Cyanotracker 2.0: Improving Cyber-Physical Monitoring Infrastructure for Harmful Algal Bloom Detection

by

Keshav Raviprakash

(Under the Direction of Fred Beyette)

Abstract

Cyanobacterial harmful algal blooms (known as CyanoHABs) are a type of phytoplankton species that have increased in frequency around the world. Their proliferation in recent years has become a major issue worldwide due to their impact on the environment, socioeconomic activities, fauna, and humans. Many methods have been developed to monitor the occurrence of CyanoHAB events including satellite remote sensing, in-situ testing, and citizen monitoring applications. However, these methods fail to provide an accurate, real-time data for interested parties to verify the presence of CyanoHABs within water bodies. Thus, there is need to implement an early detection and warning response system for monitoring CyanoHABs. For this work, we propose that improvements can be made to the design of the original Cyanotracker system through development of an in-situ sensor system, social sensing approach, and use of satellite operations (communication and remote sensing) to monitor CyanoHAB events.

Index words: [Remote Sensing, Cyber-Physical Systems, Environmental Monitoring, Cyanobacterial Harmful Algal Blooms (CyanoHABs), Social Sensing, Data Mining]

Cyanotracker 2.0: Improving Cyber-Physical Monitoring Infrastructure for Harmful Algal Bloom Detection

by

Keshav Raviprakash

B.S.C.S.E, The University of Georgia, 2023

A Thesis Submitted to the Graduate Faculty of The
University of Georgia in Partial Fulfillment of the Requirements for the Degree

Master of Science

Athens, Georgia

2024

Cyanotracker 2.0: Improving Cyber-Physical Monitoring Infrastructure for Harmful Algal Bloom Detection

by

Keshav Raviprakash

Major Professor: Fred Beyette

Committee: Fred Beyette
Lakshmish Ramaswamy
Deepak R. Mishra

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2024

Dedication

I would like to first dedicate this work to my parents for the sacrifices and time that they time that have put into helping me succeed and supporting all of my endeavors. I would also like to dedicate this work to my brother for his help and support during the course of my time working on this research. Finally, I want to dedicate this to all of my family, friends, peers, and well-wishers for their support of me throughout my academic journey.

Acknowledgments

During the course of my time at the University of Georgia, I have had a number of people who have impacted my academic journey and made this work possible. First, I would like to thank my research advisor, Dr. Deepak Mishra, who took a chance on me and has helped provide me not only with great advice throughout the time I have known but also supported me in a number of ways. I also want acknowledge my major professor, Fred R. Beyette Jr. He has known since my first year as an undergraduate student and mentored me throughout most of my time at UGA. I am grateful for all the support he has provided me over the years. Additionally, I would like to Dr. Ramaswamy for being apart of my committee and providing advice as needed for this research. Furthermore, I would like to thank all of the members of the Cyanotracker project (Chintan Maniyar, Abhishek Kumar, Isabella Rose Fiorentino, and Nathan Tesfayi) and Sydney Whilden for making this work possible and helping me with any problems that I had throughout the past two years. Moreover, I would like to thank all of the members from the Cyanosense 2.0 Capstone team (Mark Seferian, Jonathan Crowell, Keshav Raviprakash, Aladdin Al-Khatib, Maggie Kurdle) and Dr. Peter Kner for helping bring the idea of Cyanosense 2.0 to life and providing their help to support this work. Finally, I would like to thank everyone within the Small Satellite Research Laboratory, the UGA College of Engineering, and the Georgia Space Grant program for their continued support of me and this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cyanobacteria are photosynthetic microorganisms that have a long evolutionary history on Earth (Bettina et al., 2015). Cyanobacteria have developed abilities that allow them to thrive in variety of environmental conditions. Consequently, when they start to proliferate at rapid rate, they can form cyanobacterial harmful algal blooms or CyanoHABs. This results in a loss of oxygen within water bodies through a process known as eutrophication and/or the release of poisonous toxins known as cyanotoxins depending on the type of cyanobacteria species present as displayed in Figure 1.1. CyanoHAB events have increased in frequency around the world in recent years as they degrade aquatic habitats, harm or kill local fauna and species within water bodies, and negatively impact recreational and economic activities that are dependent upon the use of different water bodies. This is shown by events such as the sudden deaths of elephants in Botswana (Veerman et al., 2022), the deaths of a large fish populations within Chile (Mardones et al., 2023), and the increased presence of harmful algal blooms within Lake Erie (Tewari et al., 2022).

While large amount of information has been collected to verify CyanoHAB events, issues with accuracy and precision have resulted in either the misclassification of these events and/or lack of data for specific time periods during data collection. To fix these problems, a cyber-physical system approach was created via the CyanoTRACKER project that utilizes on-site monitoring, social sensing (through community observations and multimedia data mining), and remote sensing to improve the monitoring of CyanoHAB events across the world. However, each of the components of this system needs to be further developed to improve detection accuracy for CyanoHAB events while reducing the waiting period for interested users of the system. Thus, in this work, we introduce the idea of CyanoTRACKER 2.0 and demonstrate improvements that were made to different components of the system to meet these requirements.

1.1 CyanoTRACKER 1.0

The CyanoTRACKER project (referred to as CyanoTRACKER 1.0 within the rest of this work) was a funded National Science Foundation initiative created by a group of researchers at the University of Georgia to address the cyanobacterial harmful algal blooms within water bodies across Georgia. The project used a multi-disciplinary approach to integrate community observations, remote sensing, and multimedia

Figure 1.1: Cyanobacterial Harmful Algal Bloom in Binder Lake, Iowa Credit: USGS Environmental Health Program

data analysis for effective monitoring of CyanoHABs and robust data collection (D. R. Mishra et al., 2020). The project created monitoring infrastructure for early detection of CyanoHABs across the state of Georgia and the rest of the world while providing a framework for the creation of similar infrastructures on a global scale. The system design for CyanoTRACKER 1.0 consisted of three components: a sensor cloud made up of wireless sensor systems deployed at various Georgia water bodies in conjunction with satellite remote sensing, a social cloud for collecting data from social media, news feeds, and/or a mobile application reports regarding suspected CyanoHAB events, and a computational cloud for analyzing data collected from the other clouds. From each of these of these components, data was collected and analyzed for dissemination to the public and researchers across the world through the implementation of unique mechanisms in data collection, noise reduction, incentivization, etc. While CyanoTRACKER 1.0 was able to achieve many of the objectives that it was originally created for, a number of improvements were noted for each of the components in order to improve operations of the proposed system and for scaling the system to be used across the United States and the rest of the world. Some of the suggested improvements include the development of on-site monitoring for improved data collection and communication as well as improvements in the processing and grouping of data collected from social media or new aggregator services.

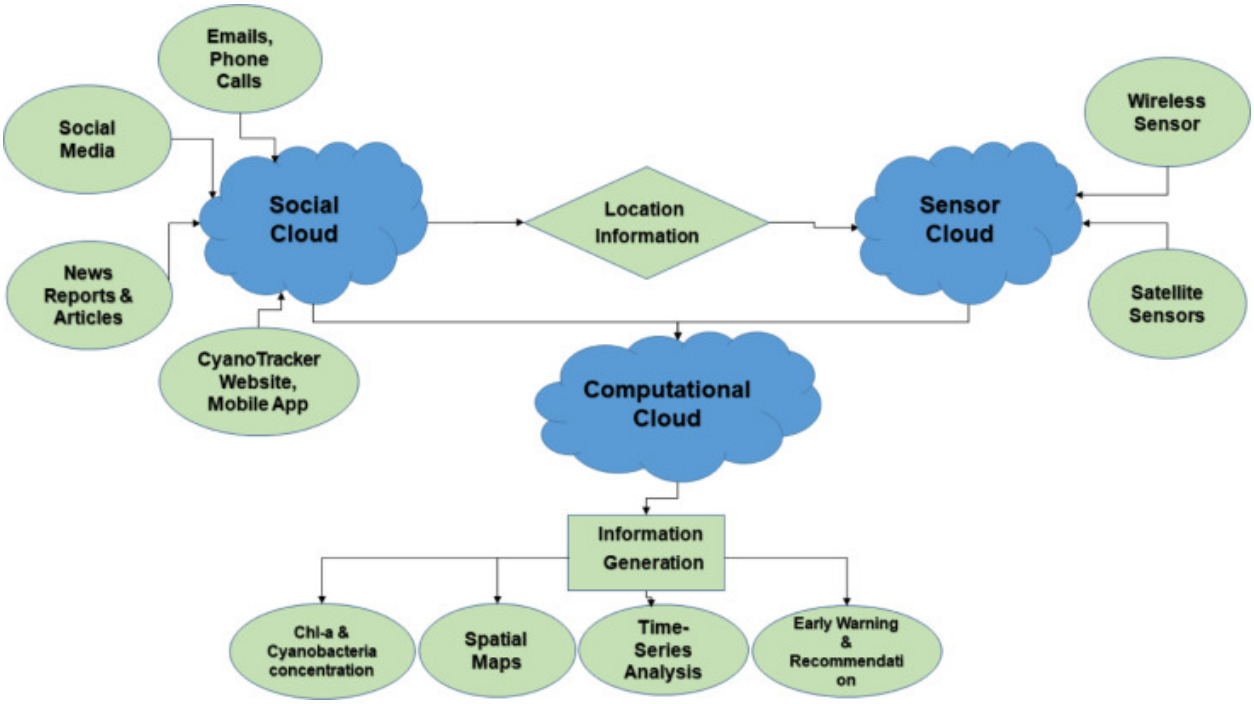

Figure 1.2: CyanoTRACKER 1.0 General Architecture for Monitoring CyanoHABs (Image Source - Borrowed with permission from Mishra et al. 2020 CyanoTRACKER: A cloud-based integrated multi-platform architecture for global observation of cyanobacterial harmful algal blooms

1.2 Research Objectives

In this thesis, we aim to explore how the existing cyber-physical infrastructure (CyanoTRACKER 1.0) can be improved by enhancing the detection accuracy for CyanoHAB events, reducing the waiting period for interested users of the system, and extending the field of study to include locations across the rest of the world. In order to achieve this, the following objectives are proposed for completion:

1. Develop an improved on-site spectrometer monitoring system that will have reduced energy consumption, low footprint, and system costs compared to Cyanosense 1.0 while integrating autonomous data transmission.

2. Develop software approach for collecting and processing results from social media and news services to improve the detection accuracy of CyanoHAB events across the world in a more timely manner.

1.3 Outline of Thesis Work

The rest of the thesis is organized as follows. Chapter 2 describes an overview of on-site monitoring techniques including Cyanosense 1.0 and citizen science/social sensing approaches for CyanoHAB events such as the CyanoTRACKER 1.0 Social Cloud. Chapter 3 describes the methodology used to develop the architectures of Cyanosense 2.0 and CyanoTRACKER 2.0 social sensing framework. Finally, Chapter 4 gives the conclusion of the analysis we have performed on the results extracted from both components.

Chapter 2

Background and Literature Review

2.1 Introduction

In this chapter, we briefly introduce existing cyber-physical monitoring projects and systems for monitoring CyanoHABs that are similar to CyanoTRACKER 1.0. We also explore existing methods for detecting CyanoHAB events using on-site monitoring and reported community observations. Finally, we compare these methods with those used in CyanoTRACKER 1.0 such as Cyanosense 1.0 or the methods used within the Cyanotracker 1.0 Social Cloud.

2.2 Review of Approaches for Developing Cyber-Physical Infrastructure for Monitoring CyanoHABs

While CyanoTRACKER 1.0 was one of the few successful projects that helped to develop an early warning system to monitor CyanoHAB events across a large spatiotemporal scale, other projects and/or systems have also been developed in recent years that have relied on a multi-domain approach to monitor CyanoHAB events. For instance, the Cyanobacteria Monitoring Collaborative Program is program created by the EPA to provide data and/or educational resources on harmful cyanobacterial algal blooms at various locations within the United States as requested by stakeholders within the City of Worcester (of Health, n.d.). The program utilizes on-site monitoring and citizen science to track CyanoHAB events in local and global water bodies through the use of various technologies such as fluorometers or shared photographs taken by citizens across the world. The Harmful Algal BloomS Observing System (also referred to as HABSOS) Program run by the Florida Fish and Wildlife Conservation Commission and the National Oceanic and Atmospheric Administration provides information about harmful algal blooms within the Gulf of Mexico ("About HABSOS", 2024). The project uses both in-situ results and satellite imagery to predict and/or determine conditions at areas of interest in the Gulf of Mexico to inform the

public and/or other organizations. This is similar to the approach used by the NOAA Great Lakes Environmental Research Laboratory which utilizes satellite remote sensing, monitoring buoys, and advanced genetic techniques to monitor harmful algal blooms present within various water bodies such as Lake Erie, Lake Huron, and Saginaw Bay ("Great Lakes Harmful Algal Blooms (HABs) and Hypoxia", n.d.). Additionally, a multi-tiered cloud-based infrastructure for monitoring CyanoHABs within inland water bodies in India was developed (Maniyar et al., 2022). This infrastructure relies on the use of crowd-sourced Twitter data and spectral analysis conducted on remote sensing imagery datasets within Google Earth Engine. Finally, the Comprehensive Management Program of Cyanobacteria established within Argentina is an early warning system that utilizes participation of municipalities and communities to confirm the presence of CyanoHABs (Aguilera et al., 2023) (O'Farrell et al., 2019). CyanoHABs are reported through various reports and/or the online map provided on the project website.

2.3 On-site Monitoring of CyanoHAB Events

While a number of detection methods are available to determine the presence of CyanoHABs, the use of on-site monitoring is still a common practice for determining the presence of CyanoHABs within water bodies and providing validation of information obtained from satellite remote sensing. Measurements can be collected through a variety of methods such as the analysis of water samples to the measurement of light emitted from water bodies. Each of these detection methods are primarily categorized based on the location of where data analysis occurs, parameters being measured, scalability, and the amount of time between data collection and when final results can be determined (Johnasen et al., 2023). In the following subsections, we will discuss the importance of each of these categories.

