MINING SOCIAL MEDIA TO ASSESS PUBLIC PERCEPTION OF WATER QUALITY

by

Ha Do, Prashank Mishra, Longyi Yang

Dr. James Heffernan

April 28, 2021

Master's project submitted in partial fulfillment of the requirements for the
Master of Environmental Management degree in
the Nicholas School of the Environment of
Duke University

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. ABSTRACT

Social media provides potentially new sources of information for the detection and management of undesirable water quality events such as harmful algae blooms (HABs) in surface waters. Current methods for identifying HAB include field sampling and laboratory tests which are time-intensive and can cause a delay in the issuance of warning advisories, resulting in public health consequences. The potential strengths of social media as water quality indicators are that social media data can be collected in real-time using Application Programming Interfaces (APIs) and is less expensive compared to traditional water quality sampling methods. But the challenge lies in understanding what water quality parameters the public perceives and responds to. To address this challenge, we explored tweets (2016 – 2020) expressing negative sentiment related to the water quality of Utah lake which is well-known for its algae blooms. We used sentiment analysis, natural language processing, spatial interpolation, and count regression modeling to evaluate temporal correlations of social media posts obtained using Twitter API and water quality data collected by the Utah Department of Environmental Quality. We found that the negative tweet counts were significantly and positively associated with many of the perceivable water quality parameters studied such as turbidity, chlorophyll-a, phytoplankton cell count, phytoplankton biovolume, cyanobacteria cell count, and cyanobacteria biovolume. Surface samples for algae concentration and population were also significantly related to the negative tweet counts while the composite samples were not significant, thereby supporting the idea that the public perceives and responds to the toxic water quality near the water surface. Our work serves as a preliminary study that highlights the potential of using social media for identifying water quality events in lakes. To achieve the ultimate goal of developing a real-time public warning system, further studies should be conducted to develop metrics that can translate social media sentiment and activity to a quantitative measurement of water quality health.

2. EXECUTIVE SUMMARY

Recent years have witnessed several promising studies using social media for various environmental applications such as flood inundation mapping and air quality monitoring. This project aims to investigate the suitability of social media in detecting water quality issues such as harmful algal blooms (HABs) in Utah Lake. Utah Lake is a shallow surface water lake popular for recreational activities such as boating and fishing. However, public health concerns of frequent harmful algal blooms in the lake caused by nutrient rich runoffs have resulted in the closure of Utah Lake State Park several times in the past years. The current HAB detection methodology of physically collecting and testing lake water samples takes a few days to issue the warning advisories during which the public might be unaware of the HABs. Social media uses people as sensors and has the potential to provide real-time and inexpensive water quality monitoring. The goal of the study was to assess whether tweets can be an effective proxy measure to identify undesirable water quality events, such as HABs, in the Utah Lake.

To understand whether tweets expressing negative sentiment towards Utah Lake are related to the actual water quality of the lake, we focused on water quality parameters that are perceivable to the human eye. We selected visual water quality properties that affect surface water clarity or the greenness of the lake such as turbidity, chlorophyll-a, Secchi disk depth, total suspended solids, phytoplankton cell count (surface and composite samples), phytoplankton biovolume (surface and composite samples), cyanobacteria cell count (surface and composite samples), cyanobacteria biovolume (surface and composite samples). Due to most tweets not being georeferenced, we reduced the spatial variability of water quality data using spatial interpolation techniques to a single representative value for the whole lake on any given day. We also performed text classification on tweets with keyword “Utah Lake” and its variants to identify tweets talking about water quality with a negative sentiment. Count regression analyses was then performed with negative tweet counts as the response variable and various water quality parameters as predictor variables in separate regression models.

We found that the frequency of tweets talking about water quality of Utah Lake with a negative sentiment was statistically significant and positively associated to several water quality parameters thereby indicating that exacerbating water quality increases the number of negative

tweets. Significant predictors included turbidity, chlorophyll-a, phytoplankton cell count (surface), phytoplankton biovolume (surface), cyanobacteria cell count (surface), cyanobacteria biovolume (surface). Only surface samples of cyanobacteria and phytoplankton variables were statistically significant with negative tweets as opposed to the composite samples suggesting that public does perceive the surface water quality. Moreover, biovolume measurements were significantly associated to the negative tweets posted before water sampling were performed indicating that the tweets may be used to predict algal biovolume in the lake. For automatic text classification of tweets, we found that Stochastic Gradient Descent and Logistic regression algorithms performed the best to identify tweets talking about water quality with a negative sentiment.

Based on our results we would recommend our client to conduct further studies focusing on the water quality variables that were found to be statistically significant in predicting the incidence of total number of negative tweets. For improving text classification accuracy, the number of water quality tweets can be increased by installing poster boards around Utah Lake and promoting the use of Twitter to express opinion on the water quality of the lake. To improve capturing public opinion of water quality on Twitter, text classification can be used on users' bio to only include tweets made by general public and not the government agencies/news organizations. Citizen science initiatives and consistent water quality sampling practices can be used to improve the sample size of the water quality data for more robust statistical modeling.

Our study provides preliminary yet important insights into a novel, relatively quick and cost-effective approach of using social media to identify water quality issues in Utah Lake. Further research should be conducted to overcome the limitations of this study and explore other statistical designs for translating the public sentiment to a quantifiable criterion that can be used to estimate the water quality health of Utah Lake.

3. INTRODUCTION

3.1. Water quality issues in the US

Nutrient pollution is one of the major causes of freshwater degradation in the US (EPA, 2019; Khanna & Shortle, 2017). It is caused when excessive amounts of nutrients such as nitrogen and phosphorus flow into surface waters due to land use and human activities (Smith, 2016). In an unaltered ecosystem, nitrogen and phosphorus provide food and energy to aquatic plants and microorganisms such as algae which, in turn, sustain fish and other life forms. However, high concentration of nutrients in freshwater ecosystems results in a disproportionate growth of algae. Such events are known as harmful algal blooms (HAB). An urgent need is to develop approaches to water quality monitoring that can promptly identify HAB events and issue real-time warning for the recreational use of water bodies.

HABs affect the ecological and recreational value of surface waters because of reduced water clarity and toxins. Such negative impacts can also lead to public health issues. HAB events can reduce penetration of sunlight into water, which hinders the plants' ability to produce oxygen, leading to the increase of hypoxic regions, making uninhabitable zones where fish or other life forms are unable to survive (EPA, 2016). Certain strains of algae can produce aquatic toxins, such as cyanotoxins, which can be harmful to human beings and animals. The contact with such toxins can cause various symptoms including respiratory irritation, headache, diarrhea and can even negatively impact kidneys and liver (CDC, 2020). The negative impacts of HAB events on public and ecological health lead to considerable economic loss to the society as well.

Freshwater degradation due to HABs is estimated to cost $4 billion each year to the US economy. Such a significant loss comes from harm to aquatic food industry, drinking water supply, property values and recreational and tourism activities (Ho et al., 2019; Jakus et al., 2013). According to another estimate, federally authorized water pollution programs provide water quality benefits worth of $11 billion to US economy. EPA estimated that economic benefits of boatable, swimmable and fishable waters alone in Willimette River, Oregon ranged from $120 to $260 million per year (Kauffman, 2018). Another study conducted by the State of Utah found that Utah households were willing to pay $70 million to $271 million to protect

water from nutrient runoffs and spend about $1.4 to $2.4 billion on recreational trips to water bodies, thereby contributing to the state economy (Jakus et al., 2013). To prevent such huge economic loss, there is a serious need from the communities to develop prompt water quality detection methods that can prevent public health consequences due to polluted water.

3.2. HAB detection methods

Traditional methods to identify bad water quality events and issue public advisories do not protect the public from polluted waters quickly enough. HAB detection performance is essential for authorities to take prompt mitigative actions and provide timely warnings to visitors / water users about the water quality issues. The common procedure to issue public warnings on recreational activities in lakes experiencing HAB events takes a minimum of 5 days to more than a week after the event. This lag is because the state entities rely on field surveys and laboratory tests to detect a possible HAB event and issue public advisories. However, people who visit the waterbody within the lag time are not aware of any ongoing HAB issues which can jeopardize the health of the visitors. Thus, a HAB detection method with shorter lag time is necessary to protect the visitors who may potentially recreate with a water body before a HAB warning is issued.

To overcome the delays of HAB identification process, EPA developed Cyanobacteria Assessment Network Mobile Application (CyAN app) which uses satellite imagery data of about 2000 large river and lakes in the US. It allows water managers to perform weekly monitoring on water bodies for issuing quicker public health advisories to pause recreational activities in case of algal blooms. However, large spatial resolution can prevent the use of CyAN app on smaller water bodies and long revisit cycles of satellites can limit the applicability of remote sensing methods to alert officials of HAB in real-time (Li et al., 2018). Social media data mining can potentially provide faster and less expensive detection of HABs than that by traditional methods and can be expanded to lakes and rivers not currently analyzed through CyAN app.

3.3. Social media data mining and its applicability

Social media provide an extraordinary multitude of data on people's feelings, behaviors and sentiments at a granular spatiotemporal scale (Atefeh et al., 2013). Due to its fine and real-

time spatial and temporal extent (Yu et al., 2019), social media data mining, defined as the extraction and determination of actionable insights from social media (Zafarani et al., 2014), has been widely studied and utilized in multiple sectors such as public health, criminology, natural disaster management, and pollution management to identify the historical trends and produce timely forecasts.

In the water sector, social media data mining is considered a promising approach to provide well-timed and cost-effective predictions (Yu et al., 2019) on water quality and quantity issues and supplement water surveying activities. Based on posts shared by users on social media platforms, policy makers and water body administrators can be promptly informed about potential events in their water bodies. Various applications of social media data have been studied such as to map flood events (Middleton et al., 2014, de Bruijn et al., 2019).

However, no application of social media to identify water quality issues in surface waters has been studied yet. This could be because using social media in the context of surface water quality can present unique challenges. First, there is a need to identify specific surface water quality parameters that the public can perceive and respond to out of the multitude of surface water quality parameters that are measured by state agencies. This study aims to understand the public perception of such water quality variables in Utah Lake. Second, there can be an associated difficulty to identify social media messages that are really related to water quality issues. The variability of language can be misleading as a message that contains the keyword may actually be totally irrelevant information (Hang Zheng, 2017). To solve this issue, we employ manual labeling and natural language processing techniques. Third, the posts lacking spatial information also makes it difficult to confirm the location of water quality issues (Hamstead, 2018). To identify useful water-quality-related information from massive amount of data can be challenging. This study aims to overcome these challenges and highlight the potential of using social media data to develop a timely and cost-effective tool to monitor the water quality in critical recreational water bodies needed for water quality management.

Twitter is chosen as the social media platform and the source of public perception data. Twitter is one of the most popular and active social media platforms in the world and in the US (with 22% of adults being Twitter users, accounting for 25% of total social media users in the

country by 2019) (Omnicore, 2020), generating 600 million tweets everyday worldwide. Twitter has also been used in many studies to provide real-time predictions in the US (Jordan, et al., 2019) and the world (de Bruijn, 2019; Middleton, 2014).

4. OBJECTIVES

The objectives of our Master Project are as follows:

i. Determine potential social media data sources relating to public perception of water quality (Twitter) and acquire public perception data on water quality using data mining tools (Twitter API).
ii. Analyze public perception data for the Utah lake using techniques including sentiment analysis and time-series analysis.
iii. Compare public perception data to quantitative water quality data from USGS Water Quality Portal for Utah Lake and propose recommendations to improve the understanding of our statistical analyses.

The analysis of different data types and sources will help us to decide what data to use for our regression model and data-mining tool. The ideal data should be: (1) easy and inexpensive to extract from the data source; (2) include sentimental text with key words that suggest water pollution type, posted date, (and/or spatial information) so that the data-mining tool can use the data to identify the HAB event.

For the hypothesis, we would like to test if the real-time public perception data obtained from social media can reflect the real quality of studied water bodies at a significant level and, if so, what specific kind of data can best provide us the water quality information we want. The fact that some water quality parameters such as turbidity are more perceivable can help us to choose the potential water quality parameters for our study.

The results of our study will provide the scientific evidence for developing a new water quality detection system based on public perception data. If we successfully prove that the public perception data has significant relationship with water quality, then the idea of new detection method is possibly practicable. And if not, it is equally important to understand and identify water quality variables that cannot be used with the public perception data.

5. STUDY AREA

Our study focuses on the water quality of Utah Lake due to its frequent algal bloom issues. Utah lake is a large freshwater lake spread across 150 square miles located in the north-central region of the Utah state. It is 23 miles long and 9-feet deep shallow lake which drains into the Great Salt Lake through the Jordan River. Utah Lake is highly eutrophic, and its shores support more than a quarter million residents (Britannica, n.d.; Fuhriman et al., 1981). It also provides the site for Utah Lake State Park which is popular for recreational activities such as boating and fishing.

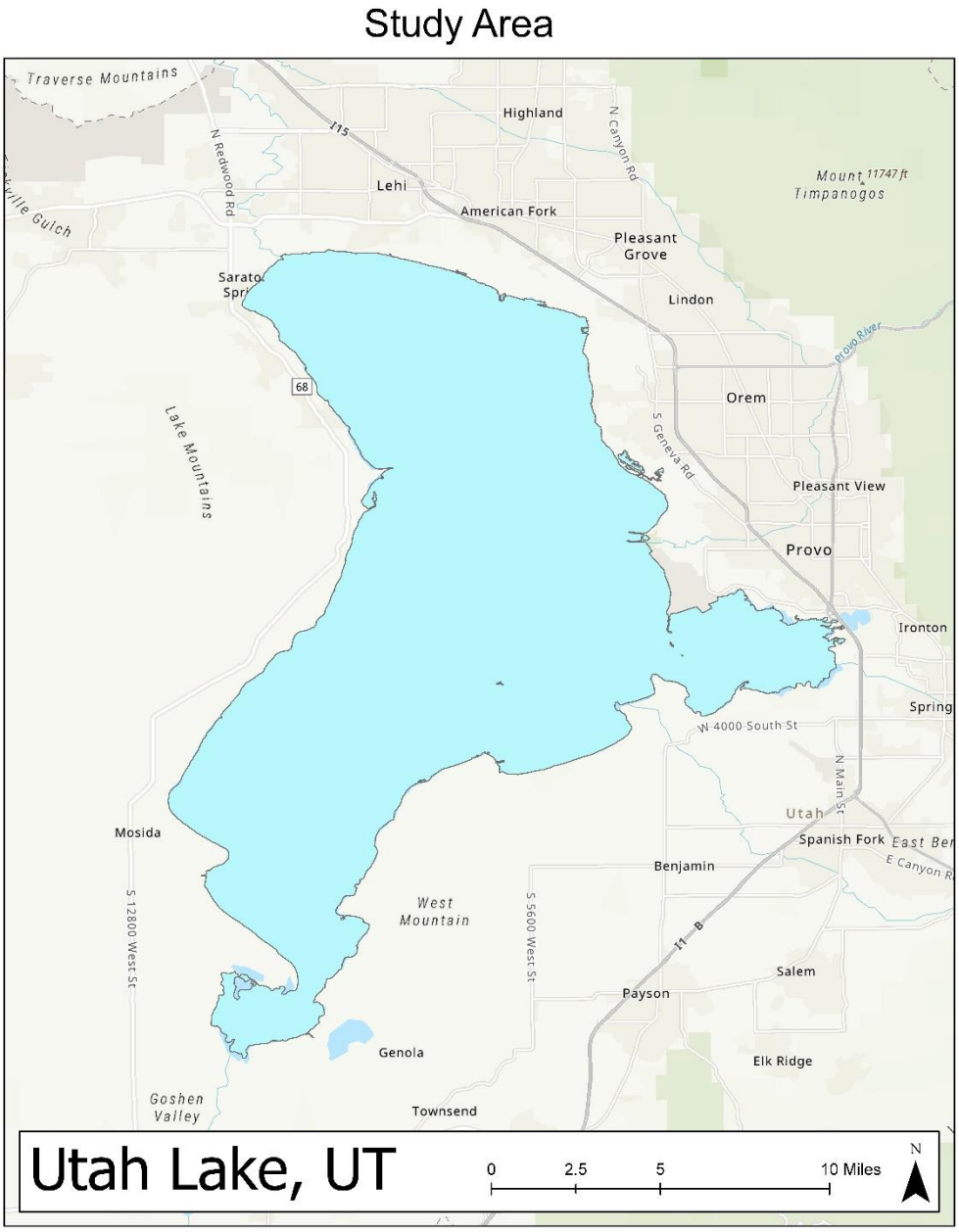


Figure 1. Location of the study area

Due to the semi-arid climate in the region, Utah lake loses a large amount of water due to evaporation, especially during the summer months resulting in an increased concentration of total dissolved solids in the lake (Fuhriman et al., 1981). Utah Valley on the east side of Utah Lake is undergoing rapid urbanization with a current population of more than 500,000 which is predicted to double by the end of 2050 (Randall et al., 2019). Due to increased inflows of nutrient rich run-offs from urbanized and agricultural areas, shallow depth and increased net evaporation loss, Utah

Lake is vulnerable to harmful algal blooms (HAB). Public health concerns due to HABs have led to the closure of the Utah Lake State Park several times over the past years. Utah Department of Environmental Quality (Utah DEQ) monitors the HABs in the lake from May to October.

