

# Examining drivers of cyanobacteria blooms in two shallow, eutrophic bays in Lake Champlain

*Wilton Burns, Herman Njoroge Chege, and Mahalia Clark*

*12/2/2019*

## 1. Introduction (WB)

GENERAL NOTES:

- why important to study?
- cite a few other studies doing similar work
- end intro with “in this study” paragraph to set up reader to how we plan to attack the problem
- hypotheses?
- Figs for this section:
  - map of study area
  - bar graph of cyanobacteria bloom timing being different every year (long-term monitoring)

Anthropogenic eutrophication, meaning increases in nutrients entering aquatic systems due to human activity, of natural water bodies is a global phenomenon that often leads to decreased utility of affected ecosystems (eg. Reynolds 1987; Paerl 1988; Paerl et al. 2011). One problem associated with eutrophication is that harmful cyanobacteria blooms are becoming more prevalent in coastal oceans and thousands of inland lakes and ponds (Paerl et al. 2001, 2011; Paerl and Huisman 2008). Phytoplankton play an integral role as the base of aquatic food webs in both freshwater and marine systems, however, when cyanobacteria growth is unchecked, blooms can be detrimental to organisms living in and around freshwater ecosystems and often cause problems that propagate up the food web (add REFS). The climate is changing, and environmental conditions are becoming more favorable for bloom-forming harmful cyanobacteria (Jöhnk et al. 2006; Paerl and Huisman 2008; Jeppesen et al. 2011; Huisman et al. 2018). Therefore, it is becoming increasingly important to understand cyanobacteria bloom onset and senescence. Understanding why and when severe cyanobacteria blooms occur will help inform mitigation efforts.

### 1.1 Factors that drive cyanobacteria blooms

**Goal with this section: Develop what is known from what is not known or what has conflicting attributions (and why), and how this leads you to your research question. This driver – that is, different watersheds and connectivity – needs to be developed in the Intro to show the reader this is a valid research question/hypothesis to pursue.**

- Eutrophication (can mention watershed bay interactions in this section)
- Warm temperatures
- Water column stratification (leads to internal loading of nutrients)
- Increased CO<sub>2</sub> in the atmosphere (some types of cyanobacteria more efficiently produce organic matter)
- Rain and wind events (mix up water column prohibiting internal loading of nutrients and also giving advantage to larger phytoplankton, like diatoms, that need turbulence to get mixed up into the photic zone)
- Cyanobacteria are physiologically diverse and have developed multiple strategies to out compete other types of phytoplankton: N-fixation, CO<sub>2</sub> concentrating mechanisms, buoyancy regulation, toxin production, predator avoidance

- Top-down grazing by zooplankton – however, many studies have shown that cyanobacteria often avoid grazing by forming dense colonies and toxin production (DeMott 1986; Lemaire et al. 2012)
- Viral lysis or fungal infection – often don’t result in long-lasting effects on cyanobacteria populations (Yoshida et al. 2008; Van Wichelen et al. 2016)
- Filter feeders like mussels - effect they have on cyanobacteria blooms is lake-specific (Reeders et al. 1989; Vanderploeg et al. 2001)
- Competition from other, non-harmful, phytoplankton

How does lake size (area/depth) affect which drivers may be relatively important (shallow vs deep)? Lake trophic status?

## 2. Methods

- 2.1 study sites (WB)
- 2.2 data used and how prepped
  - buoy sensor data (MC)
  - meteorological station data (WB): air temp, solar radiation, wind speed, wind direction (WB)
  - discharge data (WB)
- 2.3 Feature Importance

### 2.2.1 Preparing the high frequency buoy data (MC)

High frequency water quality data was collected from May through October in 2017 and 2018 by sensors on two buoys in Lake Champlain: one in St. Albans Bay, and one in Missisquoi Bay. Each buoy had sensors that took measurements at multiple depths: every 0.5m, from 0.5m below the surface down to the bottom (2.5m depth in Missisquoi Bay, 4.5m depth in St. Albans Bay). Sensors measured temperature, conductivity, pH, dissolved oxygen, chlorophyll (Chl), phycocyanin (PC), and turbidity. DeltaTemp was also calculated for each time point as a measure of lake stratification by subtracting the temperature at the bottom from that at the surface.

*Since the buoys collect data at multiple depths, we first considered whether to focus exclusively on the sensors nearest the surface, or to aggregate data from whichever depth had the highest concentration of PC at a given time, so as to track a bloom as it moves up and down in the water column. For each bay and year, we explored how PC values varied with depth. First, for each bay and year, we used R to create a correlation matrix with Pearson correlation coefficients for the PC levels at each depth. We found PC levels across depths to be moderately to highly correlated in each case (0.38-1.00), indicating that PC levels near the surface would most likely be representative of those throughout the water column. We also used R to compare the PC levels across depths at each time point, and find out at what depth the PC level was maximized. We found that for periods of low PC levels, the maximum might be found at any depth, but for periods of high PC levels, such as during a cyanobacteria bloom, maximum PC levels were most often found near the surface. For these reasons, we decided to limit ourselves to the data collected by sensors nearest the surface (0.5m depth), augmented with the deltaTemp as a measure of stratification. See the the markdown file or pdf “Exploring\_PC\_Depths” for details.*

