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, and disservices (e.g. recreation, aesthetics, and harmful algal blooms). While the relationship between nutrients and trophic state provides reliable predictions, it requires *in situ* water quality data in order to parameterize the model. This limits the application of these models to lakes with existing and, more importantly, available water quality data. To address this, we take advantage of the availability of a large national lakes water quality database (i.e. the National Lakes Assessment), land use/land cover data, lake morphometry data, other universally available data, and apply modern data mining approaches to predict trophic state. Using this data and random forests, we first model chlorophyll *a*, then classify the resultant predictions into trophic states. The full model estimates chlorophyll *a* with both *in situ* and universally available data. The mean squared error and adjusted R^2 of this model was 0.09 and 0.8, respectively. The second model (i.e. GIS only) uses universally available GIS data only. The mean squared error was 0.22 and the adjusted R^2 was 0.48. The accuracy of various trophic state classifications derived from the chlorophyll *a* predictions ranged from 69% to 87% for the full model and from 49% to 75% 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 uncertainity around those predictions. For the full model, these predicted class probabilites ranged from 0.42 to 1. For the GIS only model, they ranged from 0.33 to 0.96. It is our conclusion that *in situ* data are required for better predictions, yet GIS and universally available data provide trophic state predictions, with estimated uncertainty, that still have the potential for a broad array of applications. The source code and data for this manuscript are available from https://github.com/USEPA/LakeTrophicModelling.

1 Introduction

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

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

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

Many published models of nutrients and trophic state in freshwater systems are based on linear modelling methods such as standard least squares regression or linear mixed models (Jones et al. 2001, 2004). While these methods have proven to be reliable, they have limitations (e.g. independence and distribution assumptions, and outlier sensitivity). Using data mining approaches, such as random forests, avoids many of the limitations, may reduce bias and often provides better predictions (Breiman 2001, Cutler et al. 2007, Peters et al. 2007, Fernández-Delgado et al. 2014). For instance, random forests are non-parametric and thus the data do not need to come from a specific distribution (e.g. Gaussian) and can contain collinear variables (Cutler et al. 2007). Second, random forests work well with very large numbers of predictors (Cutler et al. 2007). Lastly, random forests can deal with model selection uncertainty as predictions are based upon a consensus of many models and not just a single model selected with some measure of goodness of fit.

To build on past work, we have identified several areas in which this research contributes. First, we build, asses, and compare two random forest models of chlorophyll *a* 1) *in situ* and universally available GIS data and then 2) universally available GIS data only. Second, we examine the important predictors for both models. Third, we examine the predictions for spatial patterns. Lastly, this paper, the code, and the data used in the models is made available as an R package from https://github.com/USEPA/LakeTrophicModelling.

2 Methods

2.1 Data and Study Area

We utilized three primary sources of data for this study, the National Lakes Assessment (NLA), the National Land Cover Dataset (NLCD), and lake 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 data were collected during the summer of 2007 and the final data were released in 2009 (USEPA

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

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

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

2.2 Predicting Trophic State with Random Forests

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

Random forests are able to handle numerous correlated variables without a decrease in prediction accuracy; however, one possible shortcoming of this approach is that the resulting model may be difficult to interpret. 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). With this method, a minimum set of variables that maximizes model accuracy is provided. This allows us to start with a full suite of predictor variables from which to select a minimum, easier to interpret set of variables. One issue with the approach in `varSelRF` is that because of the randomization inherent in random forests it is possible to get variation in the minimum selected set of variables. To account for this we repeated `varSelRF` 100 times. In our case, repeating the procedure 100 times quickly converged on a set of all possible important variables.

2.3 Model Details

Using both the `varSelRF` and `randomForest` R packages we ran models for six sets of variables and trophic state classifications (Liaw and Wiener 2002, Diaz-Uriarte 2010). These included three different combinations of the chlorophyll *a* trophic states as the dependent variables and using all variables (*in situ* and GIS variables) or the GIS only variables (i.e. no *in situ* information) as the independent variables in the random forest. A listing of all considered variables is in Appendix 1. Trophic state was defined using the NLA chlorophyll *a* trophic state cut offs and the three combinations of trophic state were used to highlight the possible error caused by misclassification of adjacent classes, such as mesotrophic and eutrophic (Table 8). 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 between models.