6. METHODS

Our project consists of the steps as shown in Figure 2. Data processes are highlighted in blue rectangles while inputs and outputs are presented in the rounded yellow boxes. Detailed descriptions of the steps are provided in this section.

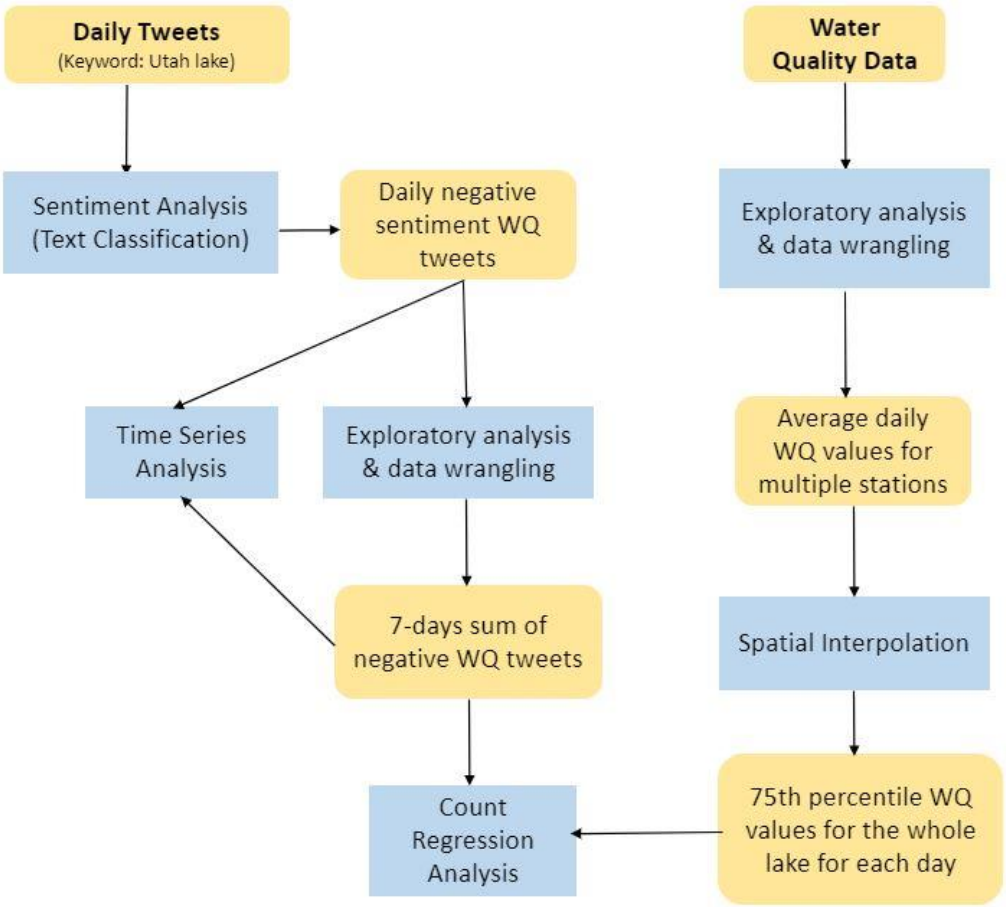


Figure 2. Project's workflow. Data processes and analyses are enclosed in blue boxes while the inputs and outputs of those processes are enclosed in yellow boxes.

6.1. Social media mining

For our study, we obtained tweets pertaining to Utah Lake containing the keywords and/or hashtag "Utah Lake" and its variants from January 1st, 2016 to December 30th, 2020. We also obtained tweet metadata such as date posted, whether a tweet was a retweet or not and latitude/longitude, if available. We removed the usernames of the people posting the tweets in accordance with our Institutional Review Board (IRB) protocols. The tweets were then divided into two groups – tweets with links and tweets without links. Our preliminary analyses of the tweets showed that generally tweets without links are made by public expressing their sentiments while the tweets containing links tend to be from government agencies or news organizations presenting a factual information and directing their followers to their websites or articles. To capture public perception, we used tweets without links. Twitter Application Programming Interface (API) was used via the Tweepy Python package to retrieve the relevant tweets for the time period of 2016 - 2020.

APIs are a set of protocols and procedures that allows a user to access data directly from an application or a web service. API's can be thought of as a waiter in a restaurant where a customer requests a meal. The waiter takes the request to the chef in kitchen and brings back the requested food when it is ready. The customer in this example can be a user interacting with a website and the kitchen as the backend server of the website from where our API brings requested data to the user. One of the benefits of using APIs is that they are official endpoints which are provided by service owners. As a result, the user can obtain data in a clean, organized JSON format separated by requested attributes and properties. Each API also comes with a detailed documentation of all its methods. However, often, APIs have limitations regarding user requests and length of data and may require payment for improved access to data. In order to use API, a user also needs to obtain API key and access token from the service provider. For our study, we obtained the API key and access token from Twitter.

6.2. Text classification

The outputs of this project are expected to contribute to an application that can help screen, analyze and categorize the contents (texts) collected from social networks, which in this

project are tweets. It is an important step to prepare inputs and metrics for the following steps and analyses.

Bag of words is one of the most popular techniques using machine learning approaches to classify text. BOWs deals with two features of a document (in this research is tweets) which are individual words and their occurrences in the tweet. To illustrate, in BOWs, a set of n unique words is generated from the input training dataset. Each tweet is, then, represented by a n-dimensional vector with each dimensional value being the occurrence of the word in the tweet. It is worth noticing that BOWs models consider only the appearance of words, but not the order and the association of the words. These characteristics may reduce the accuracy of the BOWs model. However, BOW is still preferable because of its simplicity and flexibility. Since the project focused on the Lake Utah area where the conditions are specific, BOWs models have produced prediction results with acceptable to very good accuracy (75-95%). Below, the setting and results of the BOWs models will be discussed in more detail.

Each tweet was classified and assigned with two labels: water quality (Yes - No) and sentiment (Negative - Non-negative) (Table 1). Two BOWs models, called Classifier_water and Classifier_sentiment, therefore, were built in parallel to automatically label the tweets.

Table 1. Classification of tweet counts

Label	**Water quality**		**Sentiment**	
Class	Yes	No	Negative	Not negative
If the content of the tweet is	about water quality	not about water quality	negative	not negative

The BOWs models developed in this project follow supervised learning methods, in which we used a number of labeled training documents. Input data was obtained from three main sources: (1) Tweets (without links) collected from 2016 - 2020 with keyword Utah Lake; (2) Tweets (without links) collected in 2013 with keywords Jordan Lake and (3) tweets compiled from a research conducted at Stanford University on sentiment classification (Go, A., Bhayani, R. and Huang, L., 2009). These tweets were manually labeled, of which tweets in groups (1) and

(2) were labeled by the project team, while tweets in group (3) were labeled according to the previous project. The reason for choosing the above data sources is to ensure the diversity of topics and tweets content, while still ensuring that the particular themes of Utah Lake are also included in the model. We also filtered tweets which are duplicated, basically ones retweeted. Information of the input dataset and its sources are provided in Table 2. 75% of the input dataset was automatically selected for training and 25% for validation, both stratified by classes.

Table 2. Sources of input tweet data

Source	**Classifier_water BOW**	**Classifier_sentiment BOW**
1	2000	1387
2	305	300
3	91	3399
Total	2396	5086

BOWs model development includes five main steps. Firstly, the tweets are preprocessed before feeding into the model. The processing techniques include 1) Lowercasing the tweets. 2) removing noise (ie. redundant marks associated with words) and 3) removing stop words (stop words are defined as most common ones in a language, such as I, her, he, she; in this project, the stop word list also includes “Utah Lake”, “UtahLake” and numbers which do not provide information in classifying text). Secondly, a library of all n unique words appearing in the input dataset. Third is to vectorize the tweets where each tweet is encoded by a vector consisting of n dimensions, i.e. each tweet is characterized by n features. The value per dimension is the number of occurrences of each word in the library appearing in that tweet. In the fourth step, the algorithms are developed to identify features that are highly distinguishing between classes. To compare and find the best fit algorithm for each BOWs model, four algorithms including multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistical Regression and SDG Classification were used. The fourth step is to test/validate the model and calculate the model performance evaluation model metrics such as precision, recall and f1-score and accuracy. The model with the

best accuracy was selected to classify the tweets. The formulae of evaluation metrics are provided below:

$$Precision = \frac{True\ class\ 1}{Total\ predicted\ class\ 1}$$

$$Recall = \frac{True\ class\ 1}{Total\ actual\ class\ 1}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{True\ class\ 1 + True\ class\ 2}{Number\ of\ observations}$$

6.3. Data processing

6.3.1. Processing tweet count data

After performing sentiment analysis on historical tweets from January 1st, 2016 to December 30th, 2020, Twitter data was aggregated into total counts of tweets related to water quality and containing either positive, negative or both sentiments for each day. Descriptive statistics were calculated, and time series plots were constructed for daily and weekly tweet counts to identify periods of non-zero tweets per day and to see time periods during which people tweet about water quality in Utah Lake.

Daily tweet counts were overlayed with time series plots of water quality variables to assess the days and periods when someone tweets about water quality and the days on which a water quality parameter is measured. We developed an interactive time-series plot where we can hover each data point to assess the temporal frequency of the tweets with a negative sentiment and water quality sampling.

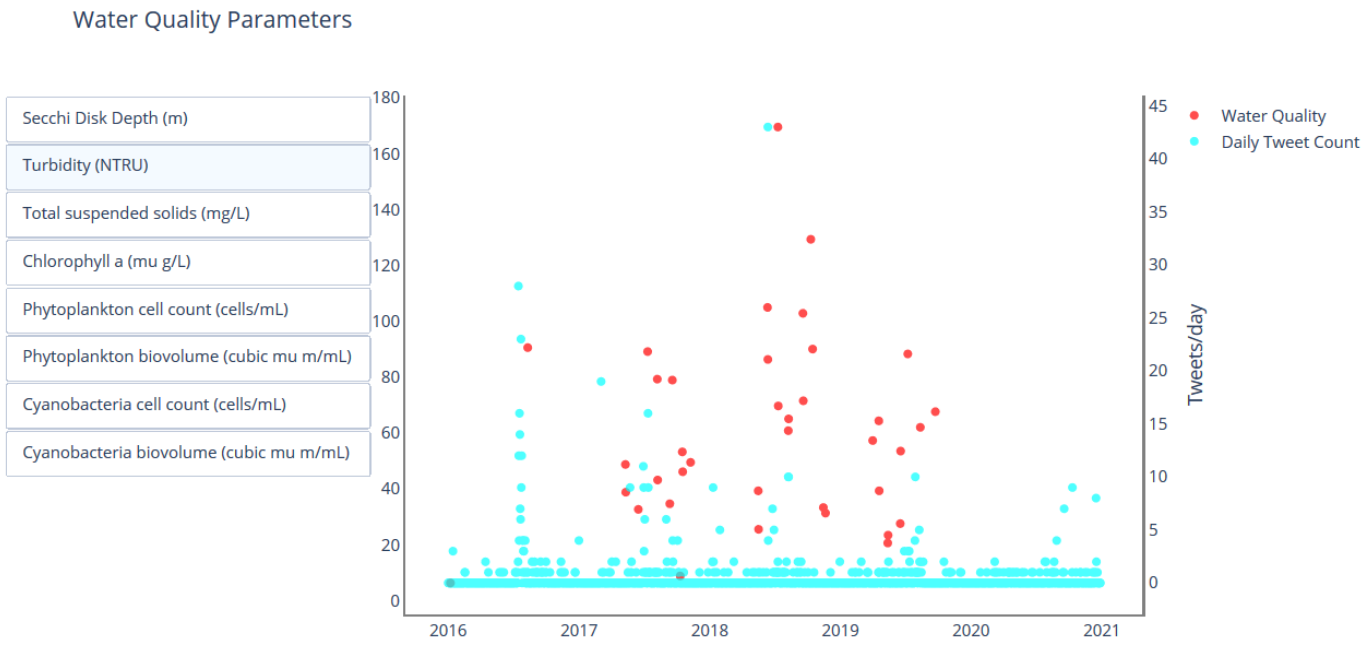


Figure 3. Interactive time-series plot overlaying tweets per day and water quality parameters data on for the days they were measured.

We observed that water quality parameters were not measured on a regular basis. For instance, Secchi disk depth might be measured twice in one month with 1 day gap in between but thrice in the next month with all the measurements being 1 week apart. Because most water quality variables are not measured on a consistent temporal basis throughout the year and because people might not tweet on the same day of a "bad" water quality event, we decided to use total count of relevant tweets within a window of 7 days around the day on which a water quality measurement was performed to maximize the utilization of water quality. Therefore, the tweet counts were aggregated over 7 days before, after and before & after a water quality measurement was made. This aggregation was performed for each water quality parameter of interest.

6.3.2. Processing water quality data

The water quality parameters that were selected from the HAB advisory data were cell counts and biovolume of total phytoplankton and cyanobacteria from surface and composite samples.

The CSV file for routine water quality data was obtained from our client in a long data format where each row represented observation for a unique water quality variable for a

monitoring location, datetime and additional characteristics. The CSV file was transformed into a wide format data frame with each column representing a water quality variable and each row representing an associated monitoring location and datetime. The columns containing water quality variables of interest were retained in the data frame and all other water quality variable columns were dropped. Multiple observations in the same day for a specific monitoring location were averaged and each water quality variable was separated into its own data frame for further analyses. The following perceivable water quality parameters were identified from the routine water quality data for the years 2016 to 2020 - Secchi disk depth, turbidity, total suspended solids and chlorophyll a (corrected for pheophytin).

For the HAB advisory data for Utah Lake, our client provided us with 2 CSV files – one for total phytoplankton population and one for cyanobacteria population. For both files, relevant columns were extracted which contained information regarding the date of sampling, cell count, biovolume, and the depth at which a water sample was taken. The data frames were further divided based on whether each observation represented a surface sample or a composite sample. Surface water samples are collected from the surface of the lake whereas composite samples are a mixture of samples at different lake depths. Processing the advisory toxin data resulted in 8 data frames – phytoplankton cell count (composite), phytoplankton cell count (surface), phytoplankton biovolume (composite), phytoplankton biovolume (surface), cyanobacteria cell count (composite), cyanobacteria cell count (surface), cyanobacteria biovolume (composite), cyanobacteria biovolume (surface) – as potential indicators of water quality in Utah Lake and were ready for further analyses.

To assess how frequently water sampling for selected water quality variables was performed and how are measurement values distributed, exploratory analysis included running descriptive statistics and plotting histograms and time series plots was performed. Water quality was mostly measured in spring, summer and fall months as the lake gets frozen in winter.

6.4. Spatial interpolation

The water quality measurements were obtained from multiple sites on different dates within Utah Lake. The spatial variability of water quality sampling was not applicable for our study because most tweets are not georeferenced. Thus, to understand the relationship between

tweets and water quality measurements of the Utah Lake, we chose to use a spatial interpolation method for each water quality parameter to obtain estimated value for all the pixels in the Utah Lake's raster layer, and then calculate the 75th percentile value as a representative value for the whole lake for each day. We decided to use 75^{th} percentile value for all the water quality parameters (except Secchi disk depth) because the public is more likely to perceive and respond to an HAB event when the concentration of a pollutant is relatively higher.

