

2020

## Exploring Seasonal Trends and Episodic Weather in the Muskegon Lake Ecosystem

Elijah Smith  
*Grand Valley State University*

Follow this and additional works at: <https://scholarworks.gvsu.edu/cistechlib>

---

### ScholarWorks Citation

Smith, Elijah, "Exploring Seasonal Trends and Episodic Weather in the Muskegon Lake Ecosystem" (2020).  
*Technical Library*. 359.  
<https://scholarworks.gvsu.edu/cistechlib/359>

This Project is brought to you for free and open access by the School of Computing and Information Systems at ScholarWorks@GVSU. It has been accepted for inclusion in Technical Library by an authorized administrator of ScholarWorks@GVSU. For more information, please contact [scholarworks@gvsu.edu](mailto:scholarworks@gvsu.edu).

Exploring Seasonal Trends and Episodic Weather in the Muskegon Lake Ecosystem

Elijah Smith

A Project Submitted to

GRAND VALLEY STATE UNIVERSITY

In

Partial Fulfillment of the Requirements

For the Degree of

Master of Science in Applied Computer Science

School of Computing and Information Systems

December 2020



The signatures of the individuals below indicate that they have read and approved the project of Elijah Smith in partial fulfillment of the requirements for the degree of Master of Science in Applied Computer Science.

---

Jonathan Leidig, Project Advisor                      Date

---

Robert Adams, Graduate Program Director      Date

---

Paul Leidig, Unit head                                      Date

## **Abstract**

Over the last 10 years, the Robert B. Annis Water Resources Institute (AWRI) at Grand Valley State University has collected buoy sensor data from Muskegon Lake. This sensor data, captured every 15 minutes, records a variety of physical and biological characteristics important to the regional biome. Ranging from wind speed to dissolved oxygen to chlorophyll, the 21 distinct metrics reveal details about the intertwined processes and seasonal variations in the ecosystem. Previous research from AWRI has shown that “episodic weather events” play a role in water column mixing and algal blooms. For example, a strong storm system will mix up nutrient-rich bottom water and distribute it throughout the lake, potentially leading to a future algal bloom. Research is ongoing to formally prove this link. The goal for this project was to aid with the analysis and visualization of this time-series data set and to provide quantitative insight into how episodic weather events are linked to other processes in the lake. Long term, the vision is to create comprehensive models that can predict information valuable to the public, e.g. the likelihood of a dangerous cyanobacteria bloom in the lake during the summer months. The data analysis focus for this project consisted of a small subset of metrics from 2011-2019, including: water temperature at various depths to measure stratification, dissolved oxygen concentration to measure hypoxia, wind speed and direction to quantify episodic storm events, and chlorophyll and phycocyanin to track different types of algal blooms. These specific metrics just scrape the surface of what is available but helped further our understanding of the cycles at play in Muskegon Lake.

## **Introduction**

The AWRI Muskegon Lake Buoy was deployed in 2011 to better understand and manage a valuable watershed in West Michigan. Annually deployed from April to November, there are 21 different metrics monitored and over 30 raw data fields (some metrics are at multiple depths). The data is openly accessible to anyone -- researchers, students, teachers, and interested citizens. In the summer of 2020, AWRI researchers came up with a number of project ideas to fill in gaps in understanding and research. These projects required a computer science background and understanding for further data cleaning, data analysis, and visualizations.

The first project involved data analysis and visualization of trends, seasonally and over multiple years, of several key metrics that were of importance. Specifically, the metrics included changing temperatures in surface and bottom waters, dissolved oxygen levels throughout the water column, and chlorophyll and phycocyanin presence to track algal blooms and eutrophication. This information helps build a picture of the ecosystem as a whole and allows for understanding of the cycles and processes at play. One of the goals of this project was to uncover hidden or unknown patterns. Another was to create new and intuitive visualizations to convey the rhythms of the lake.

The second project was to explore the time delayed effects of meteorological events, such as storm mixing of bottom nutrients to surface waters, and the effect on subsequent algal blooms. Throughout the paper, these events will be referred to as “episodic weather events.” Researchers at AWRI knew that these events occurred but had not yet attempted to quantify the effects or the specific time delay involved. One future goal of this understanding would be to predict with high

accuracy an upcoming harmful algal bloom – one that causes damage to people or the ecosystem from cyanobacteria or other toxins.

A third potential project was to create a live dashboard for current weather and water quality information to be displayed on the AWRI website. This dashboard would contain some of these same core metrics (e.g., surface and water column temperatures, chlorophyll, and dissolved oxygen) and chart these values changing over time. This project would give the AWRI researchers an ‘at-a-glance’ view of current lake metrics, and give interested citizens a place to understand more about the lake and how it changes over time.

As all three of these projects were relatively sprawling and open-ended, this master’s project iteratively provided small accomplishments from each. They were worked on in order from first to last. The first project (visualizing trends) lay much of the groundwork for an understanding of the lake necessary to work on the second project. As the full scope of work for the second project (exploring episodic weather events) became clear, it was decided to work on a subset that will aid further efforts – labeling episodic events and beginning to quantify the time delay between physical processes of the lake. Toward the end of the project period, a proof-of-concept dashboard animation was created to help understand the data pipeline necessary for the third project (live dashboard for water and weather data) and to offer a novel visualization approach for consuming historical data.

## **Background/Related Work**

This project required a strong biology and limnology background to understand the domain. Working backwards from the second project, “exploring the effects of episodic weather events,” there are several lake details and processes that are important to highlight. One of the first steps I undertook for my own understanding was breaking this into a six-step process:

1. Stratified water layers of different temperatures form in Muskegon Lake. This process occurs as a result of thermal energy from the sun, and is well accepted in literature.
2. Hypoxia or anoxia (lack of or no dissolved oxygen) occurs in the bottom layers, due to decomposition of biomass and lack of mixing. (Biddanda et al 2018, Weinke and Biddanda 2018, Weinke and Biddanda 2019)
3. Internal loading of nutrients within the bottom layer of sediment occurs (phosphorous in particular) (Weinke and Biddanda 2018). Further research is needed to prove this.
4. An episodic wind event occurs, mixing the stratified layers to some degree. (Weinke and Biddanda 2018, Weinke and Biddanda 2019)
5. Nutrients that were previously loaded in the bottom are dispersed throughout the layers (Weinke and Biddanda 2018). This is not yet confirmed. This is a crucial linking step between episodic wind events and future algal blooms.
6. Calm periods and warm temps allow algae to re-congregate at the surface, eventually leading to an algal bloom. This is well accepted in literature.

Beyond this simple sequence, recent research has shed light on the important links between hypoxia, episodic weather, and ecosystem health. One paper shows that hypoxia and episodic weather interact in a negative feedback loop throughout the summer, with cyclical loading of nutrients correlated with increasingly severe cyanobacterial blooms (Weinke and Biddanda

2018). Another shows that the timing of the late spring hypoxia is strongly linked to the lack of episodic wind events, and an earlier degradation of hypoxia in late summer is linked with an increase in episodic wind events (Weinke and Biddanda 2019). These relationships are relevant beyond the local ecosystem and are being studied across the globe as high frequency time-series data becomes increasingly available. One study in Mendota Lake (Madison, WI) investigated cyanobacteria algal bloom dynamics (Carpenter et al 2020). Of relevance to our research was a similar look at the ‘time delay’ of correlation between episodic weather and algal blooms (specifically, wind speed and precipitation). The study found statistically significant correlations ranging from 5-20 days before.

Finally, for us to fully comprehend the domain around lake mixing events during episodic storms, it was important to be aware of the lake physics. When wind blows across the surface of a lake, a back and forth momentum of the water can develop called “seiching.” This seiching can further accelerate temperature and dissolved oxygen mixing that occurs between layers, and it can affect the temperature depth readings as the layers shift internally. Winds that travel across the longest lake surface (known as the “fetch” of a lake) cause stronger seiches, which in turn causes greater mixing.



## Methods

As this project was diverse and covered a range of functional outcomes, many tools and pieces of software were used to manipulate and process the data. The buoy data provided by AWRI had already been “cleaned” (outliers excluded and raw values combined as needed). The primary software used for data visualizations was Tableau, and each sheet of input data in the Excel file needed to be joined together by the date/time field to allow for synchronized data views. When additional heavy-duty processing of data was required, Java was the programming language of choice. For example, a condensed water temperature measurement and gradient value had to be calculated for the second project. To accomplish this, the CSV file was read in by Java, calculations were performed, these calculations were exported to a new CSV by Java, imported back into Tableau, and joined on the same date/time field.

In the second project, some statistical functions available in the R Language were desired to perform time-series delay analysis. A final combined CSV file was generated from the linked Tableau sheets. Additional post-processing was done in Excel, including sorting by the date/time field and providing additional blank records between years to prevent unwanted correlations between tailing ends. This final CSV was then loaded into R Studio to run the cross correlation function (ccf) between desired measures.

In the third project, the software pipeline was a bit different as the final goal was creating a web-based dashboard of information. A few other data buoys were included in the visualization to give a broad overview of Muskegon Lake water, but these buoys had slightly different time values (off by seconds), preventing a successful join in Tableau. To overcome this, a simple find-and-replace in Excel was sufficient. As the end goal was a single ‘master sheet’ that could be imported, all of these disparate data sources were joined in Tableau and re-exported to a

‘combined’ CSV. After a bit more post-processing to convert Celcius to Fahrenheit and to add a ‘Year’ column, the data was ready to be imported.

A “client side” programming language and library was still needed to produce the final animation of historical data. Both Javascript with the D3 library and Python with the Plotly library were top contenders. Given that I had much more exposure and experience with Javascript and D3 and anticipated better future maintainability of Javascript (the Grand Valley web development team has a strong background there), it was a clear software selection choice. The D3 library is a visualization library that uses the standard document object model (DOM) combined with flexible selections to give the programmer a large number of options for traversing their data and transforming it into visualizations. An easy to use API is layered on top of this flexibility. In order to produce the final animations, two ‘layers’ of code are important to understand. First, we iterate over each time interval (row in the CSV). Second, to generate color-coded depth data, each relevant depth reading needs to be mapped to a consistent data structure. This way, the previous value can be found, and we can perform an animation. Simple Javascript “objects” with key/value pairs were used for this mapping. Given the performance requirements of animating between many color gradient ranges, all the required data structure mapping was performed at initialization.

