JOHNSTON, R. A. SMITH, AND B. MILSTEAD. 2011. Source and delivery of
398 nutrients to receiving waters in the northeastern and mid-atlantic regions of the united states. *JAWRA*
399 *Journal of the American Water Resources Association* 47:965–990.

400 PAUL, J. F., AND M. E. McDONALD. 2005. Development of empirical, geographically specific water
401 quality criteria: A conditional probability analysis approach 41:1211–1223.

402 PETERS, J., B. D. BAETS, N. E. VERHOEST, R. SAMSON, S. DEGROEVE, P. D. BECKER, AND W.
403 HUYBRECHTS. 2007. Random forests as a tool for ecohydrological distribution modelling. *Ecological*
404 *Modelling* 207:304–318.

405 RODHE, W. 1969. Crystallization of eutrophication concepts in northern europe.

406 SALAS, H. J., AND P. MARTINO. 1991. A simplified phosphorus trophic state model for warm-water
407 tropical lakes. *Water research* 25:341–350.

408 SCHINDLER, D. W., AND J. R. VALLENTYNE. 2008. The algal bowl: Overfertilization of the world's
409 freshwaters and estuaries. Page 334. University of Alberta Press Edmonton.

410 SEILHEIMER, T. S., P. L. ZIMMERMAN, K. M. STUEVE, AND C. H. PERRY. 2013. Landscape-scale
411 modeling of water quality in lake superior and lake michigan watersheds: How useful are forest-based
412 indicators? *Journal of Great Lakes Research* 39:211–223.

413 SMITH, V. H. 1998. Cultural eutrophication of inland, estuarine, and coastal waters. Pages 7–49 *in*

- 414 Successes, limitations, and frontiers in ecosystem science. Springer.
- 415 SMITH, V. H., AND D. W. SCHINDLER. 2009. Eutrophication science: Where do we go from here?
416 Trends in Ecology & Evolution 24:201–207.
- 417 SMITH, V. H., S. B. JOYE, R. W. HOWARTH, AND OTHERS. 2006. Eutrophication of freshwater and
418 marine ecosystems. Limnology and Oceanography 51:351–355.
- 419 SMITH, V. H., G. D. TILMAN, AND J. C. NEKOLA. 1999. Eutrophication: Impacts of excess nutrient
420 inputs on freshwater, marine, and terrestrial ecosystems. Environmental pollution 100:179–196.
- 421 STROBL, C., A.-L. BOULESTEIX, A. ZEILEIS, AND T. HOTHORN. 2007. Bias in random forest variable
422 importance measures: Illustrations, sources and a solution. BMC bioinformatics 8:25.
- 423 TILZER, M. M. 1988. Secchi disk—chlorophyll relationships in a lake with highly variable phytoplankton
424 biomass. Hydrobiologia 162:163–171.
- 425 USEPA. 2009. National lakes assessment: A collaborative survey of the nation's lakes. ePA 841-r-09-001.
426 Office of Water; Office of Research; Development, US Environmental Protection Agency Washington,
427 DC.
- 428 XIAN, G., C. HOMER, AND J. FRY. 2009. Updating the 2001 national land cover database land
429 cover classification to 2006 by using landsat imagery change detection methods. Remote Sensing of
430 Environment 113:1133–1147.
- 431 YUAN, L. L., A. I. POLLARD, S. PATHER, J. L. OLIVER, AND L. D'ANGLADA. 2014. Managing
432 microcystin: Identifying national-scale thresholds for total nitrogen and chlorophyll a. Freshwater
433 Biology 59:1970–1981.