In geography, interpolation is a process that uses points with specific values to estimate unknown points' values, and there are several interpolation methods available. We decide to use Inverse Distance Weighted (IDW) method for our analysis. IDW method is based on the assumption that local effect of each point will decrease as the radius increases. The points at a closer location have higher weight compared to points at far distance, thus the value of an unknown point can be estimated by the know value points around it (Interpolation Methods, n.d.) Previous studies show that IDW method outperforms other methods for the estimation of water quality parameters (Malla, 2014).

The interpolation was based on the monitoring sites' data from 2016 to 2020, the exact available amount of data points varied among parameters. ArcGIS pro's IDW interpolation tool was used for interpolation and R was used for calculating the 25%, 50% and 75% percentile for each parameter on each date.

6.5. Count regression analyses

To assess whether there was an association between tweets and water quality, count regression was decided as an appropriate statistical analysis for our study. Poisson regression and negative binomial regression are two of the most commonly used count regression models that can be used to model count variables.

Poisson regression is applicable on count data that follows a Poisson distribution. One of the key assumptions of a Poisson distribution is that the mean and variance of the distribution are equal (Coxe et al., 2009). This key assumption of Poisson regression was greatly violated, and the variance was found to be considerably higher than mean of all of our potential 7-days tweet count variables indicating overdispersion of potential response variables. Because Poisson

regression is a single parameter model, it was deemed unfavorable for our study when compared to the negative binomial regression which allows to model both the mean and overdispersion of count data. To maximize the utilization of water quality data, we used each water quality parameter as a single predictor variable to model a single tweet count variable. Therefore, we ran multiple negative binomial regression models and reported the ones with significant results.

The independent variables used were the Secchi disk depth, turbidity, chlorophyll a, and total suspended solids from the routine water quality data, and phytoplankton cell count (composite), phytoplankton cell count (surface), phytoplankton biovolume (composite), phytoplankton biovolume (surface), cyanobacteria cell count (composite), cyanobacteria cell count (surface), cyanobacteria biovolume (composite), cyanobacteria biovolume (surface) from the HAB advisory data.

The potential dependent variables were rolling sum of negative sentiment tweets pertaining to Utah Lake's water quality 7 days before, 7 days before, 7 days after, and 7 days before & after the day on which a water quality parameter was measured. For convenience's sake, we refer to the regression model with total tweets ***n*** days before (or after) the day of water quality measurement as ***n***-days-before model. We used 3 different response variables to understand how public's tweets will be associated with the measured water quality. Because water quality in Utah Lake is not measured at a regular and frequent basis, we wanted to understand how will tweet counts from different time windows will be associated with the water quality measurements.

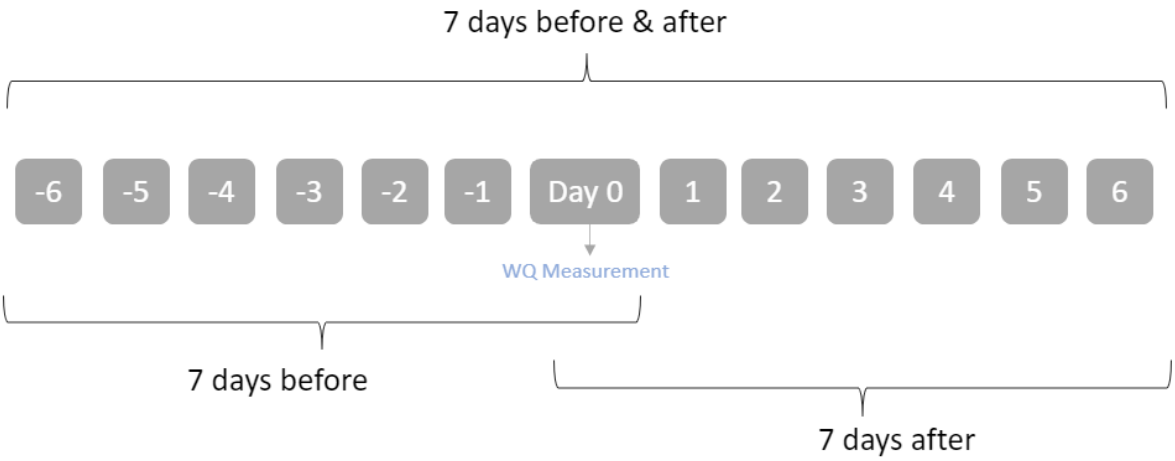


Figure 4. Visualization of 3 different time windows of total negative tweet counts used as potential response variables. Day 0 refers to the day on which a water quality parameter was measured in Utah Lake.

We used Python to automate running 36 negative binomial regression models (12 water quality variables x 3 tweet count variables). Each regression model only had a single explanatory water quality variable to maximize the extent to which we can use the water quality data. Because all water quality parameters are not measured on the same date or with a same frequency, we were not able to use multiple explanatory variables in a single regression model. For instance, using daily representative value for the whole lake, Secchi disk depth might not have a measurement for the same day as turbidity. The process for performing negative binomial regression in Python is given below.

We started with fitting the Poisson regression on cleaned data (for each water quality parameter) obtained from data wrangling process to obtain a vector of Poisson distribution parameter $\boldsymbol{\lambda}$. To calculate the overdispersion parameter $\boldsymbol{\alpha}$ for negative binomial regression, we used an auxiliary Ordinary Least Squares regression with $\boldsymbol{\lambda}$ vector as our independent variable to obtain the $\boldsymbol{\alpha}$ parameter (Cameron and Trivedi, 2013; Date, 2019). We then checked whether the $\boldsymbol{\alpha}$ was statistically significant using its t-score from OLS regression results. If $\boldsymbol{\alpha}$ was statistically significant, negative binomial regression would fit our count data better than Poisson regression models. This is because for a negative binomial regression model,

$$\text{Variance} = \text{mean} + \boldsymbol{\alpha} * \text{mean}^2$$

When $\boldsymbol{\alpha}$ is zero, variance becomes equal to the mean which is the key assumption of Poisson regression. Therefore, if there is overdispersion in our count data, the $\boldsymbol{\alpha}$ parameter would be statistically significant. We found that for most of our models, $\boldsymbol{\alpha}$ was statistically significant, therefore, implying overdispersion in our count data. Negative binomial regression was then used to run a series of count regression models. Python's Statsmodels and SciPy package were primarily used for the methodology described above.

For interpreting negative binomial regression models, we converted regression coefficients to incidence rate ratios by exponentiating them. This makes it easier to interpret the percentage change in tweet counts associated to a unit change in water quality variables. To find the percentage change in the incidence of response variable, we use the following formula,

$$\%Change\ in\ tweets\ given\ 1\ unit\ change\ in\ water\ quality = 100\left(e^{\beta} - 1\right)\%$$

where β is the model coefficient.

For models that suggested a significant relationship between water quality and multiple tweet count variables, we used model deviance as a measure for goodness of fit to select one model for that water quality parameter.

To determine goodness-of-fit of our model to the observed tweet counts, we used Chi-square goodness-of-fit test. We calculated Pearson Chi-square statistic of a regression model along with the degrees of freedom and obtained the corresponding Chi-square p-value using Excel's function CHISQ.DIST.RT. The null hypothesis of Chi-square goodness-of-fit test assumes that the model fits reasonably well and there is no significant difference between the observed and expected values. The alternative hypothesis says that there is a significant difference between the observed and expected values and therefore the model is not a good fit. Lastly, we plotted autocorrelation function (ACF) plots for the tweet counts of each WQ variable used for regression to check for autocorrelation in the tweet counts. Count regression model assumes that there should not be any autocorrelation in the count data. The ACF plots are presented in the Figure 18.

7. RESULTS

7.1. Descriptive statistics of tweet data

We found that social media activity and sentiment were significantly associated with some but not all the perceivable water quality variables. After classification, the number of tweets by categories were further analyzed. Below are the results from, i) descriptive statistics analysis and ii) time-series analysis of tweets.

7.1.1. Descriptive statistics

Table 3 provide a summary of descriptive statistics of tweet counts. From 2016 to 2020, we collected 2696 tweets with keywords Utah lake and without links (hereafter called "tweets" or "tweet"), in which 693 are tweets about water quality (called "water quality tweet"), accounting for 23%. Remarkably, nearly 90% of water quality tweets contain negative sentiment. Figures 5 and 6 illustrate the strong correlation between water quality tweet counts and negative-water quality tweet counts along the observed period. This also suggests that the majority of tweets talking about water quality in Lake Utah were negative complaints.

The daily counts of tweets, water quality tweets and negative water quality tweets with key word "Utah Lake" is over variant and highly left-skewed. The daily counts of tweets containing key word "Utah Lake" range from 0 to 73. However, on average only 1.48 tweet was generated per or more than 50% of the days have zero or one tweets. Also explained by Figure 4. Boxplot of the tweet count), the daily tweet count data is highly left-skewed and fluctuate widely with many outliers. Similarly, daily counts of water quality tweets and negative-water quality tweets ranges widely with maximum values of 46 and 43 respectively. However, more than 75% of the days have no tweets about water quality and with negative sentiment.

To deal with left-skewedness of the data which is constituted of largely zero values, we transformed the daily counts into 7-day rolling sum and monthly sum values. The transformations reduce the variation, skewedness in the data and zero values. However, using monthly tweet counts may loose the timeliness of the information that can be derived from the data. Moreover, people may visit the lake more in the weekend and are more likely to tweet on

these days. 7-day rolling sum can eliminate the effects of the day of the weeks on tweet counts. Therefore, in the following parts, we will focus on 7-day rolling sum data.

Table 3. Descriptive analysis summary of tweet data

Tweet type		**Sum**	**Min**	**25th percentile**	**Median**	**75th percentile**	**Max**	**Mean**	**Standard deviation**
Tweet count	Daily	2696	0	0	1	2	73	1.48	3.46
	7- day rolling sum		0	4	7	12	124	10.35	12.25
	Monthly		10	24.75	35	54	243	44.92	35.28
Water quality tweet count	Daily	693	0	0	0	0	46	0.38	1.92
	7- day rolling sum		0	0	1	2	87	2.66	7.64
	Monthly		0	2	5	13	153	11.55	21.83
Negative - Water quality tweet count	**Daily**	**620**	**0**	**0**	**0**	**0**	**43**	**0.34**	**1.82**
	7- day rolling sum		**0**	**0**	**0**	**2**	**87**	**2.38**	**7.41**
	Monthly		**0**	**2**	**3.5**	**11.25**	**151**	**10.33**	**21.34**