In the computing field, the choice of tools can sometimes feel arbitrary, as there are many ways to accomplish a given goal. Therefore, a judgement call must be made on which tool is the most efficient (for the programmer’s time), the most performant (for actual calculation time), or maybe even the most elegant (for future maintainability) to accomplish the given goal. Throughout this project, between Excel pre-processing, Tableau calculated measures, and full Java processing of an input CSV, software selection was not always a clear-cut decision. In the

interest of accomplishing as much as possible, preference was usually given to what would be the most efficient, with a secondary goal of future maintainability, and a tertiary of performance.

## **Results/Discussion**

### **Visualizing Trends (First Project)**

As a first step to understanding episodic events, we wanted to be able to label and distinguish them in the data. This labeling would serve a few purposes: first, it would give AWRI a large dataset of “significance events” to reference when doing future research. Second, a well-annotated dataset is a requirement for effective machine learning models. To describe it another way, we must first know the significance of the data before we can try and predict related processes. Finally, labeling is beneficial for visualization – either with static reference lines (a significant event occurred on a given date), or to accentuate with colors. Considering the five previously mentioned key metrics (temperature at depth, wind speed, dissolved oxygen, chlorophyll, and phycocyanin), we wanted to find a way both to quantify all of these (judge their significance), and compress any related or complementary metrics into a single measurement.

#### *Temperature At Depth*

The first metric we attempted to quantify and label was temperature at depth. Looking at this metric, it is quite easy to see when the lake is stratified, and when mixing events occur. Here is a reference visual from the last week in August, 2014, to the first week in September, 2014. Each line corresponds to a depth reading, so when the lines are far apart on the left side of the

chart, there is stratification, and when they come together around 70 degrees, the lake is considerably more mixed.

8/25/14 - 9/6/14: Temp at Depths

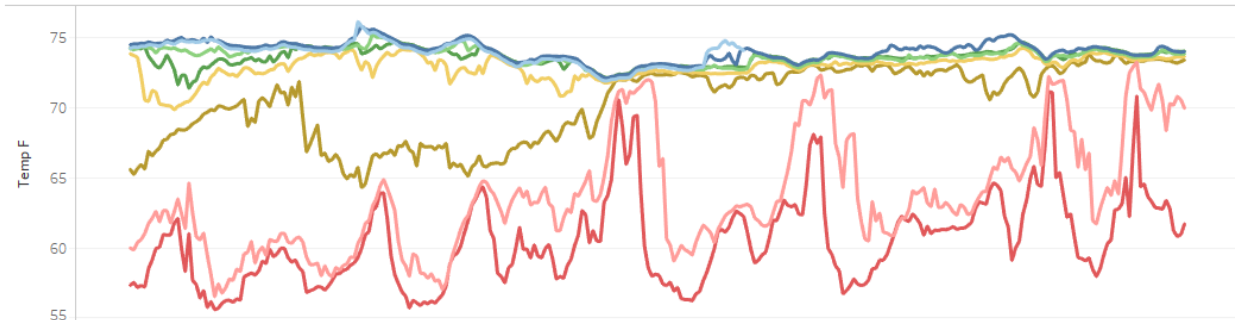


Fig 1: Line chart with each line representing a different depth shows stratification and mix events

Visually, mixing and stratification are clearly identifiable, so I developed an analytical way to capture these events. One simple calculation is the difference between the top and bottom layers. Since the very bottom layer sometimes stays stagnant while the second lowest depth reading changes, I decided to take the average of the bottom two and the top two readings when calculating this *temp diff* metric.

$$TempDiff = \frac{1.7mTempNode + 2mTempNode}{2} - \frac{10mTempNode + 11mTempNode}{2}$$

A mixing event could then be described as when the *temp diff* metric falls below 2-3 degrees F, and a very stratified lake could be described as when the *temp diff* metric is greater than 10 degrees. Testing out this method of categorization, I found that it covered many events, but not all. One good edge case to describe the limitations was during periods of slow, gradual mixing common in the fall. During this time, we would still be finding “mixing events”, even

though there was not an episodic event occurring. Digging deeper into this shortcoming, it seemed like something in the model was fundamentally wrong, so I had another idea. Where we can visually see the mixing events occur in the temperature depth readings, it is also because of the sudden “spike” of a quick change. This spike could be likened to the rates of change, or the gradient, of the temperature depth readings. Building on the previous *temp diff* metric, we could take the gradient of that difference, instead of the raw value. Using a gradient to determine a threshold label would require two things: 1) figuring out what time interval to calculate the gradient over and 2) determining what temperature threshold to use. To figure this out, I turned back to the data and created a scatter plot of what appeared to be all “significant” events and recorded both the change in temperature (in degrees) and the length of time that it occurred over. This relatively unscientific method to determine a threshold value will need to be tweaked over time, but it gave a starting point for threshold parameters. I found that time intervals usually ranged from 3-16 hours, and “drops” in *temp diff* usually were between 1-3 degrees per hour for visually clear mixing events. In the Java implementation, I accounted for this large time interval range by calculating multiple gradients for each record. If any of the calculated gradients, from 3-16 hours prior, had a gradient value (rate-of-change) of greater than 1 degree per hour, it would be labeled as a “significant” mixing event.

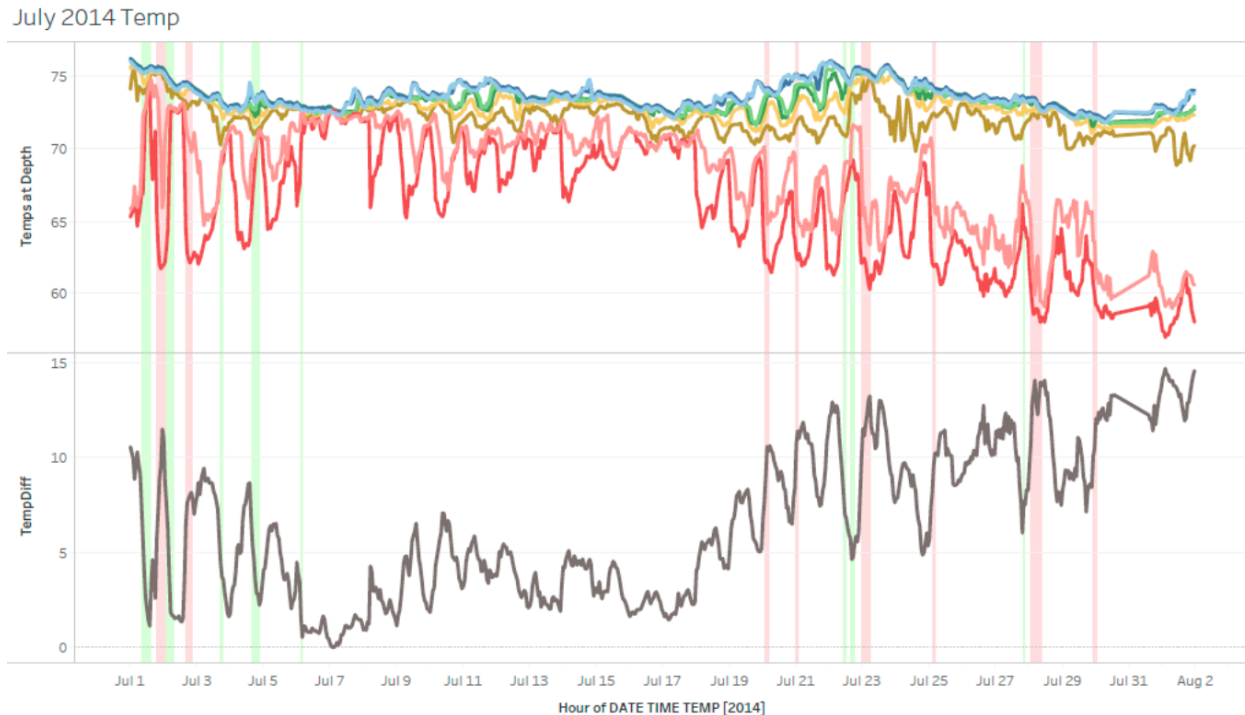
$$\sum_{i=3}^{16} \frac{TempDiff(x) - TempDiff(x-i)}{i} > 1$$

*Fig 2: Temp Diff gradient calculation. Where  $TempDiff(x)$  is the temp diff at the curent record. If this overall summation returned greater than 1 (any of the calculated slopes were greater than the threshold), than an episodic event was marked.*

Trying out this threshold, I found a vast majority of the events that appeared “significant” were accounted for (plus many more). This same idea can also be applied to find “quick stratification” events, although it remains to be verified whether this is a useful concept. For the purposes of this project, we exported both sets of significant events (rapid mixes and rapid stratifications) to a CSV file. These were then applied in Tableau as “shaded vertical reference bands,” with different colorings, to draw the eye’s attention to the date when a significant event was detected.

There may be two additional ways to build off of this in the future. First, calculating the overall *temp diff* change across an entire event may be useful. For example, a *temp diff* gradient of +3 degrees per hour clearly indicates some significant mixing or seiching, but knowing that the total change during this event was 15 degrees gives additional clarity to the significance. Second, as the water layers become more deeply entrenched in stratification, events that cause significant temperature changes may need more wind velocity. For example, in the late spring when the lake is just becoming stratified, a small wind event may cause the appearance of “episodic mixing”, whereas in late summer an extremely strong storm would be needed to produce the same effects. Our model is agnostic to these ideas, which might be desirable since it is quantifying the “mixing” of the lake. Perhaps it would be useful to give additional significance

to mixing events that occur when the lake is stratified by 15 degrees, versus reducing significance when mix events occur when the lake is stratified by only 5 degrees.