2.3.1 Equipment/Technology for Spectral On-Site Monitoring of CyanoHABs

Typically, identifying CyanoHABs requires the use of equipment and/or technologies to determine measurements from data samples collected using field-based spectra. This requires that data collection and analysis be completed at different sites where the equipment and/or technologies exists. Thus, each of the methods can be categorized into two categories: lab-based monitoring and field-based monitoring. Lab-based monitoring methods are analytical methods that use laboratory equipment and tools to determine whether CyanoHABs are present (Johnasen et al., 2023). They use an approach that requires water samples be collected from the field and analyzed under a microscope or through specific kits (Council), 2020). These methods are more accurate in determining the presence of CyanoHABs by identifying species and even the presence of cyanotoxins. However, they are more expensive and require advanced training to determine results. On the other hand, field-based monitoring is an approach in which data collection and analysis to determine CyanoHABs can occur in real-time or near real-time. Field-based monitoring typically utilizes in-situ radiometric sensors to make measurements and includes a variety approaches such as absorbance, fluorescence, and hyperspectral sensing. Field-based monitoring also offers real-time or near real-time approaches that are more effective and quicker for interested users of these systems.

2.3.2 Measuring Cyanobacteria and Related Water Quality Parameters

There are number of optical and inherent water quality parameters that can be utilized to detect CyanoHABs using on-site monitoring within water bodies. However, determining which parameters to measure is highly dependent on the type of on-site monitoring system being used (Council), 2020). For example, some parameters such as algal pigments, pH, or dissolved oxygen are considered more important for on-site monitoring to account for quality assurance or consistency concerns. Consequently, parameters such as the detection of cyanotoxins or cyanobacterial biomass are less utilized but can provide valuable insight into the extent or severity of CyanoHABs across a water body. Furthermore, some parameters are also easier to detect by various methods but can fail to provide accurate and/or comprehensive results to interested parties. For example, chlorophyll-a is easily detectable through many monitoring methods but cannot identify differences in an algal bloom or be used to distinguish plants and/or other phenomena from CyanoHABs. On the other hand, the detection of specific cyanotoxins such as microcytosin or cylindrospermopsin are much more difficult to identify and require the use of in-situ testing to determine their presence within water bodies.

2.3.3 Spatial and Temporal Resolution for Scalability

Spatiotemporal monitoring of CyanoHABs helps to solve many of the issues that might occur in order to determine CyanoHAB events. For example, CyanoHAB events can have different severity levels or areas of distribution throughout a water body depending upon various factors including but not limited to weather conditions, water depth, or water flow within a water body(Council), 2020). Similarly, the time that it takes for CyanoHAB events to complete can vary depending on the presence of favorable conditions such as warm temperatures or a high nutrient influx within a water body. While there are number of benefits to the use of on-site monitoring, the consideration of variability in fields of interest along with time variations has proven difficult for many on-site monitoring methods. Typically, the use of technology that allows for movement can allow for better spatiotemporal monitoring as users of these systems are not restricted to one area. However, this can prove challenging as users are required to move across or around a water body to truly quantify the measurements. More viable solutions typically rely on autonomous movements such as drones and/or AUVs to properly account for this feature in most studies when using on-site monitoring (Garrett Bartelt & Hondzo, 2024),(Johnasen et al., 2023).

2.3.4 Turnaround Time for Data Collection and Processing

On-site monitoring methods usually vary in the amount of time that they take to determine the presence of CyanoHABs. In-situ methods usually require samples be transported and analyzed in a laboratory before determining the final result. This process can be lengthy and be problematic for users. Moreover, other methods that do not require the use of laboratory spaces for CyanoHAB detection can still require time to process data. Consequently, a research thrust to improve real-time monitoring has been identified as a possible solution to this problem. Most approaches as part of this research thrust typically rely on

communication protocols to be put into place that allow for data to be sent to other locations such as the use of 4G/5G technology or satellite communications (AquaRealTime, 2024) (Sonic, 2024). These systems can not only help to reduce data transmission overhead of CyanoHABs but also generate data results via interfaces that allow authorized users to track changes in the water within set periods of time. Other approaches to implementing real-time systems include the ability to complete computational processing and/or analysis using technology such as microcomputers, Arduinos or Raspberry Pis (Boddula et al., 2017). Finally, some systems can measure cyanobacteria concentrations and/or other properties and then provide an indicator that allows users to correctly identify CyanoHABs within water bodies during measurement such as a notification or visual measure (Council), 2020).

2.3.5 Cyber-Physical Approaches for On-Site Monitoring Systems

Cyber-physical systems are defined as systems that are built from and depend upon the integration of computation and physical components ("Cyber-Physical Systems (CPS)", 2023). The technologies that make up these systems have transformed the way people interact with different systems by allowing users While a number of on-site monitoring systems have been developed, the use of cyber-physical on-site monitoring systems for monitoring CyanoHABs has grown in recent years. The use of monitoring buoys (Miller et al., 2022) (Presa Reyes et al., 2020) or sondes (Srivastava et al., 2013) that are equipped with sensors and data communication technology have allowed on-site monitoring systems to communicate about various water quality metrics of relevance in the detection of CyanoHABs such as pH, dissolved oxygen (DO), or water temperature. Similarly, new systems have been built to provide more functionality and increase the detection accuracy for on-site monitoring. For instance, the environmental sample processors (also known as ESPs) are instruments that can collect in-situ water samples for determining the presence of toxins such as cyanotoxins when at sea (Scholin et al., 2009). These instruments can collect 60 samples and send data that was analyzed in real-time to scientists on the shore. Similarly, the use of Imaging FlowCytobots (or IFCBs) can help to identify harmful algal bloom species and/or concentrations using submersible microscopes equipped with computer vision software deployed within a water body (Gandola et al., 2016) (Kraft et al., 2021).

2.4 Cyanosense 1.0

As part of the sensor cloud, a cyber-physical approach using on-site monitoring was propsed to measure the absorption features over water bodies. To accomplish this goal, a novel sensor system was created called Cyanosense (referred to as Cyanosense 1.0 within the rest of this work). Cyanosense 1.0 was wireless hyperspectral sensor system built for proximal, on-site monitoring to validate local CyanoHAB events reported through social media and confirm the formation of harmful algal blooms over water bodies (Boddula et al., 2017). As part of the design of the system, Cyanosense 1.0 allowed for easy deployment at water bodies, reduced system power usage via an energy efficient design, performed wireless data transmissions using 3G or 4G cellular networks, and reduced overall costs for future replication by other potential users.

Additionally, Cyanosense 1.0 was able to provide results from three local water bodies (Lake Oconee, Lake Chapman, and Lake Oglethorpe) by collecting scans of the water bodies and finding the radiometric values of upwelling and downwelling from those scans (D. R. Mishra et al., 2020). From this data, remote sensing reflectance values were determined to estimate the phycocanin and chlorophyll-a concentrations using the empirically-derived formulas and satellite remote sensing validation from hyperspectral bands measured using the MERIS's sensor platform. Moreover, the sensor system was cross-calibrated with the GER 1500 to improve the accuracy of predictions for phycocanin and chlorophyll-a concentrations from the previous step. Finally, the presence of cyanobacteria within water bodies was confirmed using a portable microscope provided by the US Environmental Protection Agency (EPA). Ultimately, Cyanosense 1.0 was determined to be able to capture the onset of CyanoHABs at lower cost comparable to other on-site monitoring technology and without the need for field site verification.

2.5 Citizen Science and/or Social Sensing Technology

In this section, the use of technology for citizen science and/or social sensing is explored. First, citizen science and social sensing are defined. Next, the use of these approaches within environmental monitoring are explored through the development of different technological approaches created. Finally, the use of citizen science and/or social sensing technology applications for CyanoHAB monitoring are discussed in more detail.

2.5.1 Citizen Science

Citizen Science is defined as a collaboration between scientists and citizens are motivated to make a difference due to their curiosity or concern for a particular issue ("Citizen Science", 2024). Citizen science can help to encourage citizens to participate in research efforts being conducted, increase the sample size and spatial variation of data samples for study in various research projects, and increase the participation of the general public in public policy. Examples of successful citizen science projects include classifying images of different species (Willi et al., 2018) (Soltani et al., 2023), reporting individual health conditions such as diet and exercise (Han et al., 2011) (Chen et al., 2016), and helping with disaster management/response during environmental events (Ottinger, 2022). However, with the increase in technology and increased interactions between people around the world, the use of phenomena known as citizen sensing is more commonly used to allow for citizens to collect sensor data and become more engaged with citizen science projects. Additionally, citizen sensing can allow citizens and researchers to analyze data collected from the general public without their direct involvement in a research study. In particular, the use of citizen sensing for monitoring specific events of interest has grown in recent years in a process known as social sensing. The following subsection discusses the use of social sensing in citizen science projects and introduces the frameworks commonly used to attain accurate and precise results.

2.5.2 Social Sensing

The increased use of technology by society has led to the usage of different media platforms for broadcasting information by different individuals related to their personal lives, the local environment, and/or common topics of interest with users across the world. Consequently, a large amount of data has been generated on these platforms that can be analyzed for determining relevant information for a variety of topics of the physical world in process known as social sensing. Social sensing is more advantageous than citizen science in that extracted data comes from users over large areas such as different states or continents, extracted data can be collected over set time periods, and users of these media platforms do not need to consent in order to participate in research studies (as long as data extraction is performed using publicly available data). A number of research studies have incorporated the use of social sensing from extracted social media data to improve various socioeconomic problems faced in modern society such as market trends, disaster management, and disease monitoring (Wang et al., 2019) (Galesic et al., 2021) (Rashid et al., 2023) to identify various patterns and/or relevant information for use within different scenarios. However, each of these studies faces challenges with validation of data for analysis, proper preprocessing of data for improved data analysis results, and problems with commonly used human language processing technologies. The rest of this section aims to solve these problems in the context of technological applications for environmental monitoring and CyanoHAB monitoring.

2.5.3 Citizen Science and/or Social Sensing Technology for Environmental Monitoring

A number of environmental monitoring studies utilize citizen science and social sensing to solve various research problems. In particular, a number of the approaches proposed within these studies has relied on the use of technology in order to involve communities and individuals with various aspects of a research study. For example, the Marine Debris application was created by the NOAA Marine Debris Program and University of Georgia College of Engineering program to allow citizen scientists to provide data about the presence of plastic pollution across the world (Ogle, 2024). Users can report marine debris items on the application with only the GPS signal of their mobile phones and then share data for users to view on public map located on the website. Additionally, the use of citizen science technology for environmental monitoring can be seen in the National Weather Service (NWS) SKYWARN program where volunteers can help spot severe weather events particularly severe local thunderstorms (US Department of Commerce, 2024). By communicating through tools such as HAM Radio, the NWS can help to provide more accurate and timely warnings to concerned citizens during severe weather events as a result. Furthermore, the use of social sensing is seen as useful tool in monitoring disasters in real-time like flood events. In particular, the use of Twitter was seen as great way to crowd source information about flood events in the United Kingdom by helping to improve geolocation and relevancy of the information used for detecting these events (Arthur et al., 2018). Finally, the use of social sensing can help to monitor the environmental impact of humans such as refugee or displaced persons camps using open-source tools such as geoportals

(Programme, 2024-01). By using these tools, the UN has been able to monitor the impact of refugees on deforestation and soil degradation in Tanzania and impact international policy around the world.

2.5.4 Review of Technology Applications for Monitoring CyanoHABs using Citizen Science and/or Social Sensing Technology

The use of citizen science and/or social sensing technology to improve environmental monitoring and early warning systems has led to development a number of projects related to monitoring CyanoHABs. For instance, the Cyanobacteria Tracker is a project created by the Lake Champlain Committee in conjunction with the Vermont Department of Health for volunteers to collect data from local water bodies in Vermont and report it for users to access (of Health, n.d.). These citizen science observations are mapped onto a tracking application that allows users to receive alerts and reports from each of the reported test sites as needed. Additionally, the US Environmental Protection Agency through the Cyanobacteria Monitoring Collaborative Program has created three crowd sourcing applications to identify various programs to identify CyanoHABs by gathering information about their presence and/or developing a better understanding of factors leading to CyanoHAB events ("Cyanobacteria Monitoring Collaborative", 2024). Each of these applications requires the use of citizen scientists to report observations collected with different types of technology available. Finally, the EyeonWater Australia campaign run by the Commonwealth Scientific and Industrial Research Organization has created an application for identifying potential CyanoHAB events through the use of the Forel-Ule color scale, Sechi disk, and/or probe (Malthus et al., 2020). Citizen scientists can report water quality, clarity, pH, and other metrics to determine whether CyanoHABs are present in the water and potentially identify events in real-time.

2.6 CyanoTRACKER 1.0 Social Cloud

One of the original research thrusts of CyanoTRACKER 1.0 was to develop a methodology for obtaining and analyzing data from communities-at-large. This was achieved by using citizen science and social sensing for collecting, storing, and analyzing suspected CyanoHAB events from social media, dedicated media applications, and/or news aggregator services while geolocating water bodies of interest. For example, the use of Twitter via social sensing was seen as useful in providing relevant information for study into the type of users reporting suspected events, the relevancy of collected data to environmental monitoring of CyanoHAB events, and analyzing spatiotemporal trends and/or distributions of collected data points (Patil, 2018) (Joshi, 2015). Furthermore, the use of Google News articles was explored to identify the use of news articles in identifying CyanoHAB events (Jadhav, 2018). In this study, Google News articles were used to extract textual data for relevancy classification to CyanoHAB monitoring using natural language processing techniques such as Latent Direchlect Allocation (LDA) and Named Entity Recognition (NER). Finally, the use of citizen science results from dedicated media applications such as an online Qualtrics survey, a CyanoTRACKER mobile application, dedicated CyanoTRACKER user accounts on Facebook and Twitter, and from emails/phone calls allowed users to report suspected CyanoHAB events

by providing information about CyanoHAB events in the Southeastern United States (Chanda Das, 2017). As part of this study within the social cloud of CyanoTRACKER, shared images of CyanoHAB events were extracted for analysis while a system for gamification and incentivization were promoted to improve the citizen science reports shared within the study as well.

2.7 Natural Language Processing (NLP)

Natural language processing is a subfield of computer science and artificial intelligence that involves the use of various systems to understand human languages such as computers (Harrison & Sidey-Gibbons, 2021). A number of organizations utilize NLP to improve their operations for a variety of socioeconomic issues. In particular, natural language processing can help to identify CyanoHAB events that occur within extracted citizen observations for CyanoTRACKER 2.0 by improving the processing and analysis of extracted text observations. The following subsection highlight various NLP operations that will be utilized to improve the results obtained for this study.

