Understand the drivers of cyanobacteria blooms in Missisquoi and St. Albans Bays in 2017, 2018, and 2019 using high-frequency buoy sensor data

Wilton Burns & Mahalia Clark & Herman Njoroge Chege

11/1/2019

Research goal

To better understand the drivers of cyanobacteria blooms in Missisquoi and St. Albans Bays in 2017, 2018, and 2019 using high-frequency buoy sensor data. And make beautiful documents in R markdown.

Personal research goals

- Learn ecological forecasting model techniques and more about machine learning
- Have help (and fun) intellectually interpreting the findings of the study and what data to include/how to prepare it for the model
- Use machine learning approaches to understand drivers of cyanobacteria blooms
- Practice working with data in R and working with satellite imagery/data

Available Data

- High-frequency buoy monitoring data: Dissolved oxygen, pH, temperature, stratification, nutrient concentrations, salinity, cyanobacteria, etc.
- Long term monitoring: Similar data over longer time periods but lower frequency
- Weather (solar radiation, wind speed & direction) and stream discharge
- Satellite imagery: Imagery of the bay is available and there are indices for mapping cyanobacteria & blooms. Satellite data: intensity of different spectral bands (visible, NIR) and indices combining bands for multiple time points.
- We found volunteer monitoring data with visual assessments of algae blooms for the two bays.

Project ideas/questions/sub-sections

(These will depend on our interests and rate of progress)

- What environmental factors were driving the blooms in 2017, 2018, and 2019?
- Can we use the long-term monitoring data (since 1992 but less frequent time steps, no cyanobacteria levels) to forecast bloom dynamics in the bays?
- Can we use high frequency data to forecast blooms?
- How many days out can we forecast blooms (the best I've seen is a few days)? Does it change from year to year and bay to bay?
- We could compare PC and CHL data from buoy sensors to what the satellites pick up: How well does the satellite spectral index for cyanobacteria correlate with buoy data on cyanobacteria concentrations? Is there a bloom threshold for the satellite index? Can we use this to determine a bloom threshold for the buoy sensor?
- Combine high frequency buoy data (daily averages?) with available satellite imagery for buoy locations. See if we can use all this data to predict/forecast buoy cyanobacteria concentrations and/or satellite cyanobacteria index.

- See how well our model/forecast matches long term monitoring data: if we have our model trained with the buoy data, and we put in the long term data on water temp & concentrations, will it predict blooms when there were blooms? Or would it only work for short term forecasting with high frequency data? In that case, Wilton could still try to forecast blooms next summer.

Results

1) St AB both years

- 0 day lag: pH, spCond
- 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, dishcarge, 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: dishcarge, 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: which direction(s) are correlated with increased PC?
- Should we run feature importance without pH and/or ODO?
- What's the significance of the difference in cluster numbers??