*Fig 3: Final stacked chart containing temperature at depths, temp diff, and episodic events marked in red (rapid mixes) and green (rapid stratifications)*

### *Wind Speed*

The wind speed metric was our primary indicator for episodic storm events. One related metric that we knew would also factor in, due to the physics of lake mixing, was wind direction. The raw wind direction data was complex, because it came in as a 0-360 degree measurement. Therefore, trying to calculate correlations or build a model with this metric would find a starkly different outcome between 359 and 2 degree winds, even though these wind directions are only 3



degrees different. To handle this, it is important to take into account which wind speeds are the most important to us. Going back to the lake physics, the fetch of Muskegon Lake should be what determines the “most significant” wind direction. Looking at a map, this appeared to be ~35 degrees ENE. If we use this angle in a trigonometric function, we can come up with a function that gives the most weight to angles nearest 35 degrees (and 180 degrees opposing), and the least weight to angles 90 degrees perpendicular. This seemed ideal from a physical standpoint, as winds 180 degrees reversed from the fetch would have equal significance. I did this by creating an Excel formula that transposed a sin function 35 degrees, coming up with an “adjusted” wind direction metric ranging from 0 – 1.

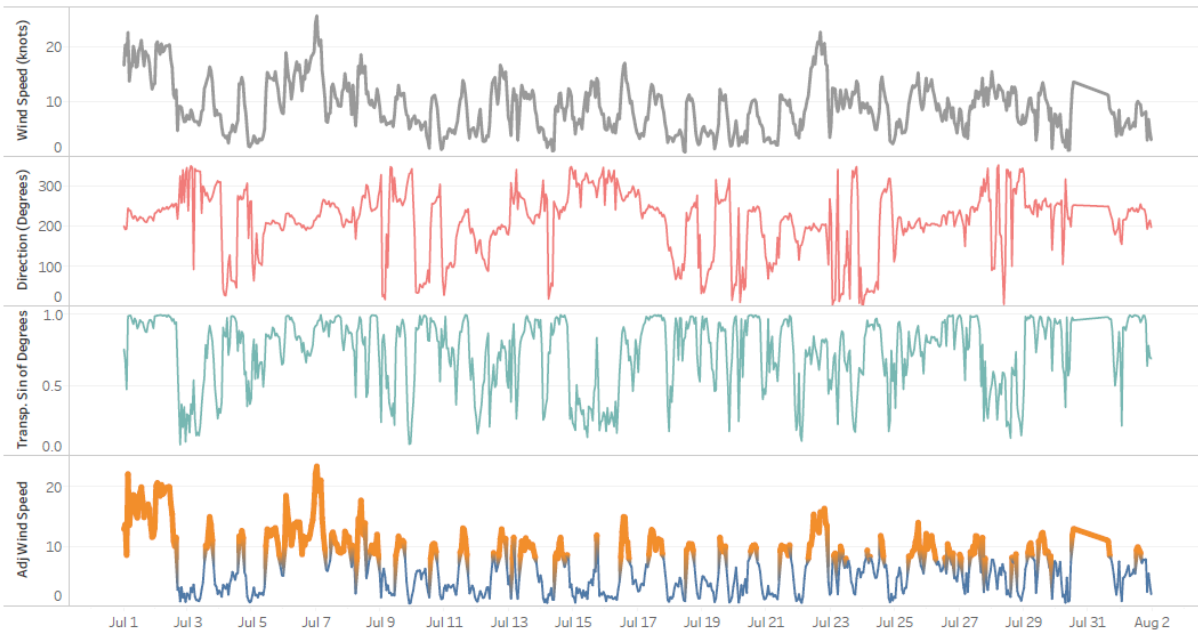
$$\text{AdjustedDirection} = \sin((\text{direction} - 36)\%180) * 0.0174533)$$

*Fig 4: Adjusted direction calculated by subtracting the fetch of the lake from the recorded direction, taking the modulus of 180, converting to radians, then taking the sin function.*

Taking this one step further and getting back to our core metric of wind speed, I wanted to try applying this “adjusted” wind direction metric to the wind speed metric, to come up with a final “combined” wind speed significance that took direction into account. I did this using another Excel formula, (simply taking the product of the two) and applied it to the entire dataset so the new metric could be visualized in Tableau.

$$\text{AdjustedWindSpeed} = \text{AdjDirection} * \text{WindSpeed}$$

July 2014 Wind Speed



A final coloring and visualization threshold of  $> 8$  (adjusted knots / hr) was applied to

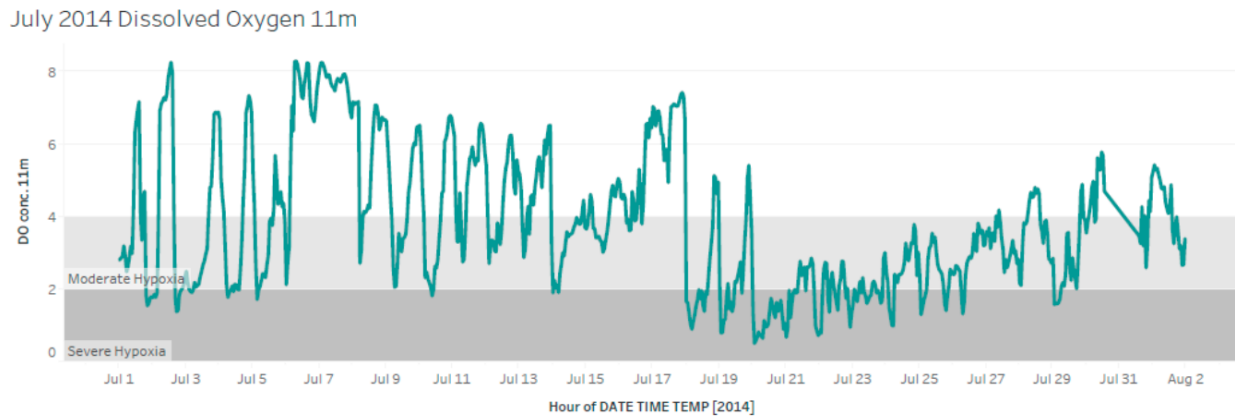
color episodic wind events. Given that this combined metric was unverified, I took some time to statistically analyze it later on in the project.

*Figure 5: Combination chart showing wind speed, raw wind direction, adjusted wind direction, and adjusted wind speed, respectively.*

### *Dissolved Oxygen*

We used dissolved oxygen data to analyze the hypoxic state of the lake. For our analysis, we primarily used the “bottom dissolved oxygen” concentration, i.e., the data from the 11m depth. For quantification and visualization, we used standard accepted hypoxia values – anything less than 4 mg/L (dissolved oxygen concentration) is considered moderate hypoxia, less than 2 mg/L is considered severe hypoxia, and 0 mg/L is considered anoxia or no dissolved oxygen.

These levels have correspondingly severe effects on aquatic life. To label this data, the master dataset was processed through the same Java program used for temperature depth readings. Additionally, we thought to calculate the “rate of change”, or gradients, for this dissolved oxygen data, so we could label “rapid oxygenations” and “rapid hypoxia events.” This code was very similar to the temperature depth gradient calculations. Although these labels were not specifically requested by AWRI, they may be useful for further studies, and they should closely track with episodic temperature change (“mixing”) events.



*Figure 6: Dissolved oxygen at 11m, with reference bands highlighting hypoxic zones.*

### *Chlorophyll / Phycocyanin*

These two metrics were used to track algal bloom progressions in Muskegon Lake. Harmful algal blooms contain cyanobacteria that produce phycocyanin (a pigment protein). On

the other hand, both harmful and general algal blooms contain chlorophyll. Both the chlorophyll and phycocyanin sensors are benchmarked each year using known concentrations of pigments.

The dynamics and relationships of both “general algal blooms” and “harmful algal blooms” were important to understand and visualize, as they affect each other and have different effects on the ecosystem. One data nuance of measuring algal blooms from these sensors is they

*GeneralAlgalBloomThreshold = (NormChl + NormPhyco) > 1SD*  
have a large hourly variance, so daily averages are essential. Using Tableau, we calculated intra-annual means and standard deviations. We then compared each daily average to the mean and standard deviation to find out if there was an “above average” algal bloom. To track “general” algal blooms, we wanted to use both chlorophyll and phycocyanin and see if either were above 1 standard deviation. To do this, we had to first normalize the data to get them onto the same scale (chlorophyll is measured in micrograms per liter and ranges from 0 to 20, whereas phycocyanin is measured in cells per milliliter and ranges from 0 to 40,000). Once the data was normalized, we were able to combine these two measures and see when values exceeded one standard deviation.

We colored these events as “general algal blooms”, and we were able to use Tableau to export this selection of items for future research. To determine “harmful algal blooms”, we wanted to highlight timeframes with only high phycocyanin levels, excluding high chlorophyll values. To do this, we subtracted normalized chlorophyll from normalized phycocyanin creating a new combined measure. Then, we only considered values in this combined metric that were above 1 standard deviation and additionally had a normalized phycocyanin value that was greater than the mean (0 standard deviation). This latter part of the check would make sure periods of

extremely low chlorophyll levels (negative standard deviation) but average phycocyanin did not trigger a categorization as a harmful algal bloom.

$$\text{HarmfulAlgalBloomThreshold} = (NormPhyco - NormChl) > 1SD \wedge (NormPhyco > 0SD)$$

Finally, thinking about the rate of change and the importance of being able to detect sudden algal bloom spikes, we investigated the rate of change of both of these combined metrics. Due to the daily aggregation requirement, we were not able to run individual record data through our Java program similar to water temperature and dissolved oxygen metrics. Instead, a Tableau “last difference” calculated measure was used to track changes from the previous value. We proposed that a daily increase of 50% or greater in either of these respective combined measures could be considered a “general algal bloom spike” or a “harmful algal bloom spike”. In the future, understanding why algal bloom spikes such as these occur would be important for predictive models, but being able to categorize them as such would allow for real-time warnings to be disseminated to the public.