The six model combinations were:

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

morphometry, and landscape)

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

Our modelling work flow was as follows:

1. Use `iterVarSelRF` in the `LakeTrophicModelling` R package to identify a minimal set of variables that maximize accuracy of the random forest algorithm (Diaz-Uriarte 2010, Hollister et al. 2014). This subset of variables, the reduced model, is calculated for each of our 6 models.
2. Using R's `randomForest` package, we pass the reduced models selected with `iterVarSelRF` and assess model performance (Liaw and Wiener 2002).

2.4 Measures of Model Performance and Variable Importance

We assessed the performance of the random forest models by comparing the total prediction accuracy and the kappa coefficient of the final confusion matrix. For each of the models, the final predictions were compared to the original data via a confusion matrix. A confusion matrix shows agreement and disagreement with predicted values forming the columns of the matrix and observed values, the rows. The total accuracy (i.e. percent correctly predicted) was calculated. Since some agreement can be expected by chance alone, it is also useful to take this type of error into account. For this we calculated the kappa coefficient from the confusion matrix for each model as well (Cohen 1960, Hubert and Arabie 1985). The kappa coefficient can range from -1 to 1 with 0 equalling the agreement expected by chance alone. Values greater than 0 represent agreement greater than would be expected by chance, with values greater than 0.61 considered "substantial" agreement (Landis and Koch 1977). Negative values are rare and would indicate no agreement between the predicted and observed values. Additonally, 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).

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

3 Results

Our complete dataset includes 1148 lakes; however 5 lakes did not have chlorophyll *a* data. Thus, the base dataset for our modelling was conducted on data for 1143 lakes. As chlorophyll *a* is used to create the trophic state classifications, it was necessary to remove these data because no chlorophyll *a* trophic state could be determined for these lakes. The lakes were well distributed both across the four trophic state categories (Table 8) and spatially throughout the United States (Figure **??**).

3.1 Models

Accuracy for the models built with all predictors ranged from MSE and R2 to MSE and R2. Trophic state results were .... The GIS only models had a total accuracy between 'from MSE and R2 to MSE and R2. Trophic state results were .... The importance of variables Details for each model are discussed below.

3.1.1 All Variables

The all variables model built was built using 1080 total observations. The variable selection process for this model produced a reduced model with (Figure **??**). The most important variables were ecoregion, growing degree days, and percent evergreen. (Figure **??**). MORE HERE.

3.1.2 GIS Only Variables

The gis only variables model built was built using 1138 total observations. The variable selection process for this model produced a reduced model with (Figure **??**). The most important variables were ecoregion, growing degree days, and percent evergreen. (Figure **??**). MORE HERE.

4 Discussion

4.1 Trophic State Probabilities

One of the powerful features of random forests is the ability to aggregate a very large number of competing models or trees. Each tree provides an independent prediction or vote for a possible outcome. In the context of our trophic state models, we have 10,000 votes for each lake. These values may be interpreted as the probability that a lake is in a given trophic state. For instance, for a single lake (National Lake Assessment ID = NLA06608-0005), the vote probabilities for Model 1 were “r 100*round(lakeVotes[lakeVotes$COMID==23491387,][,4],2)“% for oligotrophic,”r 100*round(lakeVotes[lakeVotes$COMID==23491387,][,5],2)”% for mesotrophic, “r 100*round(lakeVotes[lakeVotes$COMID==23491387,][,6],2)“% for eutrophic, and”r 100*round(lakeVotes[lakeVotes$COM for hypereutrophic. This suggests little uncertainty in the predicted oligotrophic state.