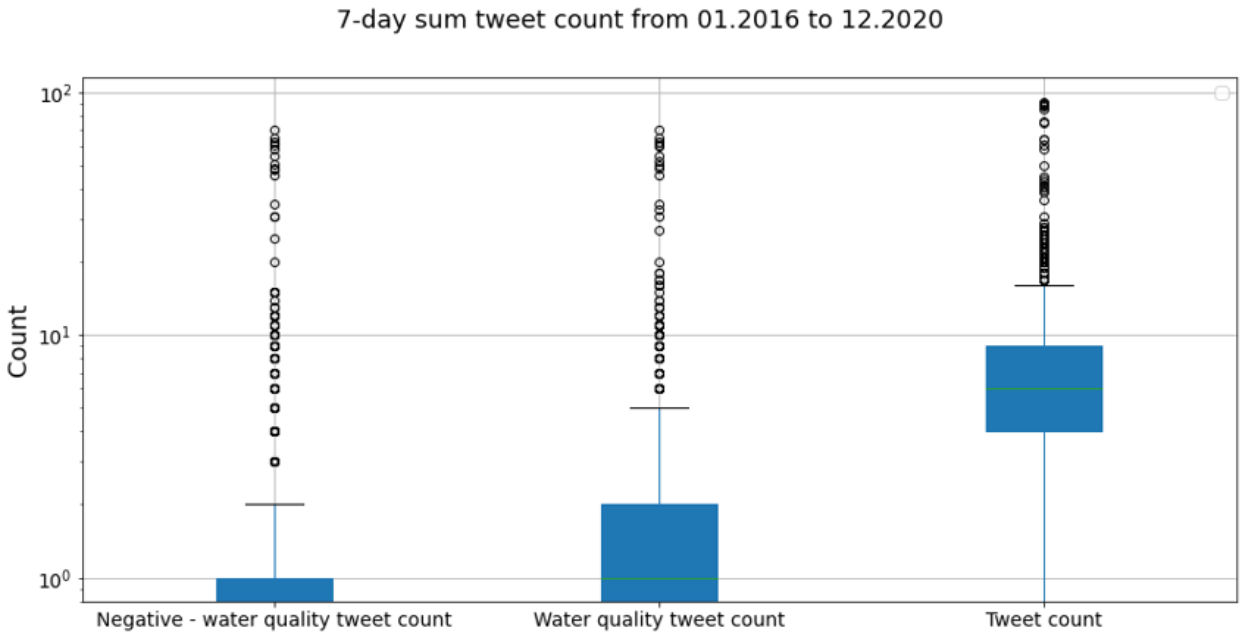


Figure 5. Boxplots of tweet counts

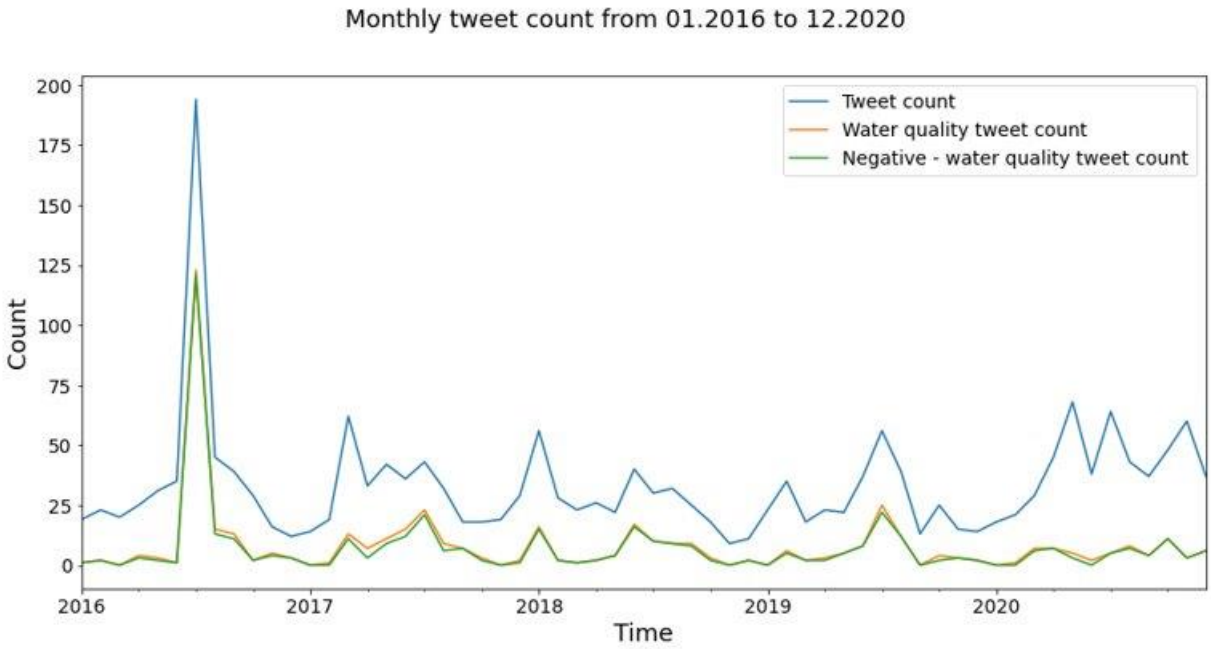


Figure 6. Monthly sums of tweet counts over the period from 2016 – 2020

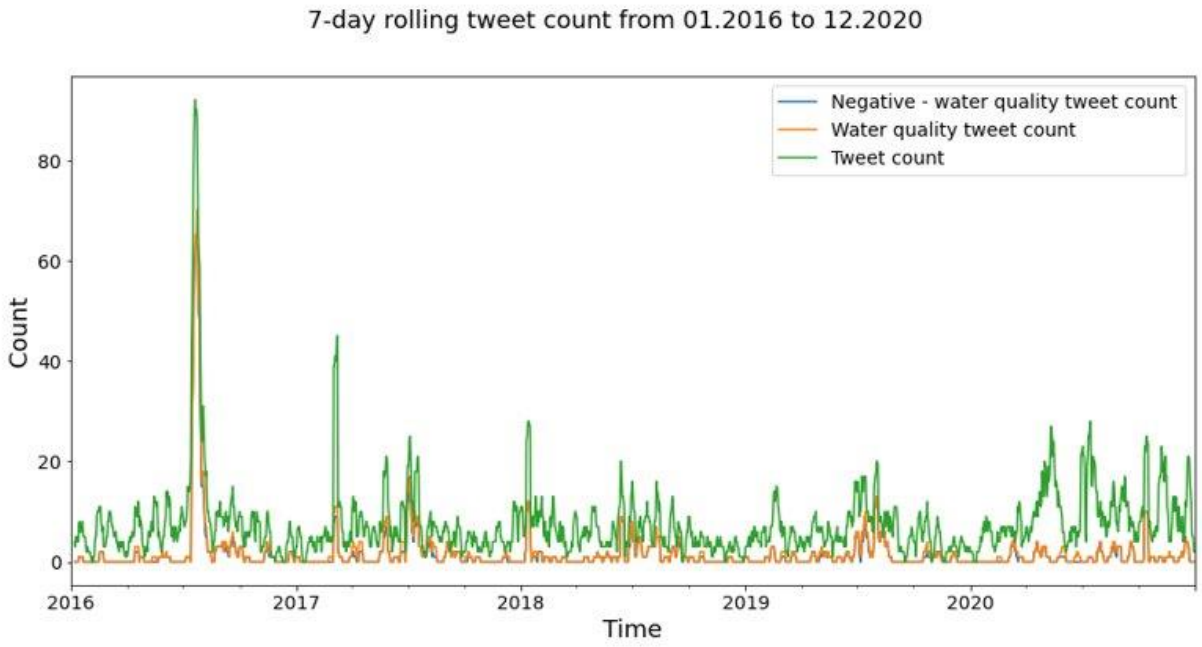


Figure 7. 7-day rolling sums of tweet counts over the period from 2016 – 2020

7.1.3. Time-series

To account for potential seasonality and autocorrelation in the 7-day sum and monthly tweet counts, we used ACF plots to assess temporal patterns. Figures 7 and 8 show that monthly data of both tweet counts and negative-water quality tweet counts are unseasonal and not autocorrelated (with AR1 of 0.2 and the autocorrelations within the confidence intervals of 95% - blue zone). On the other hand, the 7 day rolling sum of tweet counts did exhibit autocorrelation with an AR1 value of nearly 1.0 and ranges of about 16 days for tweet count sum and 14 days for negative-water quality tweet counts (Figures 9 and 10). In the next part of the project, we analyze the correlation between water quality data and tweet data (particularly 7-day rolling negative-water quality data). However, as the water quality measurements are temporally scattered with a frequency of about once to twice a month for most of the studied variables, the autocorrelation issue of the tweet data will not affect the analyses. Figures 9 and 10 also show that there is no seasonality in the 7-day rolling tweet data, therefore, will not be factored into further analyses.

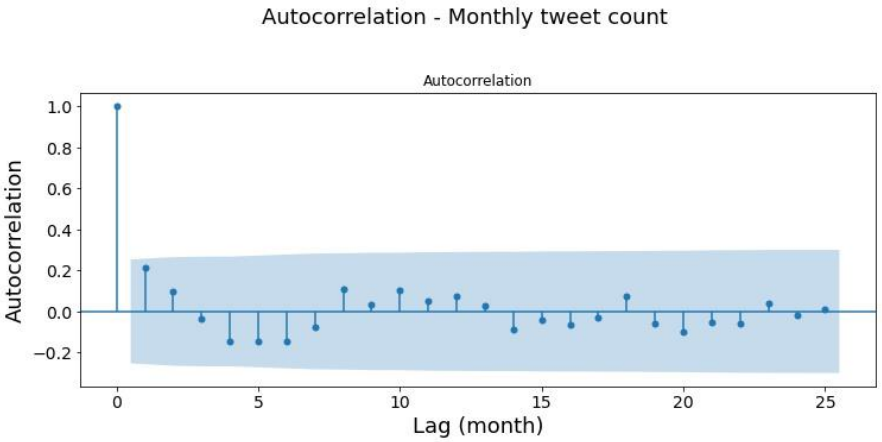


Figure 8. ACF for Monthly sums of tweet counts. The shaded region represents the boundaries of 95% confidence interval. Autocorrelations lying outside the shaded region are significant.

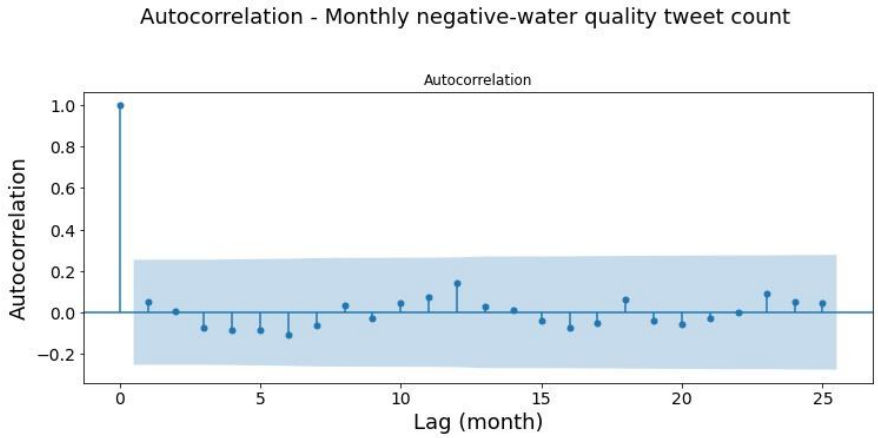


Figure 9. ACF for monthly sums of negative water quality tweet counts. The shaded region represents the boundaries of 95% confidence interval. Autocorrelations lying outside the shaded region are significant.

7.2. Descriptive statistics of water quality variables

We chose 4 perceivable water quality parameters as our independent variables from routine water quality data - Secchi disk depth, turbidity, chlorophyll a, and total suspended solids. We selected 8 parameters from HAB toxin advisory data - phytoplankton cell count (composite), phytoplankton cell count (surface), phytoplankton biovolume (composite), phytoplankton biovolume (surface), cyanobacteria cell count (composite), cyanobacteria cell count (surface), cyanobacteria biovolume (composite), cyanobacteria biovolume (surface). Below is a table providing descriptive statistics of each independent water quality variable after reducing spatial variability to one representative value for the lake for each day of sampling. The

distribution for most water quality variables is highly positively skewed with a long right tail (Table 4).

Table 4. Descriptive statistics of each water quality variable

Water quality parameters	**Minimum**	**25th Percentile**	**Median**	**75th Percentile**	**Maximum**	**Mean**	**Range**	**Standard deviation**	**No. of observations**
Secchi disk depth *(m)*	0.10	0.20	0.25	0.31	0.85	0.28	0.75	0.14	**38**
Turbidity *(NTU)*	6.40	35.77	55.48	79.25	169.56	60.16	163.16	33.06	**38**
Chlorophyll a *(µg/L)*	3.09	12.70	23.92	53.00	305.67	47.28	302.59	59.90	**37**
Total suspended solids *(mg/L)*	7.00	35.16	61.27	96.86	334.36	74.09	327.36	57.59	**63**
Phytoplankton cell count - Composite *(thousand cells/mL)*	0.00	28.88	61.21	289.73	45,202.32	1,517.86	45,202.32	6,551.60	**59**
Phytoplankton cell count – Surface *(thousand cells/mL)*	0.89	110.79	1,089.09	3,729.99	49,118.25	4,665.59	49,117.36	9,328.55	**53**
Phytoplankton biovolume - Composite	0.90	9.20	13.76	30.33	521.40	47.56	520.51	91.93	**41**

(million $\mu m^3/mL$)									
Phytoplankton biovolume - Surface *(million $\mu m^3/mL$)*	12.00	112.53	1,042.06	2,513.63	9,168.87	1,998.82	9,156.86	2,485.74	**29**
Cyanobacteria cell count - Composite *(thousand cells/mL)*	0.16	11.04	46.14	240.95	45,202.32	1,253.99	45,202.16	5916.08	**72**
Cyanobacteria cell count - Surface *(thousand cells/mL)*	0.04	57.90	726.34	3,852.80	103,588.36	6,035.32	103,588.32	15315.15	**61**
Cyanobacteria biovolume - Composite *(million $\mu m^3/mL$)*	0.01	1.66	5.09	21.17	505.02	34.01	505.01	83.37	**58**
Cyanobacteria biovolume - Surface *(million $\mu m^3/mL$)*	0.00	42.90	797.45	2,920.35	6,565.07	1,808.69	6,565.06	2,204.54	**37**

7.3. Text classification

As the validation results show, all four algorithms performed very well for classifier_water, with accuracy ranging from 84 – 90%, among which SGD was the most suitable one (accuracy of 90%). Meanwhile, algorithms for classifier_sentiment produce accuracy at acceptable levels, ranging from 69-72% and logistic regression is selected with 72% accuracy.

Table 5. Performance evaluation metrics – Valuation results of classifier_water

Model	**Class**	**Precision**	**Recall**	**F1-score**	**Accuracy**
Naïve Bayes	Water	0.66	0.74	0.7	0.86
	Non-water	0.93	0.89	0.91	
Logistic Regression	Water	0.92	0.56	0.69	0.89
	Non-water	0.89	0.99	0.94	
Bernoulli NB	Water	0.81	0.33	0.90	0.84
	Non-water	0.84	0.98	0.47	
SGD	**Water**	**0.83**	**0.69**	**0.75**	**0.90**
	Non-water	**0.92**	**0.96**	**0.94**	

Table 6. Performance evaluation metrics – Valuation results of classifier_sentiment

Model	**Class**	**Precision**	**Recall**	**F1-score**	**Accuracy**
Naïve Bayes	Negative	0.71	0.83	0.77	0.71
	Non-negative	0.70	0.54	0.61	
Logistic Regression	Negative	**0.75**	**0.77**	**0.76**	**0.72**
	Non-negative	**0.68**	**0.65**	**0.66**	
Bernoulli NB	Negative	0.68	0.92	0.78	0.70
	Non-negative	0.79	0.41	0.54	
SGD	Negative	0.74	0.73	0.73	0.69
	Non-negative	0.64	0.65	0.64	

The low accuracy of classifier_sentiment, particularly lower than that of classifier_water, can be explained by four main reasons. First, water quality topics, especially in the Utah Lake area, are typically specific and revolve around major themes such as algae bloom (Figure 11a), while in sentimental analyses, topics are much more diverse. For this reason, studies on sentimental analysis can use up to millions of tweets to train the model for example in (Go, 2009). However, to ensure a representative model of the Utah lake area, we deliberately maintained a high portion of tweets from Utah Lake in the input dataset (low number of observations). The second reason is that as tweets were manually labeled, the classes can vary by individual's understanding and judgement. Thirdly, many tweets have metaphorical and sarcastic expressions that are difficult for both human and machine to learn and classify the opinion underneath. Below is an example of sarcasm from the collected dataset. We identified it as about water and having negative sentiment, however, the appearance of words without context, makes it very challenging to classify the tweet.

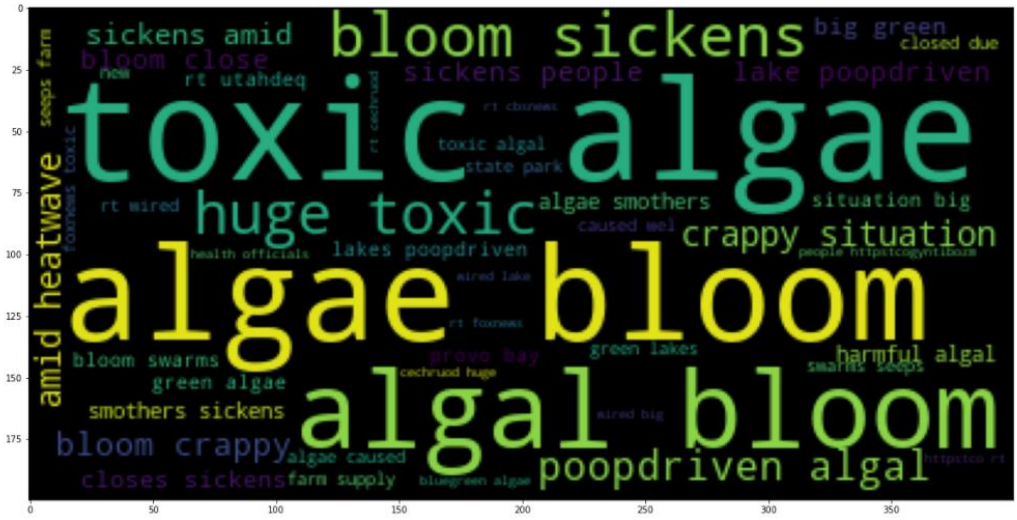
"If you've ever swam in Utah Lake, you don't need to worry about what's in the vaccine."

Thirdly, text can contain multiple aspects where each of them has a different sentiment. Each tweet may contain both positive and negative opinions at different intensity. we acknowledge that by simplifying the method and classifying the tweet into two sentimental groups, we may omit the sentiment in between. An example of words with different aspects is as follows:

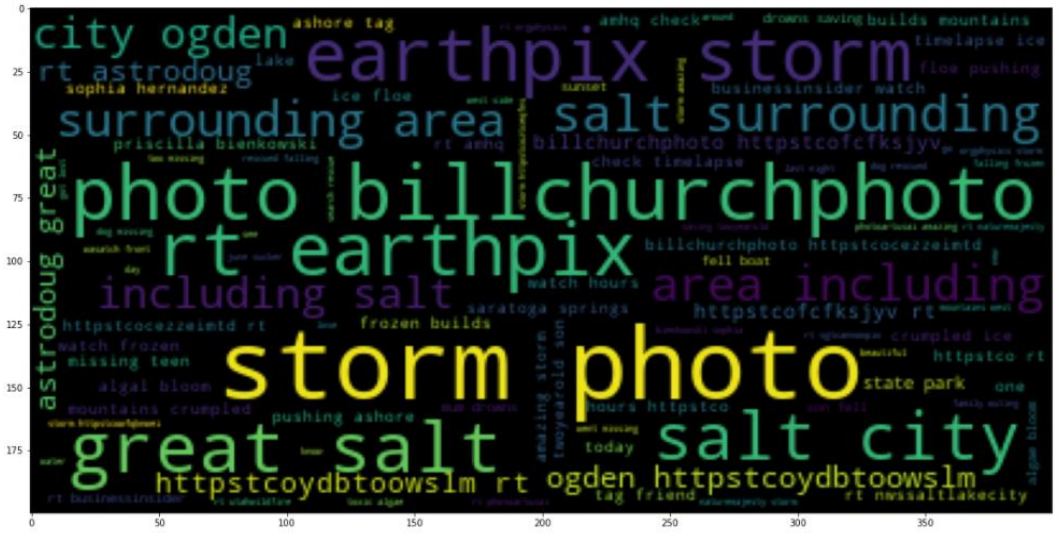
"@deqdonna That is our hope. Obviously, we would love to have a year without a bloom, but @UtahDEQ and Utah Lake Commission will be ready."

Fourthly, negation words play an important role in defining the sentiment of a document (for example: "bad" = "not good"). However, due to the characteristics of BOWs, negation words are analyzed as an individual word but not in association with specific words. Therefore, the accuracy of the text classification model, particularly for sentimental analysis can be improved by 1) increasing the size of the input dataset, 2) using more advanced NLP techniques which consider association and order of words and 3) classifying the tweets into more than two classes.

After classifying the tweets, we used word clouds visualization to identify the most common words appearing in each classes in the tweets collected from 2016 -2020 with keyword “Utah Lake”. As can be seen in Figures 11 a and b, tweets about water quality consistently include words such as “toxic algae”, “algae bloom”. Moreover, in 693 tweets about water quality, 347 tweets contain words such as “algae” or “algal”. This suggests that algae bloom is a most concerned topic related to water quality in Utah Lake.



a. Water quality tweets



a. Non-water quality tweets

Figure 10. Most popular words for water vs non-water quality tweets

7.4. Spatial interpolation

Due to the number of parameters, only interpolation results for parameters that are proven to be significant by count regression are displayed in appendix A. The displayed parameters include turbidity, chlorophyll a, phytoplankton cell count at lake surface, phytoplankton biovolume at lake surface, cyanobacteria cell count at lake surface, and cyanobacteria biovolume at lake surface. To ensure that enough information can be included in the maps, we used stratified sampling to select 6 samples for each parameter.