July - Aug: Phycocyanin & Chlorophyll

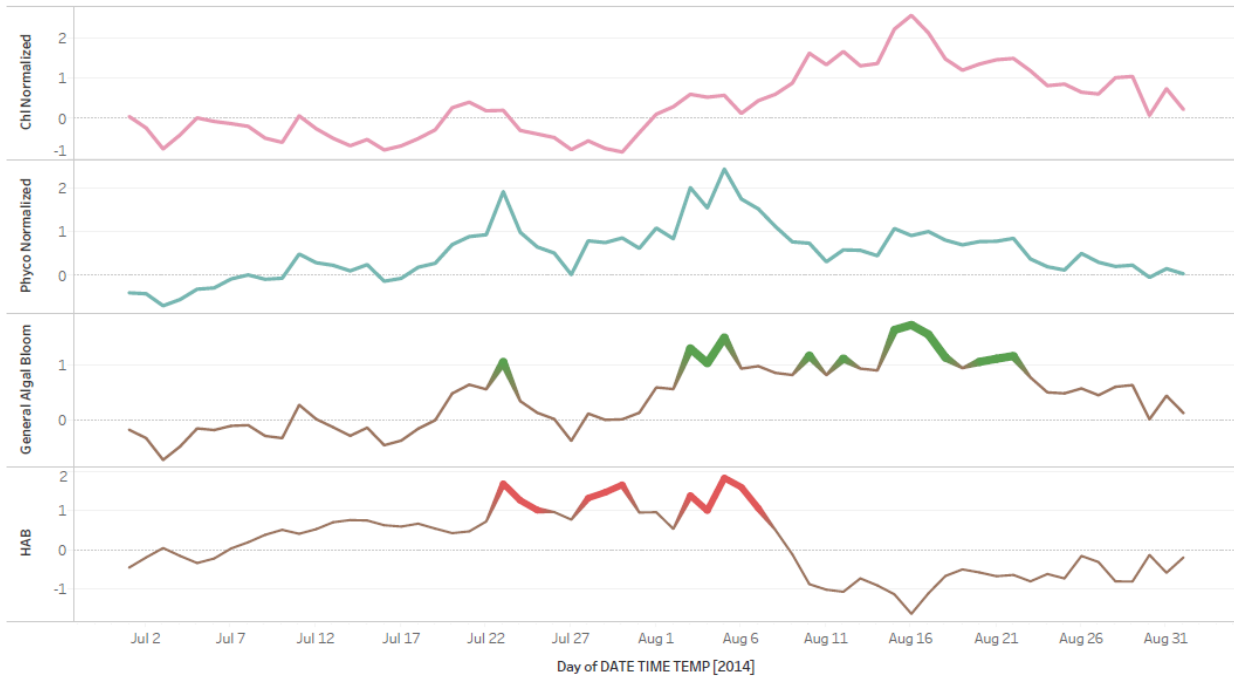


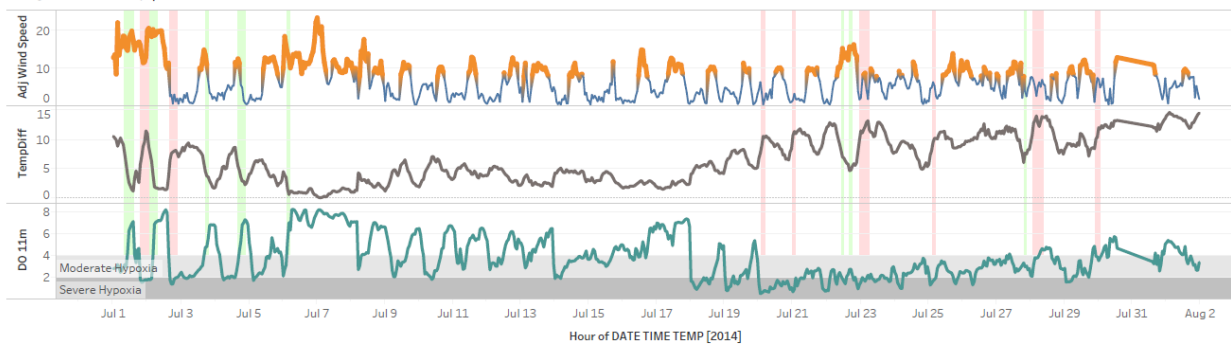
Figure 7: Stacked chart showing normalized chlorophyll and phycocyanin (respectively), the combined algal bloom measure with coloring of significant dates, and the harmful algal bloom measure with coloring of significant dates.

### Creating a Combined Visualizations

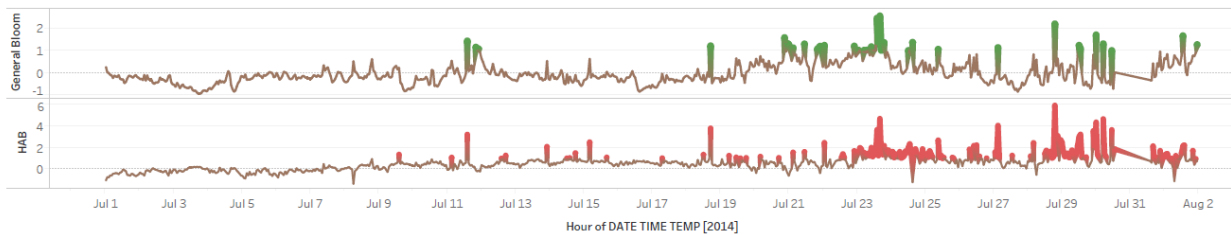
One of the benefits of quantifying and compressing these measures is having a quick way to discern trends. By providing colors or reference bands to values that fall outside of the norm, we can accentuate latent patterns. Further, by stacking multiple metrics and looking at them over the same period of time, we can begin to see how they might affect one another. Seeing this visually, and having the idea of a potential relationship, can open the door for further scientific analysis to be done in the future. We tried to provide some of this to AWRI by generating our

own “combined” chart featuring all of the metrics previously discussed. We came up with a very powerful visual tool to summarize key metrics of the ecosystem. This was conducted by adding color to episodic wind events in the “adjusted” wind speed measure, vertical reference bands marking rapid changes to the temperature diff metric, horizontal reference bands to mark when bottom dissolved oxygen became hypoxic, and color to both the general algal bloom and harmful algal bloom normalized metrics.

July 2014 Temp/DO



Phycocyanin & Chlorophyll



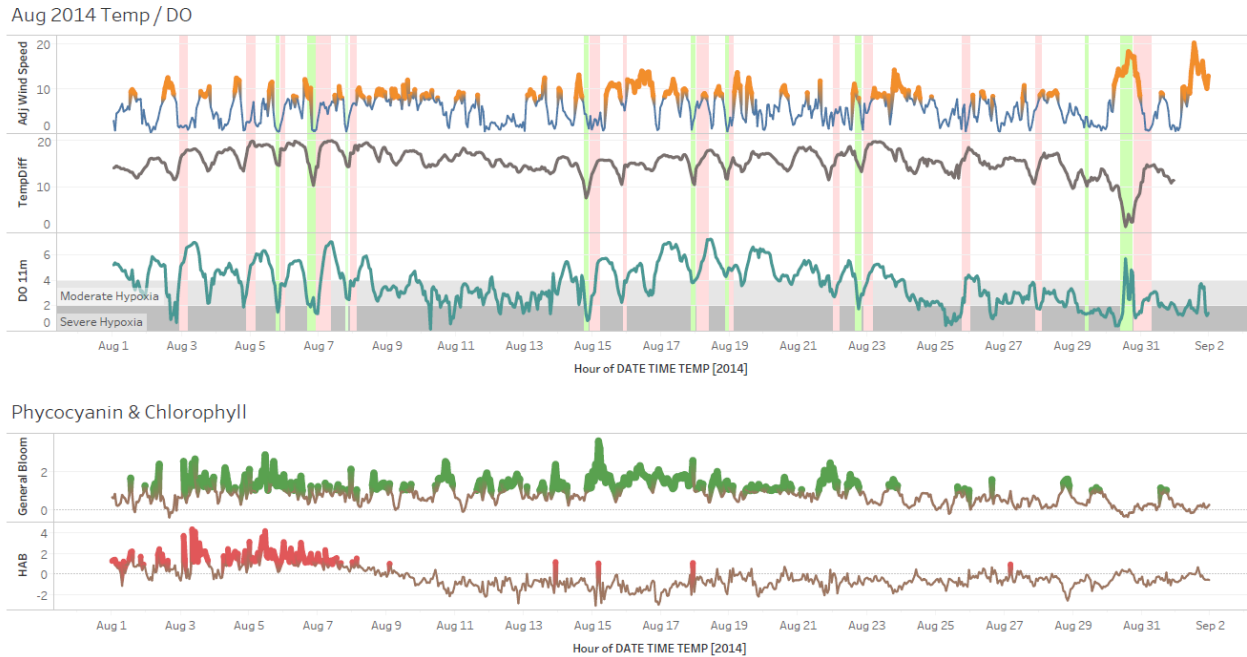


Figure 8 and 9: Combination charts showing adjusted wind speed, temp diff, bottom dissolved oxygen, general algal bloom measure and harmful algal bloom measure. July and August of 2014.

### Other Visualizations and Patterns

After presenting and discussing with AWRI, their team had a few other ideas for insights that might be useful to visualize. First, *looking at seasonal trends from all years*. This would set expectations for how means change throughout each season and also provide baseline values for how seasonal trends are changing year over year. For example, maybe the last four years have had dissolved oxygen values below average in August. Second, *exploring daily variations in a few of these most important metrics*. The daily cycle of the sun accounts for many of the patterns in the data that can be observed, from thermal stratification of the lake to cyclical wind to algal bloom spikes. Providing new ways to visualize and understand these short-term cycles will help the AWRI team and give ideas for further research.



After exploration in Tableau, I developed a few charts that were particularly succinct at capturing the patterns that were of interest. For the first chart, by calculating a daily mean over the nine years of data and grouping by month, we were able to identify several clear patterns that validated what we already knew but communicated visually. In this first chart, I included average air temperature and average water temperature (across all sensors) in the top two rows to orient the viewer. The third row uses the *temp diff* metric, previously described, wherein stratification begins to develop every year in early July and begins to break up again toward the end of August. Finally, in the bottom chart, we observe a similar pattern with bottom water hypoxia that seems to follow the inverse of the *temp diff* (although slightly delayed).

