**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 that is largely 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 and disservices. Thus, models of trophic state might be used to predict things like cyanobacteria. In the preliminary work reported here, we:

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

**Methods**

Data

We utilize four primary sources of data for this study.

1. National Lakes Assessment (NLA) 2007: Using consistent methods and metrics, the NLA collected data from ~150 lakes across the contiguous United States, on biophysical measures of lake water quality and habitat (Map 1). For this analysis we primarily examined the water quality measurements from the NLA NLA(2007).

2. National Land Cover Dataset (NLCD) 2006: The NLCD is a national land use/land cover dataset. We calculated total land covered and total percent impervious surface within a 3 kilometer buffer of each lake to examine larger landscape-level effects (Bhaner et al. 2004; Xian, Homer & Fry 2009).

3. Modelled lake morphometry: Various measures of lake morphometry (i.e., depth, volume, etc.) are important in understanding lake productivity, yet are often difficult to obtain for large numbers of lakes. Modelled estimates solved this problem. (Hollister & Milstead 2010; Hollister, Milstead, & Urrutia 2011; Hollister; Hollister & Milstead in prep).

4. Estimated Cyanobacteria Biovolumes: Measuring of cyanobacteria dominance is best done with biovolume as there is great size variability within and among taxa. Breslaw et al. (2011) used literature values to estimate biovolumes for the time in the NLA. They shared these data with our colleagues and we have summed that information on a per-lake basis.

**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 is recursively partitioned according to a given random subset of predictor variables and completely grown without pruning. With each new tree, both the sample data and predictor variable subset is randomly selected.

Random forests are able to handle numerous correlated variables without a decrease in prediction accuracy. The large number of included variables can reduce accuracy and lead to over-fitting. This problem, faced in gene selection, has been addressed with a variable selection method based on random forests (Diaz-Uriarte & De Andres 2006). Using 100 iteration of varSelRF it is, how to determine now our variables are included in a final model (Law & Witten 2002). From these random forests we collect a consensus prediction and calculate a confusion matrix and summary stats.

**Model Details**

Using a combination of the varSelRF and randomForest we test our models for six combinations of variables and trophic state classifications. These combinations included different combinations of the Chlorophyll a trophic states (Table 2) with all variables or with the GIS variables (i.e., as in situ information). The six model combinations were:

1. Chlorophyll a trophic state - 4 class = All variables (i.e., in situ water quality, lake morphometry, and landscape)
2. Chlorophyll a trophic state - 3 class = All variables (i.e., in situ water quality, lake morphometry, and landscape)
3. Chlorophyll a trophic state - 2 class = All variables (i.e., in situ water quality, lake morphometry, and landscape)
4. Chlorophyll a trophic state - 4 class = GIS variables (lake morphometry and landscape)
5. Chlorophyll a trophic state - 3 class = GIS variables (lake morphometry and landscape)
6. Chlorophyll a trophic state - 2 class = GIS variables (lake morphometry and landscape)

**Model Results - All Variables**

![Figure 1. Trophic State (4 Classes) - All Variables](image1)

**Model Results - GIS Variables**

![Figure 2. Trophic State (3 Classes) - GIS Only](image2)

Map 1. Geographical Distribution of Samples in the 2007 National Lakes Assessment.

**References**


**Poster Source on GitHub**

All of the materials that make up this poster are available via GitHub. Included in this repository are all R Markdown documents, and R Package with data, and the final poster layout as a .svg or .pdf. The repository is available at http://github.com/USEPA/hkm2014ESA.