According to the interpolation results (Appendix A), the magnitude distribution shows different patterns among the water quality parameters. For turbidity, the distribution tends to concentrate at the southern part of Utah Lake (Goshen Bay). For chlorophyll a, the distribution tends to concentrate at the east part of Utah Lake, (Provo Bay). For total biovolume density, the distribution tends to concentrate at the south-eastern part of Utah Lake (Lincoln Point). These results indicate that there may be multiple sources of pollution that are causing HAB events, and the sources may have different characteristics that lead to HAB events. The higher concentration of contaminants on the southern and eastern part of the lake suggests that the pollution could be because of increased human population in the Utah Valley. The maps also show that the HAB events keep happening in Utah Lake from 2016 to 2020 with a seasonal pattern.

The quality and accuracy of interpolation depends on the amount of data points used, the more points used for interpolation, the more accurate the result can be. In our study the available number of data points on a day vary from 1 to 7, so some interpolation results may seem unnatural. But given the total amount of interpolation results, the water quality data for modelling was reliable.

7.5. Count regression analyses

Explanatory variables that were statistically significant in predicting tweet counts at a significance level of greater than 95% are marked with asterisk symbol(s) in table 7 and table 8. We can see that Secchi disk depth, turbidity, chlorophyll a, total phytoplankton cell count in surface samples, phytoplankton biovolume in surface samples, cyanobacteria cell count in surface samples and scyanobacteria biovolume in surface samples indicated a significant

relationship with tweet count variable(s). Incidence rate ratios from table 8 are used to quantify the percent change in negative tweets for each unit change in water quality. For a full summary of each negative binomial regression model, please refer to Appendix B.

Table 7. Negative binomial regression coefficients of negative sentiment tweet counts in response to each water quality parameter with their associated level of significance.

Water Quality Parameters	**Sum of negative WQ tweets - 7 days**		
	Before	**After**	**Before & after**
Secchi disk depth	-1.687	-3.963	-2.408
Turbidity	0.015	0.031***	0.023***
Chlorophyll a	0.003	0.012**	0.010**
Total suspended solids	-0.001	0.005	0.001
Phytoplankton cell count (Composite)	-0.001	-0.001***	-0.001**
Phytoplankton cell count (Surface)	1.53E-08	8.88E-08	5.99E-08
Phytoplankton biovolume (Composite)	2.39E-10	4.39E-09	3.67E-09
Phytoplankton biovolume (Surface)	2.39E-10**	3.55E-10***	3.01E-10***
Cyanobacteria cell count (Composite)	5.31E-09	-0.001	-0.001
Cyanobacteria cell count (Surface)	4.73E-08	6.68E-08***	3.71E-08*
Cyanobacteria biovolume (Composite)	-0.001	2.78E-09	2.64E-09
Cyanobacteria biovolume (Surface)	1.77E-10**	3.44E-10***	2.79E-10***
NOTE: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$			

Table 8. Incidence rate ratios of tweet counts (exponentiated regression coefficients) in response to each water quality parameter with their level of significance.

Water Quality Parameters	**Sum of negative WQ tweets - 7 days**		
	Before	**After**	**Before & after**
Secchi disk depth	0.185	0.019	0.090
Turbidity	1.015	1.031***	1.023***
Chlorophyll a	1.003	1.012**	1.010**
Total suspended solids	0.999	1.005	1.001
Phytoplankton cell count (Composite)	0.999	0.999	0.999
Phytoplankton cell count (Surface)	1+1.53E-08	1+8.88E-08***	1+5.99E-08**
Phytoplankton biovolume (Composite)	1+2.39E-10	1+4.39E-09	1+3.67E-09
Phytoplankton biovolume (Surface)	1+2.39E-10**	1+3.55E-10***	1+3.01E-10***
Cyanobacteria cell count (Composite)	1+5.31E-09	0.999	0.999
Cyanobacteria cell count (Surface)	1+4.73E-08	1+6.68E-08***	1+3.71E-08*
Cyanobacteria biovolume (Composite)	0.999	1+2.78E-09	1+2.64E-09
Cyanobacteria biovolume (Surface)	1+1.77E-10**	1+3.44E-10***	1+2.79E-10***
NOTE: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$			

7.5.1. Turbidity

Negative tweet counts pertaining to Utah Lake's water quality were found to be significantly and positively related to the turbidity of Utah Lake for 7-days-after and 7-days-before-after time periods. Using the lowest deviance method, we determined that the model predicting tweet counts 7 days after water sampling performed the best. Based on the negative binomial regression model, there was a 3.15% increase of tweet counts for one NTU increase in turbidity. The Chi-square goodness-of-fit test (Pearson χ^2 = 25.91, df = 36, p=0.89) indicated that the model's expected values were not significantly different form the observed values suggesting that the model performed reasonably well. No autocorrelation (p<0.05) in the tweet counts was observed (Figure 17).

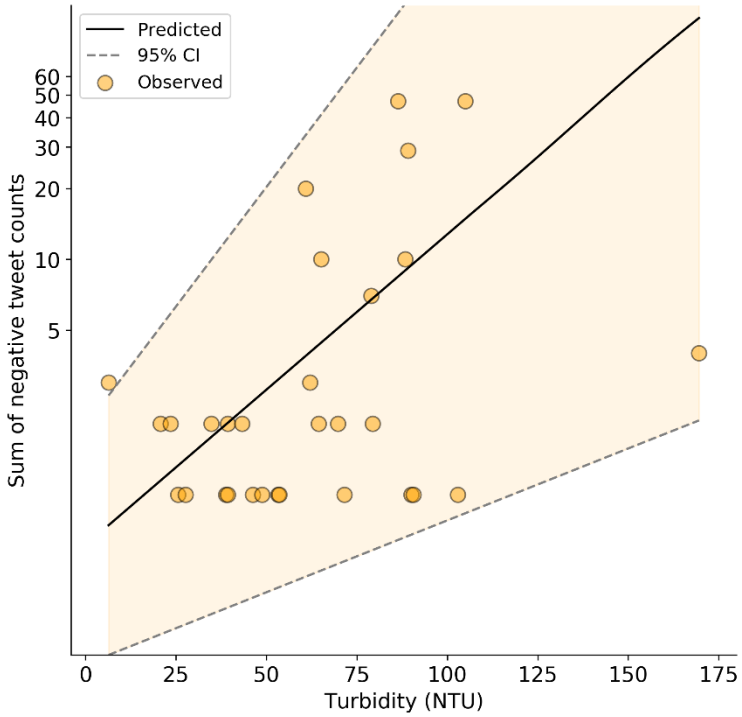


Figure 11. Observed and predicted sum of negative tweet counts pertaining to water quality 7 days after each turbidity measurement (n=38)

7.5.2. Chlorophyll-a

Chlorophyll a was found to be a significant positive predictor of the rolling sum of negative tweets for 7-days-after and 7-days-before-and-after time periods. The model with 7-days-after tweet counts had lowest deviance among the others. For one µg/L increase of chlorophyll concentration, there was a 1.2% increase in average tweet counts in the next 7 days after the day on which chlorophyll concentration was measured. However, Chi-square goodness of fit test revealed that the model did not fit well with significant differences in observed and expected tweet counts (Pearson χ^2 = 81.75, df = 35, p<0.001). There was no autocorrelation (p<0.05) in the tweet counts as well (Figure 17).

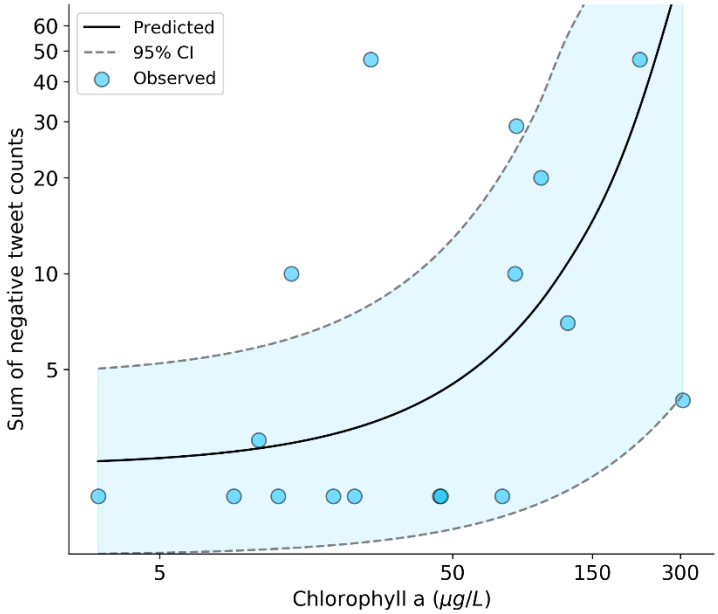


Figure 12. Observed and predicted sum of negative tweet counts pertaining to water quality 7 days after each chlorophyll a measurement (n=37)

7.5.3. Phytoplankton cell count – Surface

Tweet counts for both the 7 days time periods (i.e. 7 days after and 7 days before & after the day of sampling for total phytoplankton cell count) had a significant positive association with phytoplankton cell count measurements. Using deviance, we identified the model with tweet counts in 7-days-before-and-after period performed the best. According

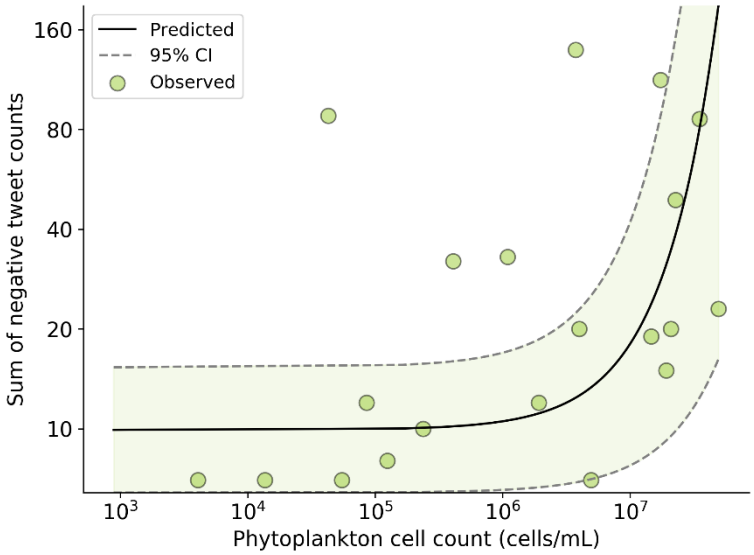


Figure 13. Observed and predicted sum of negative tweet counts pertaining to water quality 7 days before and after each phytoplankton cell count measurement at lake surface (n=53)

to the count regression model, there was a $5.9x10^{-6}$ percent increase in tweet counts on average for every 1 cells/mL increase in phytoplankton cell count. In other words, every 1 million cells/mL increase in cell count was associated to an increase in negative tweets by 5.9%. However, Chi-square goodness of fit test revealed that the model did not fit well (Pearson $\chi^2 = 97.64$, df = 51, $p<0.001$). Both the 7-days models had significant differences between the observed and expected tweet counts. Slight autocorrelation was observed in the tweet counts until lag 3 ($p<0.05$) for the 7-days-before-after model (Figure 17).

7.5.4. Phytoplankton biovolume – Surface

Phytoplankton biovolume at lake surface was a significant positive predictor of negative tweet counts for all the time periods. Using model deviance, the 7-days-after model was selected as the model with the best fit. The percentage change in tweet counts associated to one $\mu m^3/mL$ increase in phytoplankton biovolume at the surface was 3.5 X 10^{-8} %. In other words, a 1 million unit increase in biovolume translated to a 0.035% increase in negative tweet counts. Using the chi-square goodness of fit test, we found that there was a strong evidence that the model was a good fit (Pearson $\chi^2 = 29.38$, df = 27, p = 0.34). Both the 7-days models performed well with no significant differences between the observed and the expected tweet counts. The 7-days-before tweet counts were also significantly associated with the phytoplankton biovolume and passed the goodness of fit test. No autocorrelation ($p<0.05$) of the negative tweet counts was observed (Figure 17).

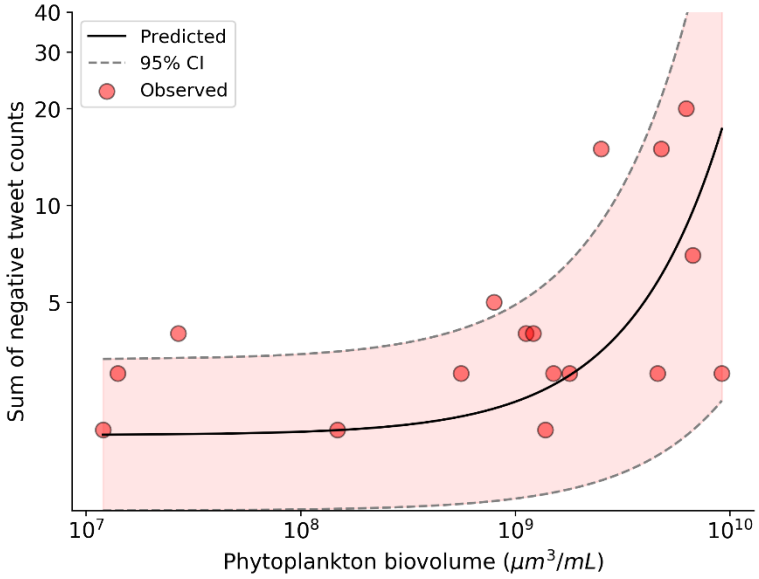


Figure 14. Observed and predicted sum of negative tweet counts pertaining to water quality 7 days after each phytoplankton biovolume measurement at lake surface (n=29)

7.5.5. Cyanobacteria cell count – Surface

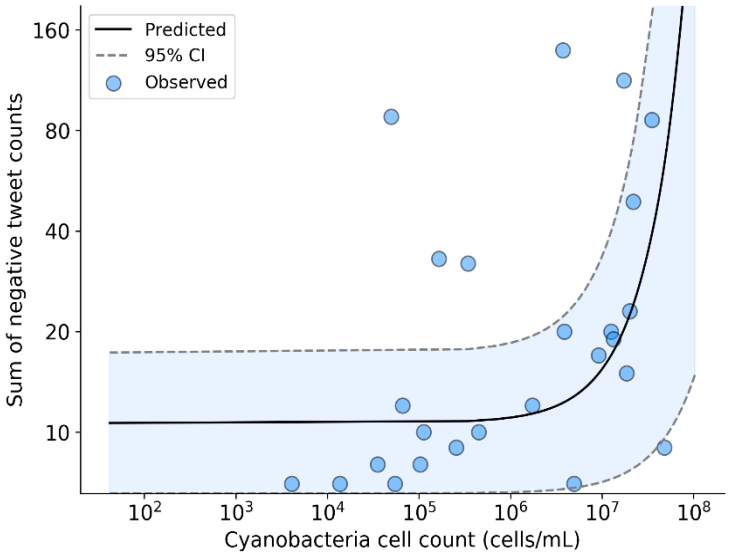


Figure 15. Observed and predicted sum of negative tweet counts pertaining to water quality 7 days before and after each cyanobacteria cell count measurement at lake surface (n=61)

There was a strong evidence that cyanobacteria cell count at lake surface was a significant predictor of the rolling sum of tweets for all the time periods i.e., 7 days after and 7 days before and after the day of water sampling for cyanobacteria cell count. The model with 7-days-before-and-after tweet counts had the lowest deviance among the others. For one cells/mL increase of cyanobacteria cell count, there was a $3.7x10^{-6}$ % increase in mean tweet counts. In other words, every 1 million cells/mL increase in cell count was associated to an increase in negative tweets by 3.7%. Chi-square goodness of fit test supported the model performance with no significant differences in observed and expected tweet counts (Pearson χ^2 = 66.18, df = 59, p = 0.24). All other models were also a good fit to our data. Slight autocorrelation was observed in the tweet counts until lag 3 ($p<0.05$) for the 7-days-before-after model (Figure 17).

7.5.6. Cyanobacteria biovolume – Surface

Based on our count regression models, there was a strong evidence to suggest that cyanobacteria biovolume at lake surface was a significant predictor of tweet counts in both 7 days periods. The model with 7 days after tweet counts had the lowest deviance. The percentage change in tweet counts associated to one $\mu m^3/mL$ increase in cyanobacteria biovolume at the surface was $1.8x10^{-8}$ %. In other words, a 1 million unit increase in biovolume translated to a 0.018% increase in negative tweet counts. Using the chi-square goodness of fit test, we found that there was no significant difference between the observed and the expected tweet counts (Pearson $\chi^2 = 37.29$, df = 35, p = 0.36). Both the 7-days models performed well with our data. The 7-days-before tweet counts were also significantly associated with the cyanobacteria biovolume and passed the goodness-of-fit test. There was no autocorrelation in the negative tweet counts (Figure 17).