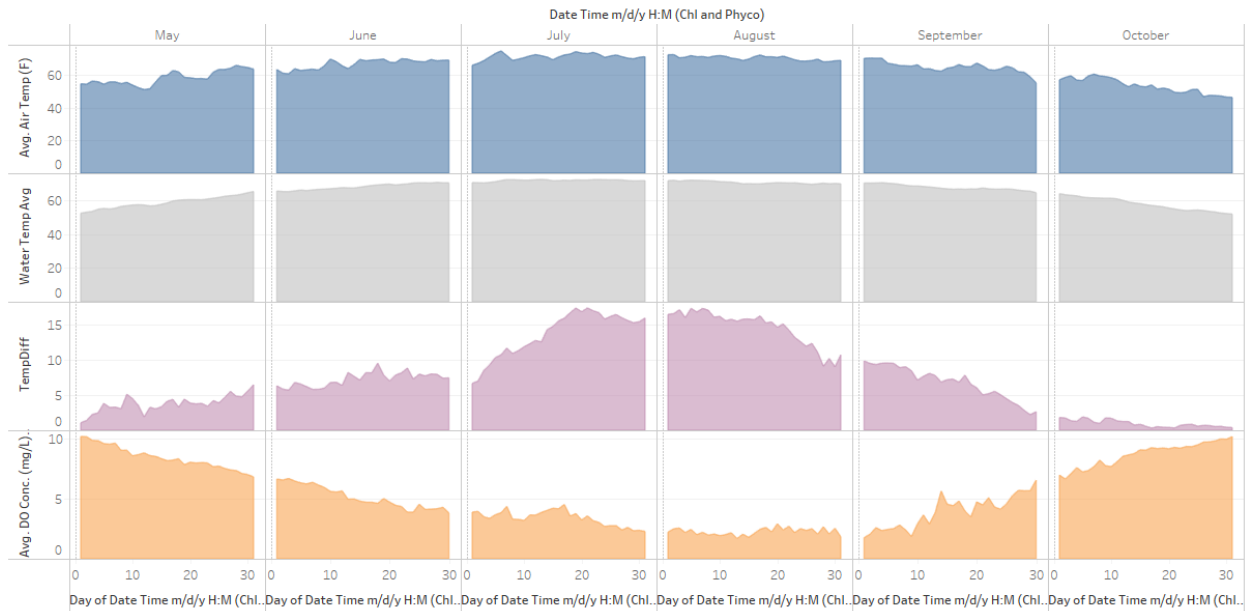
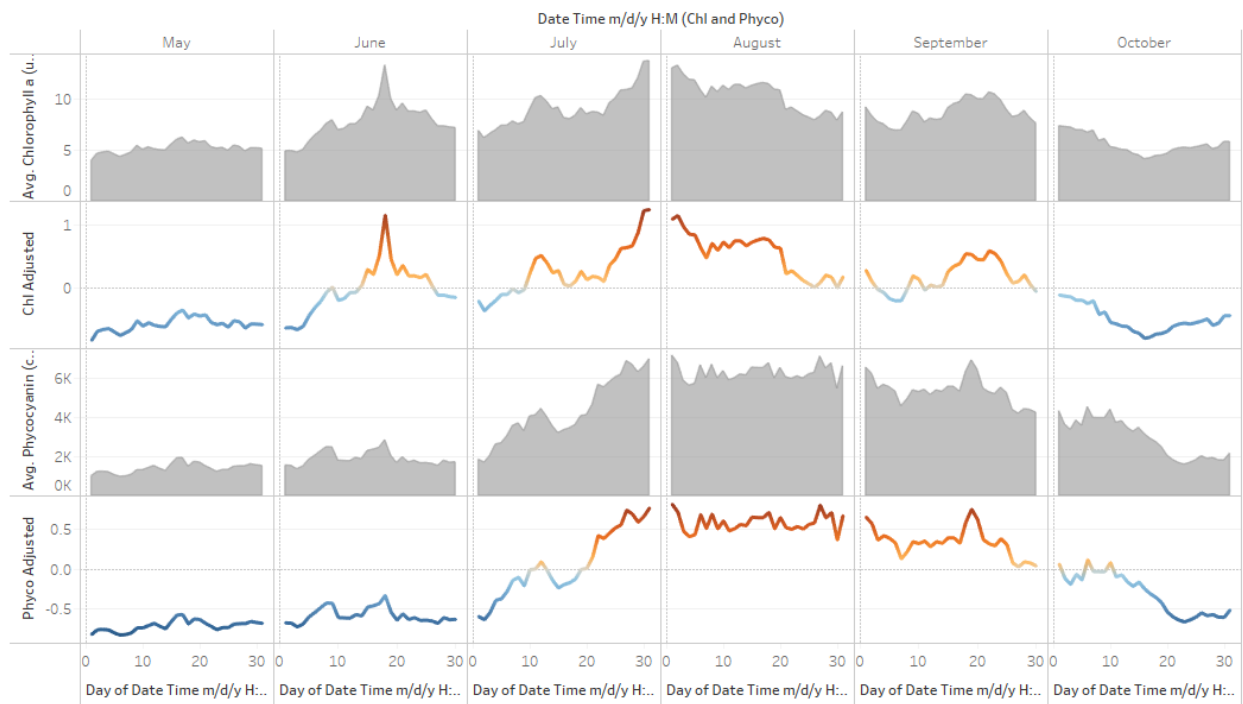


Figure 10: Combination chart showing seasonal trends averaged over all years of data, 2011 to 2019, with months clearly marked.

In this second chart, also looking at seasonal trends, we take a closer look at the algal

bloom metrics. The top two rows cover chlorophyll, looking at the raw data and then the normalized values (respectively). The bottom two rows cover phycocyanin, following the same pattern of raw and normalized data. Chart colors visually highlight normalized values around 1 standard deviation to give additional emphasis. On the chlorophyll rows, we can see three clear “peaks” that seem to occur year over year, in June, July-August, and September. Further research and investigation is needed to confirm if this is a scientifically valid phenomena or just a result of a limited dataset (nine years might not be sufficient for verifying trends). The phycocyanin rows show a similar but more sustained peak period from the end of July to mid-September. This trend seems to be more consistent with expectations.



*Figure 11: Combination chart showing seasonal algae metrics averaged from 2011 to 2019.*

The astute reader may be wondering where wind speed is on these charts. When plotted annually over each day, a non-descript trend line was seen without clear peaks or valleys. Past research has shown increasingly severe storms in the Spring and Fall (Weinke and Biddanda 2019), but perhaps this pattern gets muted by taking a daily average over many years.

Looking at daily patterns, a few other interesting trends emerge that may not otherwise be clear. Following a similar pattern at the other charts, I conducted additional groupings by month because certain daily trends are only observed or observed with greater intensity during the summer months. Going over each row, we can describe clear patterns that may occur over the course of a day. As the sun warms the earth and average air temperature increases (row 1) wind speed increases as well as wind speed is caused by a difference in pressure that can occur between warm and cold areas (row 2). This increased wind speed causes the *temp diff* metric to decrease as the lake seiches and mixes more (row 3). Finally, we can see the bottom dissolved oxygen metrics responding accordingly to the *temp diff* peaks and valleys, as the lake becomes oxygenated from mixing (row 4). Although this story may feel exaggerated, it represents nine years of data and fits with our understanding of lake physics.

Finally, we consider the daily patterns of algal blooms. For chlorophyll, we can observe a pretty clear “valley” throughout the day time hours, whereas for phycocyanin, it appears to be more of a “peak.” Perhaps this can be accounted for by the mechanisms of the sensor, which react to pigments in the water and may be affected by light and dark. Alternatively, perhaps something else is going on biologically that accounts for this daily variation.

As previously mentioned, these analyzed metrics are a subset of available datasets. We anticipate that these ideas and visualizations can inspire future work and lead to greater understanding of our ecosystem.

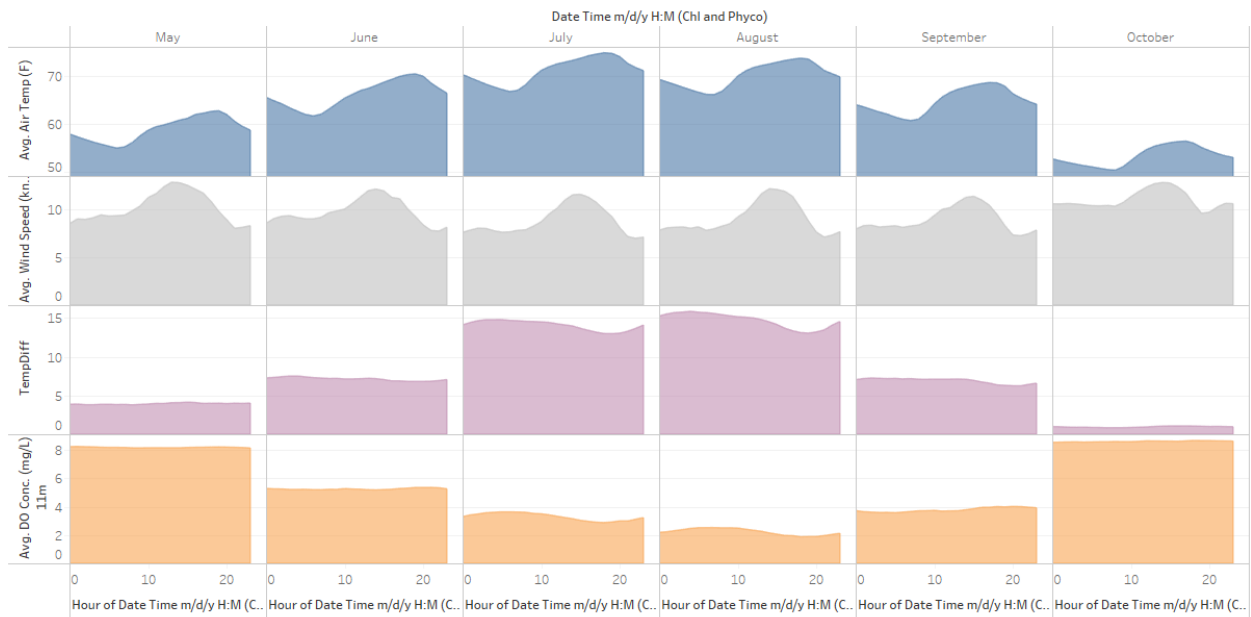
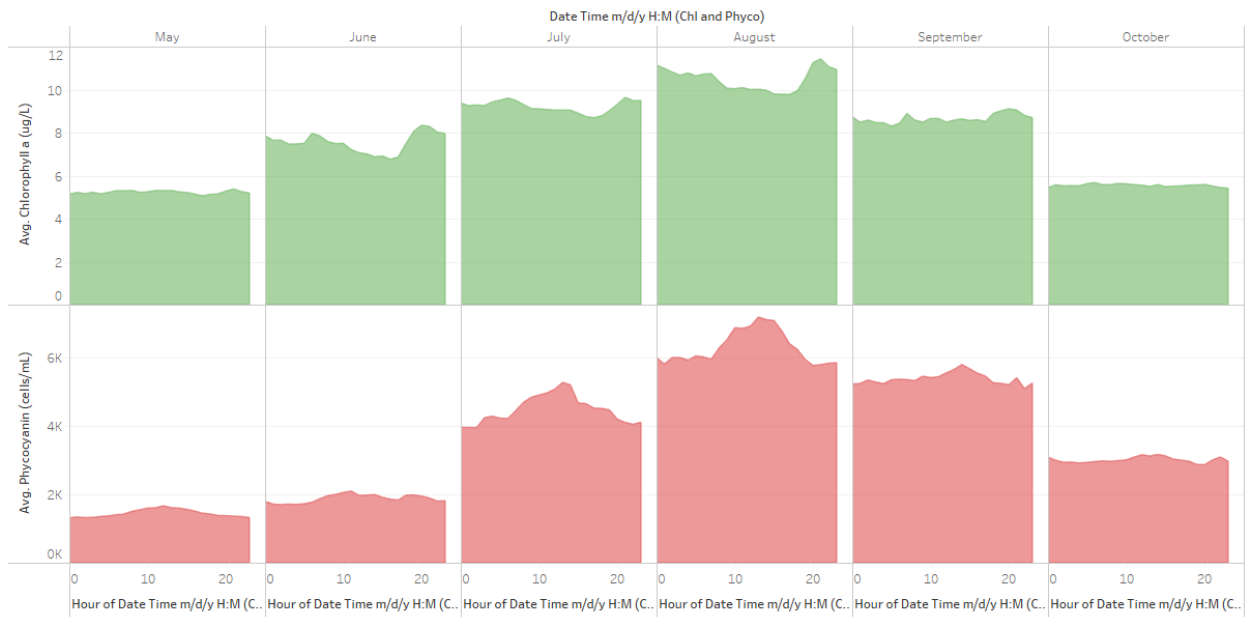