2.7.1 Morphological Analysis

Morphology is a branch of linguistics which is concerned with the study of word structure (GeeksforGeeks, 2024). When it is applied in natural language processing, it looks at the computational preprocessing of word structures and/or forms in a process called morphological analysis (Zarkar, 2022). This is achieved by breaking down words into their original forms called morphenes to understand their meanings and/or roles within textual data. One key technique of morphological analysis that is used in this study is lemmatization. The process of lemmatization is used within NLP text preprocessing by reducing words to their base form by grouping together different forms of a word together and then reducing them to their lemma (base) form. By implementing lemmatization, the parts of speech and context of individual words within text can be established. Additionally, the use of techniques such as tokenization and/or the conversion of text to lowercase can help to make text more standardized for further NLP analysis by converting text into small parts called tokens and reducing issues with ambiguity or complexity within textual data. Finally, the removal of characters and/or words (called stop words) are used to improve the extracted results during analysis.

2.7.2 Semantic Analysis

Semantic Analysis is the field within natural language processing that aims to look at the meaning of human language via machine systems such as computers (GeeksforGeeks, 2021). This is done by either individually analyzing words or by looking at the meaning of words in context to the sentences that they are being used in. As part of the semantic analysis used within this study, a technique known as named entity recognition is utilized to determine the presence of suspected CyanoHAB events within the extracted textual data. Named entity recognition identifies predefined categories of objects within a

text. The extracted data objects are referred to entities and are classified into categories based on names, locations, quantities, and so on.

Chapter 3

Methodology

In this chapter, we introduce the development of two new components of the CyanoTRACKER 2.0: Cyanosense 2.0 and the CyanoTRACKER 2.0 Social Cloud. We first discuss the development of improved version of Cyanosense known as Cyanosense 2.0 and present a design methodology of a prototype created to satisfy this requirement. Additionally, a methodology is introduced for the collection and analysis of suspected CyanoHAB events from social media and news platforms using social sensing to determine the presence of CyanoHAB events across the world.

3.1 Cyanosense 2.0 Goals

While there were a number of successes for Cyanosense 1.0 that allowed it to be a more viable option in comparison to traditional CyanoHAB monitoring methods, there were a number of improvements that were highlighted in previous works regarding the system that could help to meet the original objectives but also solve issues of concern with regards to performance in the field, differing field conditions for interested users, and system design. To meet both the new and original goals, the following objectives were identified as being important:

1. Reduce the overall cost of the system to ensure that the system is cheaper than existing sensor systems on the market and cost-effective in detecting CyanoHABs.

2. Improve accuracy and precision of readings for potential users of the system to determine the presence of CyanoHABs and validate satellite remote sensing data of these events.

3. Provide communication methods that can reduce human intervention if needed and/or provide alternatives based on the use case for interested users.

4. Protect system from water damage and heat resistance to increase the durability of the system.

5. Reduce net power consumption for system to reduce environmental footprint of system and/or conserve energy during periods when low amount of power is generated by a solar panel.

6. Reduce maintenance and/or serviceability of the system in order to allow systems to be functional for longer time periods and/or reduce any associated costs for the repair of different components.
7. Reduce the overall volume of the system to increase durability and portability of the system.
8. Comply with any regulations and/or laws set forth by different regions across the world to allow the system to be used in different areas where interested users are present.

The next two sections provide the methodology used to achieve these objectives.

3.2 Proposed System Architecture

Similar to Cyanosense 1.0, Cyanosense 2.0 utilizes a number of components in order to achieve the objectives mentioned in the previous section. The following section provides a detailed overview of the sensor system design as well as the electrical, software, and mechanical design process.

3.2.1 Cyanosense 2.0 System Overview

To attain the proposed objectives, the following components were identified as important to the design of the system (Note: Some components are not listed due to their importance during specific design processes.)

- **Spectroradiometer Sensors** - Spectroradiometers are used to measure the spectral characteristics of various targets based off light intensity. For this system, two Hamamatsu C12880MA Spectrometers were chosen due to their small size, low cost, low resolution range, and compatibility with the Arduino IDE.
- **ESP32** - ESP32 is a single chip microcontroller designed with low power technology to achieve the best RF and power performance in variety of applications. It was chosen over the Arduino for Cyanosense 2.0 because it has improved processing and memory storage capabilities, is compatible with the Arduino IDE, and contains the required number of GPIO pins needed for communication with the rest of the system.
- **Photorelay** - Photorelays are electric switches that utilize light as an input signal without contact. The TLP3556A photorelay is used in the design of Cyanosense 2.0 to turn the spectroradiometers sensors off and on to conserve battery power.
- **Voltaic solar panel and battery pack combo** - To meet the design objectives for Cyanosense 2.0, the combination of a solar panel and battery pack need to be taken into consideration. The V25 Voltaic from the Voltaic ecosystem was included in the Cyanosense 2.0 design due to the availability of solar panels that are already configured to a 6400mAh Li-Ion battery along with its abundant power capacity from configuration to power Cyanosense 2.0 for longer periods of time.

- **Iridium Satellite Modem** - To transmit data in remote areas, the ROCKBLOCK Iridium 9603 Satellite Modem can be used to send data through the Iridium satellite constellation to users with access to the Iridium interface. For Cyanosense 2.0, spectra data stored on the ESP32 from the spectroradiometers is converted into bytes for transmission via the satellite modem.
- **Switch** - The use of manual switch is important as it allows the system to be set to one software transmission mode at a time. The 1P2T SPF switch was used to switch between the different software transmission modes for sending data.

To operate the system, the RTC is used to turn on the ESP32 microcontroller to boot up the system for operation (Note: While the ESP32 has an RTC module provided, the use of an external RTC for the system was chosen instead due to performance issues discussed in more detail in the electrical design process). Afterwards, to perform manual scans, the ESP32 activates the photorelay which in turn powers the sensors to take measurements. Once the measurements are taken, the photorelay is turned off and the collected data is stored on the ESP32 to be uploaded either through a USB connection to a connected computer or the Iridium satellite modem for transmission via the Iridium satellite network.

3.2.2 Cyanosense 2.0 Electrical Design

To meet the above objectives stated above, a design was chosen that reduced the number of components and reduced power consumption by the system. As part of the electrical design, the following pieces of information needed to be considered:

1. The GPIO Pins of the ESP32 can only receive 3.3 Volts.
2. The Iridium satellite modem can only receive 5 Volts.
3. The photorelay typically needs a forward current of 10 mA to be activated.
4. The ESP32 RTC (Real Time Clock) experiences a daily drift of a couple of minutes.

To account for these facts, special electrical design features were created on two breadboards. For example, wires were color-coded to identify where they send current to. This included the use of red wires to indicate a connection back to the 5V power source, purple wires that connect to the Real Time Clock, and green wires connected to the video or the output of the C12880MA. Each of the color schemes used in the wiring are shown in Figure 3.1 below. Furthermore, the use of GPIO pins were decided based on the functions that they were needed for by the system or by the placement of physical components as part of the system design. For example, GPIO pins 1 and 2 were each used respectively to send video (or data) from the sky and water sensors as they were ADC pins, while GPIO pins 6 and 7 are connected to the RTC and start pins of the C12880MA sensors. Similarly, GPIO pins 17 and 18 were selected to act as serial input and output for the ROCKBLOCK modem as well as allow another serial port to be created for sending data through the ROCKBLOCK Iridium modem. Moreover, to meet the voltage and current requirements listed above, different electrical design configurations were created. For instance, the voltage entering

into the ESP32 from the sensors is set equal to 3.3 V from 5 V through a voltage divider configuration using four resistors (R9, R10, R11, and R12) after which the original voltage of 5 V is used to turn off the photorelay by sizing Resistor R6 for a forward current activation of 10 mA. Similarly, the voltage from the ESP32 was changed from 3.3 V to 5 V when waking up the ROCKBLOCK Iridium modem using a voltage gain generated from a N-channel MOSFET. Finally, changes were made to the overall electrical design to improve operation of Cyanosense 2.0 and reduce the need for future maintenance of the system. For example, an external RTC module was added to reduce the drift being experienced by the ESP 32 ESP module, resulting in the daily drift being reduced from minutes to milliseconds and allowing for more precise wakeup times. Likewise, all connectors used in the electrical design were labeled and keyed on the physical system to identify any potential issues during operation and allow for the system to be easily maintained in the future.

3.2.3 Cyanosense 2.0 Software Design

As part of the design of the sensor system, a number of software components were created as part of the design of the system. This includes the software for setting up the general operation of Cyanosense 2.0, for executing each of the transmission modes, and for performing data processing and analysis. The software written for the system contains a number of software programs and/or files that help to control the power and signals generated from the ESP32, perform storage and/or mathematical calculation operations for values collected by the sensor system, and allow for changes to the overall operation of the sensor system based upon user input or environmental variations at field sites. Each of these three operations are highlighted in more detail in the rest of this section.

Cyanosense 2.0 General Operation Software Design

To allow for Cyanosense 2.0 to properly function, embedded software written onto the ESP32 was used to perform the general operations described in the Cyanosense 2.0 System Overview subsection. The embedded software written to perform these general operations contains software programs and/or files that help to control the power and signals generated from the ESP32, perform storage and/or mathematical calculation operations for values collected by the sensor system, and allow for changes to the overall operation of the sensor system. To accomplish all of these operations, the embedded software design is controlled by a software program called `ESP32_Dual_Spectrometer_SPIFFS_Sleep_Serial_01.ino` which provides the functionality mentioned previously. As part of the software program, both the pins and variables that will be used are defined and set to a designated predefined value depending upon their use case. Afterwards, a number of operations are completed to setup the system for data collection such as setting pins to the right state based on their use case, powering on or off components such as the ROCKBLOCK Iridium modem, and initializing or configuring system changes such as alarms or API calls. Once all of these operations are completed, an embedded file system was mounted using the LittleFS software package and the calibration files are read into memory. Finally, the local time is printed using the `Time.ino` to indicate successful completion of general software operation.