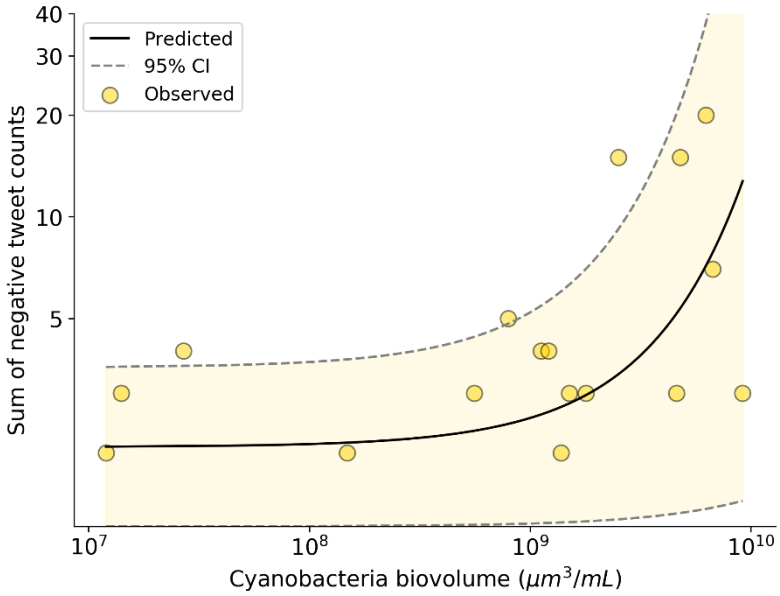


Figure 16. Observed and predicted sum of negative tweet counts pertaining to water quality 7 days after each cyanobacteria biovolume measurement at lake surface (n=37)

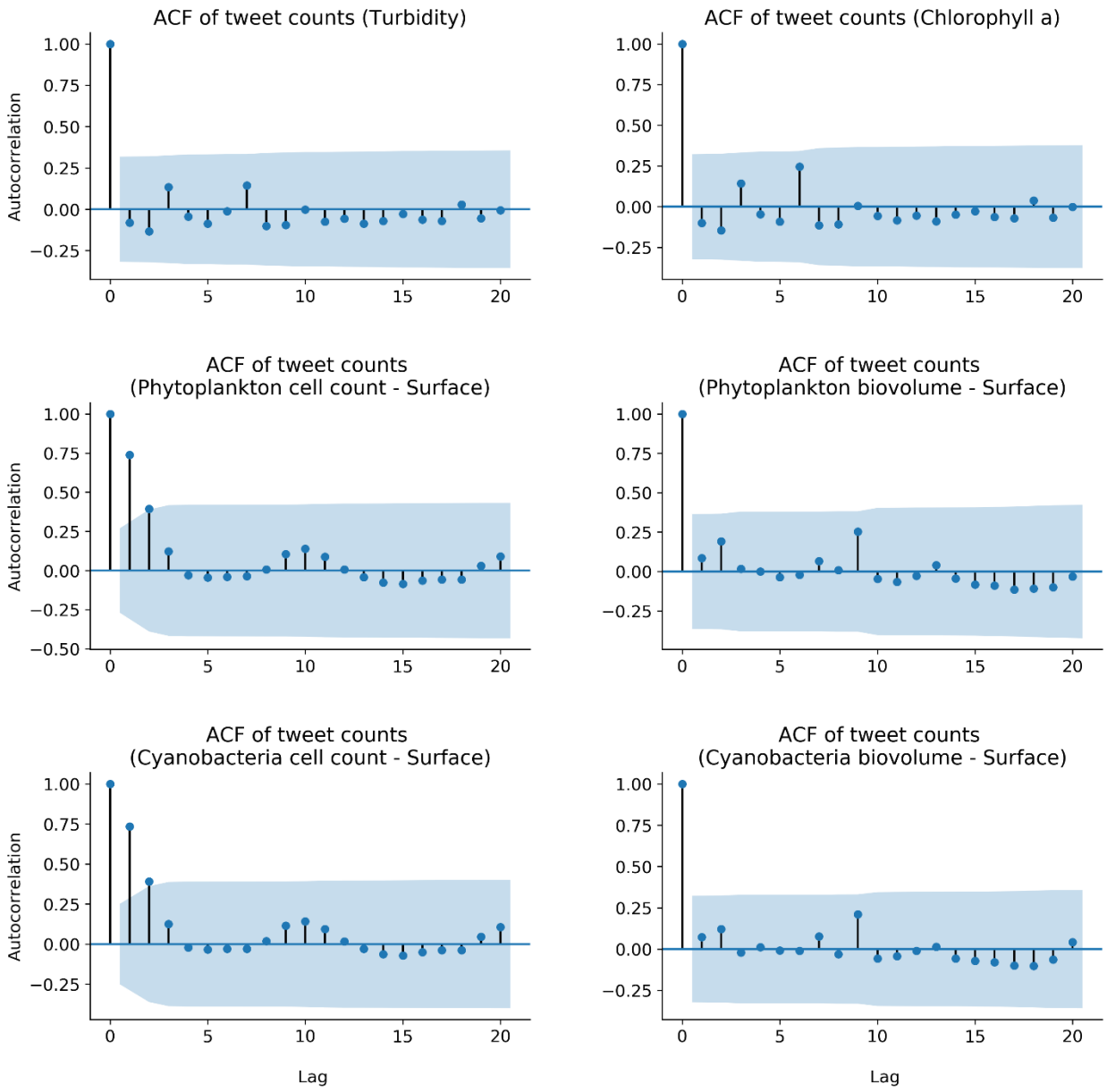


Figure 17. Plots showing autocorrelation function of negative tweet counts for each water quality parameter's best regression model based on the lowest deviance. The shaded region represents the boundaries of 95% confidence interval. Autocorrelations lying outside the shaded region are significant.

8. DISCUSSION

The central goals of this study were to assess whether the number of tweets expressing negative sentiments about Utah Lake's water quality is related to the field measurements of perceivable water quality parameters of Utah Lake, and to quantify the extent to which the number of tweets expressing negative sentiments about Utah Lake's water are associated to

perceivable water quality measurements measured by Utah DEQ in Utah Lake. We used water quality data provided by Tetra Tech and obtained historical tweets regarding Utah Lake using Twitter API for the years 2016 to 2020. We found that turbidity, phytoplankton biovolume at lake surface, cyanobacteria cell count at lake surface and cyanobacteria cell count at lake surface had significant predictive power for incidence of negative tweets with reasonably well model fitness. Other water quality parameters such as chlorophyll-a and phytoplankton cell count were also significant predictors but expected values of tweet counts were found to be significantly different from the observed values implying that the negative binomial regression models did not perform well in those cases. The direction of association between various water quality parameters and negative tweets was found to be consistently positive implying that worsening water quality conditions in the lake is more likely to increase the number tweets expressing negative sentiment towards the lake.

We used 3 different response variables with varying time windows relative to a water quality measurement to understand whether and how the public tweets in response to water quality in Utah Lake. Because water quality variables are measured at an inconsistent temporal frequency, we hypothesized that water quality measurements and the actual water quality in the lake might be independent from each other, meaning that a bad water quality event might start before it is measured. To test our assumption, we decided to use the 7-days-before tweet counts to determine whether negative tweets can be used as a predictive way to capture bad water quality before actual lake samples are measured. We also used the 7-days-after tweet counts because algal growth in aquatic environments is not symmetrically distributed around their peak time and can have a gradual increase over time with lagged ecological effects in response to growth stimuli (Duarte, 1990; Giudice et al., 2021). The 7-days-after tweet counts was also used as a response because it was observed that the number public tweets would increase after a government agency issues an HAB warning advisory. We wanted to see if the water quality tweets are in response to an already existing water quality issue. Lastly, we used the full-time window of sum of negative tweet counts 7 days before and after a water quality measurement to see if the results would differ from the previous two windows of tweet counts.

Based on our negative binomial regression models, the 7-days-after and the 7-days-before-after tweet counts were found to be significantly associated to turbidity, chlorophyll,

phytoplankton cell count and biovolume at surface and cyanobacteria cell count and biovolume at surface. Although the models predicting tweet counts in the span of 7-days-after or 7-days-before-and-after the day of water quality measurements had the highest number of significant results, total tweets in the 7-days-before timespan were also found to have significant relationship with two water quality parameters - phytoplankton biovolume and cyanobacteria biovolume. The 7-days-after tweet counts had the best fit for 4 out of 6 water quality parameters, while the 7-days-before-after tweet counts were best fit for 2 out of 6 water quality parameters. Phytoplankton and cyanobacteria biovolume were the only two water quality variables with significant associations to the 7-days-before tweet counts, thereby suggesting the potential of using tweets to predict the biovolume measurements in Utah Lake.

Surface measurements for algae population and concentration – phytoplankton cell count and biovolume, and cyanobacteria cell count and biovolume – were found to have a consistent positive relationship with the negative sentiment tweets. None of the regression models with composite measurements from HAB advisory data were found to be associated with the tweets. This finding supports that the tweets portray at least some degree of public perception of water quality because people can only notice water quality at the surface of the lake.

Previous studies that apply social media to environmental applications have used georeferenced tweets for spatial analysis. However, only 3% of global tweets are georeferenced with longitude and latitude (Hahmann et al., 2014) and inaccurate tweet locations in the context of an event or phenomenon of interest have contributed to inaccuracies in estimating the phenomenon (Brouwer et al., 2017; Li et al., 2017). In our study, less than 5% of the tweets related to Utah Lake's water quality were georeferenced and none of them were posted from the lake. This could be because people might not tweet while they are recreating in Utah Lake and might tweet later when they return to their homes. Therefore, we used Inverse Distance Weighted (IDW) interpolation to collapse the spatial variability of Utah Lake's water quality data and used higher percentile water quality values (75^{th} percentile for all parameters except Secchi disk depth where 25^{th} percentile was used because higher magnitude translates to clearer water) to obtain measurements representative of more perceivable water quality for the whole lake for each day of available data.

Like any count regression model, negative binomial regression assumes that each instance of the count variable is independent of each other. For our count regression models, generally autocorrelation of negative tweet counts in chosen time windows was not an issue (Figure 17). However, this might not certainly be true for future analyses because, in some cases based on our observation, the number of tweets on any given day can be influenced by the number of tweets on a previous day. In some cases, we observed that the number of people tweeting with negative sentiments towards Utah Lake's water quality increased after someone else's negative tweet. Therefore, for further studies, autocorrelation in tweet counts should be considered if count regression models are employed.

We used tweets without links as a way to obtain tweets posted by the general public. Our study included retweets into the total sum of tweets to capture the negative sentiment tweets posted by the public interacting with Utah Lake and not the tweets from the government or news agencies disseminating factual information. An alternative approach to ensure counting of tweets posted by the public could be to extract the bio of the user who tweeted. Using bag-of-words methodology, the user can be classified into one of the following categories – public, government, news agency. The tweets can then be filtered by the group of interest.

Consistent water quality measurement methodologies can be helpful for developing better statistical models. For instance, some phytoplankton and cyanobacteria measurements were from lake surface while others were a composite of different lake depths. Since surface measurements tend to have higher algae population, we separated the surface and composite samples. This resulted in a decrease of sample size for algae water quality measurements. Further studies can also consider incorporating citizen science data currently hosted on the Utah Water Watch platform. Utah DEQ can promote the use of citizen science for increasing their water quality sampling efforts. It can also install poster boards near the lake to promote Twitter as a platform where people can express their sentiments about Utah Lake with a specific hashtag. Such initiatives can increase the sample size for water quality measurements and the number of tweets that can be used for future statistical analyses.

We also considered using Facebook as potential sources of the public sentiment towards Utah Lake's water quality. However, while the Twitter API is structured around tweets being the

most basic form of data, the Graph API used by Facebook is structured around users and not the posts that users make. So we can easily obtain posts made by a specific user but we cannot filter all the Facebook posts on a specific topic. The challenges to obtaining Facebook posts directly without having to know a user who created that post made Facebook an inapplicable social media source to obtain public sentiments for Utah Lake.

Our study showcased that social media can be a promising, resource-efficient and real-time data source to identify worsening water quality conditions. Social media mining has also been used previously for various other environmental applications such as flood mapping and air pollution detection. A study on 2015 South Carolina floods used georeferenced tweets related to the flood event with other remote sensing and hydrologic inputs to develop a kernel-based flood inundation model (Li et al., 2017). Another study used tweets to create deterministic and probabilistic flood maps for the city of York with slight overestimation but close to real flooding extent (Brouwer et al.,2017). In another study researchers found that the frequency of air quality-related tweets posted by individuals were highly correlated to the Air Quality Index (Jiang et al., 2015). Such studies have shown great potential of using tweets as a "citizen sensor" tool to address environmental issues on ground.

However, none of the studies employing social media data for environmental applications have focused on water quality issues in surface waters. This study is a first of its kind to overcome some of the challenges related to applying public perception data to identify water quality issues by using text classification, spatial interpolation and count regression modeling to understand the association between the frequency of tweets expressing negative sentiment and field measurements of perceivable water quality in a lake, specifically Utah Lake in our case. This study provides important insights and highlights the potential of using social media for water quality of surface waters. This approach provides the opportunity to promptly identify possible water quality issues that can be perceived by the public and inform the government agencies which can then issue timely warning advisories to prevent the public health impacts. Therefore, our results lay the groundwork towards the ultimate goal of developing a tool that can help provide automatic near real-time warnings of surface water quality events based on information from social media networks.

9. CONCLUSION

Preserving surface water quality provides important environmental and socio-economic benefits. Surface water quality issues such as harmful algal blooms (HABs) poses risk to public health and can cause economic and biodiversity loss as well. Currently, the HABs are identified via field surveys by government entities and consultants. Such surveys can be resource intensive and can lag in identifying HAB events by a few days to weeks which can cause a delay in issuing warning advisories to the public. To cut costs and speed up detection of any possible water quality issue, we explored tweets expressing negative sentiment related to water quality as a quick and real-time data source to identify water quality issues in Utah lake (2016 – 2020) which is well-known for its algae bloom issues. We found that the negative tweet counts were significantly and positively associated to many of the perceivable water quality parameters we studied such as turbidity, chlorophyll a, phytoplankton cell count, phytoplankton biovolume, cyanobacteria cell count and cyanobacteria biovolume. Surface samples for algae concentration and population were also significantly related to the negative tweet counts while the composite samples were not significant, thereby supporting the idea that the public perceives and responds to the toxic water quality at surface levels. Our results show that social media can be a promising data source to obtain prompt and cautionary insights into any potential water quality issues such as HAB. These insights can be explored further to develop a tool that can issue timely and quick public warning compared to the traditional methods currently in place. Our work serves as a preliminary study that highlights the potential of using social media for water quality events. Using the water quality variables found significant to negative sentiment tweets, further studies should be conducted for applying social media data for water quality applications into a scalable alerting system.

10. REFERENCES

Atefeh, F. &. (2013). A survey of techniques for event detection in Twitter. *Computational Intelligence, Volume 0,* 133–164.

Berk, R., & MacDonald, J.M. (2008). Overdispersion and Poisson Regression. Journal of Quantitative Criminology, 24(3), 269-284. doi:10.1007/s10940-008-9048-4

Britannica, T. Editors of Encyclopaedia (Invalid Date). *Utah Lake. Encyclopedia Britannica.* https://www.britannica.com/place/Utah-Lake

Brouwer, T., Eilander, D., Loenen, A. V., Booij, M. J., Wijnberg, K. M., Verkade, J. S., & Wagemaker, J. (2017). Probabilistic flood extent estimates from social media flood observations. *Natural Hazards and Earth System Sciences*, 17: 735-747.

Cameron, A.C. & Trivedi, P.K. (2013). *Regression Analysis of Count Data: Second Edition*. Cambridge University Press.

Coxe, S., West, S.G., & Aiken, L.S. (2009). The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Statistical Developments and Applications*, 91(2). doi:10.1080/00223890802634175.

