REMOTE SENSING OF HARMFUL ALGAL BLOOMS:

DATA MODELING AND BUSINESS ASPECTS

by

GEORGE SKOUFIAS

A thesis submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE

in

Data Science

Thesis Supervisor: Dr. Gregory Yovanof

DEREE - The American College of Greece

2022

Abstract

The present work accomplished a spherical investigation of remote sensing techniques for harmful algal bloom (HAB) detection. The areas covered by the research are: (1) environmental science aspects of the problem, (2) devising a pipeline and a model for HAB prediction based on artificial neural networks (ANN) from acquired satellite imagery, (3) a high-level design for the system capable of integrating data from various types of sensors and (4) a preliminary business analysis for a HAB detection service proposed herein.

The modeling activity resulted in a robust pre-processing pipeline able to transform acquired satellite products (set of images) to feature vectors for model training purposes. The feature vector components were carefully selected according to literature review findings. The resulting ANN model showed indications of overfitting. Further analysis of the dataset though revealed the limitation of the current research owed to the fact that no in-situ sensor data for HAB detection were available.

The service was conceptualized as an integrator of HAB related environmental data collected from various type of sensors. Issues concerning the technical requirements of the service were analyzed with the scope of enabling future system architecture activities. The combination of satellite, drone and sensor in-situ data streams led this analysis to the selection of a micro-services based architecture with well identified core application components.

The businesses analysis for the proposed HAB service yielded a well-defined target market (i.e. governmental environment agencies). Solid hypotheses for the customer profile are formed and based on that, a viable business model is framed. A cost analysis investigation resulted in a very competitive marginal operating cost that amounts to \$0.42 per km^2 for HAB monitoring using the proposed service. This cost figure indicates that the service can be marketed at very competitive prices allowing for high profit margins and economies of scale.

Acknowledgements

Firstly, I would like to express my appreciation for my thesis supervisor, Dr. Gregory Yovanof, for his continuous guidance throughout this research. His inputs assisted me in developing many aspects of the present work and his previous experience on the matter was invaluable.

I would like also to thank Dr. George Drakakis for the consultation he provided with respect to image processing issues that I faced during my work.

Last, but certainly not least, I would like to thank my dear classmate, Ms. Eleni Ntokou for the constructive discussions we had on data modeling matters that helped me understand the limitations of the current work.

Contents

Project Scope, Aims and Objectives 1

Methodology 2

Introduction 4

Literature Review 6

Overview of HAB phenomenon 6

Severe HAB incidents and economic impact 8

HAB: Prevention, Remedy-Mitigation and Monitoring Practices 11

HAB Prevention 12

HAB Remedy-Mitigation 13

HAB Monitoring 18

HAB: Processing and Algorithms for Satellite Remote Sensing 28

Satellite Data HAB Modelling 30

Data Acquisition 30

Modelling Strategy and Feature Extraction 33

Dataset Used and Labeling 36

Processing Pipeline and ANN model 42

HAB Remote Sensing Service 50

Service Description and Features 50

Service Design Issues 53

HAB Service: Business Analysis 56

State of Market 56

Customer Profile and Business Modelling 59

HAB Service: Pricing and Cost Issues 63

Discussion 67

Conclusion and Future Work 70

References 72

Appendix A – Sentinel hub API Example Request Payload 78

Appendix B: Value Proposition Canvas (VPC) 79

Appendix C: Business Model Canvas (BMC) 80

Project Scope, Aims and Objectives

The scope of the present project focuses on the design and development of a novel Harmful Algal Bloom (HAB) detection and alertness system that will serve as an extra asset in the technological toolbox for the digital transformation of water monitoring activities.

The aim of the project is two-fold. Namely:

1. Design, develop and validate a minimum viable product for HAB sensing and alertness
2. Conduct a preliminary techno-economic analysis for the solution in question by considering all the current market parameters

The Objectives of the project analyzed per aim are as follows:

Aim #1 (Design, develop and validate a minimum viable product for HAB sensing and alertness):

1. Review literature for HAB phenomenon and relevant detection methods and models
2. Conduct MVP development activities for the system proposed in the current project
3. Conduct preliminary system design activities (requirements gathering, infrastructure etc)

Aim #2 (Conduct a preliminary techno-economic analysis for the solution in question by considering all the current market parameters)

1. Conduct product market research
2. Stakeholder analysis and business design activities
3. Develop pricing models

Methodology

Herein, the methodology followed in the project is going to be presented. The tactic here is to draw a viable methodology that uses as input the objectives presented in the previous chapter. The analysis of these objectives yields the steps that need to be taken towards their accomplishment. Thus, the methodological steps discussed here are presented in a per objective basis.

Objective #1 (Review literature for HAB phenomenon and relevant detection methods and models)

1. Review literature to understand the HAB phenomenon and its underlying causes
2. Report literature findings for severe incidents of HAB and understand its economic impact
3. Report on current remedy and/or mitigation practices
4. Outline major in-situ (on-site) chemical detection techniques
5. Discuss literature review findings on remote sensing of HAB
6. Present major algorithms found in literature utilized for remote sensing of HAB by focusing on the processing and classification pipeline of satellite imagery

Objective #2 (Conduct MVP development activities for the system proposed in the current project)

1. Finalize and present a pre-processing pipeline that is triggered by data acquisition of satellite imagery
2. Implement HAB classification algorithms
3. Evaluate HAB detection models to incorporate best performing algorithm in the MVP

Objective #3 (Conduct preliminary system design activities)

4. Translate system description to requirements (hardware, software and integrational)
5. Consider all necessary system components for real-time (or near real-time) HAB sensing and alertness

Objective #4 (Conduct market research)

6. Collect information regarding the size of the water management market

7. Conduct interviews with subject matter experts to identify: (a) potential end-buyers for the proposed HAB service and (b) future market trends

Objective #5 (Stakeholder analysis and business design)

8. Identify all the relevant stakeholders and economic buyers for the proposed solution
9. Perform stakeholder analysis by drawing value proposition canvases per stakeholder that outlines all their basic features such as customer tasks, pains and gains
10. Draw a wireframe for the business of the proposed HAB service based on the business model canvas framework

Objective #6 (Develop pricing models)

Perform pricing tasks for the developed MVP. Available methodologies are: **(i)** market research/competition, **(ii)** willingness-to-pay detection and **(iii)** cost analysis for the proposed system.

Introduction

Soil, air, and water are the three main natural resources that support life as well as economic growth. As such, all the necessary tools to steward these natural resources are welcome and considered essential for sustainable development and resource management. Nowadays, this scope is even of greater importance, and on a global scale, bearing in mind the aggravation of climate change and other related phenomena. Harmful algal blooms (HABs) are phenomena that fall in that category and adversely affect the water quality in coastal, pelagic, and freshwater regions.

Currently, environmental protection frameworks that are usually maintained by the public sector, strive to keep under close guard a wide range of phenomena with adverse repercussions to the environment and the quality of its resources.

Any tasks undertaken by the relevant stakeholders for the purpose of environmental protection aim to: (1) impose regulatory frameworks for the prevention of these phenomena, (2) install infrastructures with detection capabilities about their incidence and (3) acquire the necessary hardware and personnel for immediate mitigating actions. These tasks, beyond the fact that involve many factors and carry a significant amount of complexity in their design, deployment, and operation, usually require substantial financial resources to be kept alive. Thus, any opportunity for cost optimization of environmental protection systems and related infrastructures is beneficial for the environment, economy, and society. The same holds true for the emergence of new technologies able for disruption.

For the particular case of HABs, the current standard for detection practices includes regular water sampling and laboratory analysis as well as the installation of in-situ sensor arrays able to give real-time data regarding the levels of certain measurands that are shown to be indicative for the progression of the phenomenon. The effectiveness of these techniques is limited by their spatial and temporal resolution while they come with a significant financial cost.

A potential solution to this problem arises from the advent of mankind in space and the development of planetary exploration and observation technologies. Nowadays, man-made satellites can carry a multitude of sensors that can trace phenomena on the surface of the earth. These orbital observation stations, though extremely expensive to design, manufacture and deploy, at the end of the day perform such a wide variety of monitoring tasks that the marginal cost per observation scope is minimal. Besides, most of them are fully sponsored by multination collaboration of nation-states and the required financial resources come in the form of public investment. The satellite image products become available for the general public or to private organizations at a very low cost. HABs can also be benefited from the availability of earth observation data and the task of detecting the phenomenon can become more affordable in this manner.

Another set of recently developed and marketed technologies that might play an important role in HAB detection is drones and internet-of-things. Optical and/or chemical sensors mounted on unmanned air vehicles (UAVs) can be customized for water monitoring purposes and work in a collaborative network of systems with in-situ arrays to collect and stream multi-dimensional data to data collection systems for further computational analysis.

The present work investigates the combination of all these HAB monitoring and data acquisition techniques to arrive to a proposed system able to combine the inputs of all these datasources to deliver accurate and cost-effective predictions about HAB in a near 24/7 basis with its services being available to its customers over the internet in a Software-as-a-Service (SaaS) offering format.

Literature Review

In this chapter, critical information gathered during literature research is presented. The points that the research focuses on, and are mentioned in the methodology section, examine the HAB phenomenon from a natural science as well as economic points of view. The discussion includes also elements concerning remedy and mitigation techniques for HAB incidents as found in the literature. Finally, remote sensing techniques and algorithms utilized in HAB monitoring and early warning systems are presented.

Overview of HAB phenomenon

Harmful algal blooms (HABs) are characterized by intense growth rates of algae in seawater-coastal as well as freshwater-lake sites. This phenomenon is responsible for releasing environmental toxins having adverse effects on water quality, fishery as well as human health by means of indigestion of fishery and shellfish products that have been exposed during their lifecycle to the phenomenon [1]. In oceanic coastal line sites, the phenomenon usually manifests as a coloration of the sea surface that is commonly referred to as "red tide" whereas in lake sites the presence of HAB is evident by a shading green layer on the surface that causes oxygen depletion to the underlying water mass.

Algae are eukaryotic organisms that carry out their energy cycles either through photosynthesis (photoautotrophic) or from the metabolism of nutrient chemicals (chemoheterotrophic) that they capture from their environments [2]. The pigments contained in their cellular bodies (e.g. chlorophyl) are the reason for the coloration observed on the water surfaces that their colonies flourish. A wide range of cyanobacteria types as well as other eukaryotic species – such as Gymnodinium catenatum that is a paralytic toxin producer making its presence periodically during spring time in areas of the gulf of California and the pacific Mexican coast line [3] - are responsible for the onset of HAB phenomena.

Algal growth, that gives rise to the probability of HAB phenomena, may occur either naturally during the course of ageing of a water resource unit or as a byproduct of human activity combined with the absence of effective natural resource management practices (anthropogenic factor).

In the early 1940s, Mortimer [4] proposed the mechanism that transpires natural lake ageing and involves the release of phosphorus and nitrogen containing nutrients from lake sediments that are becoming more available in old water resources with low renewal rates. These nutrients in turn are key factors that favor eutrophication and consequent algal growth. As far as human activity is concerned, spillages and contamination events of water resources with byproducts of industrial and agricultural activities favor eutrophication that induce HAB. In January 2003, the environmental protection agency held a "roundtable discussion" reaching a set of consensual statements regarding the causes of the HAB phenomenon [5]. Amongst these it was commonly agreed that in recent years, when fertilizer use has risen to meet consumer demands, HAB incidents have increased and therefore nutrient management is imperative to control the phenomenon alongside with the implementation of prediction and detection techniques that were available at the time.

Another important anthropogenic factor that plays a key role to the apparent increase of HAB events is climate change. With the rising mean temperatures of aquatic ecosystems and erratic precipitation patterns, due to systemic global warming, algae formation phenomena are intensified reaching eutrophication conditions for maximal growth faster than ever before [6]. This mechanism is explained by the effect that climate change stressors have on the supply of nutrients that is key to HAB formation. Current climate change studies tend to include HABs as a climate change co-stressor in the attempt to build models that involve a multitude of environmental processes that are concurrent and coupled (i.e. affect each other).

The following figure helps to visualize the change in the aggression of the phenomenon over the past 47 years on a global scale. The dots/tracks in this map indicate the distribution of locations of reported HAB incidents for paralytic shellfish poisoning toxins on a global scale.