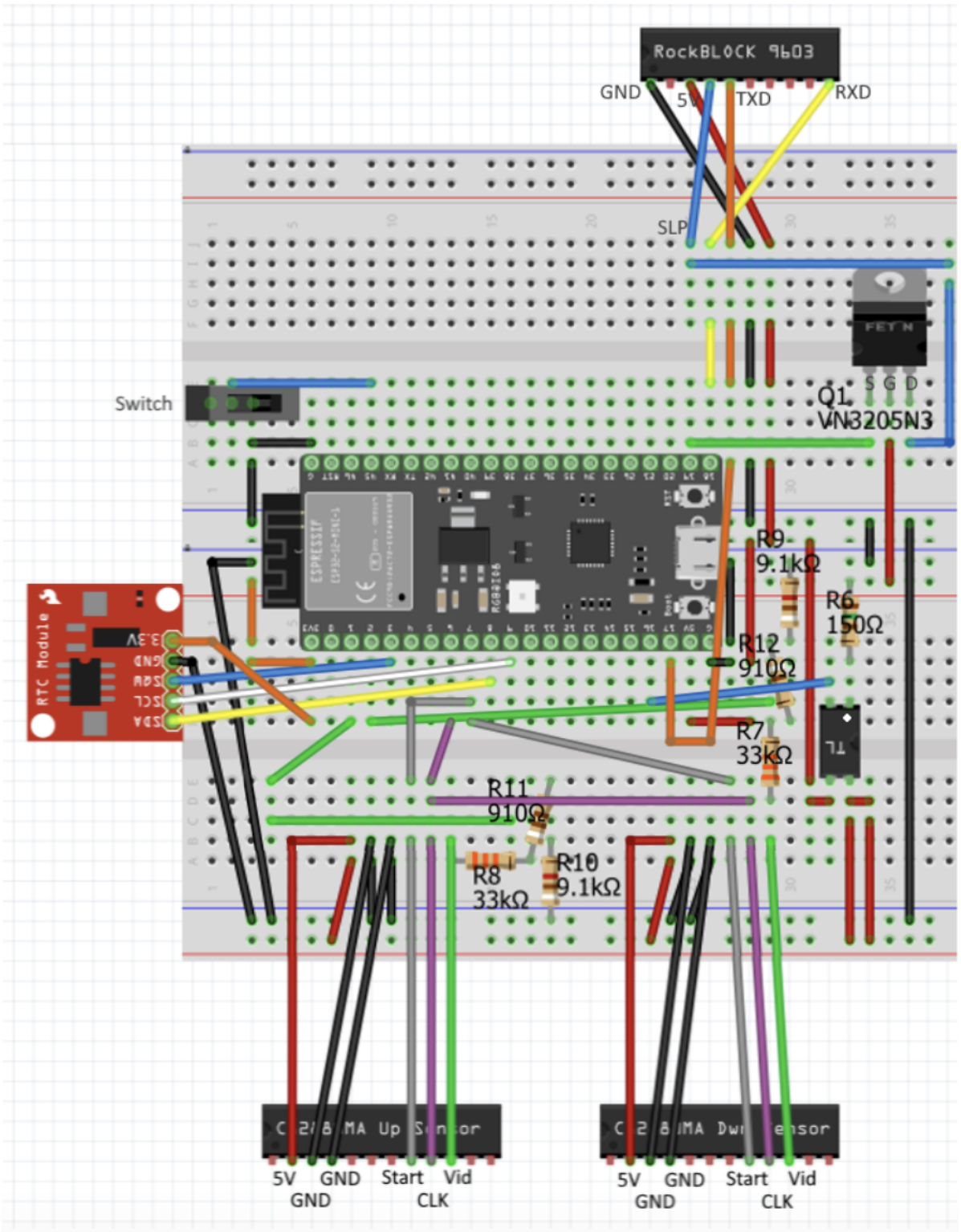

Figure 3.1: Cyanosense 2.0 Electrical Design

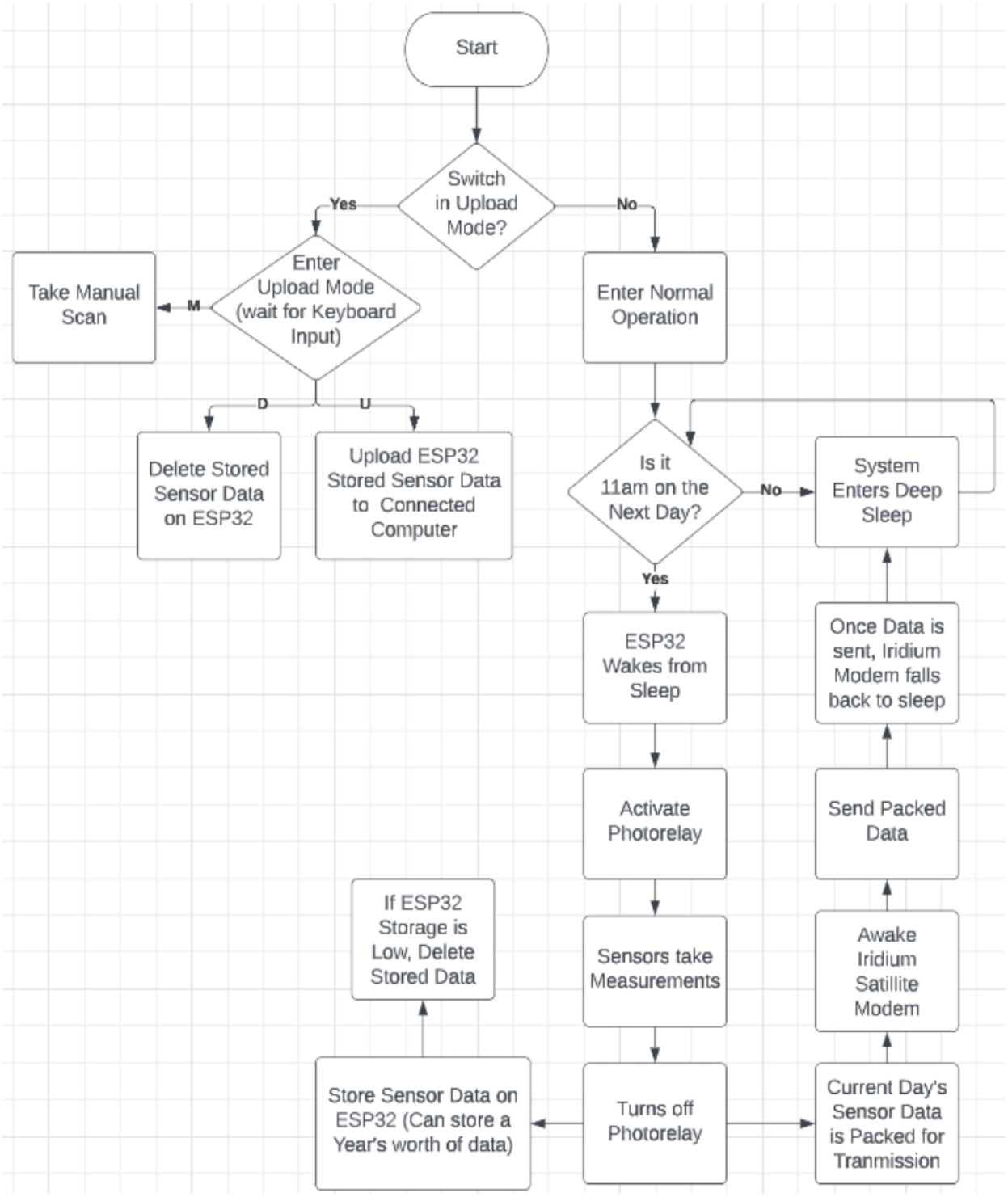

Figure 3.2: Cyanosense 2.0 General Software Architecture

Cyanosense 2.0 Sensor Operation Software

Important part of the Cyanosense 2.0 software is collecting data values from the spectrometers and saving them onto the ESP32. To accomplish this task, a number of software programs are required. After the spectrometers are powered on by the photorelay, the first software program enabled for collecting intensity readings is `ReadSpectrum.ino`. This program reads the intensity values from both of the spectrometers using the clock cycles generated from the RTC. When the first clock cycle after the start pulse goes low and 88 clock cycles have passed from powering on the sensor, the first pulse is counted for data collection. Afterwards, data can then be read using video signal acquired at the 89th trigger pulse. Finally, the integration time is determined by a start high pulse period plus 48 clock cycles (Note: The integration time can be changed by changing the ratio of the high and low periods of the start pulse) (Corporation, 2024). For the second software program, the `AGC.ino` is utilized to adjust the integration time based on the collected intensity values. Using the 288 channels present for each sensor, the integration time for both sensors is adjusted if: 1) The maximum intensity value recorded is more than 7200 nm. 2) the adjusted integration time is above 20 ms, below 0 ms, or not equal to the default integration time. Finally, the `DeepSleep.ino` helps to turn on or off components for collecting data to conserve energy by setting pins to different states or writing different alerts to the system regarding system operations (Note: The `DeepSleep.ino` helps to turn off the ROCKBLOCK Iridium modem after transmission as described in the next subsection).

Cyanosense 2.0 Transmission Software

Two software transmission methods were created to collect readings from the ESP32: an upload mode and a normal operation mode. Each of these modes is set by the 1P2T SPTF mechanical switch and completes the steps stated as part of the general architecture of the systems before starting the data transmission process. If the switch is detected to be in upload mode, code written in the file labeled `Upload.ino` is used to execute number of operations utilizing a USB connection between a computing device and the sensor system. This includes uploading new code into the sensor system, deleting previous collected sensor data samples from the ESP32, taking a manual scan using the sensor system, or uploading existing sensor data samples from the ESP32 to a directly connected computer. Operations performed in this mode require the user to interact with a terminal command program that runs through the Coolterm Capture GUI software program.

On the other hand, the switch being placed in the normal mode makes the system utilize a number of embedded software program operations to complete the data transmission process. For example, if the switch is set to normal mode at 11 AM and completes all of the general software operations, the ROCKBLOCK Iridium modem will complete two 320 byte transmissions containing all of the intensity values from the up and down sensors. As part of these transmissions, the corresponding wavelengths for each of the intensity values is excluded while their range is limited between the wavelengths of 390 nm to 720 nm. This is done in order reduce the transmission costs, the amount of data compression needed, the overall battery usage of the system, and the bit error rate for transmitted data (Note: Users can collect

the full spectra from each of the sensors by pulling data from the memory storage of the ESP32). As part of these transmissions, the collected intensity data from each of the sensors is compressed into a high and low byte via a buffer and then converted into binary for transmission. Once this action is completed, system-generated messages will be created based on the success of the transmission operations and then the ROCKBLOCK Iridium modem will be put to sleep.

Cyanosense 2.0 Data Processing and Analysis Software

As part of the software created for Cyanosense 2.0, data processing and analysis are considered an important component for Cyanosense 2.0's software for identifying CyanoHABs within water bodies. As part of this component, two software programs are written to provide users with the proper interpretation of results from Cyanosense 2.0. The first software program was created to implement a data processing protocol that will decode satellite data received from the Cyanosense 2.0 ROCKBLOCK Iridium modem into intensity values. As part of this software program, data transmitted and received from the ROCKBLOCK Iridium modem is converted into binary data and presented as hex values. This is done by converting the hex values into a high and low bytes that are then converted into decimal numbers representing intensities from the sensor. Due to the ROCKBLOCK Iridium modem not sending wavelengths, the software program compares its index range to the one stored on the ESP32 and then determines if there are any missing intensity values not received after wireless transmission. Afterwards, the second software program helps to analyze the intensity values received from the decoding software program or during upload mode by plotting these values and determining their Rrs (remote sensing reflectance) values. This is done by converting the data files received into data lists and then using Equation 1 to calculate the remote sensing reflectances collected at each water body. Depending on the needs of Cyanosense 2.0 users, they can use these values to either perform cross-validation with other on-site monitoring systems or perform additional calculations that can then be validated by satellite remote sensing techniques.

3.2.4 Cyanosense 2.0 Mechanical Design

To meet the goals stated in section three, careful consideration was taken for the design of Cyanosense 2.0. To meet these requirements, the completed design shown in Figure 6 was created. The mechanical design was split into three parts that were then combined together: sensor housing, electronics housing, and the final assembly. Each of the components of this design are described in more detail below.

Sensor Housing

The sensor housing was based on the design of a FSE PVC electric box with two modifications. For the first modification, two holes were drilled into the box to allow for the sensors to observe the outside environment within the box. Glass lens were then placed over these holes and were siliconed to waterproof the box. The second modification was completed by trimming the four corner posts by 0.15 inches. This allowed for the two 3D printed spectrometer mounts to slide onto the trimmed posts and the sensor

housing cover to form a seal to prevent water from entering into the housing (Note: The sensors were mounted to the mounts in the sensor housing using screws.)

Electronics Housing

Similar to sensor housing, the electronics housing was based off the design of an electrical junction box but incorporated different modifications to fit design needs for Cyanosense 2.0. As part of the electronics housing design, the inside of the electrical junction box was divided into a bottom layer that contains the system's battery and the ROCKBLOCK Iridium modem along with a top layer which contains the rest of the electronics. For the bottom layer of the electronics housing, the ROCKBLOCK Iridium modem was mounted using a custom 3D printed mount that was then attached to a notched plate through the use of screws onto the bottom of the box. Afterwards, standoffs were mounted onto the notched plate at the bottom layer of the electronics housing to allow for the top layer of the electronics housing to sit on top of the Voltaic battery. Finally, foam pads were then placed on the bottom and top side of the battery to clamp the battery in place and complete the bottom layer of the electronics housing. On the other hand, the top layer of the electronics housing is mainly consisting of two custom breadboards with all of the electrical components from the Electrical Design and the solar panel. Additionally, as part of the top layer, two custom breadboards were held down by thumb screws to allow users of the system easy access for unplugging the Voltaic battery. The thumb screws were then used to mount standoffs through screw holes created in the box. Six holes were then drilled into the top layer to mount an SMA connector for the external Iridium antenna, the Voltaic solar panel, and the Voltaic solar panel cable. Later, an O-ring was placed over the SMA connector to prevent water leakage and the external Iridium antenna was screwed onto the SMA connector. Afterwards, a 3D printed mount for the RTC was added to allow for the RTC to be placed on one side of the breadboard. Lastly, the Voltaic solar panel was mounted from inside the electronics housing by placing nuts onto the threads and was then siliconed along with the Iridium antenna cable to create a waterproof seal for the system.

Final Assembly

The final step of the mechanical design was to complete assembly of all components to form the design shown in Figure 6. First, a 2ft long 3/4 inches schedule 40 PVC pipe was used to help connect the sensor housing to the electronics housing. This component needed to be incorporated in order to prevent shadows from affecting the upwelling intensity readings taken by the down Hamamatsu sensor and to get the best angle of deflection across the length of the pipe. Afterwards, a hole was drilled into the electronics housing to allow for a 3/4in schedule 40 terminal connector to be attached from the inside of the electronics housing. Then, a metal fender washer was used inside the electronics housing to provide reinforcement and prevent the sidewall from flexing. Both the connector and washer were later made to be held together through the use of a metal ring while a rubber seal was used to prevent water from leaking in from outside the electronics housing. Finally, PVC primer and glue was used to connect the pipe from the sensor housing to the electronics housing (Note: As part of the final design, metal feet were installed

on the bottom of the electrical housing that could be screwed into wooden board to provide stability but were not necessary as part of the final design).

3.3 Cyanosense 2.0 Remote Sensing Reflectance Data Collection

As part of the data collected for this study, measurements were taken at various water bodies across the United States to determine the expected performance of Cyanosense 2.0 when used in the field. To do this, light intensity measurements from both of Cyanosense 2.0 sensors needed to be taken and then converted into metric that can be compared with other sensor systems and/or optical satellite remote sensing instruments such as the Copernicus Sentinel-3 Ocean and Land Colour Instrument (OCLI) or PACE's Ocean Color Instrument (OCI). The metric used as part of this study is called remote sensing reflectance and was determined using the following formula:

$$Rrs = \frac{L_w - 0.02 * L_{sky}}{1.01 * \pi * L_c} \tag{3.1}$$

where L_w is defined as water leaving radiance calculated from the spectroradiometer sensor looking downward towards water, L_{sky} is defined as sky radiance calculated from the spectroradiometer sensor looking upward towards sky), and L_c is defined as the radiance from a gray or white calibration panel calculated from the spectroradiometer sensor looking downward towards the panel selected. From this formula, chlorophyll-a and phycocanin concentrations can then be estimated using the following NDCI and PC3 formulas (D. R. Mishra et al., 2020) (S. Mishra & Mishra, 2014):

$$NDCI = \frac{R_{rs}(708) - R_{rs}(665)}{R_{rs}(708) + R_{rs}(665)} \tag{3.2}$$

$$PC_3 = (R_{rs}^{-1}(620) - R_{rs}^{-1}(665)) * R_{rs}(778) \tag{3.3}$$

Estimating the chlorophyll-a and phycocanin concentrations can help to improve the detection of CyanoHAB events as these pigments are commonly used in previous studies and can easily be validated by optical satellite remote sensing instruments. Consequently, a methodology was established that would allow the collected raw data to be converted to remote sensing reflectances for analysis. This was done first by conducting preliminary tests at various sites near UGA such as Lake Herrick, Memorial, and Chapman where Cyanosense 2.0 was deployed and took light intensity measurements from water bodies and a control panel reading at various points of interest. Afterwards, the same process was applied at six water bodies where CyanoHAB events were commonly found: Green Bay, Lake Erie, Lake Okeechobee, Clear Lake, the San Luis Reservoir, and Lake Pontchartrain.