Figure 12: Combination chart showing daily patterns of air temperature, wind speed, temp diff, and bottom dissolved oxygen, grouped by month.



*Figure 13: Combination chart showing daily patterns of algae metrics, grouped by month.*

## **Exploring Episodic Weather (Project 2)**

### *Cross Correlation Function to Quantify Time Delay*

Knowing that a relationship exists between these processes brings us back to one of our original goals – how can we model this complex intersection of physics and biology? We quickly realized that predicting a process such as an algal bloom was going to have to be completed by future work. In the biological realm, there are many contributing sources and eliminators of an algal bloom, and this combined with the “random” dynamics of a bloom cycle can make a full model seem daunting at best (Carpenter et al 2020). Even considering only the physical properties, there are many unknowns about how specific processes interact in a lake ecosystem. Thus, we decided to spend our efforts on physical properties and focus on quantifying these relationships. For example, it is known that episodic storm events cause mixing, but what type of delay is involved between high wind events and increased dissolved oxygen levels from mixing? Gathering inspiration from the Carpenter paper, we decided to perform a statistical analysis of data to quantify some of the delays that we had observed from our visualizations. As exploring “episodic weather” was the focus of the project, we tested two relationships following wind events. Specifically, 1) the delay between high wind and the *temp diff* metric and 2) the delay between high wind and bottom dissolved oxygen. To explore these delays, a statistical function called a cross correlation function was used to “shift” the datasets against each other, finding the periods of highest correlation. Running this function on the first relationship revealed a very significant short delay of about 4 hours between high winds and declining *temp diff* values. This

can be seen on the chart as a “lag” of 16, given that each period in the data is 15 minutes. The second relationship was also found to be significant but much longer, i.e., over a period of 13 hours.

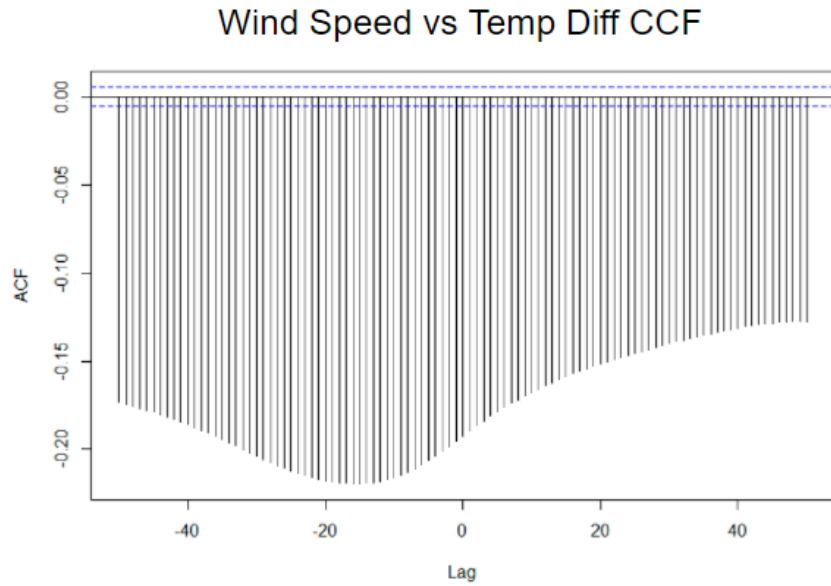
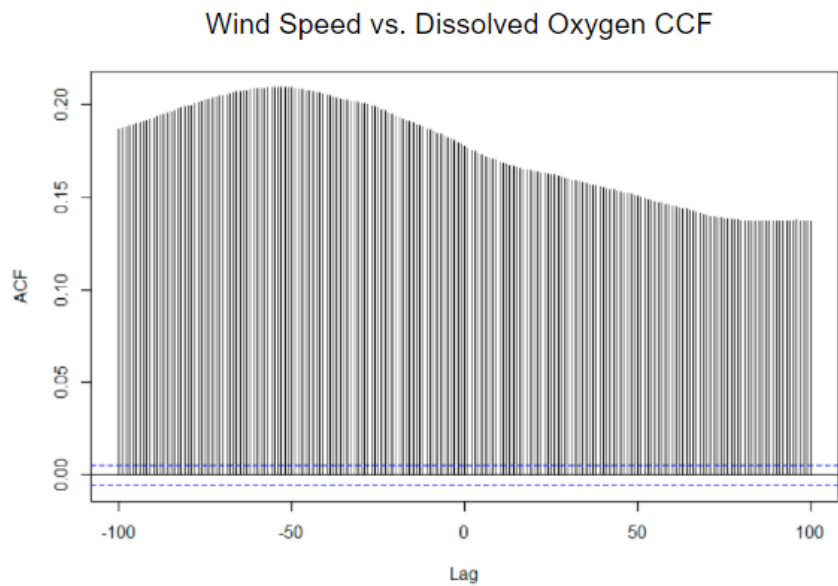


Figure 14: Cross correlation function of wind speed and temp diff metric. Peak lag for ACF correlation is seen at about 15-16 periods, or 4 hours.

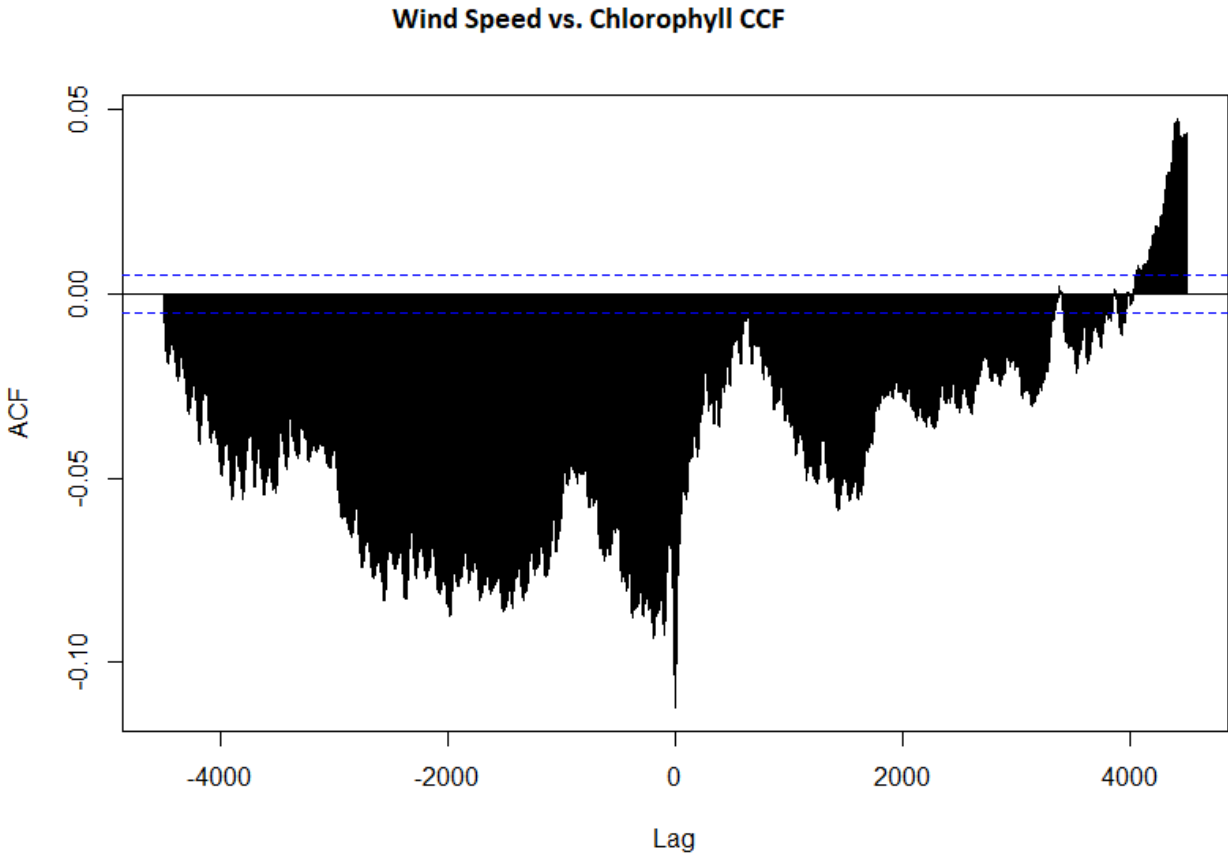


*Figure 15: Cross correlation function of wind speed and bottom dissolved oxygen. Peak lag for ACF correlation is seen at about 52 periods, or 13 hours.*