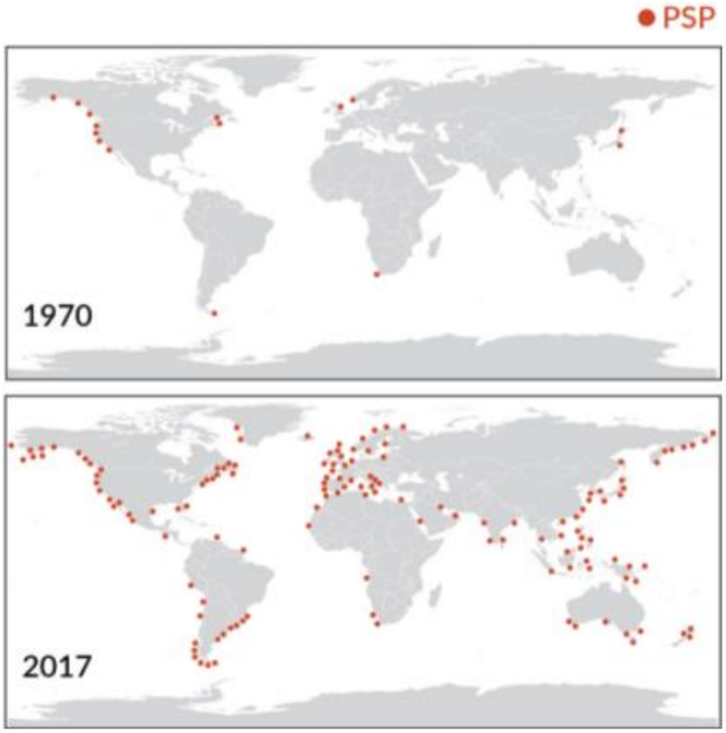


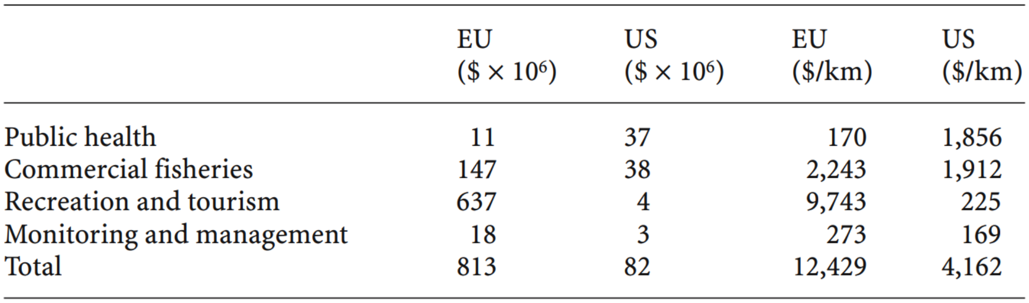
Figure 1: Global distribution of PSP toxins recorded in 2017 (bottom panel) compared with 1970 (top panel). (Credit: US National Office for Harmful Algal Blooms, Woods Hole Oceanographic Institution) [7]

Severe HAB incidents and economic impact

In this section, a series of severe HAB events is presented to serve as case studies that facilitate the evaluation, translated in financial terms, of the damage occurring for the environment, the economy and the society in the aftermath of the phenomenon.

As far as the societal impact is concerned, video literature was found containing testimonies of fishermen and business owners active in the recreation sector in the US, claiming that the periodic occurrence of HAB incidents is threatening their jobs and sometimes even leading them to unemployment [8, 9]. The toxins released in the water during severe HAB incidents are absorbed by fish rendering them unfit for consumption. Moreover, HAB induced hypoxia in seawater diminishes fish populations in fish farms. Thus, the supply of fishery is disrupted due to the reduced supply and/or to avoid any further health risks for the end-consumer having a negative impact on the fishing industry. In turn, the revenue of recreational businesses is affected. Another societal aspect with great economic consequences is the public health issue associated HAB induced diseases to humans. The entry routes of these pathogens can be either through digestion of sea products that have been contaminated or through the respiratory tract via inhalation of aerolyzed pathogens [10].

From a purely economic viewpoint, what is of great concern is detecting changes in the values of assets as well as any alterations in the financial output of economic activities provided that a severe HAB incident has occurred [11]. This discussion facilitates the selection of potential products and services that address the issues of remedy, prevention, mitigation and control for HAB since the expended capital for these types of systems should not outweigh the financial damages suffered from the phenomenon. On a decision-making drawing board, “welfare economics” provides the appropriate methodology to perform cost analysis on (1) the economic consequences of HAB and (2) the implementation of new policies and installation of infrastructures to battle the problem. The scope of this cost analysis activity is to recognize optimal financial points of operation where the sum of these two cost components minimizes. Table 1 summarizes the annual economic effects of coastal HAB in US and EU as reported in 2005 (money figures in 2005 dollar value – Euro 1,13 = $ 1,00) and calculated from the averaging of cost analysis and gathering of data regarding the loss of value for the past 10 years from the date that the study was conducted. The final two columns normalize this figure to the length of the coastlines (EU and US).



	EU ($ × 10^6)	US ($ × 10^6)	EU ($/km)	US ($/km)
Public health	11	37	170	1,856
Commercial fisheries	147	38	2,243	1,912
Recreation and tourism	637	4	9,743	225
Monitoring and management	18	3	273	169
Total	813	82	12,429	4,162

Table 2: Annual economic effects of coastal HAB in EU and US in absolute numbers and per kilometer of coast line – year 2005. Source: Hoagland et al [11]

Although a multitude of scientific papers analyze HAB incidents, the most usable and comprehensive reporting format found in the literature are interactive electronic story maps that encompass intrinsically the temporal dimension. Thus far, web platforms for the US [12] and the English channel (including the northern French coastline) [13] were discovered in the web. Another important source where HAB events are reported and catalogued is the Harmful Algal bloom Events Dataset (HAEDAT). A multitude of organizations with regional oversight contribute to the maintenance of the dataset [10].

The current estimates for the economic impact of freshwater eutrophication in England and Wales amount to 182-237M £. Approximately 77m £ are expenses for implementing appropriate policies to handle HAB inducers. Mitigating actions include water treatments for nitrogen as well as toxins removal. A red tide incident in west Florida in the summer 1971 occurred and affected 7 different counties over a period of 50 days [10]. The event was accompanied by observations of dead fish, a sharp rise in bacterial counts collected from affected areas as well and a surge of patients visiting hospitals either with symptoms of skin and eye irritation or with problems in their respiratory tracts. The estimated damages amounted to $20M. The overall economic impact of the incident was restricted by the fact that components of the local economy located near the seaside were outperformed by other activities in the greater area and due to the fact that during that year the overall economy was growing at a steady pace. This explains the fact that the overall tax revenues for the state were not affected but it also points out the fact that in periods of slow economic growth the occurrence of HAB incidents, as they are aggravated by climate change factors, increases pressures on municipal cash flows affecting in turn the quality of services provided to the civilian populations.

The economic harm done by HAB incidents can national economy significantly. A US study estimates that the losses due to eutrophication amount $1.16 billion dollars occurring at an annual rate [10]. The area of North Carolina in specific, during the period between 1987 and 1992, suffered a $7 million loss in its tourism sector due to incidents of red tide.

A diagram depicting the “anatomy” of economic impacts due to HAB incidents was found in literature [14].

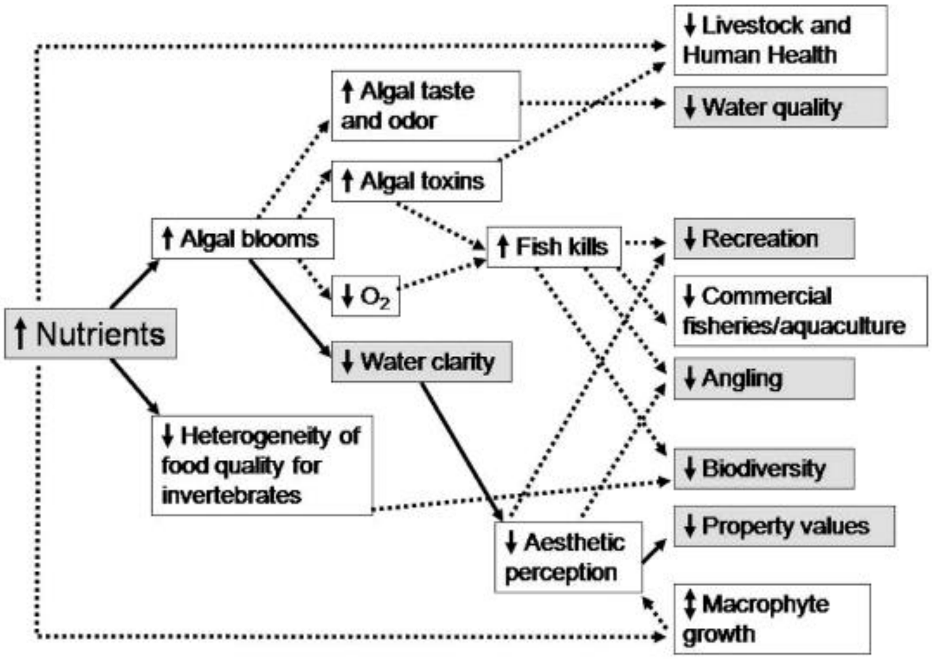


Figure 1: Economic consequences of HAB (Source: Dodds et al. 2004)

Another important cost component not shown in the diagram above stems from the necessity to channel economic resources towards HAB monitoring, prevention as well as fresh and coastal water treatment techniques to mitigate the effects on the aftermath of such incidents.

HAB: Prevention, Remedy-Mitigation and Monitoring Practices

In this section, strategies for prevention, remedy-mitigation and monitoring are presented alongside with associated cost figures for the implementation of these practices whenever relevant information is available.

HAB Prevention

All approaches for the prevention of HAB are based on some form of policy enforcement that either aims to diminish the occurrence of one or more of HAB growth parameters or intervene directly to economic activities that are shown to aggravate the phenomenon.

Elements in the literature suggest that imposing limitations on overfishing can be beneficial to battle the phenomenon since the predator chain is not disrupted [15]. Figure below shows the proposed mechanism that links the populations of "top-down" predators with the HAB growth. The decrease in the populations of fish species of high commercial interest due to overfishing results in the rise of small fishes and other species that predate in turn on herbivore organisms. These organisms are mostly dependent on the consumption of algae for their metabolism.

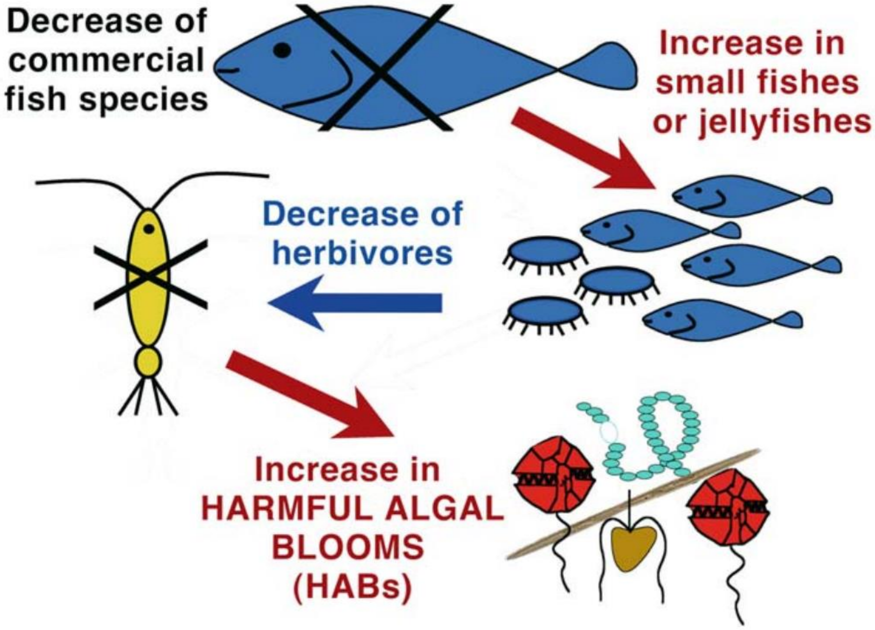

Decrease of commercial fish species
Increase in small fishes or jellyfishes
Decrease of herbivores
Increase in HARMFUL ALGAL BLOOMS (HABs)

Figure 3: Disruption of the predator chain linking overfishing with HAB occurrence (Source: [15])

Another example of controlling the adverse effects of HAB on human health is to conduct periodic checks on shell-fish harvesting. This is a direct effort to monitor the levels of paralytic shellfish poisoning toxins to avoid the societal and financial stress imposed on coastal populations that depend on this type of economic activity [16].

The onset of HAB can be regulated by controlling its growth parameters both in the case of gross environmental factors (such as pH, temperature, salinity, incident sunlight or water turbulence) or by managing elemental factors. This is achieved by managing the flow of nutrient compounds in control volumes containing elements that aggravate the phenomenon such as sulfur, iron, phosphorus and nitrogen [17]. The following figure shows how different levels of the several HAB factors mentioned here can lead to intense occurrence of the phenomenon compared to its milder version [18].

Strong legislative policies enforcing agents to take steps to control the disposal of sewage and implement water management solutions (filtering, processing etc.) can have a great impact on the prevention of the formation HAB as well as the frequency of its occurrence. These actions mostly aim to reduce the availability of nutrients. In Soto island (Japan) these practices have shown great potential to battle the phenomenon [19].