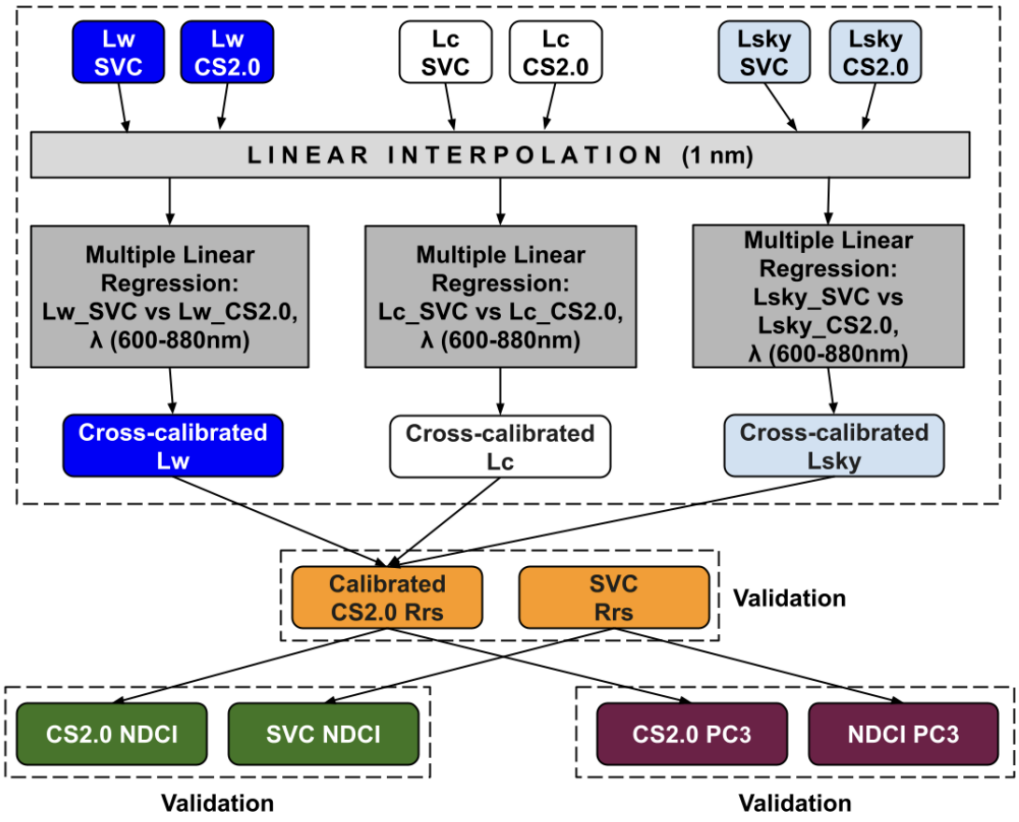

Figure 3.3: Cyanosense 2.0 Cross-Calibration Methodology

3.4 Cyanosense 2.0 Cross Calibration

Because the proposed architecture of Cyanosense 2.0 does not include the use of calibration panel, Cyanosense 2.0 was cross calibrated with another sensor system called the SVC HR-1024i (also referred to as the SuperGER). This was done by collecting results from the SuperGER and Cyanosense 2.0 simultaneously at the same sites obtained from Green Bay, Lake Erie, Lake Okeechobee, Clear Lake, the San Luis Reservoir, and Lake Pontchartrain (As part of this process, the path length, field of view, and viewing angles were kept the same for both sensor systems). The cross-calibration was done by first converting the data obtained for L_w, L_{sky}, and L_c from the SuperGER and Cyanosense into measurements measured at each nanometer using linear interpolation. Afterwards, multiple linear regression was done on each of the L_w, L_{sky}, and L_c obtained from 600 to 880 nm as these are the wavelengths where the pigments used by cyanobacteria have highest absorption. Finally, the results for remote sensing reflectance, NDCI, and PC3 collected from Cyanosense 2.0 and the SuperGER are then validated against each other to identify whether the results are correlated and match the expected results.

3.5 CyanoTRACKER 2.0 Social Sensing Goals

To meet both the objectives set forth by this study for the social cloud, the following objectives were identified as being important:

1. Collect data from different sources and create data set using data fusion.
2. Develop automated process to extract data from various social media or new sources.
3. Preprocess data to improve data extraction.
4. Identify relevant and irrelevant data from social media and news sources. Remove the irrelevant data from the dataset before analysis.
5. Accurately identify locations of interests and relevance that are CyanoHAB locations.
6. Geolocate locations and use satellite imagery to verify bloom locations.

3.6 Proposed CyanoTRACKER 2.0 Social Sensing System Architecture

As part of the proposed social sensing for CyanoTRACKER 2.0, we choose three media sources and extract textual data from each of them using different technologies. We organize the data extracted from each of these three media sources into two datasets and extract relevant locations of interest for detecting CyanoHAB events.

3.6.1 Choosing Data Sources

In this section, we introduce the different media platforms used for data collection and provide details related to user engagement on each of the platforms.

Reddit

Reddit is social media platform created in 2015 that allows users to share videos, texts, or images with other users. It has seen massive growth in recent years and has users actively engage with the platform to view, post, explore, and/or vote on shared content shared through the platform ("What is Reddit", 2024). Reddit is organized via community channels known as subreddits and allows users to post and interact with posts within each of these community channels using a message board. Reddit is unique in comparison to other media platforms in that typically has large population of younger adults (18-35) using the platform (Olivia Sidoti, 2024). The majority of users also have some college education, identify as male, and/or live in the United States in comparison to other platforms. Finally, most users of the platform

exclusively use Reddit as their only social media platform for communication and trust the content from the platform for learning about news events and/or getting product recommendations. For this study, Reddit was chosen as one of the media sources for analysis due to its use of community-generated content that can help to determine trends and/or relevant events such as CyanoHAB events while also providing easy access to posts via their API.

Instagram

Instagram is a widely used social media platform created in 2010 that allows users to upload, edit, and share images and/or videos for other users to see. The platform allows users to interact with one another by providing content from other users to their feed or profile based on user activity, preferences, usage patterns, and/or permissions set by a user. The simple user interface and different features for engagement (such as Reels) has allowed the social media platform to become the third most popular media application in the world with over 1 billion users (Shepherd, 2024). Instagram is similar to other platforms as shown by survey of US adults in that it is popular with younger users and tends to have users from with some college education. (Gottfried, 2024) However, Instagram differs from other platforms in that has almost equal split in usage by males and female users across the world and has wider distribution of usage across countries with close to 90% of users residing outside the United States (Gottfried, 2024). Additionally, users of the platform tend to use it with other media platforms such as Facebook, YouTube, and TikTok ("Global Social Media Statistics", 2024). Instagram was chosen as one of the media sources for analysis in this study due to its use of micro-blogging for user posts and its promotion of hashtags to classify media posts during searches of the media platform.

Google News

Google News is a news aggregator service that provides worldwide news from a a variety of news outlets. It was officially released in 2006 and has since been widely used by a number of users. Its provides real-time news by aggregating news articles from a number of different sources, categorizing each of the aggregated news articles into variety categories using a system set up by Google, and allows users to search for relevant news articles based on provided search preferences and/or extracted information from previous searches. Google News was found to be have some credibility in a survey conducted of United States adults in February 22 (Watson, 2023). Additionally, the platform scans over thousands of sources and provides user-friendly content in over 50+ languages for its users to read (Filloux, 2013). Each of these unique traits allow the Google News to present relevant searches related to CyanoHAB events based on locations, search topics of interest, and relevant publishes for this study.

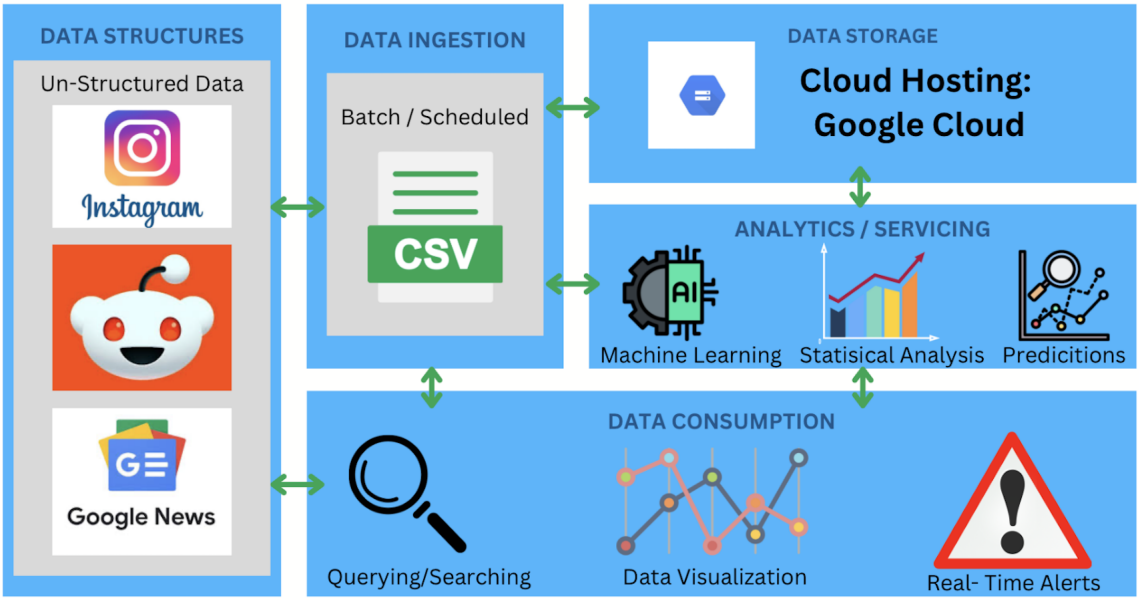

Figure 3.4: CyanoTRACKER 2.0 Social Sensing Proposed Architecture

3.7 Suspected CyanoHAB Event Location Extraction Process

The following section provides the framework for determining suspected CyanoHAB locations of relevance for this study. As part of this section, the data ingestion process is introduced where the different tools and data extraction procedure are introduced.

3.7.1 Dataset Ingestion Process

To extract data from each of the selected media sources into two datasets, a protocol was established that would allow the maximum amount of unstructured textual information from each media source using various tools. This subsection aims to highlight the tools and the protocol used in more detail. Afterwards, the location extraction process is introduced and provides information about the use of the spaCy NER model for the extraction of locations of interest to this study. Finally, the sections is concluded with explanation of the overall methodology used to develop the datasets for analysis and extract results discussed in the next chapter.

Reddit Data API

The Reddit's Data API is interface provided by Reddit that allows developers to access and/or edit Reddit data using software programming tools ("Reddit Data API Wiki", 2024). The API is accessible to anyone who abide by the terms of the API and Reddit's Developer Terms as well as provide authentication via a

registered OAth token and user agent (2024). Reddit's Data API allows developers to extract 100 queries (endpoints) per minute for each OAth token provided. Furthermore, developers can extract endpoints from Reddit using different criteria such as the specification of certain time periods, the use of different search terms, and/or information extracted from different sections of the platform such as a subreddits, posts, or comments.

Instagram Data Extraction

While data is extracted from Instagram via the Instagram Graph API, approval for using the API can take time. To obtain data as part of this study, Google's search engine was used to identify Instagram posts of relevance from various search queries. Afterwards, data from each of the Instagram posts provided from Google search engine were downloaded from Instagram using the curl command on the URL of each Instagram posts obtained. The curl command is command-line tool that is used to transfer data over different network protocols (Ramadhan, 2023). The curl command can be implemented in variety of ways but is implemented in this study through the subprocess library in Python.

Google News Feed

Content shown on Google News can be accessed through a method known as RSS Feeds (Rich Site Summary Feeds). The API provides content in the RSS format which is a collection of web files that can easily be read by computers through the use of XML-based formatting (Britannica, 2024). Google News Feeds allows for users and developers to access Google News data with automatic updates being provided as new articles are added onto the website. The data accessible from the Google News RSS Feeds include titles, timestamp, web URL link to the aforementioned news articles, and a brief description of the aforementioned news articles.

BeautifulSoup Library

BeautifulSoup is a library that allows developers to scrape website content written in HTML and XML format (Richardson, 2015). The library allows users to download all or portions of the content from website by loading the content into a BeautifulSoup constructor. Then, the content is converted into Unicode and back into readable HTML or XML content for developers to use. By using this library, content from news articles on Google News can be downloaded and used for analysis within the study.

Data Extraction Procedure

As part of the dataset creation process, a protocol was created that used the tools mentioned above to could create one dataset for analysis. To extract data from the three selected media sources, the following procedure is performed.

1. A search query is chosen from the following list of query keywords: 'algal+bloom', 'algae+bloom', 'toxic+algae', 'cyanobacteria', 'blue-green+algae', and 'red+tide'. These search terms were chosen

as they found most relevant to identifying CyanoHAB events from previous conducted as part of CyanoTRACKER 1.0 (Boddula et al., 2015).

2. The selected search query is inputted into software program that interfaces with Reddit's Data API. This allows one hundred data points to be collected with information about a Reddit post related to the search query such as its title, body text, and timestamp.

3. Next, the selected search query is inputted into software program that interfaces with the Google News RSS Feeds by downloading the title and timestamp of thirty Google News articles related to the search query. The web links extracted from the Google News RSS Feeds are then imported into the another software program that uses Beautiful Soup to download all of the article text.

4. Then, the selected search query is inputted into software program that interfaces with the Google News RSS Feeds and searches for related Instagram posts using the following search query keywords: 'instagram: [Selected Search Term]' (ex. instagram:algal+bloom) via Python. Twenty web links are then extracted from the Google News RSS Feeds and both the titles and timestamps of each of these web links is extracted. Finally, the links extracted from the Google News RSS Feeds have their web content downloaded using the curl command. From running this command, the text content of the Instagram post is extracted from each of the web links.

5. All extracted posts are then combined together are and labeled with the media source they were extracted from.

6. The process is repeated for the rest of the list of query keywords until all query keywords have been created. A data file is then created containing all the information extracted in one dataset.

3.7.2 Location Extraction Process

To extract relevant water bodies of interest, the text needs to be analyzed to identify words of importance. The use of named entity recognition can help identify and classify locations of relevance by tagging each of the words as entities of interest. This is achieved in this work by using the Spacy NER models.

spaCy NER

spaCy is free, open-source library for assisting in performing natural language processing (NLP) in Python ("spaCy 101: Everything you need to know", n.d.). spaCy allows developers to use production-level software for natural language processing tasks such as information extraction, natural language understanding, etc. One of the features provided by spaCy is the use of named entity recognition (NER) to label objects within imported text by converting the text into tokens. The SpaCy NER model allows developers to quickly assign labels to extracted tokens from imported text inserted in the model. The tokens are labeled based on tags such as GPE type (for identifying countries, states, etc.) or PERSON type (for identifying real or fake people). In particular, this study looked at tagging tokens extracted from the text with the

FAC type and LOC types for extracting water bodies and/or nearby landmarks of interest for determining the occurrence of CyanoHAB event. These tags are done for all data points extracted within each data set and then added as another column for further analysis in the future.

3.7.3 CyanoTRACKER 2.0 Social Sensing Data Collection Methodology

For this work, the data and location extraction processes mentioned are used to analyze data collected on June 20th over the course of one year and three months. First, the datasets were developed by using the data extraction process mentioned above for the time range of one year when using the Reddit Data API, Google Search Engine, and Google News RSS Feeds. The results were then combined together to create the dataset with 900 data points in total. The same process was completed but done by selecting data over the time range of a month from the Reddit Data API and Google News RSS Feeds. Once both datasets were completed, each of the datasets were run through code where locations of interest were then extracted using the location extraction process running on UGA's Juncus server. Finally, the results extracted were added as separate columns for analysis afterwards.

Chapter 4

Results and Discussion

After the data collection process established in the previous chapter, this section aims to provide a detailed account of the results collected from Cyanosense 2.0 and the social sensing approach as part of the CyanoTRACKER 2.0 Social Cloud.

4.1 Cyanosense 2.0 Architecture Preliminary Test Results

As part of the proposed architecture design of Cyanosense 2.0, tests were completed to help in determining whether Cyanosense 2.0 achieved some of the aforementioned goals established in the previous chapter. The following subsections highlight each of these tests in more detail.