(CDC) Centers for Disease Control and Prevention. (2020). *Avoid Harmful Algal Blooms.* CDC website https://www.cdc.gov/habs/be-aware-habs.html

Date, S. (2019). Negative Binomial Regression: A Step by Step Guide. Towards Data Science website https://towardsdatascience.com/negative-binomial-regression-f99031bb25b4

Dang-Xuan, S. S. (2013). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems,*, 217-248. doi:DOI: 10.2753/MIS0742-1222290408

de Bruijn, J. d. (2019). A global database of historic and real-time flood events based on social media. *Scientific data*, 1-12. doi:doi:http://dx.doi.org.proxy.lib.duke.edu/10.1038/s41597-019-0326-9

Duarte, C. M. (1990). Time lags in algal growth: generality, causes, consequences. *Journal of Plankton Research*, 12(4): 873-883.

(EPA) Environmental Protection Agency. (August, 2016). *Indicators: Water Clarity*. EPA website https://www.epa.gov/national-aquatic-resource-surveys/indicators-water-clarity

(EPA) Environmental Protection Agency. (February, 2019). *Nutrient Pollution: The Issue*. EPA website https://www.epa.gov/nutrientpollution/issue

Florida Keys National Marine Sanctuary. (2011). Phytoplankton are Microscopic Marine Plants. NOAA website https://floridakeys.noaa.gov/plants/phyto.html

Fuhriman, D.K., Merritt, L.B., Miller, W., & Stock, H.S. (1981). Hydrology and water quality of Utah Lake. *Great Basin Naturalist Memoirs*, (5), 43-67. Retrieved April 21, 2021, from http://www.jstor.org/stable/23376546

Giudice, D. D., Fang, S., Scavia, S., Davis, T.W., Evans, M. A., & Obenour, D. R. (2021). Elucidating controls on cyanobacteria bloom timing and intensity via Bayesian mechanistic modeling. *Science of The Total Environment*, 755(1).

Go, A., Bhayani, R. & Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1(2009)*, p.12.

Hahmann, S., Purves, R. S., & Burghardt, D. (2014). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. Journal of Spatial Information Science 9(2014): 1 – 36.

Ho, J.C., Michalak, A.M., & Pahlevan, N. (2019). Widespread Global Increase in In Intense Lake Phytoplankton Blooms since the 1980s. *Nature, 574*, 667-670. https://doi.org/10.1038/s41586-019-1648-7

Li, Z., Wang, C., Emrich, C. T., & Guo, D. (2017). A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 South Carolina floods. *Cartography and Geographic Information Science*, 45(2).

Jakus, P., Kealy, M.J., Loomis, J., Nelson, N., Ostermiller, J., Stanger, C., & Stackelberg. (2013). Economic Benefits of Nutrient Reductions in Utah's Waters. State of Utah.

Jiang, W., Wang, Y., Tsou, M. H., & Fu, Xiaokang. (2015). Using social media to detect outdoor air pollution and monitor air quality index (AQI): A geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter). *PLoS ONE*, 10(10).

Jordan, S., Hovet, S., Fung, I. ..-H., Liang, H., Fu, K.-W., & Tse, Z. (2019). Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. *Data*.

Khanna, M. & Shortle, J. (2017). Preserving Water Quality: Challenges and Opportunities for Technological and Policy Innovations. *Choices, 32(4)*, 4 pages.

Middleton, S. E. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intellectual Systems*, 9-17.

Omnicore. (2020, 10 2). Retrieved from https://www.omnicoreagency.com/twitter-statistics/

Randall, M.C., Carling, G.T., Dastrup, D.B., Miller, T., Nelson, S.T., Rey, K.A., Hansen, N.C., Bickmore, B.R., & Aanderud, Z.T. (2019). Sediment potentially controls in-lake phosphorus cycling and harmful cyanobacteria in shallow, eutrophic Utah Lake. *PLoS One 14*(2): e0212238, https://doi.org/10.1371/journal.pone.0212238

Reza Zafarani, M. A. (2014). *Social media mining.* Cambridge University Press.

Takeshi Sakaki, M. O. (2010). Earthquake Shakes Twitter Users:Real-time Event Detection by Social Sensors. *: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010.*

Yufeng Yu, Y. Z. (2019). Applications of Social Media in Hydroinformatics: A Survey. Retrived from https://www.researchgate.net/publication/332960654_Applications_of_Social_Media_in_Hydroinformatics_A_Survey on 10.02.2020.

APPENDIX A – Spatial Interpolation

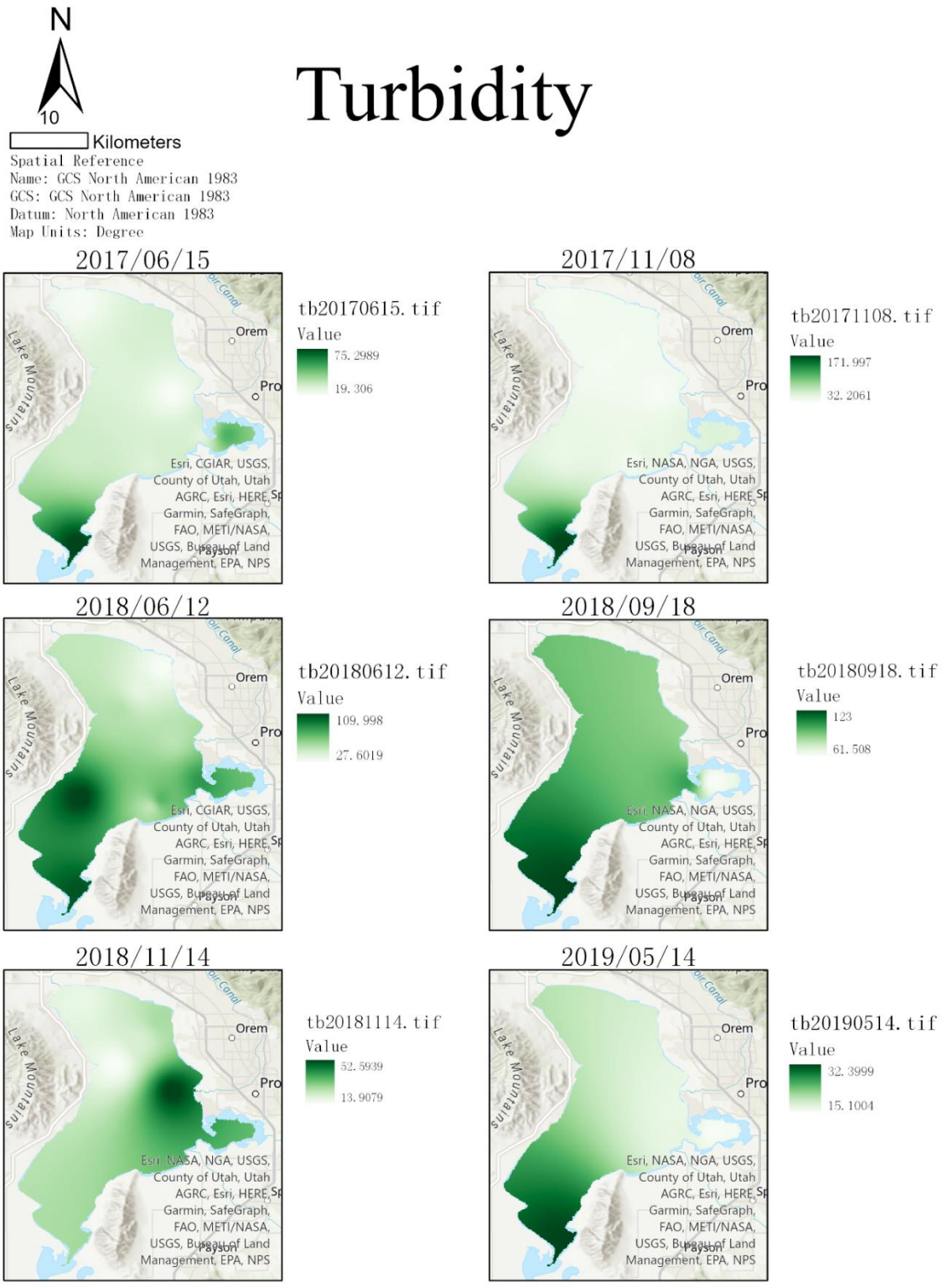


Figure 18. Maps showing the results of interpolating turbidity measurements for specific days for the whole lake. Darker color represents higher magnitude of turbidity.

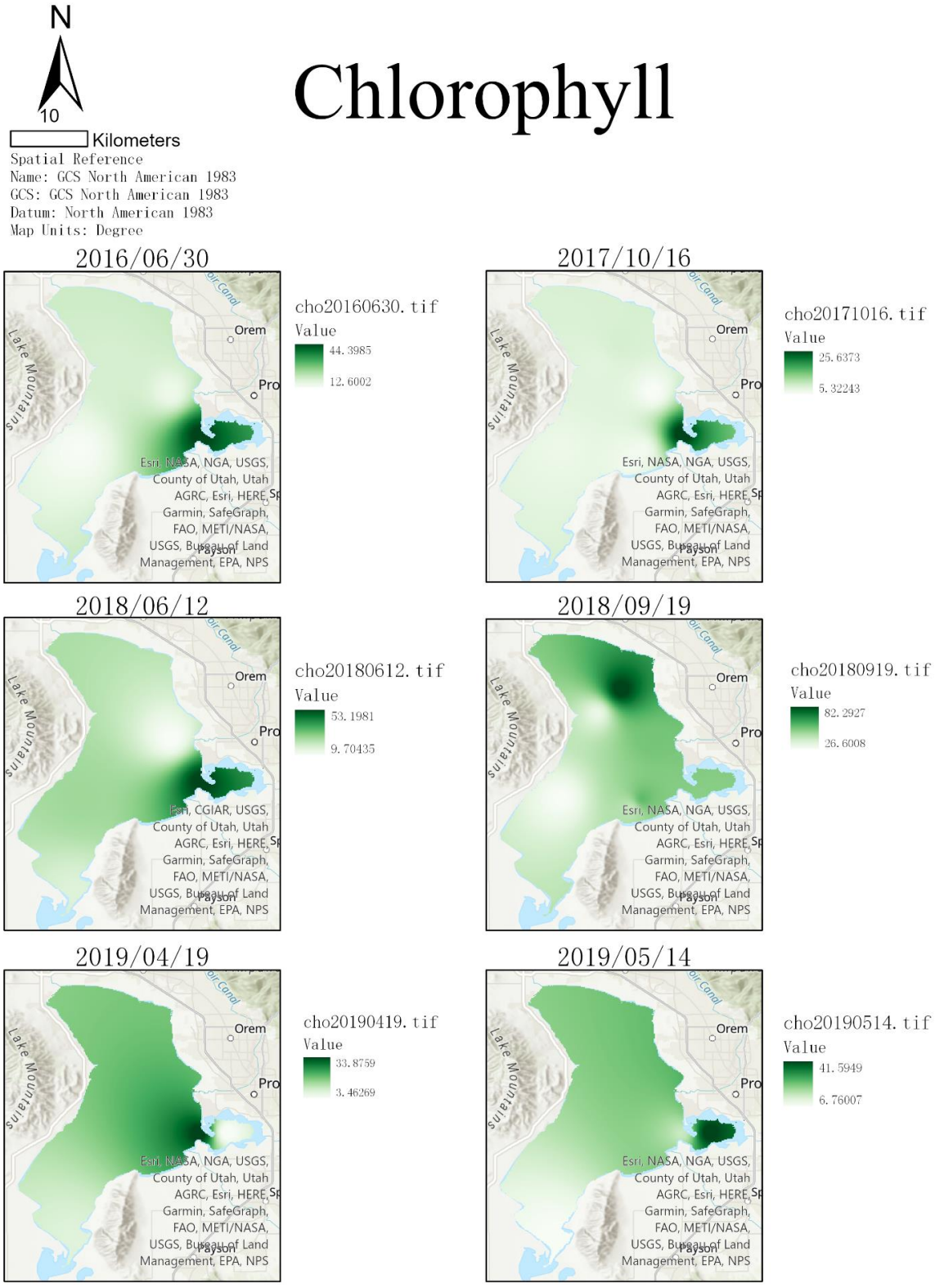


Figure 19. Maps showing the results of interpolating chlorophyll a measurements for specific days for the whole lake. Darker color represents higher magnitude of chlorophyll a.

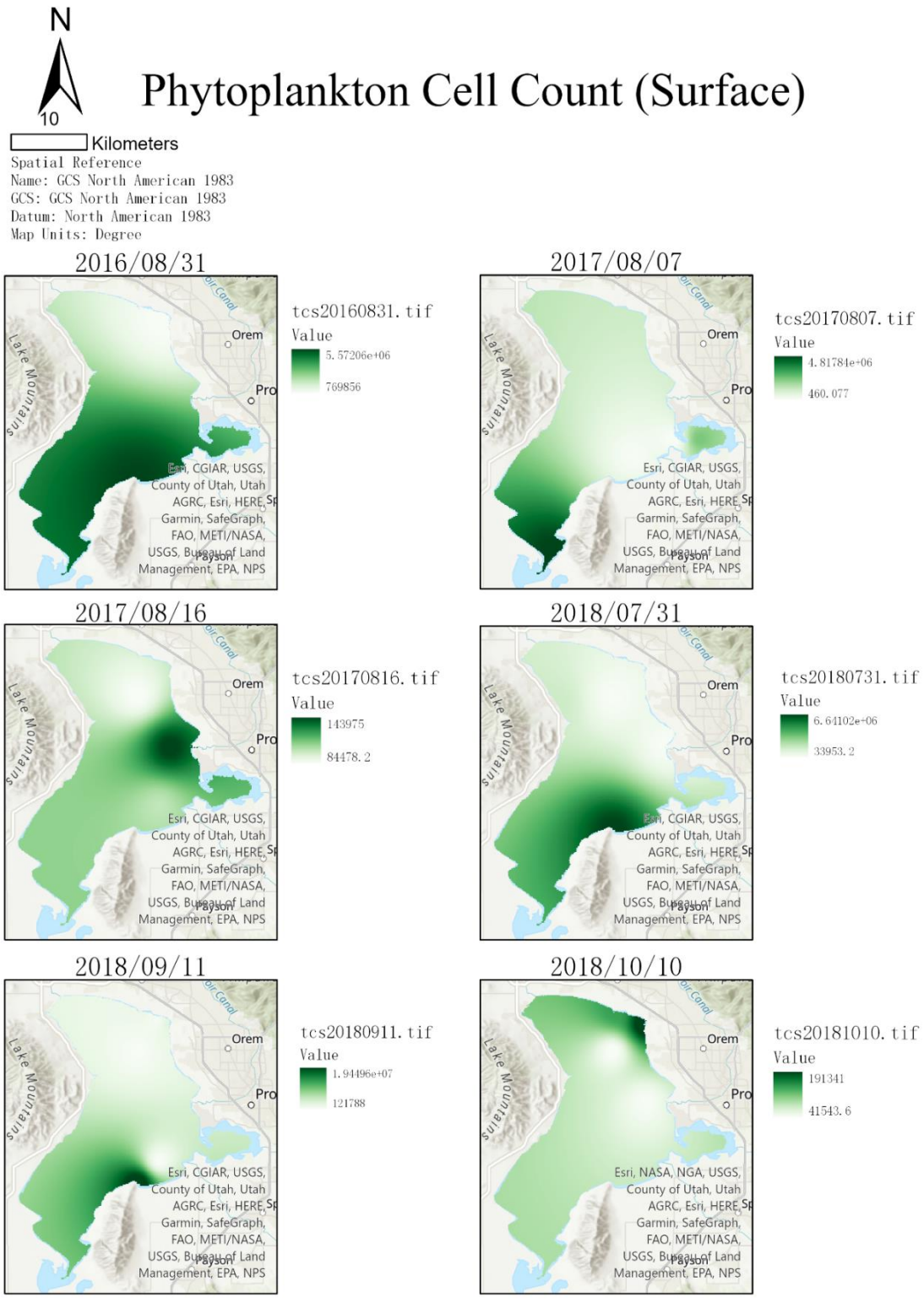


Figure 20. Maps showing the results of interpolating phytoplankton cell count measurements at lake surface for specific days for the whole lake. Darker color represents higher magnitude of phytoplankton cell count.