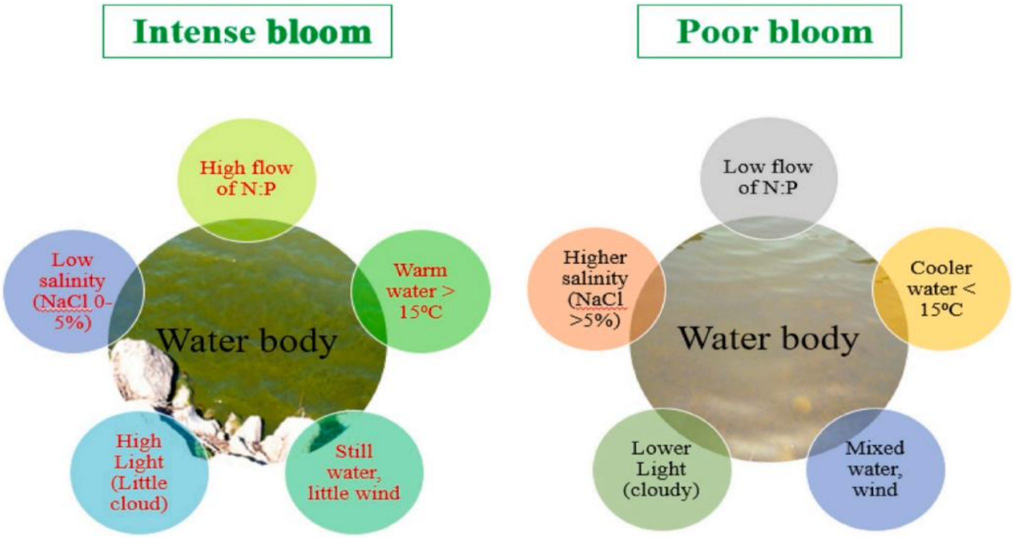


Figure 4: Impact of environmental conditions and nutrients on different aggression levels of the phenomenon (Source: [18])

HAB Remedy-Mitigation

Here we deal with several types of actions/steps taken after the occurrence of a HAB incident. Common approaches for controlling harmful species, and that can also be implemented in the case of HAB, fall into one of the following categories:

1. Mechanical
2. Biological
3. Chemical
4. Genetic

Techniques of the last category (i.e. genetic) naturally raise public concern despite the fact that have already be proven successful for land uses and plant crops [20].

One example of the chemical approach to HAB mitigation is the dispersion of copper sulphate. This technique was adopted to battle HAB incidents over the coast of Florida during 1970s. The method involved spraying of copper sulphate from airplanes over 51,5 kilometers of coastline. Although this methodology derived some short-term relief, the fact that copper is lethal for many species, renders this approach inappropriate for frequent use [21].

The most widely used method to remedy HAB incidents is the application/spraying of clay particles. The mechanism involves binding and flocculation of clay particles to HAB cells. The combined particulate matter forms a sediment that deposits on the waterbed. The combination of clay to HAB cells causes death to the latter usually via membrane rupture [22].

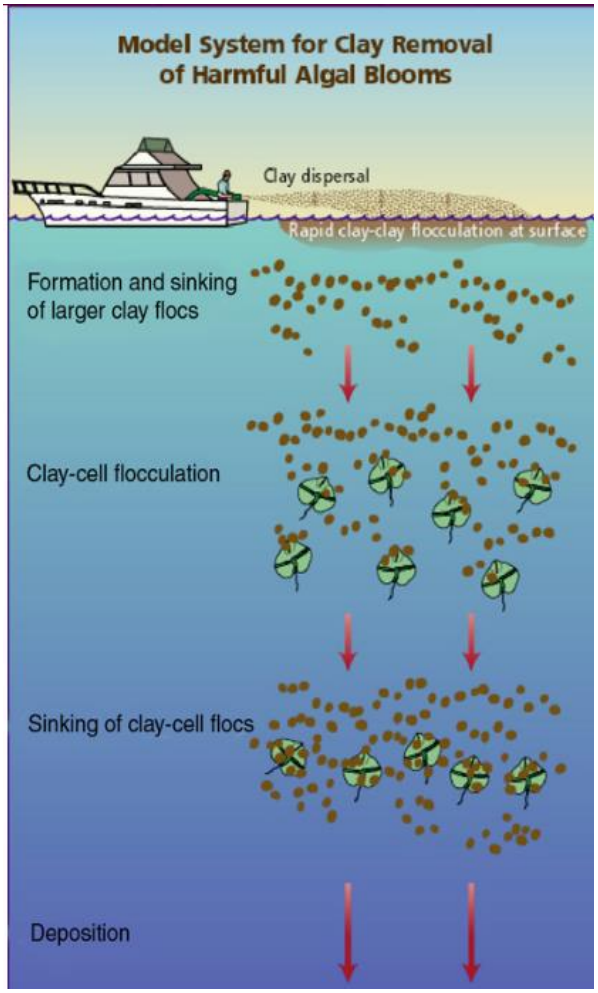

Model System for Clay Removal of Harmful Algal Blooms
Clay dispersal
Rapid clay-clay flocculation at surface
Formation and sinking of larger clay flocs
Clay-cell flocculation
Sinking of clay-cell flocs
Deposition

Figure 5: Mechanism of HAB removal using clay spraying methodology (Source: [22])

There are several types of clay mixing and particulate dispersion systems that are most commonly mounted on vessels platforms (see figure below) [23].

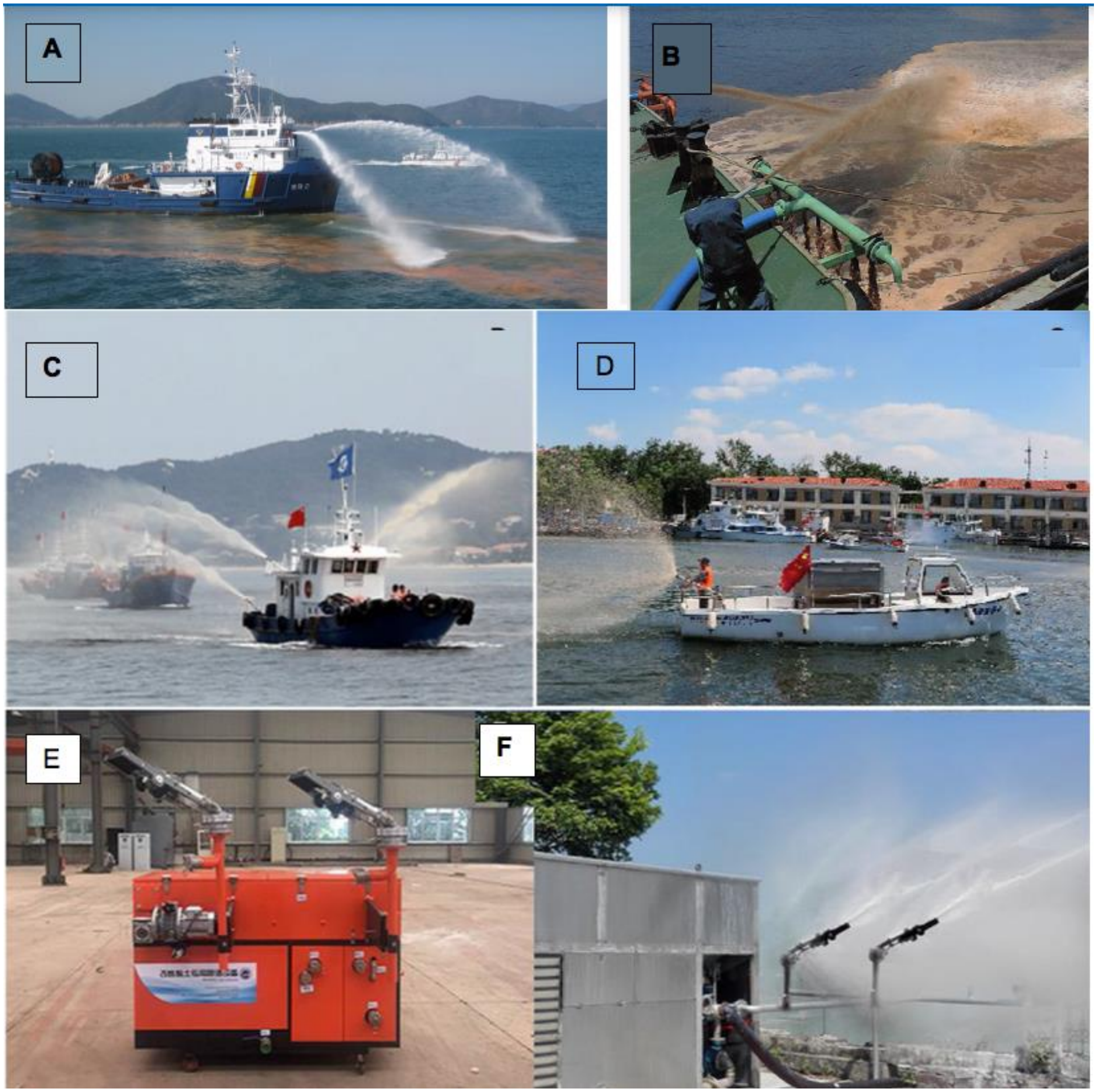


Figure 6: Ships and machines for dispersing clay. A. South Korean vessel spraying clay; B. South Korean vessel with large manifold at the rear of the ship for clay dispersal; C. Chinese vessels spraying clay; D. Small boat designed specifically for clay dispersal (Source: [23])

Studies suggest that although the dispersion of clay is considered environmentally inert (benign), high frequency of application and high total dosage delivered to the ecosystem can harm some marine species such as clams [24].

In order to improve the flocculation efficiency of a clay dispersal HAB treatment, novel clay types have been developed. These modified clay types are produced by increasing the surface positive charges of clay. This surface modified clay types are shown to be a dozen times more efficient, in terms of flocculation and clay-to-HAB cell binding. Thus, when compared to its unmodified clay counterpart, much less clay masses are required (10-400 tones/km^2 for unmodified Vs 4-10 tones/km^2 for modified) to mitigate the effect of a HAB incident with much less stress imposed on the environment. Application of modified clays is the standard in East China coastline. Figure below shows the effect of clay dispersal to a HAB infected area as well as the measured cyanobacteria concentration within a specific time frame [25].

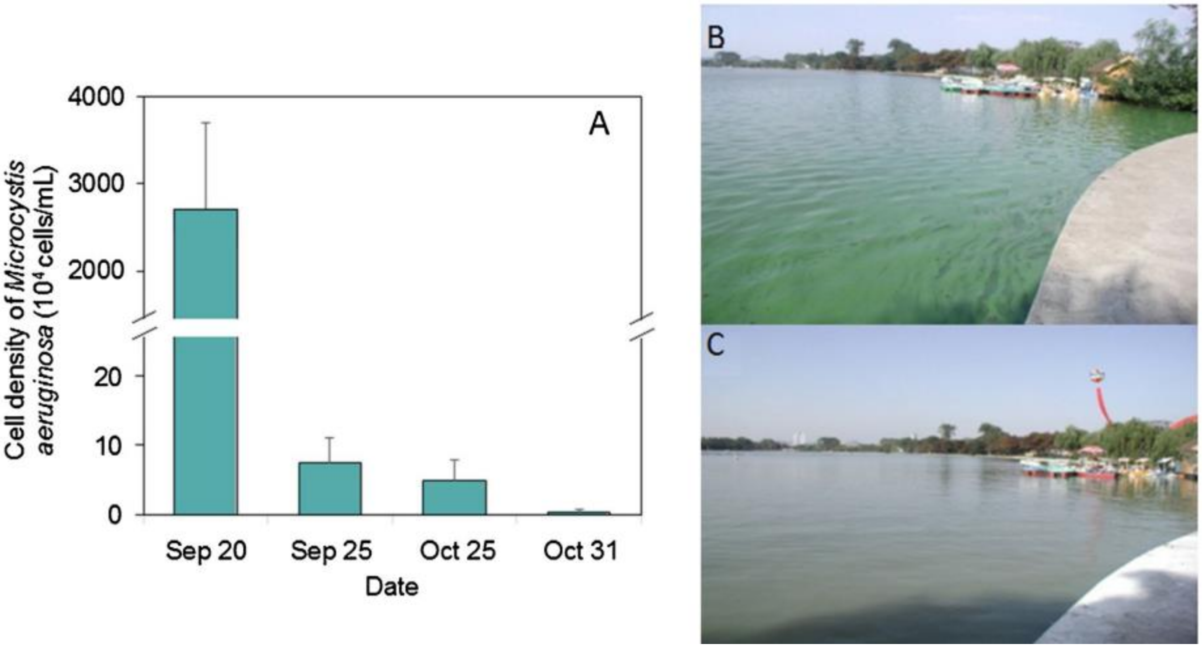


Figure 7: Mitigation of cyanobacterial blooms in Xuanwu Lake. (A) The density of Microcystis decreased during HAB mitigation using modified clay; (B) & (C) Appearance of the water before and after HAB mitigation (Source: [25])

One electronic literature source was found that addresses the issue of clay dispersal in terms of its economic cost. The study involves the application of modified clay in China. The dosage ranges from 4 to 10 tones/km^2 yielding thickness of the as-deposited layer between 12-40 micrometers (assuming uniform layer deposition). The cost of modified clay is approximately $500 per tone. Thus, for a modified clay dosage/application rate of 7 tones/km^2, around $3,500 are required for the purchase of the modified clay alone (not including the cost of vessel mobilization and personnel).