4.1.1 Iridium Satellite Communications Test

As part of the proposed architecture, the use of ROCKBLOCK Iridium modem allows Cyanosense 2.0 to interact with the Iridium satellite network and send data autonomously from any place on Earth where satellite connectivity was present. To ensure that this process works as intended, a test was conducted that transmitted data from Lake Herrick to a local Iridium account that the CyanoTRACKER team could access. The confirmed messages on the Iridium platform along with software code to download the data are shown below as proof that this test was successful.

Date Time (UTC)	Device	Direction	Payload	Length (Bytes)
23/Apr/2023 15:58:15	RockBLOCK 210580	MO	07e7085d095e0a870bb30c860d1a0de10e6b0ebd0eef0ed40e9a0e260da10d800d7c0dc30e040e310e600ee90f210f6e0...	320
23/Apr/2023 15:58:02	RockBLOCK 210580	MO	0a9a0ad20b100bb00d170ebd1082117f124c132b13d814581476141213ce12df123211e111b611b9120e1203122112481...	320

Figure 4.1: Iridium Platform Transmission Messages

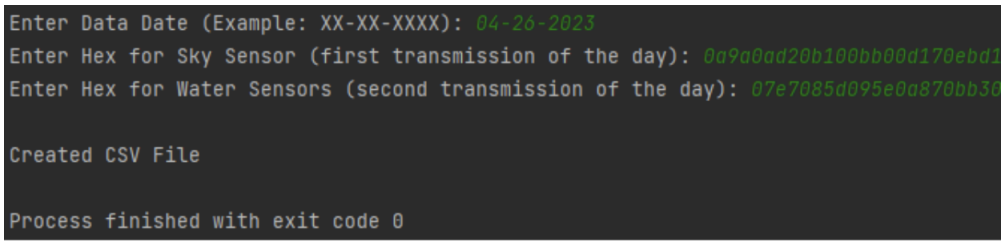

Figure 4.2: Python Program Interface for Extracting Iridium Results

4.1.2 Cost Breakdown

One of the most important considerations as part of the design was the cost of creating the Cyanosense 2.0 prototype. The following table below highlights the cost of each of the components used when building the architecture which is estimated to be about $1,300 (Note: This cost was computed in May 2023 and many components have since changed in value. Additionally, some components may no longer be available due to supply chain issues and/or manufacturer choice as well. Finally, costs for sending messages using the Iridium satellite network were not include in this estimate). Through this analysis, it was determined that Cyanosense 2.0 was able to cost close to 1/3 of the cost of Cyanosense 1.0. Thus, the system was determined to be more affordable for potential users of the system across the world.

4.1.3 Power Consumption Test

Tracking power consumption can help to ensure that net energy consumption was reduced by 20% each day. In order to ensure this goal was met, we conducted a test on the total power draw of the system per day for either software transmission mode used. If energy is created by the solar panel each day, the total battery gain of the system per day would be close to 810 mAh as shown in Figure 4.3. Additionally, if the system is not charged by the solar panel and continuously relies on the battery model used within the system, Cyanosense 2.0 can last up 25 days.

4.1.4 System Temperature Test

One of the goals for building Cyanosense 2.0 was to make the system more heat resistant in relation to changes in the internal system environment and for external environment near water bodies. To determine whether this goal was met, a temperature test was conducted to measure the temperature of the system when operational. This was done by attaching two K-Type thermocouples both inside and outside the sensor housing. A battery was then plugged into the system to simulate operation after which the system was placed on a wooden platform and placed in a certain manner to avoid shadows being cast during the test. Table 4.2 below show the results collected from on March 31st, 2023 to ensure that this objective was met.

Mode	Part	Current Consumption (mA)	Time in Mode (seconds/1 Day)	Charge Consumed (mA*s)
Idle				
	ESP32-S2	1		
	C12880mA (QTY: 2)	0		
	RockBLOCK Irdium 9603 Moden	0.1		
	TLP3556A Photorelay	0		
	Always on Battery	7		
	Total Idle:	*8.1*	*86099*	*697401.9*
Data Collection (Once Per Day)				
	ESP32-S2	32		
	C12880mA (QTY: 2)	40		
	RockBLOCK Irdium 9603 Moden	0.1		
	TLP3556A Photorelay	10		
	Always on Battery	7		
	Total Data Collection:	*89.1*	*1*	*89.1*
Transmission (1 Data Cycle Per Day)				
	ESP32-S2	32		
	C12880mA (QTY: 2)	0		
	RockBLOCK Irdium 9603 Moden	145		
	TLP3556A Photorelay	0		
	Always on Battery	7		
	Total Transmission:	*184*	*300*	*55200*
	Total:	*281.2*	*86400*	*752691*
	Average Current Consumption Per Day (mAh):	*8.71*	*Average Power Per Day (mW):*	43.56

Battery				
Model	mAh	Battery Life (Days) No Recharge	Battery Draw Per Day (mAh)	
Voltaic V25	6400	30.61	209.08	
Solar Panel				
Model	Volts	Peak Watts	mA	Battery Return per Day (mAh)
Voltaic P102C	6	2.2	340	1020
			Battery Return Per Day (mAh)	1020
			Battery Draw Per Day (mAh)	209.08
			Battery Gain Per Day (mAh):	810.92

Figure 4.3: Cyanosense 2.0 Power Draw Table

Table 4.1: Cost breakdown of Cyanosense 2.0 Individual Components

Part(s)	Cost (USD)
Hamamatsu Spectrometer (2)	$700.00
Photorelay	$4.20
ROCKBLOCK Iridium Satellite Modem	$299.95
ROCKBLOCK Adaptor Cable	$4.95
Iridium Passive Antenna	$64.95
Iridium Antenna Cable	$3.63
ESP32-S2 Saola Dev Kit	$14.50
Real Time Clock (RTC)	$3.80
Switch	$0.09
Voltaic 2 Watt Solar Charger Kit	$59.99
Micro USB to USB Cable	$4.99
Breadboards (2)	$17.98
Jumper Wires	$30.96
Electronics Housing Box	$19.99
Sensor Housing Box and Cover	$11.26
Conduit Components	$3.05
Mechanical Fasteners	$12.06
Microscope Cover Glass Slips	$8.99
Silicone Sealant (Recommended)	$14.99
Miscellaneous	$20.00
TOTAL	**$1,300.33**

4.1.5 Raw Sensor Data from Cyanosense 2.0

An important component of Cyanosense 2.0 was collecting the radiometric measurements from both sensors during testing to ensure that the system was operational when used at various water bodies. Figure 4.4 show a sample measurement taken from both sensors on Cyanosense 2.0 as part of the testing of proposed architecture.

4.2 Cyanosense 2.0 Cross Calibration Results and Analysis

As part of the analysis conducted with the Cyanosense 2.0 system, the system was cross-calibrated SVC HR-1024i (or Super GER) in order to establish the relationship between the two systems. This was done by removing data that did not correspond between both systems, plotting the corresponding data points

Table 4.2: Temperature Test of Cyanosense 2.0

Time (EST)	Sensor Housing Temperature (°C)	Outside Ambient Temperature (°C)	Outside Conditions
10:30 AM	24.0	18.9	Cloudy
11:00 AM	25.2	19.5	Cloudy
11:30 AM	26.9	20.9	Cloudy
12:00 PM	29.8	22.2	Partially Sunny
12:30 PM	31.5	23.0	Partially Sunny
1:00 PM	32.6	23.7	Partially Sunny
1:30 PM	30.1	24.4	Cloudy
2:00 PM	31.5	24.6	Partially Sunny
2:30 PM	32.1	24.8	Partially Sunny
3:00 PM	28.9	24.3	Cloudy
3:30 PM	26.4	24.0	Cloudy
4:00 PM	32.1	23.9	Partially Sunny
4:30 PM	27.1	23.5	Cloudy
5:00 PM	31.5	23.2	Partially Sunny
5:30 PM	29.0	23.0	Cloudy
6:00 PM	27.0	22.7	Cloudy

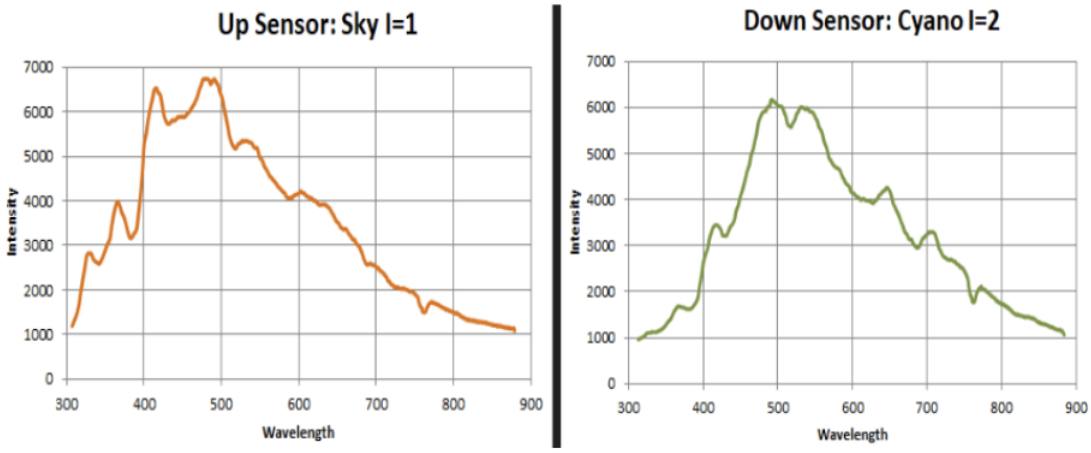

Figure 4.4: Sample Spectra Data from Cyanosense 2.0 Sensors

and metrics with each other, and creating a line of best fit to determine the correlation between the two systems. For example, the L_w, L_c, and L_{sky} for each systems was cross calibrated after removing any points that were not shared between the two systems. Consequently, there were 8,711 data points shared between the two sensor systems that were used to determine both the cross calibration and validation of

each of the sensor systems. From the results, it was determined that the L_w, L_{sky}, and L_{panel} showed strong agreement in terms of the cross calibration and validation correlations extracted. Additionally, the remote sensing reflectance were determined for each of the extracted points and had a very high cross validation correlation between results. Finally, the NDCI and PC3 formulas shown in Equations 2 and 3 were used to cross-validate both sensor systems by using 31 corresponding metrics between the two systems. Both validation correlations for each formula were found to be high with values of 0.84 and 0.81 for R^2, respectively (Note: Vicarious calibration was used to correct the original validation correlation determined for NDCI as shown in Figure 4.9).