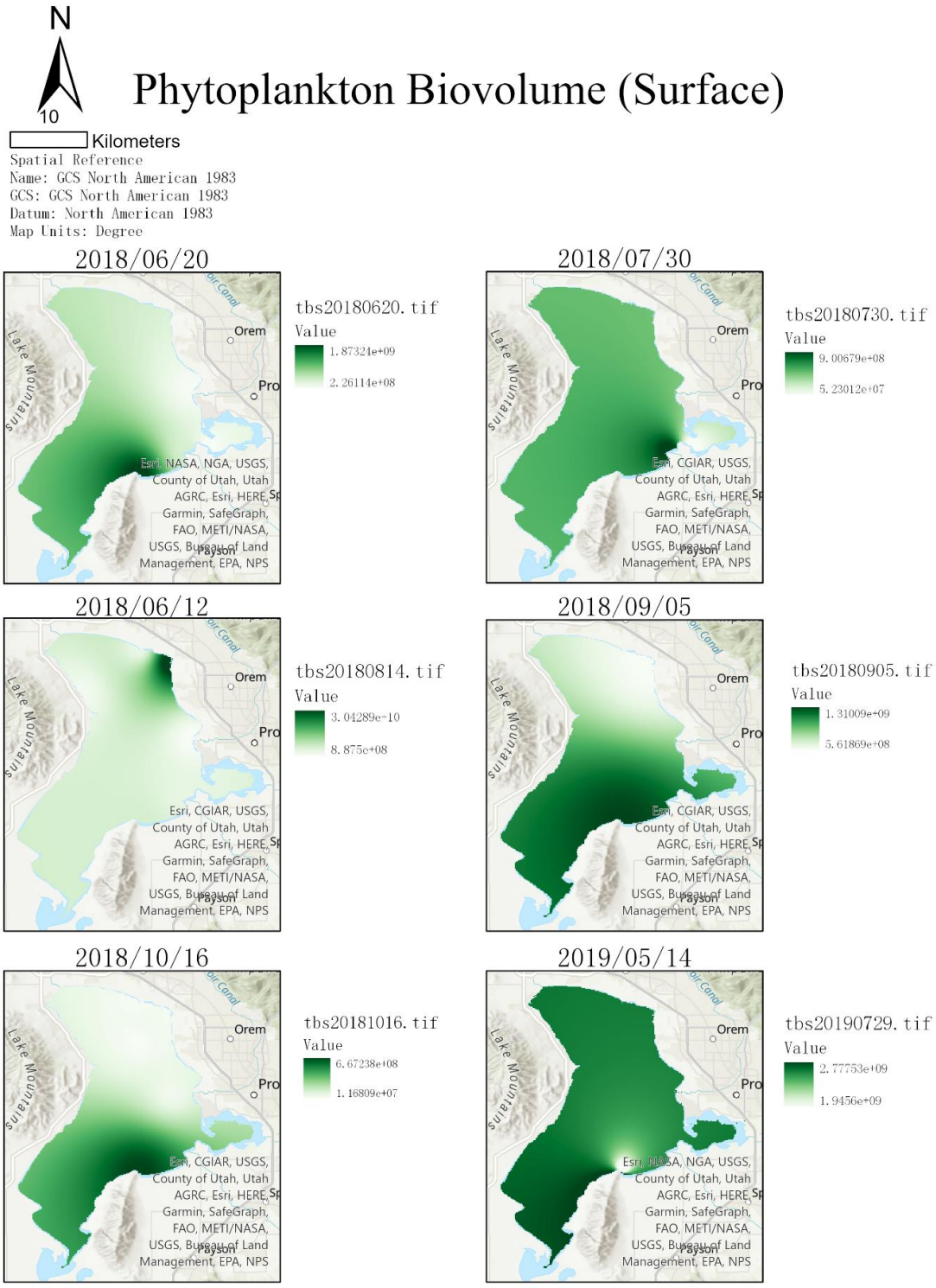


Figure 21. Maps showing the results of interpolating phytoplankton biovolume measurements at lake surface for specific days for the whole lake. Darker color represents higher magnitude of phytoplankton biovolume.

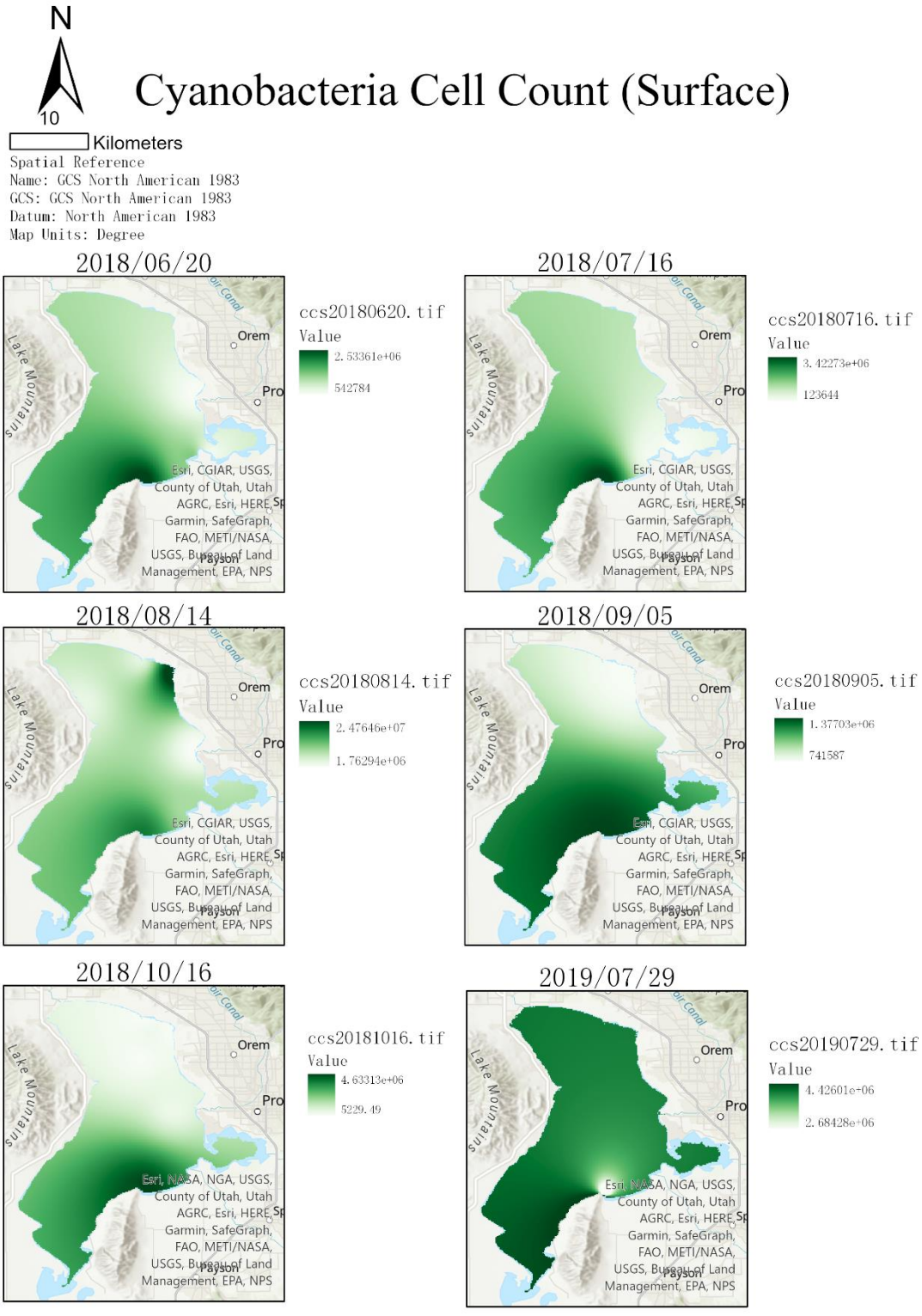


Figure 22. Maps showing the results of interpolating cyanobacteria cell count measurements at lake surface for specific days for the whole lake. Darker color represents higher magnitude of cyanobacteria cell count.

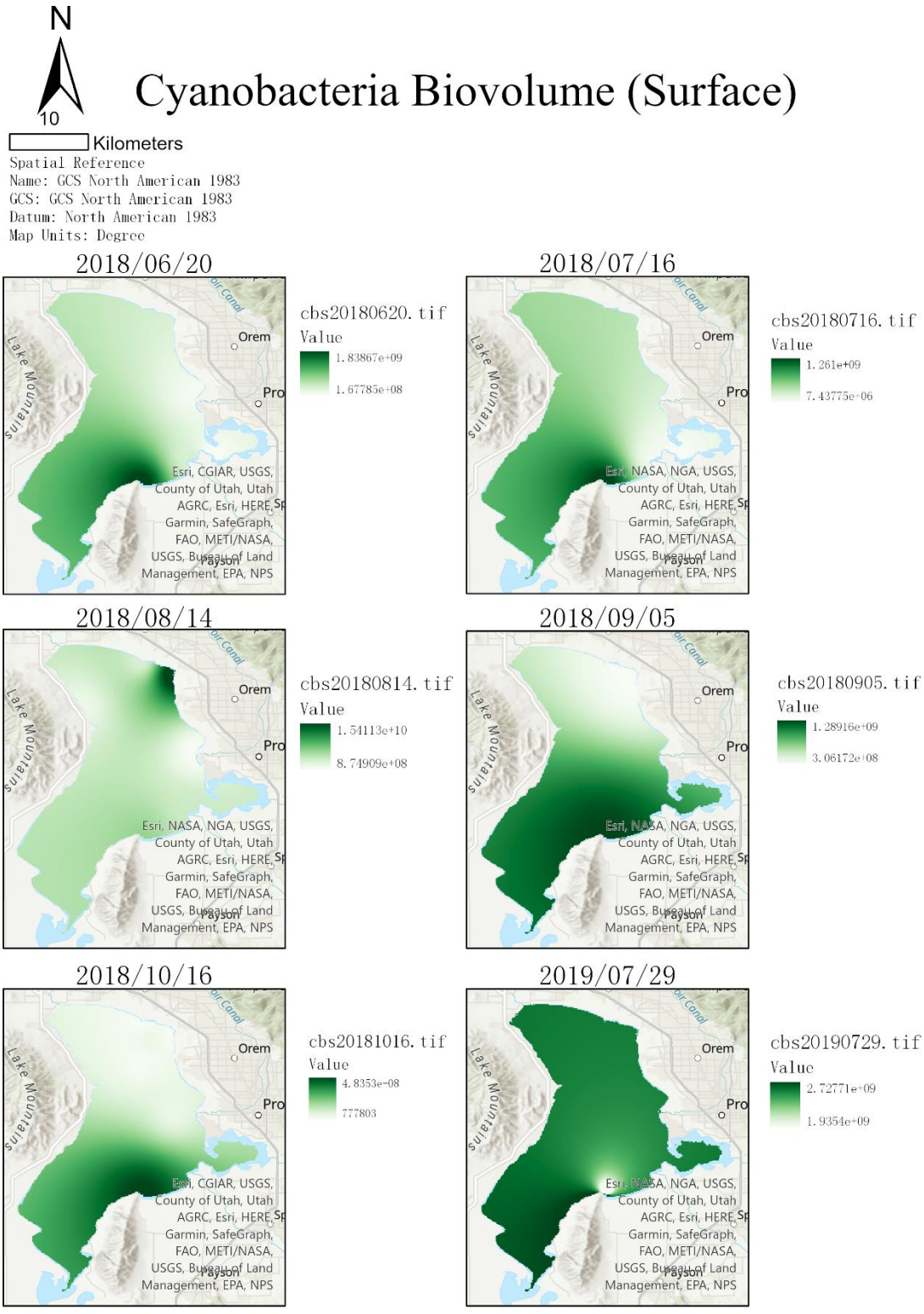


Figure 23. Maps showing the results of interpolating cyanobacteria biovolume measurements at lake surface for specific days for the whole lake. Darker color represents higher magnitude of cyanobacteria biovolume

APPENDIX B – Count regression results

Table 9. Model statistics for negative binomial regression with turbidity as explanatory variable and sum of negative sentiment tweet counts in the three time-windows as response variables

	7-days-before				*7-days-after*				*7-days-before-and-after*			
	coef	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**
Intercept	0.34	0.62	0.55	0.58	-0.49	0.59	-0.83	0.41	0.41	0.46	0.87	0.38
Turbidity	0.02	0.01	1.76	0.08	0.03	0.01	3.65	0.00	0.02	0.01	3.52	0.00
Observations	38				38				38			
Df	36				36				36			
Log-likelihood	-90.89				-94.53				-109.6			
Deviance	28.804				28.152				36.718			
Pearson Chi2	31.289				25.92				27.957			

Table 10. Model statistics for negative binomial regression with chlorophyll a as explanatory variable and sum of negative sentiment tweet counts in the three time-windows as response variables

	7-days-before				*7-days-after*				*7-days-before-and-after*			
	coef	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**
Intercept	1.20	0.42	2.89	0.00	0.91	0.33	2.77	0.01	1.41	0.27	5.30	0.00
Chlorophyll a	0.00	0.01	0.63	0.53	0.01	0.00	2.88	0.00	0.01	0.00	2.92	0.00
Observations	37				37				37			
Df	35				35				35			
Log-likelihood	-89.13				-92.8				-108.3			
Deviance	26.723				35.421				40.398			
Pearson Chi2	42.121				81.753				58.692			

Table 11. Model statistics for negative binomial regression with phytoplankton cell count at lake surface as explanatory variable and sum of negative sentiment tweet counts in the three time-windows as response variables

	7-days-before	*7-days-after*	*7-days-before-and-after*

	coef	std err	z	P>|z|	coef	std err	z	P>|z|	coef	std err	z	P>|z|
Intercept	1.89	0.33	5.76	0.00	1.47	0.21	6.99	0.00	2.29	0.22	10.33	0.00
Phytoplankton cell count	2E-08	3E-08	0.49	0.63	9E-08	2E-08	4.59	0.00	6E-08	2E-08	2.86	0.00
Observations	53				53				53			
Df	51				51				51			
Log-likelihood	-165.7				-154.6				-191.8			
Deviance	27.02				56.946				43.237			
Pearson Chi2	56.875				90.924				97.647			

Table 12. Model statistics for negative binomial regression with phytoplankton biovolume at lake surface as explanatory variable and sum of negative sentiment tweet counts in the three time-windows as response variables.

	7-days-before				*7-days-after*				*7-days-before-and-after*			
	coef	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**
Intercept	0.66	0.28	2.38	0.02	0.42	0.28	1.47	0.14	1.15	0.19	5.91	0.00
Phytoplankton biovolume	2E-10	8E-11	3.06	0.00	4E-10	8E-11	4.57	0.00	3E-10	5E-11	5.63	0.00
Observations	29				29				29			
Df	27				27				27			
Log-likelihood	-66.24				-66.82				-80.26			
Deviance	24.442				31.723				33.304			
Pearson Chi2	21.917				29.38				29.723			

Table 13. Model statistics for negative binomial regression with cyanobacteria cell count at lake surface as explanatory variable and sum of negative sentiment tweet counts in the three time-windows as response variables

	7-days-before				*7-days-after*				*7-days-before-and-after*			
	coef	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**
Intercept	1.87	0.26	7.19	0.00	2.23	0.27	8.13	0.00	2.65	0.23	11.59	0.00
Cyanobacteria cell count	5E-09	4E-08	0.12	0.90	-2E-08	5E-08	-0.35	0.73	-4E-09	4E-08	-0.11	0.91

Observations	72	72	72
Df	70	70	70
Log-likelihood	-209.1	-222	-260.2
Deviance	44.381	46.418	50.981
Pearson Chi2	72.86	70.931	71.527

Table 14. Model statistics for negative binomial regression with cyanobacteria biovolume at lake surface as explanatory variable and sum of negative sentiment tweet counts in the three time-windows as response variables

	7-days-after				***7-days-after***				***7-days-before-and-after***			
	coef	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**	**coef**	**std err**	**z**	**P>|z|**
Intercept	1.32	0.22	5.94	0.00	1.42	0.28	5.15	0.00	1.89	0.18	10.57	0.00
Cyanobacteria biovolume	2E-09	2E-09	0.84	0.40	3E-09	3E-09	0.93	0.35	3E-09	2E-09	1.37	0.17
Observations	58				58				58			
Df	56				56				56			
Log-likelihood	-143.5				-149.9				-176			
Deviance	51.393				40.26				61.113			
Pearson Chi2	71.129				59.268				66.85			