Intuitively, both of these calculated values appear reasonable, as the seiching that churns the lake does not occur immediately (4 hours appears reasonable for a lake to mix), and dissolved oxygen diffusion takes longer than water mixing (Richard 1996, Wang 1965). Finally, we took a cursory look at the high wind to algal bloom relationship via the cross correlation function. As expected, this analysis proved to be complex and not entirely illuminating. A strong negative correlation was consistently seen for both chlorophyll and phycocyanin metrics from a time range of 0 – 50 days. Considering high wind events should break up algal blooms and periods of low wind will cause more algae to congregate, this inverse relationship makes sense. However, one additional relationship we anticipated seeing was a period of *positive* correlation delayed by about 1-2 weeks. This would indicate a mix event distributing a heavy load of nutrients that *eventually* caused an algal bloom. Although this was not seen statistically, there are some easy ways to explain the discrepancy. First, the negative correlation seen by the wind events breaking up algae blooms may be more significant than the positive seen during a stir up, thus overshadowing it. Second, a cross correlation function is a simplistic look that specifically attempts to find a single amount of delay that has the highest correlation. For complex physical and biological processes, this is simply not accurate. For example, one season may have an episodic storm event that mixes bottom nutrients 14 days out, and then a period of 13 calm days before an algal bloom is seen. Another season may have an episodic event 3 days out, and only 2 calm days seen before a bloom. This is further obfuscated by the fact that high wind events can be both negatively and positively correlated with future algal blooms. Given these shifting delays



and inverting relationships, the complexity becomes too much for a cross correlation function.



*Figure 16: Cross correlation function of wind speed vs chlorophyll, showing a long range of lags (4000 lag would be 41 days).*

### *Evaluating Adjusted Wind Speed*

Given that we were able to see a very clear relationship between wind speed and the *temp diff* metric using the cross correlation function, we tested the “adjusted” wind speed metric that was explored during the visualization phase. Plugging this adjusted metric into the function instead of the raw wind speed data revealed a much lower correlation coefficient. We hypothesized that perhaps the angle was calculated wrong or the fetch was not quite right. I calculated 18 different “adjusted” metrics for each 10 degrees of a 180-degree sin function. Graphing all 18 of these revealed something interesting – there was definitely a curve trend

showing that wind directions were more effective around 30-40 degrees. However, all the adjusted metrics were still significantly lower than the raw data.

After consultation with the team, we came up with the idea that maybe the “adjustment” factor was too large. In other words, instead of adjusting from  $x_0 - x_1$  (fully discounting by the sin function weight), we instead took a smaller weighting from  $x_{0.5} - x_1$ . Knowing that the 36 degree angle seemed correct by the output’s curvature, we modified the adjustment to  $x_{0.5} - x_1$ , and this was much more successful. The modified adjustment actually outperformed raw wind speed for the correlation coefficient.

Figure

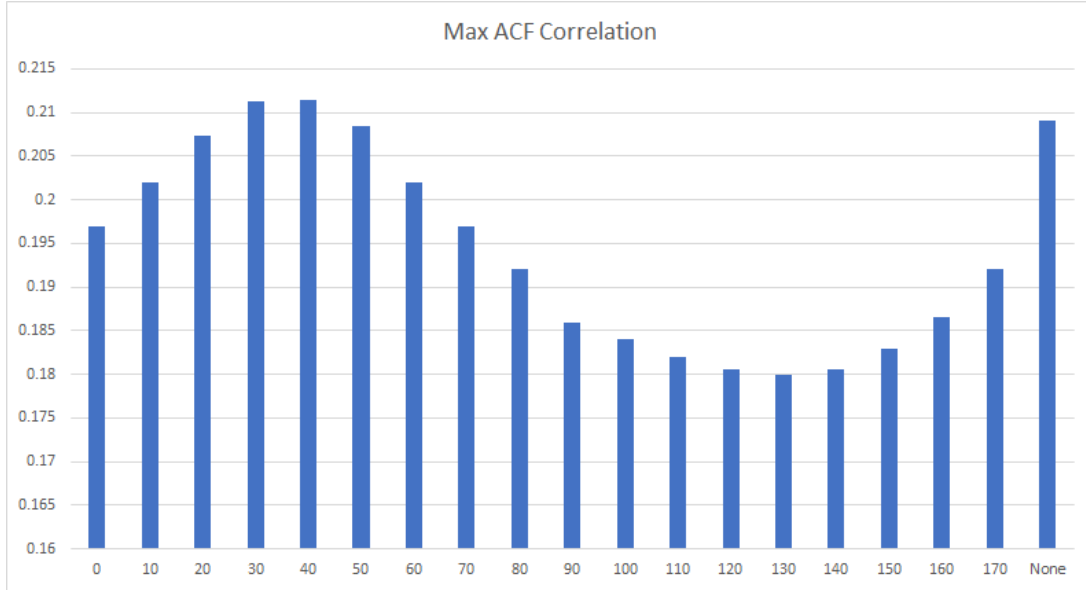


Figure 17: Max ACF correlation shown using the adjusted wind speed metric (compared with temp diff) with different targeted angles of importance.

Finally, I tested a few different adjustment strengths, using 6 different values ranging from 0-1 and 0.8-1. I found that a 40% reduction gave the highest correlation (in other words, adjusting the raw wind speed by x0.6 – x1.0 depending on the angle of the wind). In future work, it may be valuable to try a non-trigonometric function for this adjustment, e.g. a simple linear function or even a step function (on/off) may give a higher correlation between wind speed and temp diff.

$$AdjustedWindSpeed = \frac{AdjDirection + x}{1 + x} * WindSpeed$$

Figure 18: Formula for the second iteration of adjusted windspeed. X was an adjustment factor

ranging from 0.25 (20% reduction) to 4 (80% reduction).

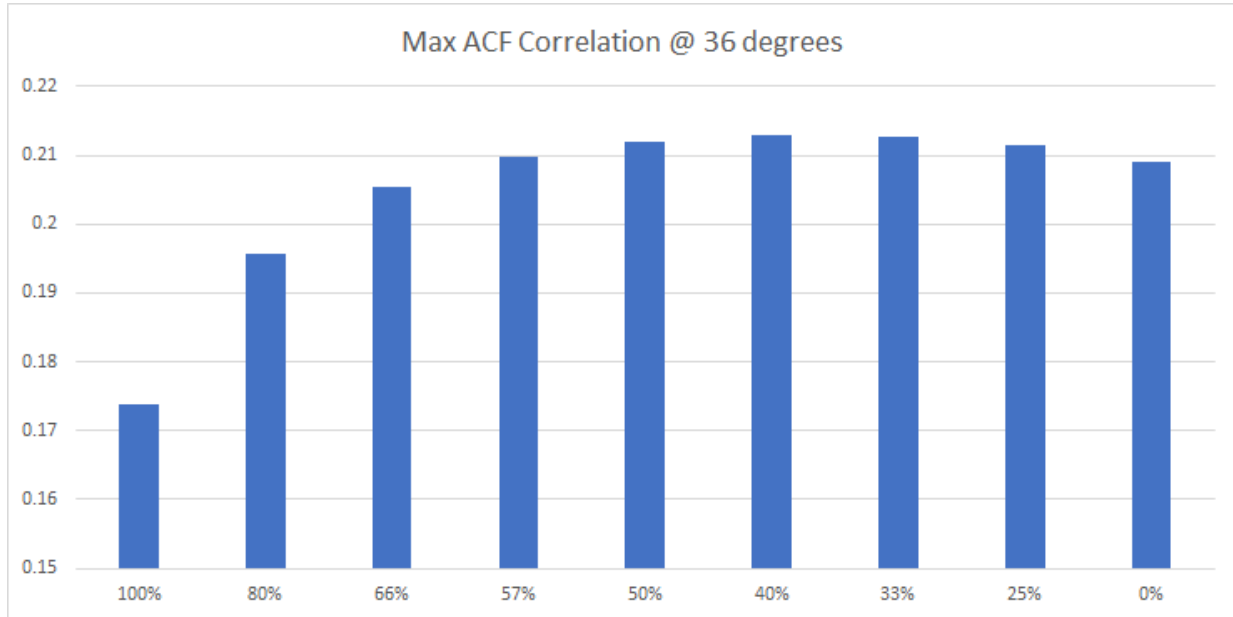



Figure 19: Max ACF correlation shown for the adjusted wind speed metric (compared with temp diff) with different adjustment factors.

## **Muskegon Lake Buoy Dashboard, and Animating Historical Data**

One of the last tasks of the project was providing a proof-of-concept dashboard for the Muskegon Lake Observatory website. At first, our idea internally was to create an animated view of some of the trends in weather and water quality as another look at the data. However, we realized that this view would tie-in with the original project request of a real-time dashboard. After researching similar visualizations of real-time water temperature charts, I knew I wanted to provide a “heat-map” view of water temperature at depths. In this style, the y-axis shows the depth of the lake and temperature readings are conveyed with coloring instead of points on an axis. This view is intuitively simpler to grasp and aesthetically pleasing, making it more suitable for a website dashboard widget where interested citizens and students learn more about the ecosystem. We also realized that we had access to buoy data from three other locations in Muskegon Lake, and it would be more powerful to visualize all four of these locations at once (potentially allowing for physical phenomena such as seiching to be seen). After some design work laying out what metrics to show and how to arrange them into “widgets,” we settled on the four buoy temperature depth readings as a heat map with dissolved oxygen shown below. To the left, a widget showing both wind speed and direction in the form of an arrow of changing lengths, and below that, the chlorophyll and phycocyanin readings. Along the bottom, a year dropdown is available to pick the specific record year that is being shown, and a date time slider allows picking the specific position within the year.