In order to run a forecasting model or look at feature importance with time lags, we needed to collapse this hourly data into daily data, so that daily cycles would not confound predictions or time lags. *We explored whether to use daily averages or daily maximums. R was used to calculate daily averages and maximums for each buoy variable, and to create time plots of the daily and hourly data. After inspecting the time plots, we decided to use daily averages rather than daily maximums, as the later were unduly influenced by sudden brief peaks or high-valued outliers in the hourly data.*

In addition to collapsing our hourly buoy data into daily averages for each environmental variable, we wanted to explore how water quality data related to PC values over different time lags. In order to include this lag, we used R to append the daily average PC levels 1-6 days in the future to the data for each time step.

See pdf/Rmd “DailyTimeLags” for details

## 2.3 Feature Importance

Feature importance is a technique to understand the factors that most contribute to the mechanisms of a system. In our study we used XGBOOST a machine learning technique that has been shown to be most accurate in tabular and structured data.

It uses gradient boosted decision trees and is built for speed and performance and hence is more accurate than alternatives. A benefit of using ensembles of decision tree methods like gradient boosting is that they can automatically provide estimates of feature importance from a trained predictive model. After we train the model we are then able to access the feature importances.

Generally, importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. This importance is calculated explicitly for each attribute in the data set, allowing attributes to be ranked and compared to each other.

Importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for. The performance measure may be the purity (Gini index) used to select the split points or another more specific error function. The feature importances are then averaged across all of the the decision trees within the model Hastie et al. [2009]. These features will help further down the road in developing a predictive model.

## 3. Results (WB)

- fig of time series of the parameters to set the scene?
- which set of feature importance do we want to use?

Next steps for today 12/2/19: - figure out way to make figure that is easy to digest -> take out Chl, pH, and ODO and then have a stacked bar graph with y axis of feature importance (all add up to 1) and x axis being days in future (0 - 6) - need to sep feature importance by bay and year AND figure out how to make a table with the results so i can make the stacked bar graph

### Time Series of the parameters to set the scene, maybe:

*I went down a rabbit hole and tried some different ways of making time series: see PrettiestTimePlot.Rmd or pdf for details. If we want to include timeplots for one or all bays, I would go for the ggplot ones, but four sets of timeplots would take up a ton of space and might not be particularly relevant - MC*

### Brief typed out results for easier interpretation (NOT GOING TO INCLUDE)

1) St AB both years

- 0 day lag: pH, spCond, temp, discharge, deltaTemp
- 1 day lag: pH, spCond, discharge, temp
- 2 day lag: pH, spCond, discharge, ODO
- 3 day lag: pH, spCond, discharge, deltaTemp
- 4 day lag: pH, spCond, deltaTemp, ws
- 5 day lag: pH, spCond, deltaTemp, wd, ODO
- 6 day lag: pH, spCond, ODO, ws, discharge

2) MB both years

- 0 day lag: pH, ODO, discharge, wd (deltaTemp last)
- 1 day lag: pH, discharge, ODO, spCond, wd (deltaTemp moves up)
- 2 day lag: pH, ODO, wd, solar, spCond,
- 3 day lag: pH, spCond, temp, wd, deltaTemp
- 4 day lag: spCond, pH, temp, ODO, ws, deltaTemp
- 5 day lag: spCond, temp, pH, ODO, deltaTemp, wd
- 6 day lag: spCond, temp, pH, ODO, wd, discharge, deltaTemp

3) 2017 both bays

- 0 day lag: discharge, spCond, pH, ODO
- 1 day lag: spCond, ODO, discharge, pH, solar
- 2 day lag: pH, spCond, discharge, ws, ODO
- 3 day lag: spCond, ODO, discharge, solar
- 4 day lag: spCond, ODO, solar, wd, temp
- 5 day lag: spCond, deltaTemp, solar, wd, temp
- 6 day lag: deltaTemp, spCond, ODO, wd, discharge

4) 2018 both bays

- 0 day lag: pH, spCond, ODO, discharge
- 1 day lag: pH, spCond, discharge, ODO
- 2 day lag: pH, discharge, temp, spCond
- 3 day lag: pH, ODO, discharge, spCond, temp
- 4 day lag: pH, spCond, discharge, temp, deltaTemp
- 5 day lag: pH, spCond, temp, discharge, ODO
- 6 day lag: pH, spCond, temp, solar, ODO

5) both bays both years

- 0 day lag: pH, spCond, year, discharge, solar
- 1 day lag: pH, spCond, year, discharge, temp
- 2 day lag: pH, spCond, discharge, year, deltaTemp
- 3 day lag: pH, spCond, discharge, year, ODO
- 4 day lag: pH, spCond, year, discharge, ODO
- 5 day lag: pH, spCond, temp, ODO, year, discharge
- 6 day lag: pH, spCond, temp, year, deltaTemp, ODO

LOOK INTO - is spCond just super highly correlated with PC? Why is it such an important feature? What is it highly correlated with? are the units different between the bays? - wd (not speed) is important for MB! let's look at correlations - should we run without pH? - what's the significance in the difference in the cluster numbers??