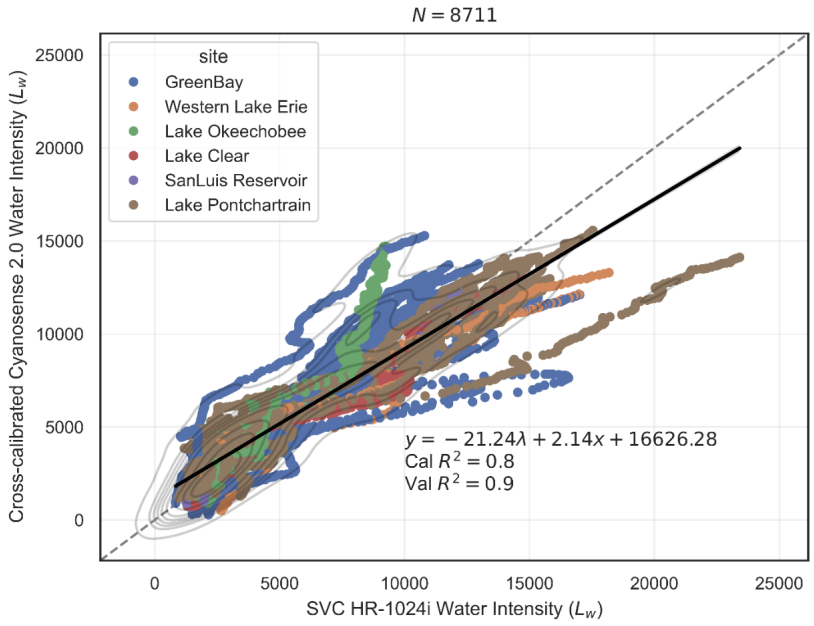

Figure 4.5: Cyanosense 2.0 and SVC HR-1024i Lw Cross Calibration

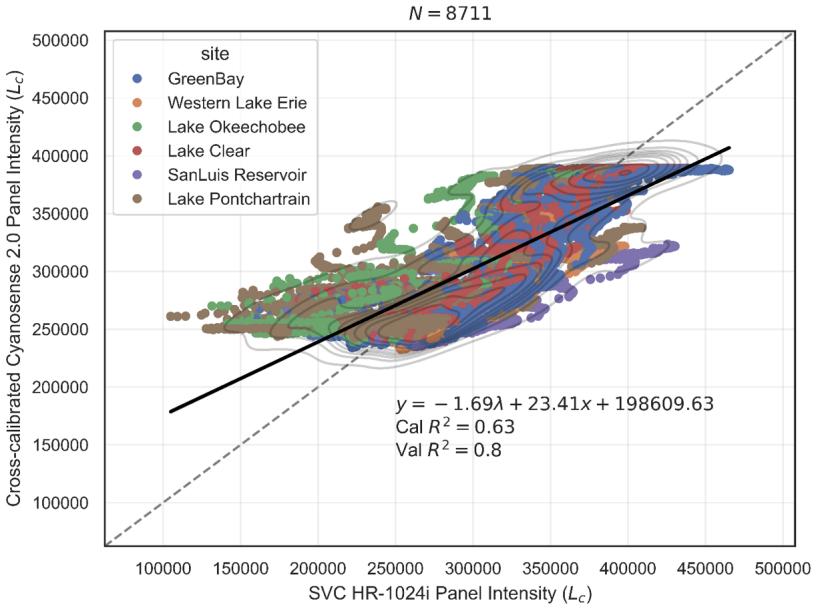

Figure 4.6: Cyanosense 2.0 and SVC HR-1024i Lc Cross Calibration

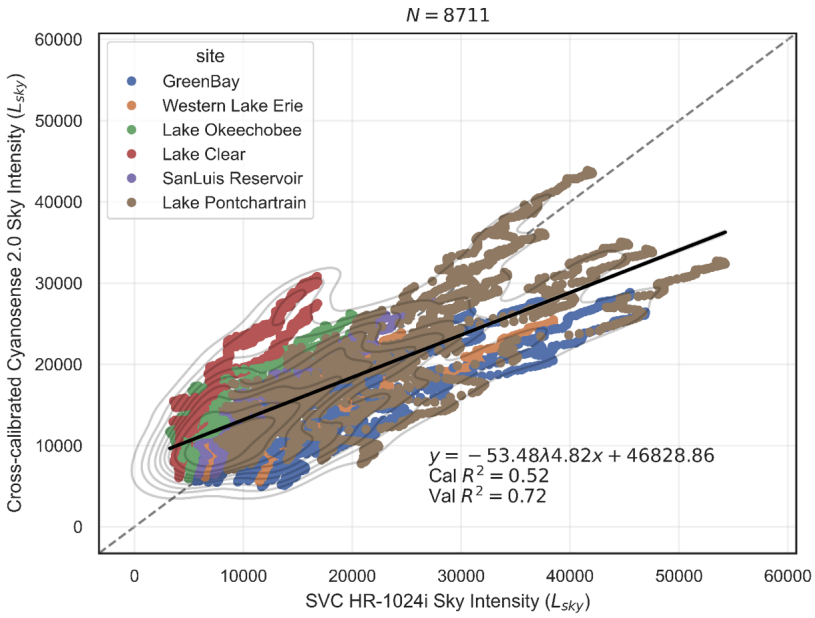

Figure 4.7: Cyanosense 2.0 and SVC HR-1024i Lsky Cross Calibration

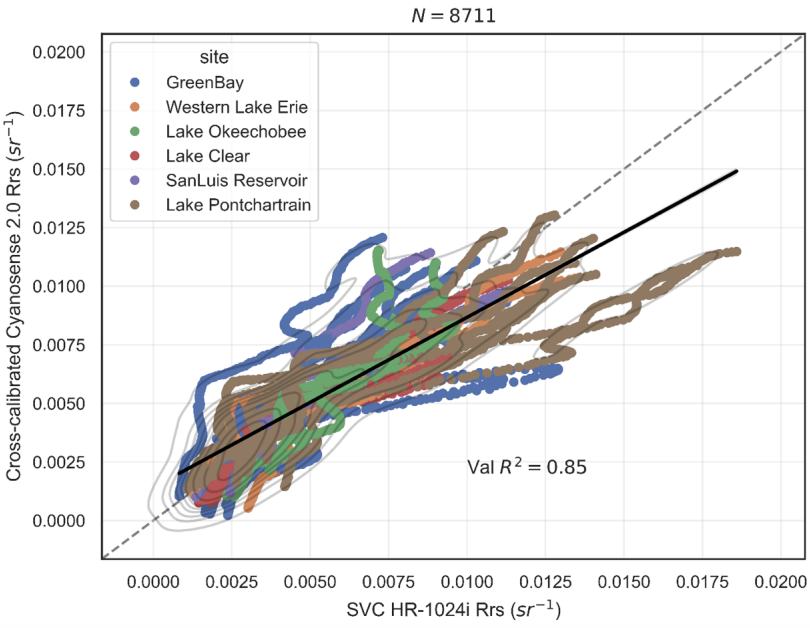

Figure 4.8: Cyanosense 2.0 and SVC HR-1024i Remote Sensing Reflectance Cross Validation

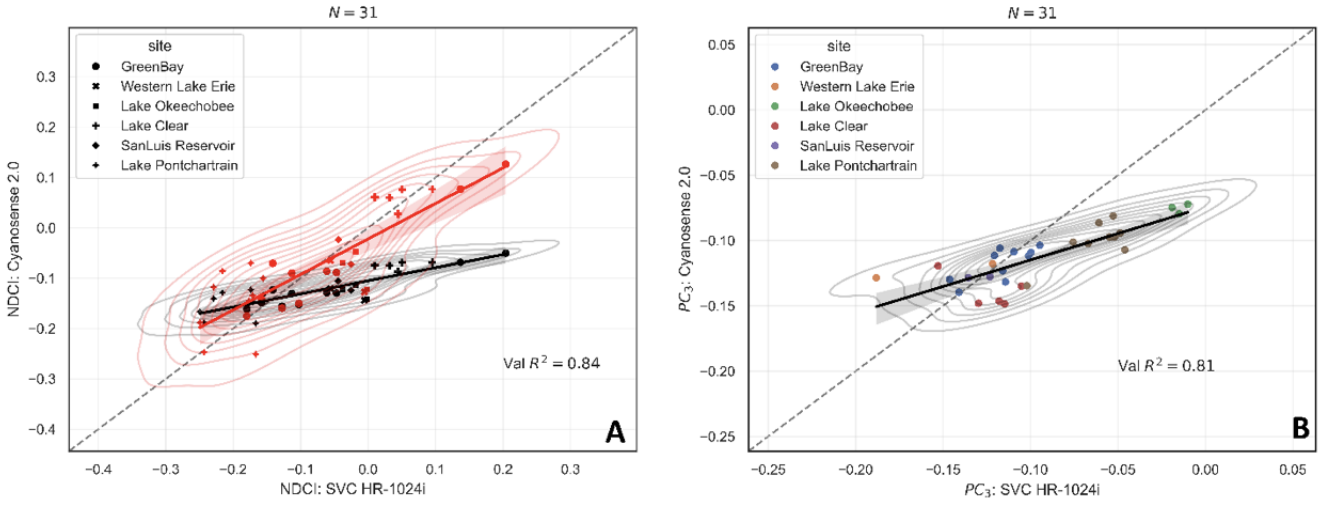

Figure 4.9: Validation of NDCI and PC_3 derived from Cyanosense 2.0 against NDCI and PC_3 derived from SVC's HR-1024i. A vicarious calibration was performed to achieve a linear fit for NDCI.

4.3 CyanoTRACKER 2.0 Social Cloud Data Collection Results

As part of the proposed CyanoTRACKER 2.0 Social Cloud study within this work, the development of two datasets were completed that collected data from each of the three selected media sources for a time period of one year and three months. An example of this dataset is provided here to show the raw data collected from each of these datasets.

https://www.palmbea	Toxic blue-green algae thickening in Calc		15 Jul 2023 13:32:1	reddit
https://i.redd.it/zs7s2	Please Help. I have a	I'm pretty sure it's cau	05 Oct 2023 9:36:30	reddit
https://www.reddit.co	Nobody know how to	I can't find any algae	02 Jul 2023 10:59:1	reddit
https://newatlas.com/	Plasma tech transforms blue-green algae i		11 Oct 2023 11:59:5	reddit
https://www.instagran	Instagram - Instagran	DUA LIPA on Instagr	12 Jun 2024 8:43:33	instagram
https://www.instagran	DUA LIPA | PULA!!!!!	DUA LIPA on Instagr	10 Jun 2024 10:10:0	instagram
https://www.instagran	Anastasia Karanikola	Anastasia Karanikola	10 Jun 2024 3:30:57	instagram
https://www.instagran	CA Dept. of Water Re	CA Dept. of Water Re #california #WaterSa Photo alt text: Hamrf	30 May 2024 7:00:0(	instagram
https://www.instagran	Simcoe Muskoka Dis	Simcoe Muskoka Dis	10 Jun 2024 21:25:2	instagram
https://www.instagran	PDSA | Blue-green a	PDSA 🐾 on Listen to PDSA Vet, [Visual description: 󠀠&#xe002(󠀠&#xe002(󠀠&#xe002(󠀠&#xe002(#PDSA #BlueGreenA	23 May 2024 8:01:0	instagram
https://www.instagran	CA Dept. of Water Re	CA Dept. of Water Re #california #WaterSa Photo alt text: Hamrf	31 May 2024 7:00:0(	instagram

Figure 4.10: CyanoTRACKER 2.0 Social Sensing Sample Dataset

4.4 CyanoTRACKER 2.0 Social Sensing Analysis Results

As part of the analysis of the proposed CyanoTRACKER social sensing, the top and overall results from the extracted locations were determined by counting the number of unique mentions of each of the locations. For the top mentions, both datasets had similar results present within their results. However, the top results for the dataset extracted for a period of one year had a number of unique water bodies mentioned unlike the dataset extracted for a period of three months. On the other hand, the opposite effect was determined in the word clouds generated from the overall results from each of the datasets as there were more unique mentions of locations of interest in the three month dataset than one taken over the course of a year. Overall, the results highlight differences in the results from data extraction for each of time period as well as a number of errors present within the results that need to be fixed.

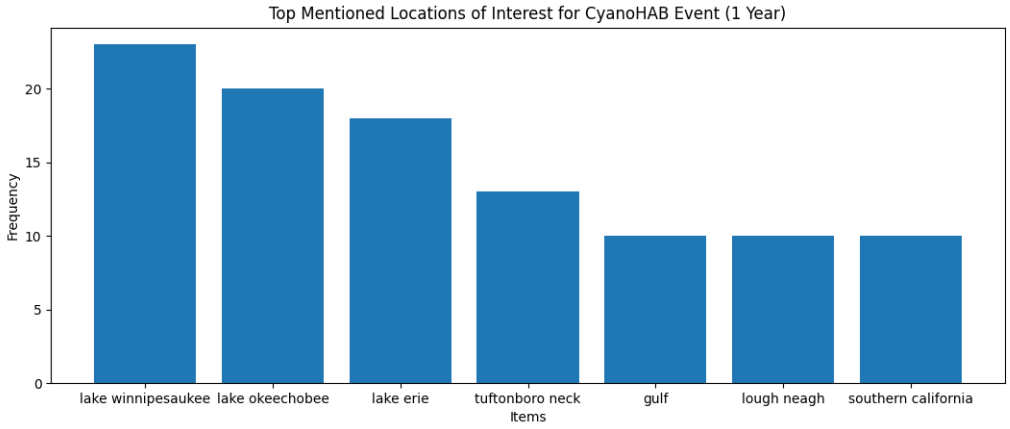

Figure 4.11: CyanoTRACKER 2.0 Social Sensing Most Mentioned Locations for 1 Year Time Period

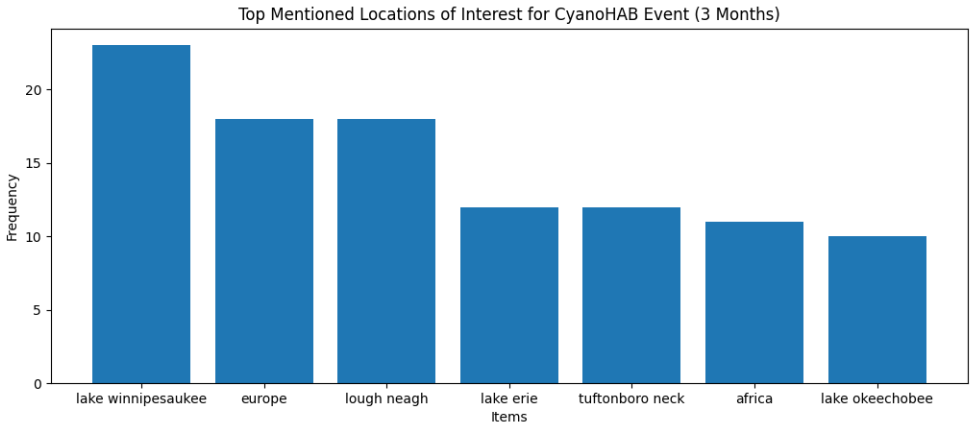

Figure 4.12: CyanoTRACKER 2.0 Social Sensing Most Mentioned Locations for 3 Month Time Period

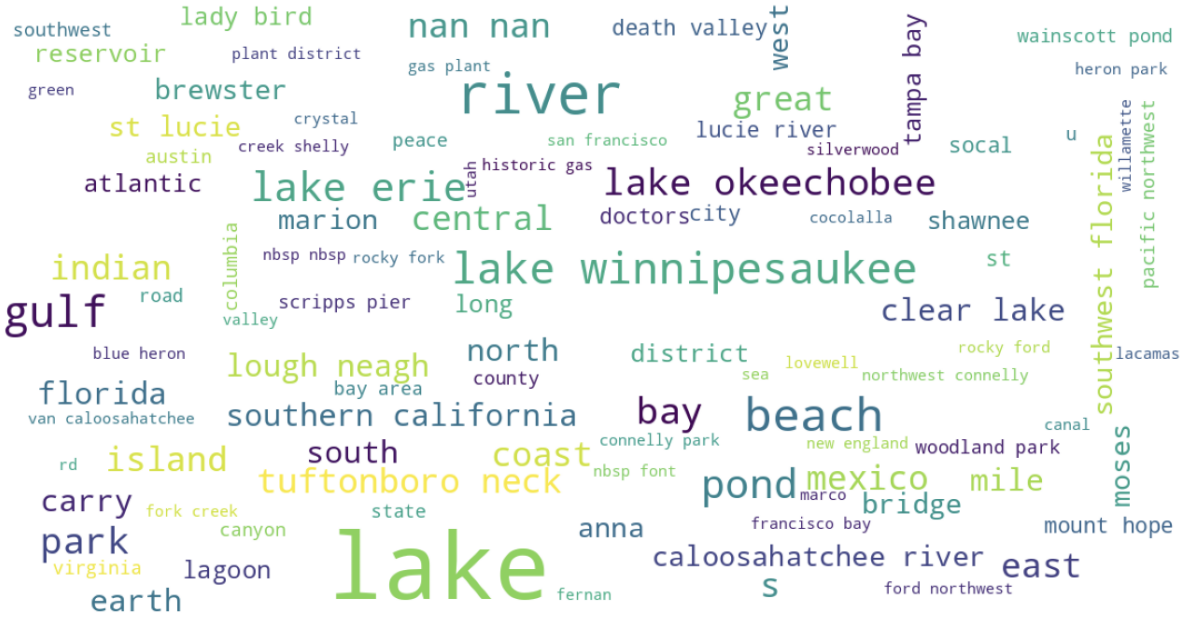

Figure 4.13: CyanoTRACKER 2.0 Social Sensing Word Cloud for 1 Year Time Period

Figure 4.14: CyanoTRACKER 2.0 Social Sensing Word Cloud for 3 Month Time Period

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this research work, we introduced cyanobacteria and the negative impact that they can have on different processes around the world when they grow uncontrollably and form cyanobacterial harmful algal blooms (also known as CyanoHABs). We then introduced the CyanoTRACKER project and talk about how the infrastructure used in this project can be improved in a new proposed cyber-physical infrastructure called CyanoTRACKER 2.0.

In Chapter 2, we explored the different cyber-physical infrastructure approaches that have been developed for detecting CyanoHABs and the different considerations that are taken with the development of both on-site monitoring and approaches in citizen science and/or social sensing for detecting CyanoHAB events. This chapter also helped to highlight the work conducted as part of CyanoTRACKER 1.0 Social Cloud and Cyanosense 1.0.