Another important purely mechanical technique of HAB removal is the use of ultrasound. Experiments have been conducted with ultrasonic irradiation showing degradation of cyanobacterial toxins (CBT) and especially microcystin-LR at 640kHz [26]. Nevertheless, these techniques require further study regarding their efficacy since they are limited due to their small effective distance from the ultrasonic source and the required energy consumption. The proposed bio-physical mechanism for the efficacy of ultrasonics suggests that the incident irradiation causes cavitational effects that in turn disrupt the algae cell wall and membrane. Thus, photosynthetic activity is interrupted alongside with the overall cell cycle (cell division) [27].

As far as biological approaches for HAB mitigation are concerned, Pal et al catalogued in their review paper several biological agents, such as bacteria, phage, fungi and zooplankton species, that inhibit algal growth [18]. It is worth noting that the interaction between the mitigation agent and algae is species-specific and its effectiveness relies either on some form of infection mechanism or on the production of algicidal compounds.

HAB Monitoring

Non-Remote Monitoring Methods

The standard method to detect chemically HAB cells is by performing periodic sampling and direct measurements of chlorophyll-a (Chl-a) and cyanobacterial cell (e.g. Microcystin) concentrations. One case found in the literature where this technique is employed is on river Charles (Boston area). These measurements, although providing with the required data to perform year-on-year comparisons, are insufficient to generate early warning signals for the onset of the phenomenon [28]. Figure below shows the distribution of monitoring stations along river Charles that are set by different institutional bodies responsible for environmental observations.

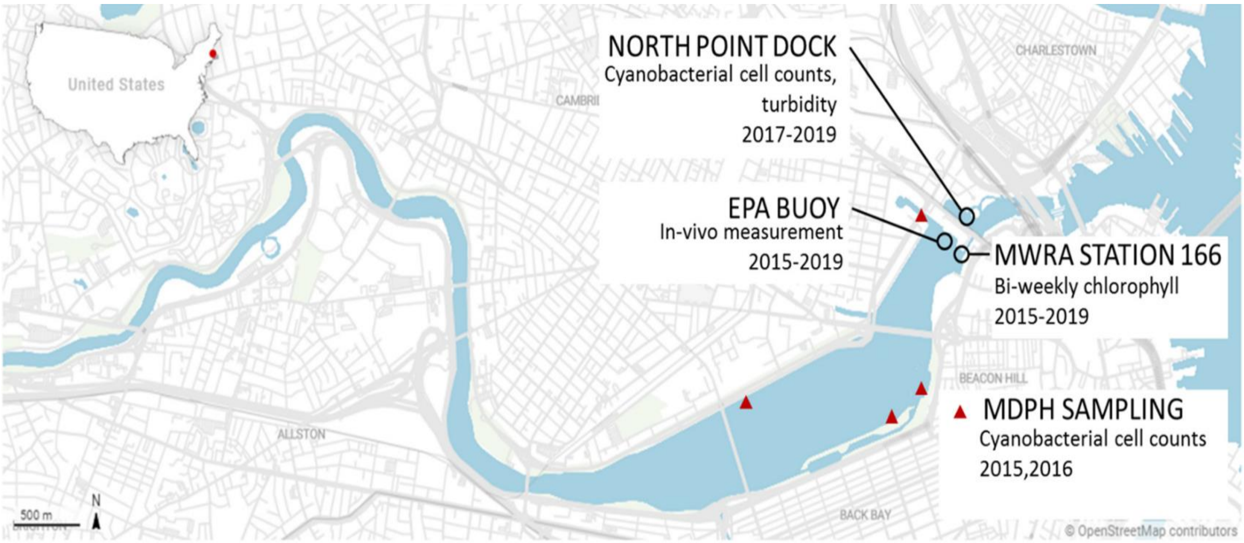


Figure 8: Monitoring stations situated along river Charles. Environmental institutions involved are: (1) MDPH – Massachusetts Department of Public Health, (2) US EPA – US Environmental Protection Agency, (3) MWRA - Massachusetts Water Resources Authority (Source: [28])

Note that the accepted cyanobacterial concentration for drinking water according to the World Health Organization is 1 μg/Liter. A 50% conditional probability for exceeding this guideline corresponds to 68 μg/Liter of Chlorophyll-a (Chl-a is considered a proxy for the existence and growth of HAB species) [29].

The sampling stage is usually followed by a cyanobacterial identification step. Most typical method for this stage is the usage of Enzyme-linked Immunosorbent Assay (ELISA) characterization. Subsequently, quantitative analysis is performed via liquid chromatography, mass spectrometry or a combination of both [30]. New, easy-to-use test kits have been developed for the detection of PSP toxins. Jellett Biotek ltd developed a rapid test, the MIST alert test strip (see image below), based on the usage of lateral flow immuno-chromatographic (LFI) test strips. The results are produced in 20 minutes and assist labs to identify negative samples in a straightforward fashion [31].

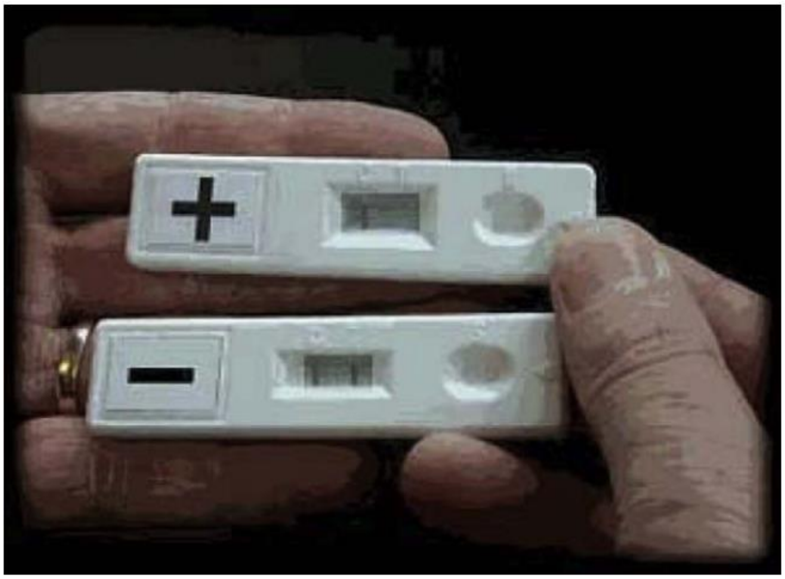


Figure 9: Jellett MIST test strip for rapid and simple PSP toxin detection (Source: [20])

Another widely spread detection system for HAB species is of optical principle. Fluorescent measurements using a fluorometer are conducted to detect the levels of chlorophyll-a. The Chl-a pigment fluoresces during photosynthesis at 680 nm (characteristic wavelength). The recorded relative fluorescence provides with a quantitative measurement that can be correlated with algal health (i.e. mild or aggressive HAB growth) [32]. Systems that are based on this type of detection technique are applicable also to in situ solutions [33]. A commercial example of this is the ROW™ algal bloom sensor, marketed by Laser Diagnostic Instruments ltd [34]. The device operates on the UV part of the light spectrum and is suitable for the detection of blue-green algae. The fluorescence algae detection technique has shown great potential for a portable and cost-effective manner to achieve continuous water monitoring. For that purpose, miniaturization of fluorometric systems has been demonstrated using materials suitable for organic opto-electronic device manufacturing and MEMS (micro-electro-mechanical systems) based micro-fluidic architectures as shown in the figure below [35].

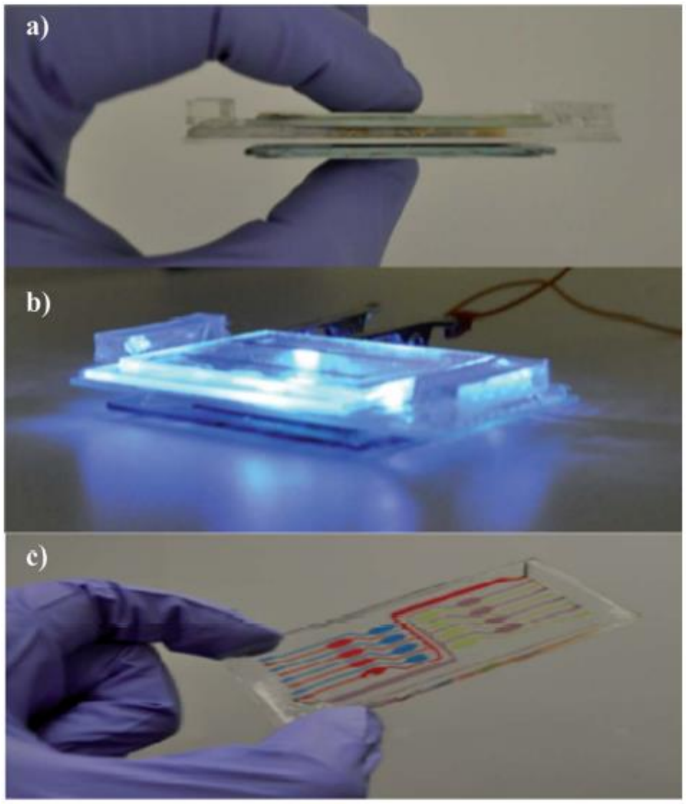


Figure 10: Pictures of the MEMS based fabricated sensor. (a) The flat device can be hold between two fingers. (b) Fluorescence sensor with illuminated OLED. (c) Microfluidic chip with 16 chambers - color dyes show the micro channels (Source: [35]).

A classical approach to HAB species detection and quantification is conventional light microscopy [36]. An alternative characterization method that also offers portability and is suitable for continuous monitoring are microscopy instruments that their operation is based on flow cytometry. FlowCam is a commercially available solution of this type [37].

Remote Monitoring Methods

In recent years, UAV technology is booming. This fact, combined with the development of even more lightweight image sensors offers the opportunity to realize a vast variety of UAV based systems. Environmental monitoring is a field where this new technology is explored. Many literature resources were found during the current research where image acquisition from UAV platforms is used for algae detection purposes.

Flynn et al, conducted 18 different missions using a low-cost (approx. $700) UAV helicopter-like platform equipped with a common GoPro Hero3 12-megapixel camera in order to obtain imagery of Clark Fork river (Montana) [38]. The visual data collected were used to characterize river algae. A challenge that had to be overcome was to employ a geometric correction technique to compensate for lens distortions using knowledge of the optical system (fisheye lens). The absence of a gimbal mechanism dictates that the dataset collected is filtered so that images only with minor tilt are taken into consideration for further processing. Adaptive cosine estimator (ACE) method was used to distinguish the algae species from its background. A supervised spectral angle mapper (SAM) algorithm was employed to identify the spectral signature in the common red, green and blue color bands. The accuracies of algae species identification with respect to its background when using the ACE and SAM techniques were 90% and 92% respectively.

Studies combining UAV image acquisition and situ sampling to derive regression model parameters were conducted in Tain-Pu reservoir (Kinmen island, Taiwan) [39]. The main measurands of this effort were chlorophyll-a (Chl-a) and total phosphorous (TP) concentrations as well as Secchi disk depth[1] (SD). Moreover, the Carlson trophic state index that is a combination of the previous three measurements was also considered. In situ data collected for over a decade in Tain-Pu reservoir found a strong anti-correlation between Chl-a and SD as well as a significant anti-correlation between SD and TP. A fixed wing UAV was used, equipped with both an RGB camera as well as a near-infrared sensor. Generally, the use of drones requires elaborate preparation by setting a network of ground control points (GCP) for UAV orientation/navigation and image correlation with respect to the ground. Chlorophyll-a absorbs in the blue as well as the red part of the light spectrum while it is very reflective in the NIR band[2]. Thus, the regression models attempted to correlate one of the water quality parameters (Chl-a, TP and SD) to the band ratio of NIR to red and blue (i.e. NIR/R, NIR/B).

UAVs equipped with NIR cameras were deployed at Centralia Lake (Kansas) during a Microcystis HAB incident in September 2012 [40]. The use of UAVs was found to be helpful in terms of providing the temporal resolution that is demanded for an early-warning system. Samples were

[1] A turbidity type of measurement where a white disk approximately 20 cm in diameter is suspended by a string and gradually submerged. SD is the limiting depth where the disk is no longer visible.