Further, the maximum probability for each lake can be used as a measure of how certain the random forest model was of the prediction. We would expect higher total accuracy for lakes that had more certain predictions. Some lakes may have many votes for a single trophic state and few votes for other trophic states and these would thus have a large maximum probability and the random forest predicitions would be more certain. Alternatively, the 10,000 votes could have been spread more equally across the trophic state classes for a lake and that lake would have a small maximum probability and the final predictions would be less certain. This should be evident by looking at the total classification accuracy of lakes given their maximum probability is above a certain point. To test this we can examine the accuracy of trophic state predictions across the full range of trophic state probabilities, similar to an approach outlined by Paul and MacDonald (2005) and implemented by Hollister et al. (2008). We utilize this approach and examine the change in total accuracy as a function of the maximum probability for each lake. As expected, lakes with higher maximum vote probabilities were more accurately predicted (Figure **??**). The increasing trend suggests that even for models with lower overall accuracy there can also be a large number of individual cases that are predicted with high accuracy.

4.2 Variable Selection and Importance

There was a great deal of agreement on the important variables for each set of models. In line with past predictive modelling of cyanobacteria abundance and not surprisingly, the *in situ* models consistently select the water quality variables (turbidity, total nitrogen, total phosphorus, and N:P ratios) as important variables (Downing et al. 2001). While there is variation in the response of cyanobacteria to changes in relative nutrient concentrations, the general pattern suggests that limiting nutrients have considerable impact once amounts increase beyond expected levels.

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

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

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

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

4.3 Associating Trophic State and Cyanobacteria

Cyanobacteria biomass should be closely associated with trophic state as cyanobacteria contribute to the chlorophyll concentration in a lake. If these associations are strong enough we may be able to expand models such as those reported here to also predict probability of cyanobacteria blooms. To test if trophic state can be used to differentiate cyanobacteria abundance, we examine distribution of cyanobacteria abundance for each trophic state and also explored linear associations between chlorohyll *a* and cyanobacteria abundance.

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

5 Conclusions

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

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

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

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

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

8 Tables

:Chlorophyll a based trophic state cut-offs with total number of possible observations.

:Random Forest confusion matrix for Model 1. Columns show predicted values and rows show observed values. Agreement indicated on diagonal and accuracy for each trophic state indicated in ‘Class Accuracy’ column.

:Random Forest confusion matrix for Model 2. Columns show predicted values and rows show observed values. Agreement indicated on diagonal and accuracy for each trophic state indicated in 'Class Accuracy' column.

:Random Forest confusion matrix for Model 3. Columns show predicted values and rows show observed values. Agreement indicated on diagonal and accuracy for each trophic state indicated in 'Class Accuracy' column.

:Random Forest confusion matrix for Model 4. Columns show predicted values and rows show observed values. Agreement indicated on diagonal and accuracy for each trophic state indicated in 'Class Accuracy' column.

:Random Forest confusion matrix for Model 5. Columns show predicted values and rows show observed values. Agreement indicated on diagonal and accuracy for each trophic state indicated in 'Class Accuracy' column.

:Random forest confusion matrix for Model 6. Columns show predicted values and rows show observed values. Agreement indicated on diagonal and accuracy for each trophic state indicated in Class Accuracy' column.

9 Appendix 1. Variable Definitions

References

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

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

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

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

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

Diaz-Uriarte, R. 2010. VarSelRF: Variable selection using random forests. R package version 0.7-3. http://CRAN.R-project.org/package=varSelRF.

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

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

Fernández-Delgado, M., E. Cernadas, S. Barro, and D. Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15:3133–3181.

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

Hollister, J. W. 2014. Lakemorpho: Lake morphometry in R. R package version 1.0. http://CRAN.R-project.org/package=lakemorpho.

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

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

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

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

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

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

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

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

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

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. biometrics 33:159–174.

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

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

Paul, J. F., and M. E. McDonald. 2005. Development of empirical, geographically specific water quality criteria: A conditional probability analysis approach 41:1211–1223.

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

Rodhe, W. 1969. Crystallization of eutrophication concepts in northern europe.

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

Schindler, D. W., and J. R. Vallentyne. 2008. The algal bowl: Overfertilization of the world's freshwaters and estuaries. Page 334. University of Alberta Press Edmonton.

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

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

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

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

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

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

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