Using On-site and Remote sensing data to Understand Key Factors Affecting Harmful Algal Blooms in Western Lake Erie

Jimmy Nguyen[1], Zesheng Chen[1], Dong Chen[2], and Veronika Meyer[2]
[1]Department of Computer Science,
[2]Department of Civil and Mechanical Engineering,
Purdue University Fort Wayne

Introduction

Objective

Lake Erie has been facing an on-going harmful algal bloom (HAB) crisis in the last decade. HABs are algae that bloom significantly and can produce toxins that threaten quality of natural waters, public health, ecological systems, and sustainable development. Our main objective in this research is to understand key factors that affect HABs in Western Lake Erie by applying various regression analysis and machine learning algorithms. We also want to identify the reliability of remote sensing platforms. We use weekly on-site buoy data, daily on-site buoy data, and remote sensing data from the MODIS Aqua and MODIS Terra satellites to identify key parameters that affect harmful algae growth. Specifically, chlorophyll-a is ubiquitous among aerobic species and is one of main indicators for HABs. Another main contributor that makes algae harmful in Western Lake Erie is known to be Microcystin, which is a toxin produced by *Microcystis Aeruginosa*, a freshwater cyanobacteria known to cause trouble in Western Lake Erie. However, on-site buoy data for Microcystin is very scarce. As a result, we focus our study on predicting chlorophyll-a and attempt to find the relationship between chlorophyll-a and other factors.

Methodology

In order to analyze key factors of HABs, we have identified four predictor input variables for our weekly on-site buoy data: dissolved organic carbon (DC), soluble reactive phosphorus (SRP), total inorganic nitrogen (TIN), and water temperature (WT), as well as a target output variable, chlorophyll-a. Our goal is to find a quantitative relationship between the four inputs and the output. As a first attempt, we applied a non-linear least squares curve fit function to find optimal parameters for the polynomial function. Moreover, in order to compare remote sensing data with on-site data, we studied the correlation coefficients of corresponding data sets collected at Western Lake Erie. Specifically, the selected data sets include chlorophyll-a, normalized fluorescence line height (nflh), particulate organic carbon (poc), sea surface temperature (sst), diffuse attenuated coefficient at 490nm (kd490), and rrs667nm (rrs667). Our main objective is to identify parameters with the high Pearson correlation coefficients between remote sensing data and on-site data, so that we can determine which data collected by remote sensing platforms are reliable in a highly eutrophic environment such as Western Lake Erie.

Data Collection

In order to apply regression algorithms and machine learning models, we have obtained feature information from on-site buoys (i.e., WE2, WE4, and WE13) and field missions at Western Lake Erie through NOAA's (National Oceanographic and Atmospheric Administration) online data catalog. For obtaining remote sensing data, we extracted data from the MODIS Aqua and Terra satellites through R programming. Specifically, we used the Open-source Project for a Network Data Access Protocol (OPeNDAP) as a source and extracted feature values from netCDF files. We targeted the latitude/longitude for our on-site buoys (WE2, WE4, WE8, WE13) and added a 5m buffer radius. The timespan for daily remote sensing data extraction was throughout June 1st to August 31st from 2014-2018. For our curve-fitting data, we used data from weekly field missions at Western Lake Erie from 2014 to 2017 in order to perform our initial attempts of predicting chlorophyll-a. We used rows where all four input features and target feature are available. Since the weekly field missions did not provide water temperature data, we used water temperature data from daily on-site buoy sensors for WE2, WE4, and WE13. The buoy locations are shown in **Figure 5.** These buoys provide daily on-site data through YSI-EXO2 water quality probes with multiple sensors. One of the main issues of remote sensing data collection is that there is scarce data depending on the time and location. From previous works, this was known to be caused by cloud coverage and/or sun glints throughout the region. Therefore, we used the data that was available.

Figure 1. Comparison of true chl-a values vs. predicted chl-a values for all buoys.

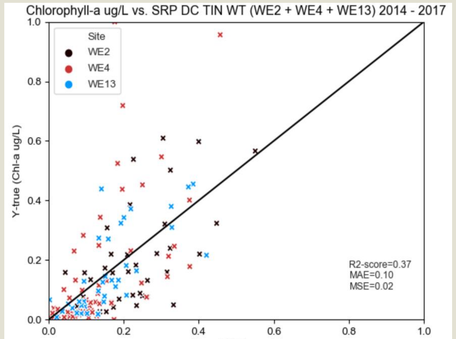


Figure 2. Subplot comparison of true chl-a values vs. predicted chl-a values for WE2 buoy.

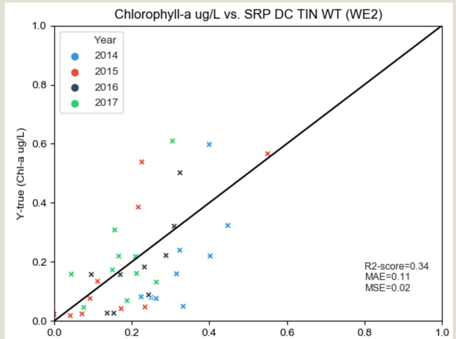


Figure 3. Subplot comparison of true chl-a values vs. predicted chl-a values for WE4 buoy.