[2] Wavelengths corresponding to central frequency of each band: Red – 660 nm, Blue – 450 nm, NIR – 850 nm

also collected, analyzed and photographed under the microscope. The measuring index of microcystin in the samples used was the buoyant packed cell volume (BPCV) while the assisting measurand employed was the blue normalized difference vegetation index (BNDVI)[3]. The significant element of this study is that UAVs had pre-planned their flight paths using a software, showing potential of this approach for automating water management activities. The results obtained by processing of the NIR imagery from UAVs followed a similar logarithmic trend between BNDVI and BPCV to the one detected in the lab using optical microscopy captures. A colored NIR orthomosaic image of a water pool in the area is shown in the figure below. The captures that are assembled to produce this image were taken at an altitude of 25 meters. Sampling markers indicate the spots where samples were collected and assist UAV-to-ground correlations. Invariant target panels are used to compensate for variances in light reflectance across different moments within a day where sun angles, irradiance and atmospheric conditions differ.

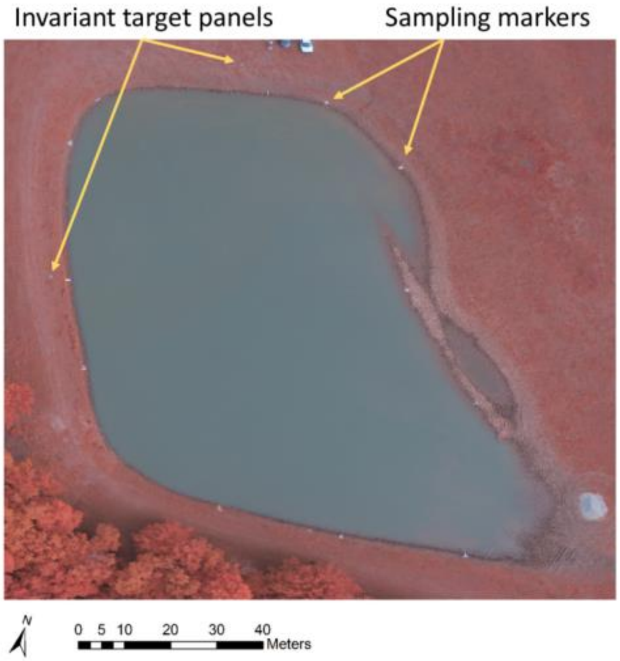


Figure 11: An averaged, UAV captured, color-infrared orthomosaic of a water pond (Source: [40])

[3] BNDVI = (NIR - blue) / (NIR + blue)

Researchers conducted HAB detection studies a conventional RGB camera in lake Taal (Philippines) [41]. The main objective was to draw regression models correlating visible color to green algae concentration levels. In this supervised learning study, the collected imagery was calibrated against laboratory processed samples where algal concentration levels are determined gravimetrically. A notable element of this study is that feature extraction was based on hue/saturation/value (HSV) color space (instead of the most used RGB). This approach is claimed to have better results in separating true color from intensity and brightness. Authors stressed out the fact that although the consensus is that UAVs provide with a mean to have a very finely grained temporal resolution, since UAV flights and image acquisition is low-cost and can be scheduled in frequent intervals, when monitoring water bodies in tropical climates is challenging since strong winds prohibit flight for lightweight aerial platforms. Again, image pre-processing techniques such as noise reduction and image smoothening were employed prior to feature extraction that happens on regions of interest (ROI) containing only algae and clear water.

Instead of using a camera or a NIR sensor, Shang et al used a pair of hyperspectral radiometers operating within 412-667 nm band mounted on a UAV [42]. The apparatus was deployed in Weitou Bay that is located in the western Taiwan strait in September 11-12 2015, during a Phaeocystis globosa dominated bloom. The sensor pair had a very fine spectral resolution (3nm). The altitude that captures of the water surface were taken was 300 meters that allowed for minimal distortion from atmospheric conditions when compared to manned flight surveillance missions (> 1000 meters) or satellite image acquisition. The study was complemented by water sampling and fluometric measurements for validation reasons. The figure of merit of interest for this study was the remote sensing reflectance (R_{rs}) that was also measured in situ using a spectroradiometer. The data collected from the UAV were corrected in order to compensate for effects such as platform tilt and cloud motion. This study effort proved that UAV mounted radiometric instrumentation can play an important role in the advent for timely warning systems of HAB incidents although data processing protocols need to be examined furtherly.

A straightforward model for mapping and identification of green tide (Ulva[4] dominated bloom) was constructed by true color RGB imagery acquired by a UAV mounted camera [43]. The site where this study was conducted was Haiyang beach of the Shandong province in China during an outbreak

[4] A type of algae also know as “sea lettuce”

of green tide in July 2017. The spatial resolution was around 8 cm from an image capturing altitude of 150 meters. The deployed system used GPS for georeferencing purposes, thus eliminating any need for preparatory steps involving setting ground control points. The algae mapping models were based on algebraic inter-relations of the as-measured RGB intensities NGRDI (Normalized Green Red Difference Index), NGBDI (Normalized Green Blue Difference Index) and GLI (Green Leaf Index)[5]. The GLI measure was able to identify algal areas in the acquired imagery with a very high accuracy (93.1 %).

Since 1978, when the launch of Nimbus-7 meteorological satellite was placed in orbit, space technology offered new methods to collect valuable oceanographic data [44]. The coastal zone color scanner (CZCS) radiometry instrument available on the satellite's sensor systems was able to detect phytoplankton with 800-meter spatial resolution from space for the first time. These measurements mark the historic beginning of satellite algae monitoring [45]. Since then, satellite sensors have become more sensitive with much higher resolutions.

The Medium Resolution Imaging Spectrometer (MERIS) satellite sensor, mounted on ESA's Envisat satellite, provides with a reliable instrument for HAB tracking [46]. Bloom events are shown to exhibit a characteristic spectral signature in the narrow range 705-709 nm that corresponds to the reflected irradiance from the water surface that signifies a high probability of chlorophyll-a [47]. A powerful figure of merit for chlorophyll and HAB detection from satellite NIR imaging data is the MERIS Maximum Chlorophyll Index (MCI) [48]. The MERIS instrument possesses a spatial resolution of 300 meters with a great spectral selectivity for the study of red-tides. MCI index has been used successfully to detect HAB events in Canada, Australia and the Antarctic area as well as other sites where significant pelagic vegetation occurs [47]. Figure below, shows three consecutive satellite snapshots taken on the 5, 21 and 27th of August 2006 that demonstrate the progression of an algal bloom phenomenon of the coast of Repulse Bay, Queensland, Australia.

[5] NGRDI = (G-R)/(G+R), GBDI = (G-B)/(G+B) and GLI = (2G-R-B)/(2G+R+B)

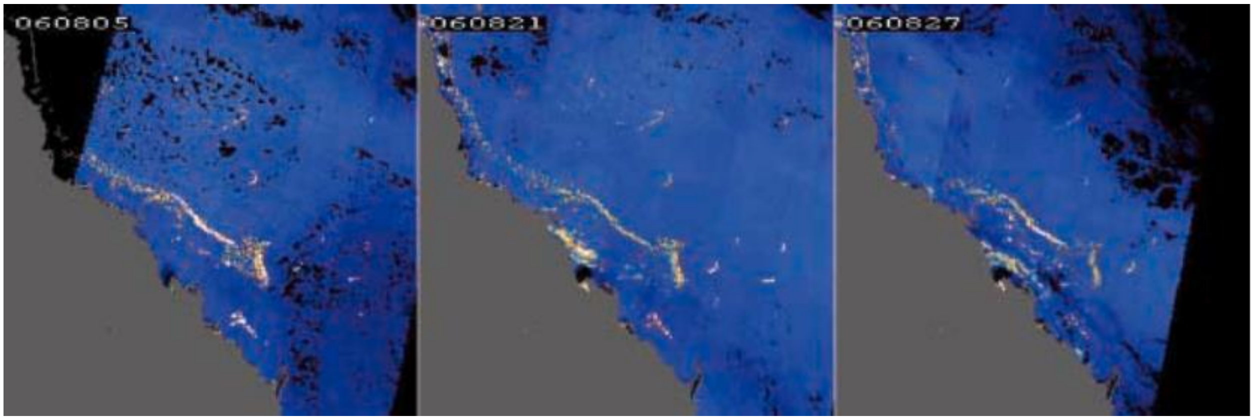


Figure 12: Progression of an algal bloom phenomenon of the coast of Repulse Bay, Queensland, Australia on the 5, 21 and 27th of August 2006 (left-to-right) [Source: [47]]

The novel Ocean and Land Color Instrument (OCLI), mounted on ESA Sentinel-3A satellite has replaced the MERIS sensor apparatus. OCLI is again employed for water quality monitoring tasks with the appropriate spectral characteristics to observe chlorophyll as well as phycocyanin pigments [49].

Figure below assists the baseline observations of algae with respect to water areas exhibiting low or no eutrophication based on reflectance measurements [50]. In the “violet-blue” realm of the spectrum (400-500 nm of wavelength), clear water reflects heavily, thus appearing blue. In contrast, areas rich in algae species show two peaks: a primary one in the “green-yellow” portion of the spectrum (around 550 nm) and a secondary one in the near infrared (approximately 700 nm). These are accompanied also by a pair of characteristic absorption peaks at approximately 500nm and 675nm.

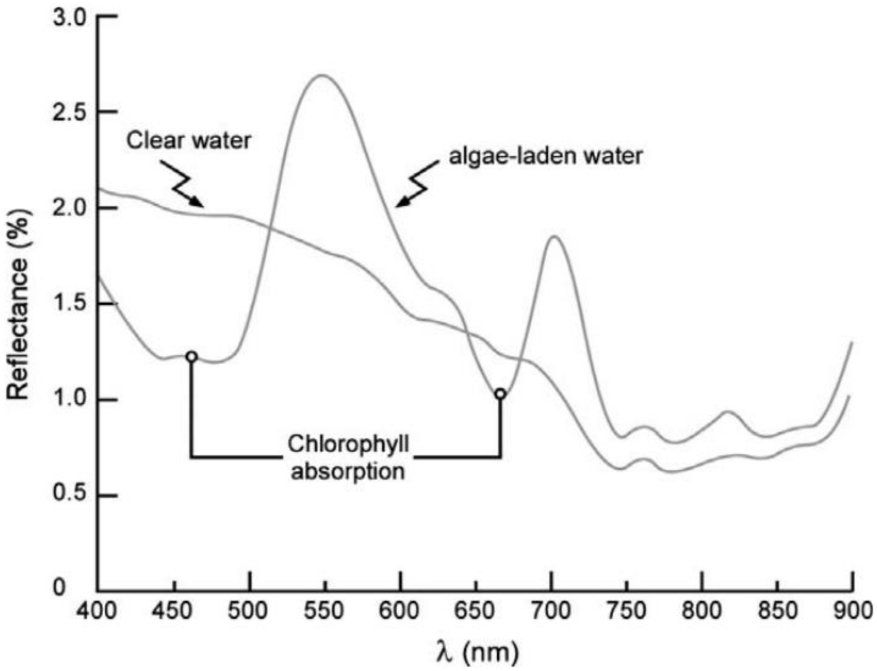


Figure 13: Comparison of clear water to algae based on reflectance spectral data (Source: [50])

The technological aspects of both remote sensing approaches (i.e. drone-assisted and satellite) offer different benefits while present many challenges. Wu et al, demonstrate these by comparison of these techniques [51].

Although satellite acquired imagery can cover vast areas, thus being able to track the evolution of large HAB phenomena (appropriate for oceanic/pelagic eutrophication incidents), this method suffers from the following drawbacks:

1. **Low temporal resolution**: In order to repeat a pass over the same region of interest twice, as much as a 30-day interval might be required
2. **Low spatial resolution**: Some sensors exhibit astounding resolution for orbital observation (e.g. SPOT6-7 sensors have spatial resolutions as low as 1.5 meters) but most systems, like MODIS and MERIS, have resolutions in the range of 250 – 1100 meters
3. **Weather restrictions**: Patches of the water surface situated under heavy clouds are non-visible
4. **Compensation during pre-processing**: Imagery needs to be pre-processed appropriately to compensate for partial cloudiness as well as atmospheric absorption effects

On the other hand, UAV technology overcomes the limitation of weather restrictions, being able to fly at altitudes under the weather, while drones can be commissioned to operate many times within a single day offering immense temporal resolution for water monitoring activities. The UAV imagery can have spatial resolutions that can reach down to 10 cm. Limitations of UAV technology for water management systems are:

1. **Weather conditions**: UAVs cannot operate in windy weather
2. **Ground preparation**: Most of the times, especially in cases where on board GPS geolocation systems are not available, a network of ground control points must be set prior to mission deployment to assist geo-referencing
3. **Limited payload**: The weight that UAVs can carry during the mission, that includes the optical sensors and other systems, can be an issue
4. **Spatial coverage**: In order to be able to monitor HAB incidents on the large scale, UAV swarms are required; increased cost of mission operation
5. **Limited distance from RF control point(s):** This renders UAVs inappropriate for oceanic studies but suitable for monitoring coastal and freshwater volumes critical for agricultural irrigation and drinking purposes

What is evident from the previous discussion is that for the design and implementation of early-warning systems that aim to limit the economic, health as well as environmental impact of eutrophication incidents, both approaches offer certain benefits and can be integrated together, ideally alongside a network of in situ sensor sites, for performant HAB and water monitoring solutions.