In Chapter 3, we present the goals for Cyanosense 2.0 and CyanoTRACKER 2.0 Social Cloud to solve the research questions posed at the beginning of this work. We then present the methodology for Cyanosense 2.0 and CyanoTRACKER 2.0 Social Cloud where we review the overall proposed architectures for both components and the tools used to implement both architectures. Afterwards, we introduce the overall methodologies used by each of the components to develop results for analysis later in this work.

Finally, we conclude by discussing the results that were attained from Cyanosense 2.0 and the CyanoTRACKER 2.0 Social Cloud via social sensing. We then highlight the success of the overall work as well as issues that affected the results collected during analysis.

5.2 Reflections/Future Work

Overall, this work was able to demonstrate how CyanoTRACKER 1.0 was improved to provide accurate detection results for CyanoHAB events, reduced turnaround time for determining results, and provided solutions that can be expanded for use throughout the rest of this world.

Some improvements that could help to improve the results collected by Cyanosense 2.0 in the future could include the implementation of electrical design on a PCB board to secure electrical components while allowing to be system to be mass produced in the future. Additionally, the use of spectroradiometers of better quality would also help to reduce issues with precision, accuracy, and saturation of data readings from the sensors within a variety of different field environments. Lastly, the software that extracts the different intensity and remote sensing reflectance readings can be improved to automatically identify any erroneous readings and/or specific trends within data through the use of machine learning or improved data analytical methods.

Moreover, the social sensing approach developed within this work could also be improved through the use of data from pictures and/or videos collected as part of further studies. This extracted content could then be used to train computer vision models that can help to improve identification of CyanoHAB events from interested users. Additionally, extracted textual data could also be improved by ensuring that the data extraction process follows protocols regarding the interested time range for data collection and/or places of interest. Lastly, the use of data from other media sources such as YouTube, TikTok, Facebook, Snapchat, or open-sourced databases should be considered for inclusion in the data collection process for future studies in order to determine variations within the collected data while also using data sources with more users and/or interest in recent years.

Finally, coordinating the use of satellite remote sensing and/or validation could help both components developed within this work. Once a framework is established, the satellite imagery collected will help to validate the results collected from Cyanosense 2.0 and social sensing of selected media sources in the future. This can then utilize the data collected to substantiate remote sensing results obtained from optical satellite remote sensing platforms.

Bibliography

(2024, June). https://support.reddithelp.com/hc/en-us/articles/14945211791892-Developer-Platform-Accessing-Reddit-Data

About habsos. (2024). https://habsos.noaa.gov/about

Aguilera, A., Almanza, V., Haakonsson, S., Palacio, H., Benitez Rodas, G. A., Barros, M. U., Capelo-Neto, J., Urrutia, R., Aubriot, L., & Bonilla, S. (2023). Cyanobacterial bloom monitoring and assessment in latin america. *Harmful Algae*, *125*, 102429. https://doi.org/https://doi.org/10.1016/j.hal.2023.102429

AquaRealTime, I. (2024). Who we are and what we do.

Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. P. (2018). Social sensing of floods in the uk. *PLOS ONE*, *13*(1), 1–18. https://doi.org/10.1371/journal.pone.0189327

Bettina, S., Gugger, M., & Donoghue, P. (2015). Cyanobacteria and the great oxidation event: Evidence from genes and fossils. *Palaeontology*, *58*. https://doi.org/10.1111/pala.12178

Boddula, V., Joshi, A., Ramaswamy, L., & Mishra, D. (2015). Harnessing social media for environmental sustainability: A measurement study on harmful algal blooms. *2015 IEEE Conference on Collaboration and Internet Computing (CIC)*, 176–183. https://doi.org/10.1109/CIC.2015.31

Boddula, V., Ramaswamy, L., & Mishra, D. (2017). Cyanosense: A wireless remote sensor system using raspberry-pi and arduino with application to algal bloom. *2017 IEEE International Conference on AI Mobile Services (AIMS)*, 85–88. https://doi.org/10.1109/AIMS.2017.19

Britannica, T. E. o. E. (2024). Rss.

Chanda Das, M. (2017). *Incentivization for citizen science: A case study on monitoring cyanobacterial harmful algal blooms*.

Chen, Y., Randriambelonoro, M., Geissbuhler, A., & Pu, P. (2016). Social incentives in pervasive fitness apps for obese and diabetic patients. *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, 245–248. https://doi.org/10.1145/2818052.2869093

Citizen science. (2024). https://scistarter.org/citizen-science

Corporation, H. (2024). Cmos linear image sensor.

Council), I. (T. R. (2020). Strategies for preventing and managing harmful cyanobacterial blooms (hcb-1). www.itrcweb.org

Cyanobacteria monitoring collaborative. (2024). https://www.worcesterma.gov/sustainability-resilience/recreational-waters/cyanobacteria

Cyber-physical systems (cps). (2023).

Filloux, F. (2013). Google news: The secret sauce. https://www.theguardian.com/technology/2013/feb/25/1

Galesic, M., Bruine de Bruin, W., Dalege, J., Feld, S., Kreuter, F., Olsson, H., Prelec, D., Stein, D., & van der Does, T. (2021). Human social sensing is an untapped resource for computational social science. *Nature*, *595*, 1–9. https://doi.org/10.1038/s41586-021-03649-2

Gandola, E., Antonioli, M., Traficante, A., Franceschini, S., Scardi, M., & Congestri, R. (2016). Acqua: Automated cyanobacterial quantification algorithm for toxic filamentous genera using spline curves, pattern recognition and machine learning. *Journal of Microbiological Methods*, *124*, 48–56. https://doi.org/https://doi.org/10.1016/j.mimet.2016.03.007

Garrett Bartelt, J. Y., & Hondzo, M. (2024). Remote cyanobacteria detection by multispectral drone imagery. *Lake and Reservoir Management*, *40*(3), 236–247. https://doi.org/10.1080/10402381.2024.2341250

GeeksforGeeks. (2021, November). Understanding semantic analysis - nlp. https://www.geeksforgeeks.org/understanding-semantic-analysis-nlp/

GeeksforGeeks. (2024, July). What is morphological analysis in natural language processing (nlp)? https://www.geeksforgeeks.org/morphological-analysis-in-nlp/

Global social media statistics. (2024). https://datareportal.com/social-media-users

Gottfried, J. (2024, January). Americans' social media use. https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/

Great lakes harmful algal blooms (habs) and hypoxia. (n.d.). NOAA % 20 - %20Great % 20Lakes % 20Enviornmental%20Research%20Laboratory

Han, K., Graham, E. A., Vassallo, D., & Estrin, D. (2011). Enhancing motivation in a mobile participatory sensing project through gaming. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 1443–1448. https://doi.org/10.1109/PASSAT/SocialCom.2011.113

Harrison, C. J., & Sidey-Gibbons, C. J. (2021). Machine learning in medicine: A practical introduction to natural language processing. *BMC Medical Research Methodology*, *21*(1), 158. https://doi.org/10.1186/s12874-021-01347-1

Jadhav, A. C. (2018). *Towards an effective approach for cyanobacteria affected locations extraction from news feeds*.

Johnasen, R. A., Katzenmeyer, A. W., Reif, M. K., & Pokrzywinski, K. L. (2023, July). A review of sensor-based approaches for monitoring rapid response treatments of cyanohabs.

Joshi, A. R. (2015). *Study of microblog activity on cyanobacterial harmful algal blooms*.

Kraft, K., Seppälä, J., Hällfors, H., Suikkanen, S., Ylöstalo, P., Anglès, S., Kielosto, S., Kuosa, H., Laakso, L., Honkanen, M., Lehtinen, S., Oja, J., & Tamminen, T. (2021). First application of ifcb high-frequency imaging-in-flow cytometry to investigate bloom-forming filamentous cyanobacteria in the baltic sea. *Frontiers in Marine Science*, *8*. https://doi.org/10.3389/fmars.2021.594144

Malthus, T. J., Ohmsen, R., & Woerd, H. J. v. d. (2020). An evaluation of citizen science smartphone apps for inland water quality assessment. *Remote Sensing*, *12*(10). https://doi.org/10.3390/rs12101578

Maniyar, C. B., Kumar, A., & Mishra, D. R. (2022). Continuous and synoptic assessment of indian inland waters for harmful algae blooms. *Harmful Algae*, *111*, 102160. https://doi.org/https://doi.org/10.1016/j.hal.2021.102160

Mardones, J. I., Paredes-Mella, J., Flores-Leñero, A., Yarimizu, K., Godoy, M., Artal, O., Corredor-Acosta, A., Marcus, L., Cascales, E., Pablo Espinoza, J., Norambuena, L., Garreaud, R. D., González, H. E., & Iriarte, J. L. (2023). Extreme harmful algal blooms, climate change, and potential risk of eutrophication in patagonian fjords: Insights from an exceptional heterosigma akashiwo fish-killing event. *Progress in Oceanography*, *210*, 102921. https://doi.org/https://doi.org/10.1016/j.pocean.2022.102921

Miller, T., Tarpey, W., Nuese, J., & Smith, M. (2022). Real-time monitoring of cyanobacterial harmful algal blooms with the panther buoy. *ACS EST Water*, *2*. https://doi.org/10.1021/acsestwater.2c00072

Mishra, D. R., Kumar, A., Ramaswamy, L., Boddula, V. K., Das, M. C., Page, B. P., & Weber, S. J. (2020). Cyanotracker: A cloud-based integrated multi-platform architecture for global observation of cyanobacterial harmful algal blooms. *Harmful Algae*, *96*, 101828. https://doi.org/https://doi.org/10.1016/j.hal.2020.101828

Mishra, S., & Mishra, D. (2014). A novel remote sensing algorithm to quantify phycocyanin in cyanobacterial algal blooms. *Environmental Research Letters*, *9*(11), 114003.

of Health, V. D. (n.d.). Cyanobacteria (blue-green algae) tracker. https://www.healthvermont.gov/environment/tracking/cyanobacteria-blue-green-algae-tracker

O'Farrell, I., Motta, C., Forastier, M., Polla, W., Otaño, S., Meichtry, N., Devercelli, M., & Lombardo, R. (2019). Ecological meta-analysis of bloom-forming planktonic cyanobacteria in argentina. *Harmful Algae*, *83*, 1–13. https://doi.org/https://doi.org/10.1016/j.hal.2019.01.004

Ogle, D. (2024). Debris tracker. https://debristracker.org/

Olivia Sidoti, e. a. (2024, January). Social media fact sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/

Ottinger, G. (2022). Becoming infrastructure: Integrating citizen science into disaster response and prevention. *Citizen Science: Theory and Practice*. https://doi.org/10.5334/cstp.409

Patil, P. P. (2018). *Classification and location extraction of harmful algal blooms from microblogs*.

Presa Reyes, M., Bogosian, B., Schonhoff, B., Jerauld, C., Moreyra, C., Gardinali, P., & Chen, S.-C. (2020). A water quality research platform for the near-real-time buoy sensor data, 287–294. https://doi.org/10.1109/IRI49571.2020.00048

Programme, U. N. E. (2024-01). The role of remote sensing and social research in monitoring the environmental impact of refugee/internally displaced persons camps - foresight brief no. 032 january 2024. https://wedocs.unep.org/20.500.11822/44774

Ramadhan, H. (2023, December). How to use curl in python (and its alternative). https://serpapi.com/blog/python-curl-and-alternative/

Rashid, M. T., Wei, N., & Wang, D. (2023). A survey on social-physical sensing: An emerging sensing paradigm that explores the collective intelligence of humans and machines. *Collective Intelligence*, *2*(2), 26339137231170825. https://doi.org/10.1177/26339137231170825

Reddit data api wiki. (2024). https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki

Richardson, L. (2015). Beautiful soup documentation. https://beautiful-soup-4.readthedocs.io/en/latest/

Scholin, C., Doucette, G., Jensen, S., Roman, B., Pargett, D., III, R., Preston, C., Jones, W., Feldman, J., Everlove, C., Harris, A., Alvarado, N., Massion, E., Birch, J., Greenfield, D., Wheeler, K., Vrijenhoek, R., Mikulski, C., & Jones, K. (2009). Remote detection of marine microbes, small invertebrates, harmful algae, and biotoxins using the environmental sample processor (esp). *Oceanography*, *22*, 158–167. https://doi.org/10.5670/oceanog.2009.46

Shepherd, J. (2024). 25 essential instagram statistics you need to know in 2024. https://thesocialshepherd.com/blog/instagram-statistics

Soltani, S., Ferlian, O., Eisenhauer, N., Feilhauer, H., & Kattenborn, T. (2023). From simple labels to semantic image segmentation: Leveraging citizen science plant photographs for tree species mapping in drone imagery. *EGUsphere*, *2023*, 1–37. https://doi.org/10.5194/egusphere-2023-2576

Sonic, L. (2024). Monitor water quality and algae with lg sonic monitoring-buoy.

Spacy 101: Everything you need to know. (n.d.). https://spacy.io/usage/spacy-101

Srivastava, A., Singh, S., Ahn, C.-Y., Oh, H.-M., & Asthana, R. K. (2013). Monitoring approaches for a toxic cyanobacterial bloom [PMID: 23865979]. *Environmental Science & Technology*, *47*(16), 8999–9013. https://doi.org/10.1021/es401245k

Tewari, M., Kishtawal, C. M., Moriarty, V. W., P. Ray, T. S., Zhang, L., Treinish, L., & Tewari, K. (2022). Improved seasonal prediction of harmful algal blooms in lake erie using large-scale climate indice. *Communications Earth Environment*, *3*, 195. https://www.nature.com/articles/s43247-022-00510-w.pdf

US Department of Commerce, N. (2024). Skywarn. https://www.weather.gov/skywarn/

Veerman, J., Kumar, A., & Mishra, D. R. (2022). Exceptional landscape-wide cyanobacteria bloom in okavango delta, botswana in 2020 coincided with a mass elephant die-off event. *Harmful Algae*, *111*, 102145. https://doi.org/https://doi.org/10.1016/j.hal.2021.102145

Wang, D., Szymanski, B. K., Abdelzaher, T., Ji, H., & Kaplan, L. (2019). The age of social sensing. *Computer*, *52*(1), 36–45. https://doi.org/10.1109/MC.2018.2890173

Watson, A. (2023, June). Google news credibility in the u.s. 2022. https://www.statista.com/statistics/1308030/google-news-credibility-in-the-united-states/

What is reddit. (2024). https://edu.gcfglobal.org/en/thenow/what-is-clickbait/1/

Willi, M., Pitman, R., Cardoso, A., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2018). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*. https://doi.org/10.1111/2041-210x.13099

Zarkar, R. (2022, January). Natural language processing. https://medium.com/@raghvendra.zarkar18/natural-language-processing-65f82c8dd7e0