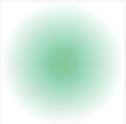
### Wind Speed and Direction



12.8 knots / hr  
259 degrees

### Algae Measures

Chlorophyll

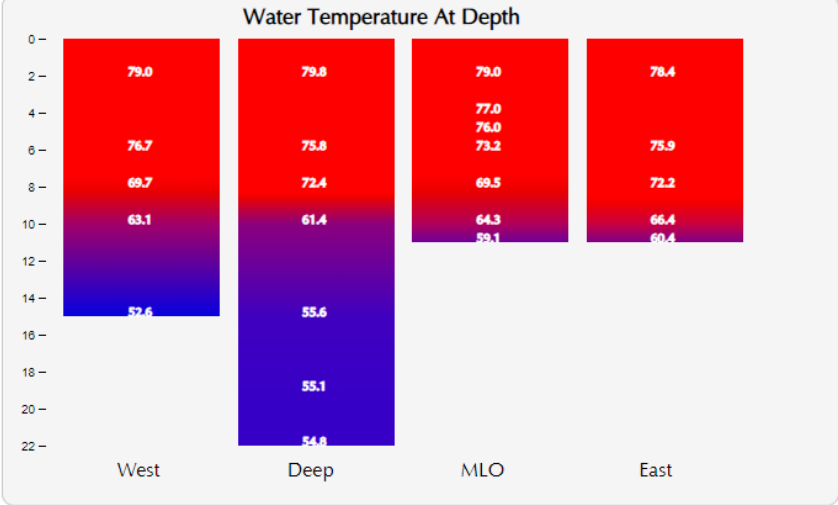


**9.7**  
ug / L

Phycocyanin



**6028.0**  
cells / mL



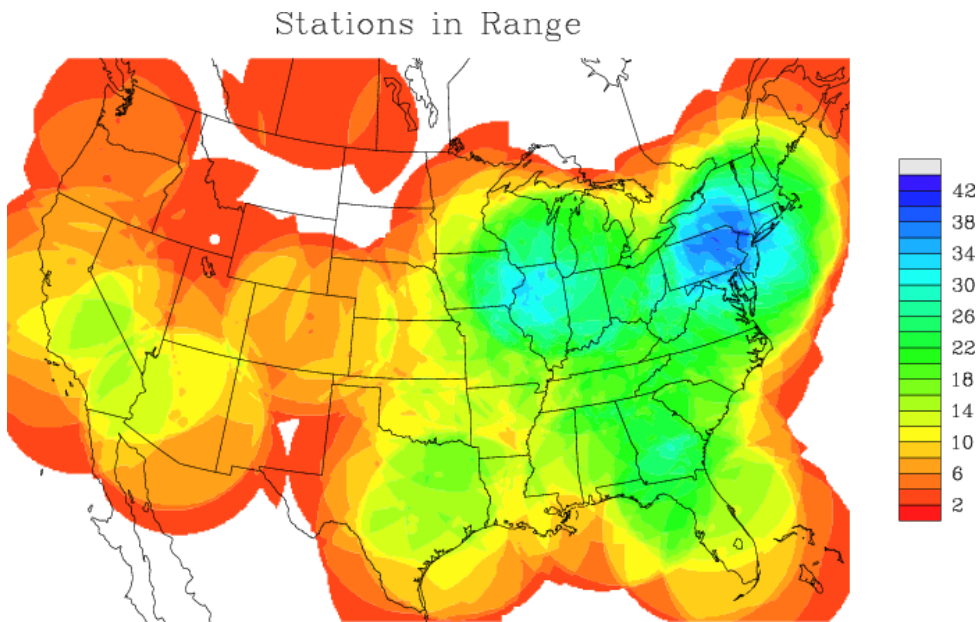
2018 ▾ Play Step Stop

8/13/2018 15:00

*Figure 20: Dashboard prototype to view historical weather and water data for Muskegon Lake.*

Color was added to all the metrics based on where values fell within the range. For example, I took the min / max of all temperature readings and used the max value as the “most red” and min value as the “most blue”. This worked well for all the widgets to highlight “above average” values, and for the water temperature heat maps, it was easy to see stratification layers during the summer months. However, due to how we perceive color, some of the nuances of the “levels” of stratification were not conveyed perfectly clearly. For example, changes of identical amounts sometimes appeared to have much different intensity differences. Although using a grayscale range would fix this, it is debatable whether or not that is ideal, given the target audience and goal to give an overview and inspire people interested in learning about the lake. Considering how visualizations such as weather radar make use of multiple color bands, this might be an effective solution to the problem.





*Figure 21: Weather radar that uses multiple color bands to communicate levels of intensity. This approach may work better for visual comprehension.*

Finally, one additional issue with the color ranges was seen for the algal bloom values. Looking over the entire dataset, there is quite a high variance, in other words the max values can be much larger than normal “above average” measures. This makes using the absolute max for a coloring range not very effective, as there may be only one time when it shows color. Once again, multiple color bands might alleviate this, however, just using the daily averages instead of raw record values would help to smooth out the high variance seen.

Overall, this web-based visualization prototype was successful, and it was a fun way to explore past historical events and lake metrics. The code written should be useful moving forward as we create the pipelines necessary to visualize live data instead of historical data. Additional thought will need to be given to how we can synchronize a live database and perform

the pre-processing automatically, but the problem is well defined.

## Conclusions

Coming back to the three original project ideas given by AWRI, there was substantial progress made on each. For the first project, visualizing seasonal trends, we provided one fully annotated and colored chart for 2014; chart components included adjusted wind speed, water temperature, bottom dissolved oxygen, and algae metrics. Additionally, we labeled all of the “significant” values for these metrics, coming up with a final list from 2011 – 2019 for significant wind events, significant water temperature change events, hypoxia threshold events along with “rapid hypoxia” and “rapid oxygenation” events, general and harmful algal bloom events, and finally general and harmful algal bloom spikes (rapid changes). By going over each metric and figuring out ways to quantify them, we have provided a base foundation for future work that AWRI will be able to take in many more directions. Additionally, we explored trends averaged over all nine years of data, which added to the existing base of scientific evidence and revealed new ideas to explore, such as three seasonal peaks for chlorophyll. Looking at these metrics on a daily scale, we observed distinct 24 hour patterns for algae (both chlorophyll and phycocyanin) during the summer months, which will need more research to understand. An introduction to the Tableau software was given to the AWRI team, which will give the domain experts the tools to explore areas of interest.

Performing raw statistic calculations on the data gave new insights into important physical time delays in lake processes, such as finding a period of 4 hours from high wind events to temperature mixing, and 13 hours from high wind events to dissolved oxygen diffusion throughout the lake. The new idea of a combined wind speed and direction metric was proven to be slightly more correlated with temperature mixing and bottom dissolved oxygen than just wind speed. Much work remains to be done in modeling episodic weather events. For example, for

better analysis of nutrient loading before and after episodic weather events, AWRI hopes to have specific sensors to detect loading of bottom nutrients. With this data in hand, we can scientifically evaluate the relationship between nutrient dispersal and subsequent bloom events. Future computer science work remains to use machine learning in predicting algal bloom events. One limiting factor during this project was a lack of “labeled” data, but moving forward with our list of significant events, we should be able to progress much more efficiently. A lack of explicit “reference” values for several metrics still limits the effectiveness of models. For example, a known value for when phycocyanin is at a “dangerous” value for a harmful algal bloom, or a known value of chlorophyll when a general algal bloom is “severe” would be the final step for a prediction that is useful to the public.

Several difficult problems were solved in the last project area, creating a live dashboard for weather and water quality in Muskegon Lake. A general design and layout was decided on, with room for more metrics to be added. Animations and transitions for a live “progression” was accomplished using the D3 library, and several programming patterns were established for adding new functionality. More design work will need to be done in tandem with the AWRI team to confirm what views and information would be valuable for them. Most of the future work will revolve around setting up a live database or a scraping mechanism for accessing this data in real-time. In this project iteration, significant pre-processing was performed on the buoy data, so these manipulations would need to be programmatically accounted for in a future pipeline or done at the source database level.

The increasing availability of time-series data in many domains makes this an exciting time for data science, ripe with new discoveries. Knowledge gained of the processes in the Muskegon Lake ecosystem will be useful for freshwater mesotrophic lakes around the globe.

## Bibliography

- Biddanda, B.A., Weinke, A.D., Kendall, S.T., Gereaux, L.C., Holcomb, T.M., Snider, M.J., Dila, D.K., Long, S.A., Vandenberg, C., Knapp, K., Koopmans, D.J., Thompson, K., Vail, J.H., Ogdahl, M.E., Liu, Q., Johengen, T.J., Anderson, E.J., Ruberg, S.A., 2018. Chronicles of hypoxia: time-series buoy observations reveal annually recurring seasonal basin-wide hypoxia in Muskegon Lake– A Great Lakes estuary. *J. Great Lakes Res.* 44, 219–229.
- Weinke, A.D., Biddanda, B.A. From Bacteria to Fish: Ecological Consequences of Seasonal Hypoxia in a Great Lakes Estuary. *Ecosystems* **21**, 426–442 (2018).  
<https://doi.org/10.1007/s10021-017-0160-x>
- Weinke, A.D., Biddanda, B.A. Influence of episodic wind events on thermal stratification and bottom water hypoxia in a Great Lakes estuary. *J. Great Lakes Res.* 45, 1103-1112.
- Carpenter, S.R., Arani, B.M.S., Hanson, P.C., Scheffer, M., Stanley, E.H. and Van Nes, E. (2020), Stochastic dynamics of Cyanobacteria in long-term high-frequency observations of a eutrophic lake. *Limnol Oceanogr*, 5: 331-336. <https://doi.org/10.1002/lol2.10152>
- J.H. Wang. Self-Diffusion Coefficient of Water. *J. Phys. Chem.* 1965, 69, 12, 4412. December 1, 1965. <https://doi.org/10.1021/j100782a510>
- T. Richard. Calculating the Oxygen Diffusion Coefficient in Water. Cornell Composting Science and Engineering. <http://compost.css.cornell.edu/oxygen/oxygen.diff.water.html>.