HAB: Processing and Algorithms for Satellite Remote Sensing

HAB detection algorithms can be classified as either optical analysis of each component present in the water or direct spectral on the captured image [52]. Optical analysis is mostly applicable in oceanic/pelagic waters whereas direct spectral analysis is suitable for coastal waters. The quality of the results strongly depends on processing that addresses atmospheric correction. Limitations that arise from these methods, when compared to in situ measurements for validation, are handled by gaining optical data from in situ optical measurements that assist in the definition of threshold values for radiance levels leaving the water surface. It is important to note that it is extremely difficult to arrive to a global prediction model due to the differences in the dominant algal species across different areas.

Groom et al [53], designed an algorithm that used thresholding at the 580-680nm band for the detection of coccolithophore by processing reflectance measurements after appropriate data filtering that compensates for interference effects caused by ozone absorption and Rayleigh scattering [54] from atmospheric particles that are much smaller than the wavelength of the incident radiation.

Ogashawara et al [49], used the normalized difference chlorophyl index (NDCI) to evaluate via regression models the levels of chlorophyl from snapshots taken from Sentinel-3 satellite sensor after atmospheric corrections were applied on the raw data for Lake Erie. For evaluation of phycocyanin pigment though the three band 3BDA algorithm was used. In a Sentinel-2 satellite HAB study conducted around the Hong Kong area, it was found that the optimum sensor bands to be used for reasons of feature engineering are band-4 (665 nm) and band-8A (865 nm) [55]. The selected figure of merit was half the relative difference between these two bands[6]. For data correction purposes, the ESA's SNAP Sen2cor utility was used that is optimized for Sentinel-2 products [56]. For a chlorophyl detection level study conducted for Sentinel-3 images of the coastal Benguela region (Angola), the features that were found more indicative were: (1) the maximum line height (MLH) for spectral peaks at 681 and 709 nm as calculated from a baseline in the region 665-753 nm and (2) the line height ratio (LHR) between these two figures [57].

[6] $(R_4 - R_{8A})/(R_4 + R_{8A})$

Satellite Data HAB Modelling

Data Acquisition

In the current project, satellite imagery is gathered using the Sentinel hub application programming interface (API) which is an ESA's cloud-based service for the provisioning of image and radiometric data from orbit. The service follows the RESTful paradigm and though being public, it requires prior acquisition of an access token according to the OAUTH 2.0 security standard - in the form of JSON web tokens (JWT) - to limit the service traffic and most importantly the number of parallel requests. Registering with Sentinel hub service provides with the appropriate OAUTH 2.0 credentials to gain service access tokens that typically expire after 1 hour from the time that were issued. According to the Sentinel hub API syntax, the specific service context for satellite data acquisition is the "Process API". The service is accessed via a POST request to the URL: https://services.sentinel-hub.com/api/v1/process. Details for the API usage can be found online from Sentinel website [58]. The payload of the REST request determines the type of the acquired image data (raster images usually in .jpg or .tiff formats) as well as some pre-processing or metadata fetching options that can be useful in further analysis. There are many methods of acquiring access tokens and finally requesting for image data. In the first phase of this study, Postman REST client software was used for acquaintance with Sentinel hub utility and performing test requests to the API while later, the service is invoked programmatically using python scripts. The body of the request, where all acquisition and pre-processing options are passed, can also be generated interactively through the request builder[7] Sentinel web utility.

An example of the payload of an example request can be found in Appendix A. The request is divided into 3 major parts: (1) input, (2) output and (3) evalscript. Within each of these parts different request and pre-processing options/fields can be defined as follows (note that in the API documentation the type and the allowed list of values per field is documented):

1. Within **input** section:
    - Bounds: The bounding box or polygon area that defines the region of interest
    - Crs: The coordinate system used

[7] https://apps.sentinel-hub.com/requests-builder/

  - Type: The collection to acquire data from. In the present work we selected Sentinel-2 L2A (S2L2A) data collection because it provides out-of-the-box with atmospherically corrected products/images[8], thus saving the effort to perform this programmatically and consequently simplifying the processing pipeline
  - Time Range: The dates for which the acquired data refer to (note that imagery is not available for all dates)
  - Data.type: Defines the used data collection
  - mosaickingOrder: Denotes the method that satellite images are going to be "stitched" to form a synthetic aperture that can capture the region of interest defined in the bounds section of the payload
  - maxCloudCoverage: The limiting cloud coverage in the captured image expressed in percent

2. Within **output** section:
   - Width/Height: These fields determine the size of the downloaded resource file
   - Resx/Resy: These fields determine the spatial resolution per axis. It is important to note that this field is used exclusively instead of defining width and height parameters and that should be set considering the resolution values per sensor channel as shown in the table below. If the resolution in the request is smaller than the one provided by the channel sensor, then an interpolation technique parameter has also to be included in the request payload
   - Format: The format of the requested file (i.e. tiff, jpeg, JSON, PNG)
3. Within **evalscript** section: Normally, this is the section where sensor band selection is done while being able to control image attributes such as brightness and contrast amongst other data manipulations

Usually, the downloaded product(s) of the REST request are a series of images that correspond to different bands or a combination of them corresponding to true color, false color (non-visible bands mapped in the visible part of the spectrum) or some other pre-defined setting available by Sentinel service. For example, considering all 13 available sensor bands, a common set of downloaded products for a specified region could be a set of 14 images in the requested file format – 13 for each of the narrow Sentinel bands and a true color image for referencing purposes. Sentinel hub

[8] https://www.sentinel-hub.com/develop/api/ogc/custom-parameters/atmospheric-correction/

provides with a wide set of examples in the form of ready requests that are amenable to field editing. This accelerates the development process significantly especially in API request customizations that require more complex manipulations as far as the **evalscript** part of the payload is concerned.

The band specifications of the Sentinel sensor are shown in the table below. These specifications include, besides the central frequency of each band, the relevant bandwidth and spatial resolution capability.

Sentinel-2 band name	Central wavelength (nm)	Bandwidth (nm)	Spatial Resolution (m)
B01	443	21	60
B02	492	66	10
B03	560	36	10
B04	665	31	10
B05	704	16	20
B06	740	15	20
B07	780	20	20
B08	833	106	10
B08A	864	21	20
B09	944	20	60
B10	1375	30	60
B11	1613	91	20
B12	2202	175	20

Table 1: Sentinel-2 sensor channels and their specifications (Source: [59])

The usage of the Sentinel hub API directly via the curl[9] system utility or any other REST client like Postman are means of acquiring satellite data which can be processed subsequently using any image analysis library – such as OpenCV – to extract features that are critical for classification purposes. Alternatively, data acquisition can be achieved programmatically, and Sentinel hub provides documentation for accessing the API service using Python [60].

There are 4 important components as far as the Python Sentinel hub client are concerned and most of them are included inside the **sentinelhub** package that can be installed using the **pip** build utility, either globally or inside the applications' virtual environment:

[9] https://en.wikipedia.org/wiki/CURL

1. **SHConfig class**: An object of this class is used to hold data about user credentials. Namely, the client ID and the client secret that are used to generate the required service API access token.
2. **SentinelHubRequest class**: An instance of this class is used to hold the request payload in the form of a python object. It is important to note that all the fields required for the formation of the request object are included in the sentinelhub package and should be imported individually in order to avoid loading the entirety of the library. This practice is common for prudent usage of the memory resources and is required for any application having external library dependencies
3. **plot_image:** This utility is imported from Python utils package, and it assists for the plotting of the acquired image(s). Of course, any other third-party library can be used for this purpose. In the present work, the OpenCV library is used to read, process and plot the image data since the plot_image functionality has a scope limited to viewing the downloaded product(s)/image(s)
4. **get_data()** method: This method operates on the constructed request object. The return variable of the method invocation is the downloaded product(s) from the Sentinel hub API service call. Since more than one product can be requested for a specific region of interest - defined either in terms of a rectangular bounding box or a polygon area – the result of this method is an **array of images and/or geodata in JSON format** (in the case of GeoTIFF Sentinel product). Image participants in this array are commonly the sensor output isolated in a per-channel basis or some other transformation defined in the **evalscript** part of the request payload.

Modelling Strategy and Feature Extraction

In this section the process pipeline is presented with the scope of producing an optimum machine learning (ML) model for HAB predictions and its application and evaluation to “unseen” satellite images. The chosen model is produced using the artificial neural network (ANN) ML methodology using the TensorFlow software library. An optimization study will determine the most efficient network architecture by designing an experiment with the number of layers and the number of respective nodes per layer as the critical design parameters. The chosen merits that are of importance for evaluating model efficacy are: (1) accuracy, (2) F_1 score and (3) area under curve

(AUC). These figures are traced when altering the experimental design parameters of the ANN architecture to arrive to an optimal model geometry.

The proposed process to build the feature vectors initiates with gathering true color images using the Sentinel hub REST API from Postman client software. From these satellite captures the dataset generated is going to be used for model training and testing purposes via manual labeling (this is discussed later in the chapter). Subsequently, using a python script this time to invoke the remote Sentinel hub service, all 13-channel narrow-band sensor images are gathered for each true color image and are converted into a single-color grey scale format using the OpenCV library for image analysis and manipulation. For each labeled pixel address of the true color images a feature vector is formed to be fed into the ANN classifier that consists of the following **19** features/vector components that are formed by the narrow-band data gathered:

1. The **12 narrow-band** reflection intensities per pixel as gathered from the Sentinel hub service (B10 not included)
2. **Normalized difference chlorophyll index (NDCI)**: This feature was used by authors reviewed in the background literature [49, 61] that states that chlorophyll concentration is proportional to the following expression that relates the sensor reflectances at 708nm and 665nm:

$$C_{chl-a} \propto \frac{[R_{rs}(708) - R_{rs}(665)]}{[R_{rs}(708) + R_{rs}(665)]} = NDCI$$

In terms of the Sentinel satellite sensor channels, the NDCI feature becomes:

$$NDCI = \frac{B05 - B04}{B05 + B04}$$

For the limiting case where the denominator of this expression is zero, the handling approach to avoid division by zero errors by setting the whole feature to a zero value

3. **Normalized difference vegetation index (NDVI)**: This feature is derived from the reflectance levels of channels B04 and B08 [62] as:

$$NDVI = \frac{B08 - B04}{B08 + B04}$$

4. **NDVI based on B08A sensor channel** (according to the feature described by Khalili et al [55]**):**

$$NDVI = \frac{B08A - B04}{B08A + B04}$$

5. **Maximum Peak Height Bloom Index (MPHBI)** – similar to the maximum line height feature reviewed in literature [57] or else according to the MPHBI custom script available from the

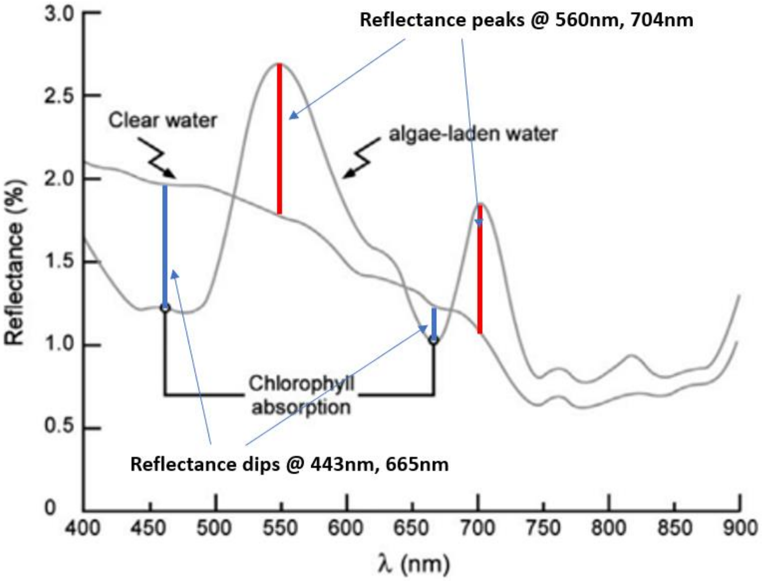
Sentinel hub script collection [63]. The feature comprises of the maximum value for thresholding parameters for the combinations of sensor channels B04, B05, B06 and B08 as:

$$T_1 = B05 - B04 - \frac{(B06 - B05) \cdot (705 - 665)}{(740 - 665)}$$

$$T_2 = B06 - B04 - \frac{(B08 - B04) \cdot (740 - 665)}{(842 - 665)}$$

And **MPHBI = max (T_1, T_2)**

6. The last set of features is extracted directly by revisiting the reflectance spectrum diagram (figure 13) that offers means of comparison between chlorophyll-a and clear water.