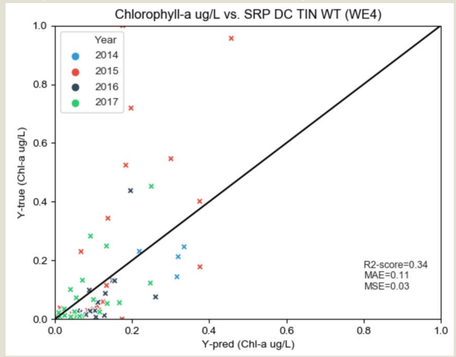
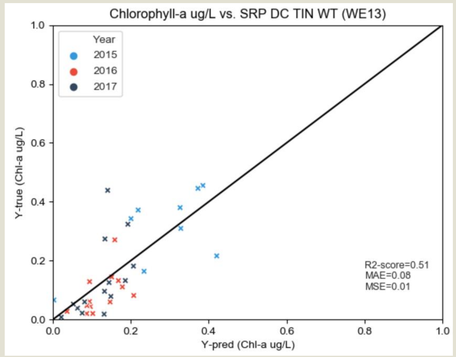


Figure 4. Subplot comparison of true chl-a values vs. predicted chl-a values for WE13 buoy.



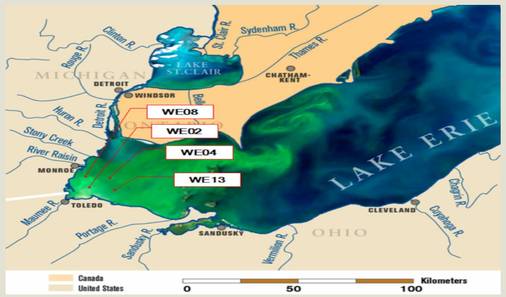


Figure 5. Location of the four buoys analyzed at Western Lake Erie. The green coloration represents algae bloom.

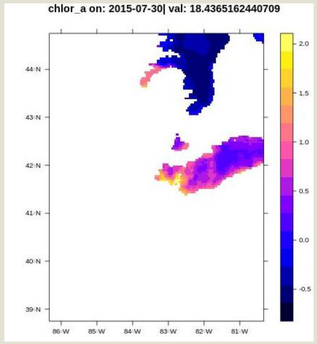


Figure 6. Extracted MODIS Aqua remote sensing image of chlorophyll-a at Western Lake Erie on 7/30/2015 with concentration value in mg/m^3

Observations

After applying our curve fitting function and obtaining optimal parameters, we found that chlorophyll-a is approximately described by a third-order polynomial function from our four predictor variables. **Figure 1** displays the comparison of true chlorophyll-a vs. predicted chlorophyll-a between all buoys from 2014-2017. For individual buoys **(Figures 2-4**), the outputs with R^2-scores of 0.34, 0.34, and 0.51 for buoy WE2, WE4, and WE13, respectively. With the values changing depending on the location and time, we think that predicting chlorophyll-a based on nutrient features are highly dependent on location, time, and environmental events. Western Lake Erie is the most shallowest and nutrient abundant lake out of all the Great lakes. Therefore, it is reasonable to suspect that nutrient factors such as dissolved/particulate organic carbon, phosphorus, and nitrogen play an important role in algae growth. **Figure 6** shows a heat map of chlorophyll-a concentration during algal bloom season. Run-off from farmland and rivers can dump excess nutrients and can therefore cause eutrophication, stimulating uncontrolled algae growth and production of toxins. We also observe that there would have to be slightly different models depending on the location and time of the bloom season. This is due to the unpredictable nature of algae growth and nutrient scattering throughout the lake. We have not found a "one-size fits all" solution for addressing algae growth. For our remote sensing data observation, we attempted to compare MODIS Aqua and MODIS Terra satellites at Western Lake Erie when correlating with daily on-site data. We observe that MODIS Aqua does perform slightly better than MODIS Terra in the Pearson correlation coefficient and that sst is the most reliable parameter when comparing MODIS Terra vs MODIS Aqua. Nflh does show consistent negative correlation, and chl-a's performance is nearly identical to kd490. Remote sensing chl-a is highly inconsistent when comparing the two satellites. This may be due to the everchanging nature of aerobic organisms and the fact that the satellites differ by 2-3 hours each day.

Conclusion/Future Works

We analyzed our collected data and performing various accuracy tests (i.e. R^2, MAE, MSE, etc.). We concluded that predicting chlorophyll-a, based on weekly on-site buoy data using polynomial fitting and regression, will vary based on location and time. One of our main challenges was to obtain a large amount of data and features. For remote sensing data, we found that certain features such as particulate organic carbon and normalized fluorescent line height do show some correlation with daily on-site chlorophyll-a data. This does depend on the location of the buoy and the specific year, but the results do show some consistency. The most reliable remote sensing feature is sst, which is known to be highly maintained and provided more data points than other features. Our main challenge in using remote sensing data was the availability of data on most days. Since satellite platforms use reflectance band data and pre-defined algorithms to generate parameter estimates, having cloud coverage or sun glints can render the data unavailable on a particular day. In summary, our observations provide a better understanding to key factors that affect HABs in Western Lake Erie.

Acknowledgements
This work was supported by the 2019 PFW IRSC Collaborative Research Grants and Microsoft Azure Grant.