## Correlations

*I made correlations matrices for pretty much every environmental parameter with the 7 time steps of PC data to look at trends of how correlations with PC change with increasing time lag... Sometimes they show similar trends as feature importance, but sometimes not. In theory I could summarize these correlations coefficients into a chart with environmental parameters by PC time steps, which could be interesting for looking at trends. I could even graph them! But I haven't yet. See All\_the\_correlations.Rmd/pdf for all the correlations. - MC*

- Made lots of correlation tables
- Haven't summarized trends yet: match feature importance sometimes, not always
- Discharge and salinity are NOT correlated in MB, but ARE correlated in St. Albans.

## Discussion

- what we know right now is that 6 days out, pH is IMPORTANT for PC but the next step is figuring out if it's PREDICTIVE of a bloom → need to build model
- We got kinda hung up on the fact that there were parameters in the feature analysis results that were highly correlated with PC (response variable) but Easton reminded us that if our main goal is to predict a bloom, Chl or PC levels 6 days before is important to include if we are building a predictive model because that's still useful info ! Like it'd be great if we could predict if a bloom would start in 6 days just by knowing PC or Chl levels.
- IMPORTANT POINT: there is a difference between mechanistic models and predictive models. Depending on how we decide to move forward (ie. which model we choose) we might not really need to understand the mechanisms driving the blooms because we might just try throwing all the data into a machine learning algorithm that tells us the relative predictive power of each of the parameters (I think...).

## Future Work (MC)

We could build on this work in a number of ways, by expanding our data set, correlating our phycocyanin measurements with satellite data or volunteer observations, identifying a bloom threshold, and most importantly, by using our high frequency data to create a forecasting model that could predict cyanobacteria blooms.

First, we could expand our data set by including cumulative degree days as another environmental variable, indicative of temperatures experienced throughout the season up to a given time point. We could collect mean daily temperatures for 2017 and 2018 for the nearest weather stations to the two bays from wunderground.com, and calculate the cumulative degrees above freezing (or above a biologically relevant temperature threshold such as 4°C) for each day. We could then include this variable in analyses such as correlations, feature importance, or a forecasting model.

Second, we could correlate the buoys' measure of PC levels with other indicators of bloom presence such as volunteer observations and satellite data. There is an online data set publicly available with volunteer observations at the two bays: biweekly observations of bloom presence or absence throughout the 2017 and 2018 seasons. Comparing these observations with the daily average PC levels from our buoy sensors on the days of the observations could help us identify a threshold for what PC levels indicate a cyanobacteria bloom for management purposes.

There is also satellite imagery available covering both bays, and it can be used to calculate a spectral index that's indicative of cyanobacteria presence. We could calculate this index at the position of each buoy for the time points of the available satellite imagery. We could then see how the index correlates with the daily average PC levels measured by the buoys, and examine whether there is a simple, consistent conversion between the two. If there are literature thresholds for what value of the spectral index constitutes a cyanobacteria bloom, we could then use a correlation or conversion to calculate an analogous bloom threshold for our buoy-measured PC levels. Having an accurate bloom threshold would be valuable for investigating bloom drivers and creating a forecasting model. It would give us the option to investigate continuous PC values directly, or convert them to categorical 'Bloom'/'No Bloom' data in case the later is easier to forecast or more strongly correlated with certain drivers.

Finally, our long term goal is to use this high frequency data to develop a forecasting model using machine learning techniques, in the hopes of predicting cyanobacteria blooms a few days before they occur. We would

begin by inputting all available high frequency buoy, weather and discharge data, in order to predict PC levels (or a categorical variable for bloom presence or absence). If that proved successful, we would then begin to remove input variables one by one, to see what effect that has on the forecast accuracy. The goal would be to have a functional forecasting model with as few input variables as possible.

## References

INFO ON HOW TO CITE in .Rmd from [https://rmarkdown.rstudio.com/authoring\\_bibliographies\\_and\\_citations.html](https://rmarkdown.rstudio.com/authoring_bibliographies_and_citations.html) :

- Citations go inside square brackets and are separated by semicolons. Each citation must have a key, composed of ‘@’ + the citation identifier from the database, and may optionally have a prefix, a locator, and a suffix. Here are some examples:
  - Blah blah [Isles et al., 2017].
- Then in Zotero create a .BibTex file by going to File -> Export Library -> change to BibTex

## References

T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848846. URL <https://books.google.com/books?id=eBSgoAEACAAJ>.

Peter D. F. Isles, Donna M. Rizzo, Yaoyang Xu, and Andrew W. Schroth. Modeling the drivers of interannual variability in cyanobacterial bloom severity using self-organizing maps and high-frequency data. *Inland Waters*, 7(3):333–347, July 2017. ISSN 2044-2041. doi: 10.1080/20442041.2017.1318640. URL <https://doi.org/10.1080/20442041.2017.1318640>.