It can be understood that although the differences in reflectance at the characteristic chlorophyll-a wavelength peaks (560nm, 704nm) – red lines – are significant, the same does not apply to the reflectance dip in the vicinity of 665nm. From this observation three new features can be extracted with the aid of clear water statistical modes (i.e. most frequently occurring values) at the characteristic wavelengths 443nm, 560nm and 704nm. Thus, the three pixel-based extracted reflectance difference features are:

- RD_1 = statistical_mode_of_Clear_Water_B01 – B01
- RD_2 = statistical_mode_of_Clear_Water_B03 – B03
- RD_3 = statistical_mode_of_Clear_Water_B05 – B05

Dataset Used and Labeling

During the current study no ready datasets were discovered. The term "ready" here refers to the existence of true color and narrow-band satellite images that are labeled and geo-referenced. To serve the purpose of the current HAB identification model attempt, such a dataset requires to be labeled (i.e. HAB/No HAB) in a pixel-by-pixel basis since the phenomenon occupies certain areas that correspond to a fraction of the water control volume inside the imagery and not its entirety.

Normally, labeling involves as discussed in the literature section, in-situ detection techniques that are providing the necessary chlorophyll and toxin concentration measurements. This data helps to verify the occurrence of HAB phenomena and assign to a specific region a certain level of severity. Having these kinds of tools at one's disposal, allows to produce refined HAB detection satellite imaging models since there is sufficient information to train classifiers for the detection of eutrophication, mesotrophication, and oligotrophication. In the absence of these kind of verification techniques, that require elaborate work in setting up their operation and human as well as monetary capital for their deployment, the HAB classification model developed herein is binary and able to identify image pixels with HAB regardless of the associated severity level.

Due to all the reasons explained before, the dataset used in this study was produced by looking into web resources for severe HAB incidents and acquiring the respective satellite data from the Sentinel hub service. One of the places with the highest occurrence of HAB incidents is the western side of Lake Erie (coastal line belonging both to Ohio and Michigan states). The National Centers for Coastal and Ocean Science (NCCOS) provides with data for past HAB events on Lake Erie as well as with future forecasts to assist the communities to adjust their precautionary actions accordingly [64]. A period with very severe occurrence of HAB was during September 2021 with a severity index of 6.0. The phenomenon was intensified by warm waters and the mildness of water currents during that period. The image below was produced from Copernicus Sentinel hub satellite data produced on the September 6th, 2021, showing the distribution of cyanobacteria in the region [65].

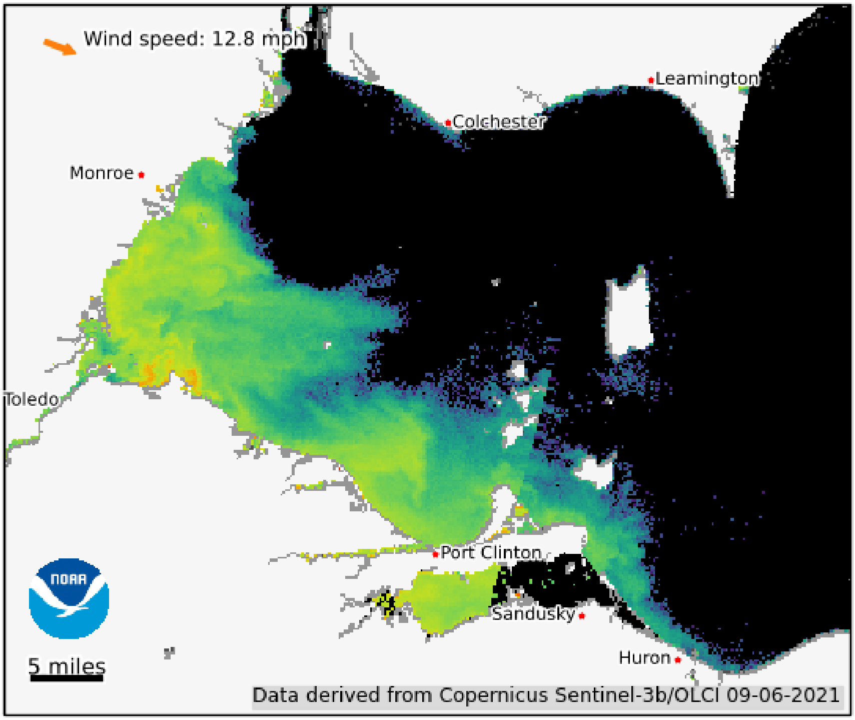


Figure 14: Distribution of cyanobacteria in the western part of Lake Erie on September 6th, 2021 (source: [65])

Within the NCCOS web resources, mean HAB severity indices (SI) were found on a yearly basis. Figure below shows these SI averages. In order to assist data labeling activities, it is normal to have also data for very mild HAB incidents or satellite images where the absence of the phenomenon can be observed. This type of data serves for benchmarking the no-HAB state for the classification model and in order to have a balanced dataset that will be unbiased with respect to one of the two classes to be detected by the classifier. It is worth noting in this point that although imagery collected during either year 2002 or 2005 would seem optimal for this purpose, scoring the lowest SI as the figure indicates, the earliest available time for image acquisition from the Sentinel hub service is in November 2016 - bearing in mind that the Sentinel-2A satellite was launched in June 2015 [66].

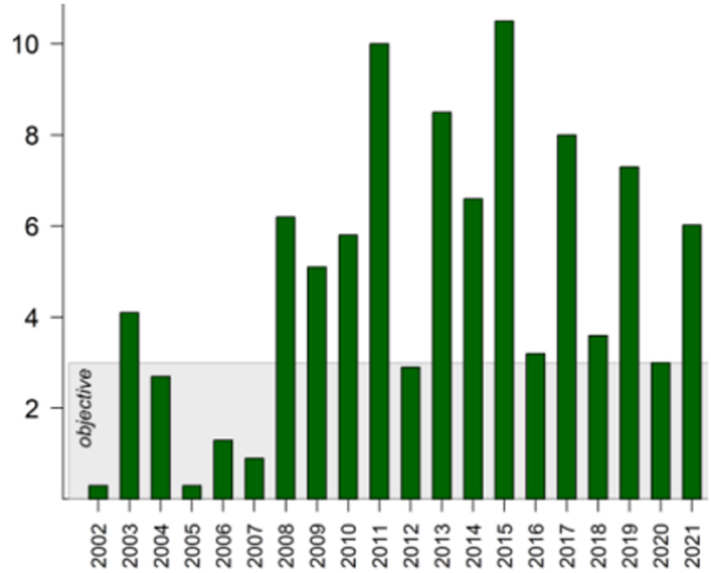


Figure 15: Yearly average HAB severity index for Lake Erie (source: [64]). The objective for SI is set to the benchmark levels of year 2002.

The 6th of September 2021 was selected as the benchmark date for data acquisition of high/severe HAB, whereas the corresponding date for low/mild HAB is the 22nd of August 2020. Data acquisition was made using the bounding box having upper-left/lower-right corners at coordinates: [-83.721252, 41.182788, -81.748928, 42.167475]. Figure below shows the area of interest as drawn from the Sentinel hub request builder web utility and its location with respect to Lake Erie. The region under considerations amounts to a large surface area of approximately 17976 km^2.

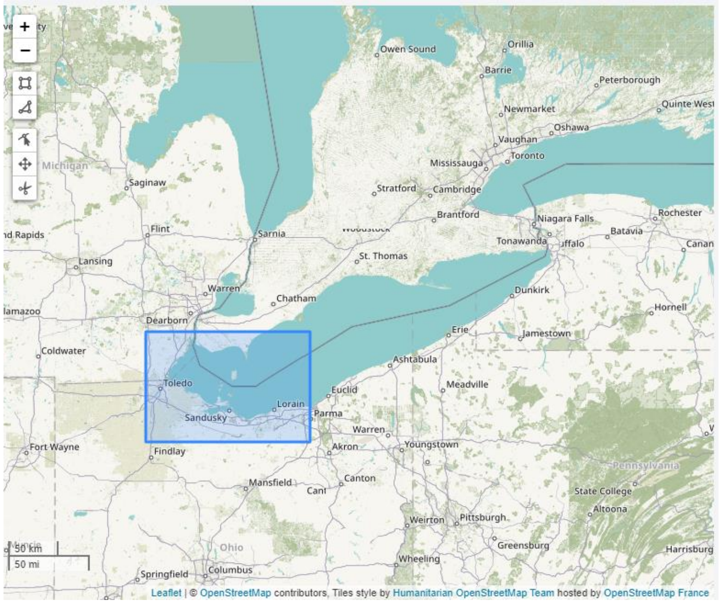


Figure 16: The western side of Lake Erie highlighted. Corresponding to the region of interest defined by the bounding box: [-83.721252, 41.182788, -81.748928, 42.167475]. The surface area of interest amounts to 17976 km^2

The resolution of the imagery was set to just over 60 meters that corresponds to a image width of 2500. This is done to match the maximum resolution achieved by B01 sensor as it is can be seen in Table 1. Cloud coverage was set to 10% or less. Sample images for true color and data products for narrow sensor bands B01, B04 and B08 for the high HAB (6th of September 2021) case region of interest can be shown in the figure below. The narrow band sensor products downloaded from Sentinel hub are acquired in greyscale format. This simplifies further processing by removing the need to convert images from the common three-color channel format (i.e. red, green, blue) to a single greyscale. In figure 5, the same collection of images is shown for the low HAB case (i.e. 22nd of August 2020).

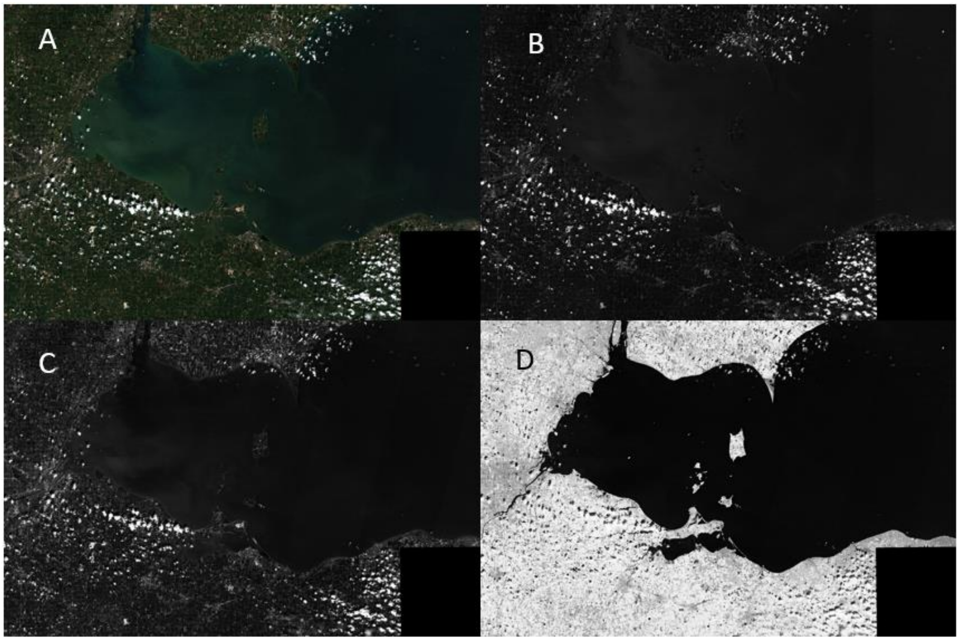


Figure 17: Collection of satellite images for high HAB case – September 6th, 2021. A – true color image, B – B01 Sentinel sensor, C – B04 Sentinel sensor and D – B08 Sentinel sensor. The black rectangular area at the bottom right of each inset corresponds to cloud coverage exceeding the prescribed limit (i.e. 10%).

