Modeling Lake Trophic State: A Data Mining Approach

Jeffrey W. Hollister[*,a], W. Bryan Milstead[a], Betty J. Kreakie[a]

[a] *US Environmental Protection Agency Office of Research and Development National Health and Environmental Effects Research Laboratory Atlantic Ecology Division 27 Tarzwell Drive Narragansett, RI, 02879, USA*

Abstract

Productivity of lentic ecosystems has been well studied and it is widely accepted that as nutrient inputs increase, productivity increases and lakes transition from low trophic state (e.g. oligotrophic) to higher trophic states (e.g. eutrophic). These broad trophic state classifications are good predictors of ecosystem health and ecosystem services/disservices (e.g. recreation, aesthetics, fisheries, and harmful algal blooms). While the relationship between nutrients and trophic state provides reliable predictions, it requires *in situ* water quality data in order to paramterize the model. This limits the application of these models to lakes with existing and, more importanly, available water quality data. To expand our ability to predict in lakes without water quality data, we take advantage of the availability of a large national lakes water quality database, land use/land cover data, lake morphometry data, other universally available data, and modern data mining approaches to build and assess models of lake tropic state that may be more universally applied. We use random forests and random forest variable selection to identify variables to be used for predicting trophic state and we compare the performance of two models of trophic state (as determined by chlorophyll a concentration). The first model estimates trophic state with *in situ* as well as universally available data and the second model uses universally available data only. For each of these models we used three separate trophic state categories, for a total of six models. Overall accuracy for the *in situ* and universal data models ranged from xx% to xx% and xx, xx, and xx described the most variation in trophic state. For the universal data only models, Overall accuraccy ranged from xx% to xx% and xx, xx, and xx described the most variation in trophic state. Lastly, it is believed that the presence and abundance of cyanobacteria is strongly associated with trophic state. To test this we examine the association between estimates of cyanobacteria biovolume and the measured and predicted trophic state. Expanding these preliminary results to include cyanobacteria taxa indicates that cyanobacteria are significantly more likely to be found in highly eutrophic lakes. These results suggest that predictive models of lake trophic state may be improved with additional information on the landscape surrounding lakes and that those models provide additional information on the presence of potentially harmful cyanobacteria taxa.

[*] Corresponding author
Email address: hollister.jeff(at)epa.gov (Jeffrey W. Hollister)

Introduction

Cyanobacteria are an important taxonomic group associated with harmful algal blooms in lakes. Understanding the drivers of cyanobacteria presence has important implications for lake management and for the protection of human and ecosystem health. Chlorophyll a concentration, a measure of the biological productivity of a lake, is one such driver and is largely, although not exclusively, determined by nutrient inputs. As nutrient inputs increase, productivity increases and lakes transition from low trophic state (e.g. oligotrophic) to higher trophic states (e.g. hypereutrophic). These broad trophic state classifications are associated with ecosystem health and ecosystem services/disservices (e.g. recreation, aesthetics, fisheries, and harmful algal blooms). Thus, models of trophic state might be used to predict things like cyanobacteria.

We have three goals for this preliminary research:

1. Build and assess models of lake trophic state
2. Assess ability to predict trophic state in lakes without available *in situ* water quality data
3. Explore association between cyanobacteria and trophic in order to expand models.

Data and Modeling Methods

Data

We utilize four primary sources of data for this study. These are outlined below and in Table 1.

1. National Lakes Assessment (NLA) 2007: The NLA data were collected during the summer of 2007 and the final data were released in 2009. With consistent methods and metrics collected at 1056 location s across the conterminous United States (Map 1), the NLA provides a unique opportunity to examine broad scale patterns in lake productivity. The NLA collected data on biophysical measures of lake wat er quality and habitat. For this analysis we primarily examined the water quality measurements from the NLA [1].
2. National Land Cover Dataset (NLCD) 2006: The NLCD is a nationally collected land use land cover dataset. We collected total land use land cover and total percent impervious surface within a 3 kilo meter buffer surrounding the lake to examine larger landscape-level effects [2,3].
3. Modeled lake morphometry: Various measures of lake morphometry (i.e. depth, volume, fetch, etc.) are important in understanding lake productivity, yet many of these data are difficult to obtain for large numbers of lakes over broad regions. To add this information we modeled lake morphometry [4–7].
4. Estimated Cyanobacteria Biovolumes: Cyanobacteria biovolumes are a truer measure of Cyanobacteria dominance than abundance as there is great variability in the size within and between species. To a ccount for this, Beaulieu *et al.* [8] used literature values to estimate biovolumes for the taxa in the NLA. They shared this data and we have summed that information on a per- lake basis.

References

1. USEPA (2009) National lakes assessment: a collaborative survey of the nation's lakes. ePA 841-r-09-001.

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

3. Xian G, Homer C, Fry J (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.

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

5. Hollister JW, Milstead WB, Urrutia MA (2011) Predicting maximum lake depth from surrounding topography. PLoS ONE 6: e25764. Available: http://dx.doi.org/10.1371/journal.pone.0025764. Accessed 28 Jun 2013.

6. Hollister J (2013) lakemorpho: Lake morphometry in r. Available: http://www.github.com/USEPA/lakemorpho.

7. Hollister JW, Milstead WB (In Preparation) National lake morphometry dataset v1.0.

8. Beaulieu M, Pick F, Gregory-Eaves I (2013) Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes data set. Limnol. Oceanogr 58: 1736–1746.