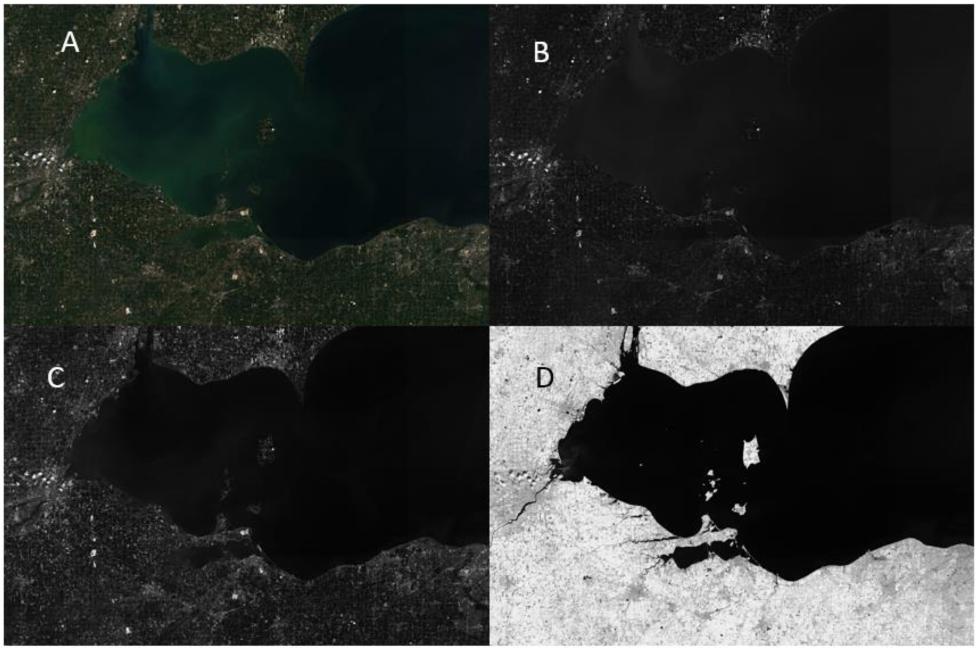
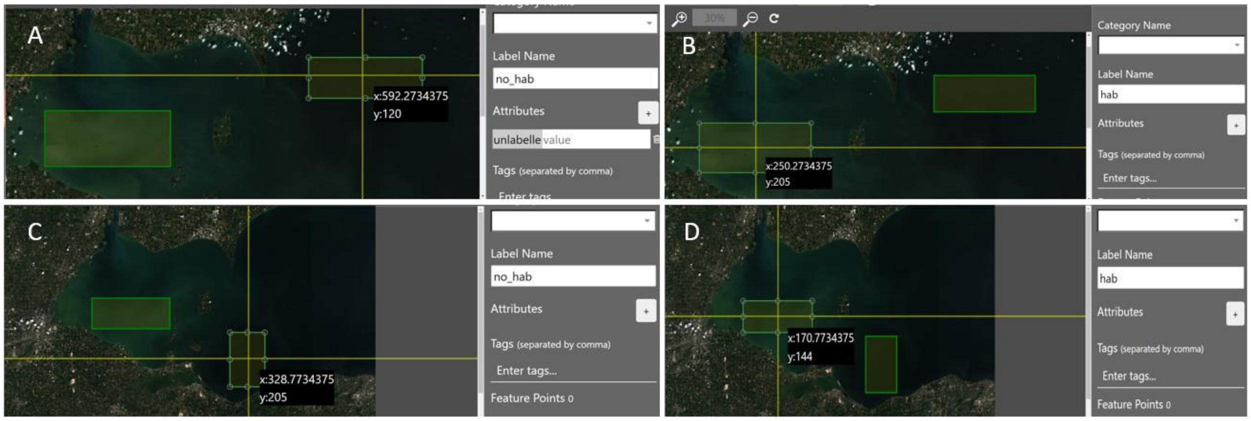


Figure 18: Collection of satellite images for low HAB case – August 22nd, 2020. A – true color image, B – B01 Sentinel sensor, C – B04 Sentinel sensor and D – B08 Sentinel sensor. There are no areas where the cloud coverage limit is exceeded.

In order to perform manual labeling of the dataset, the ImgLab online annotation tool was used [67]. Images below shows the true color image for the severe and mild HAB imagery alongside the labeled areas that correspond to algal bloom as well as clear water patches that no algae growth is observed.



A: Label Name: no_hab; Attributes: unlabelle value

B: Category Name; Label Name: hab

C: Label Name: no_hab

D: Label Name: hab

Figure 19: Satellite imagery for the high HAB (insets A, B for date: September 6th, 2021) and low HAB (insets C, D for date: August 22nd, 2020). Respective HAB and NO_HAB labelled patches in each inset correspond to areas that are in the crosshair.

The manual labeling process in this online tool involves drawing rectangular boxes on images and associating them with specific classes. The output of the labelling process is an XML file containing information about the pixel addresses that each label occupies in the picture. In order to verify the fact that the correct pixel sets are matched with the appropriate labels, the information from the XML export were parsed and assisted in color coding the original pictures using OpenCv image processing library. Figure below shows this comparison and verifies that after labelling in ImgLab web tool, the correct pixels are associated with the prescribed labels. The region of interest studied corresponds to 12,435,000 total pixels. For the case where the phenomenon was at its peak – September 6th,2021 – the manual labeling process using ImgLab resulted in 141,000 pixels labeled as “HAB” and 93,840 pixels labeled as “NO_HAB”. The respective pixel counts per label for the satellite imagery acquired during a period where the phenomenon showed signs of attenuation – August 22nd, 2020 – are 107,624 and 85,774 respectively. These data concerning the pixel sets collected are described in the table below.

Date for Satellite imagery	Label: "HAB" – pixel count	Label: "NO_HAB" – pixel count
September 6th, 2021 (High HAB Case)	141,000	93,840
August 22nd, 2020 (Low HAB case)	107,624	85,774

Table 2: Summary regarding labelized pixel sets gathered from acquired satellite imagery

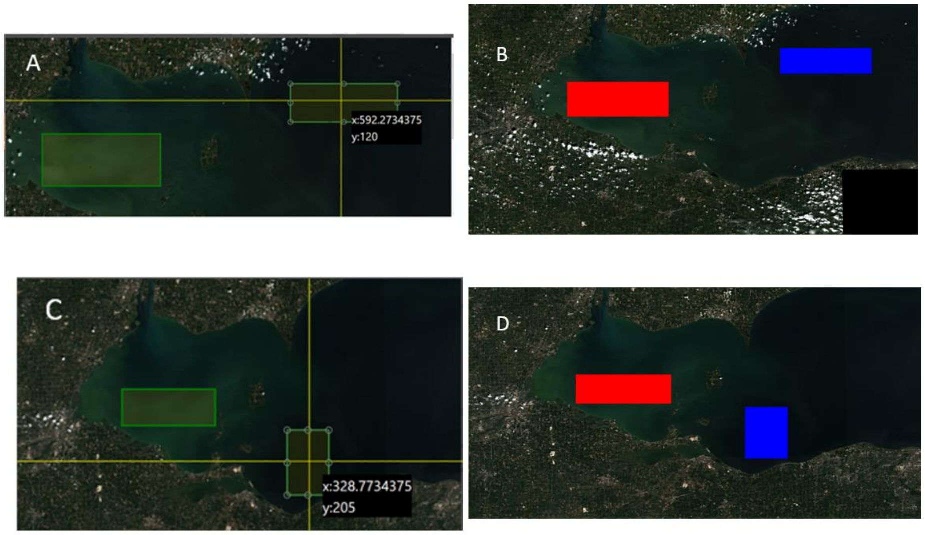


Figure 20: Verification using OpenCv regarding the pixel sets generated during manual labeling in ImgLab. The insets B, D have been color coded (red: HAB areas, blue: no HAB areas). A, B -> High HAB image (products downloaded for date: September 6th, 2021) AND C, D -> Low HAB image (products downloaded for date: August 22nd, 2020)

Processing Pipeline and ANN model

In this section the data processing pipeline is presented[10]. For the two sets of acquired products (High HAB occurrence: September 2021, Low HAB occurrence: August 2020) from the Sentinel hub web service, the steps describe herein are repeated in order to produce the final dataset. This consists of pixels that are labeled as HAB and NO_HAB, with the aid of ImgLab webtool as described

[10] The code for the modelling work done in this research can be found on a public GitHub repository in the URL: https://github.com/giso360/hab_thesis_code

before, and with the associated feature vectors consisting of 19 components - 12 components acquired via narrow band products and 7 that are generated and described in the previous section.

The first step of the processing pipeline is to parse correctly the XML export of ImgLab labelling tool. This yields a set of pixel addresses per label. Subsequently, these pixel collections are randomized and afterwards a predefined subset is selected for further processing. This was done in order to reduce the computational overhead that is associated with transforming a big number of pixels (herein > 450,000) to their respective feature vectors and feeding them to the modeler. The selected limit in the present work was set to 5,000 per label and per set of acquired products; this strategy yielded a final dataset that is fully balanced between the HAB and NO_HAB labels with 20,000 pixels. The image below show the originally labeled images as well as the 20,000 selected pixels super-imposed on the images (shown as a speckle pattern that is designated by the blue and red oval shapes).

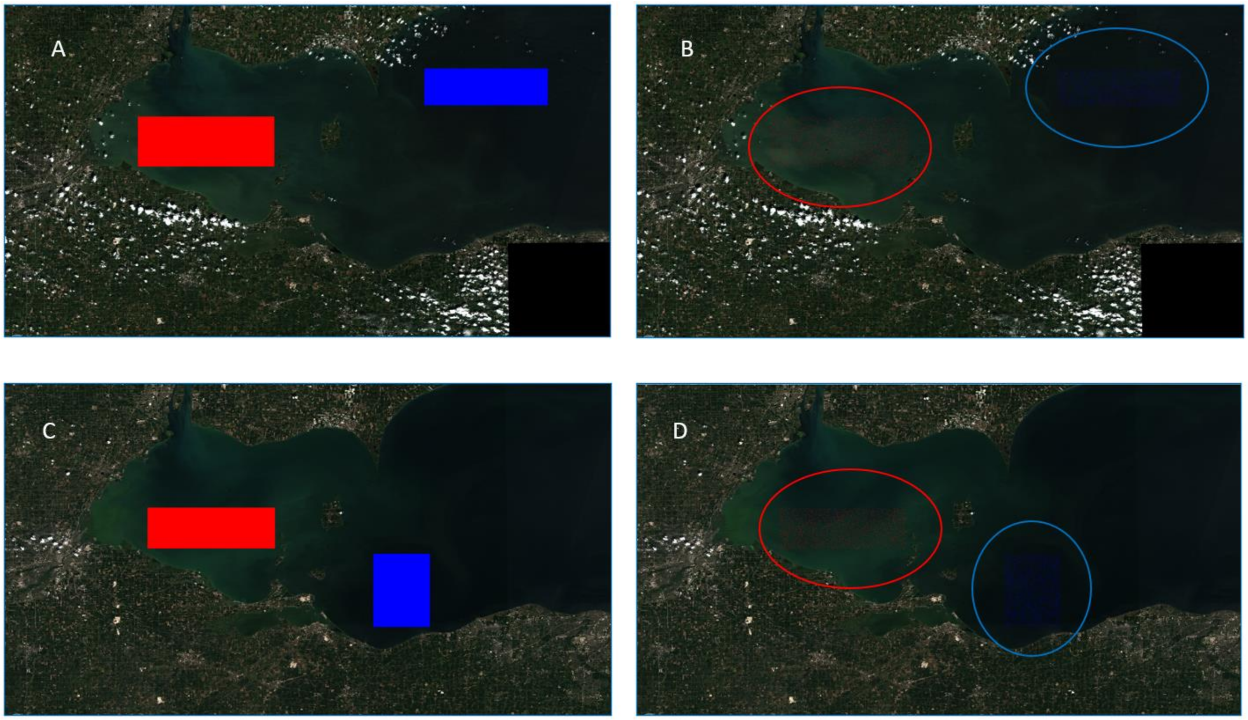


Figure 21: The original labeled areas in ImgLab and the speckled patterns formed by the selected pixels that partake in the final dataset. (A – true color labeled image acquired on September 6th 2021, B – 10,000 selected pixels labeled, C - true color labeled image acquired on August 22nd 2020, D – 10,000 selected pixels labeled)

Prior to embarking in the effort of constructing feature vectors, a quick review of the selected features that participated in the model brought forward the necessity of pre-calculating the statistical mode for narrow-band channels B01, B03 and B05. This is done only for the pixels that are labeled as NO_HAB and it is repeated per Sentinel hub acquired product in order to set the baseline for the calculation of features RD1, RD2, RD3. Table below summarizes these findings. Note that the selected pixels are save to csv files. This forms a mode of permanent storage for the pixel addresses since re-running the code that generates them will result in a different set due to the randomization that is performed.

	Mode B01	**Mode B03**	**Mode B05**
High HAB (Sep., 2021)	14	18	10
Low HAB (Aug., 2020)	11	12	4

Table 3: Summary of statistical modes for channels: B01, B03 and B05 for pixels of "Clear water" (i.e. labeled as NO_HAB).

The next step in the processing pipeline is to generate the feature vectors for the stored pixels. For that purpose, each narrow-band component of downloaded products is read sequentially using OpenCv library for python. The images are fed in greyscale. This simplifies further processing since for every pixel address a single intensity value is acquired instead of the common red, green, blue tri-channel. In this manner, a collection of pixels is formed whereby querying pixels addresses sequentially in each narrow-band image the base components of the feature vector are populated. Using these values, the other features are defined (i.e. NDCI, NDVI, NDVI_B8A, MPHBI, RD1, RD2 and RD3). Note that during this task the labels are also appended in the feature vector. The final action in this step is to store these collections to csv files and perform a manual merge to produce a single CSV that contains all feature vectors.

With the final dataset formed the next step is to perform data inspection/exploration and visualization actions. The CSV with the labeled feature vectors is read using pandas data processing library for python and stored in a dataframe. This strategy simplifies further processing steps since the dataframe API is versatile and provides a wide range of utilities for the activities of data inspection and manipulation. The shape of the dataframe was queried and verified that 20,000 data points were contained in the dataset having 19 feature components and a single label. The HAB pixels were associated with the numeric value 1 whereas NO_HAB pixels with 0. It was also verified that during the feature collection and generation process no records contained missing values. The datatypes of most features were of type int64. Float types were employed only for features that required division operation for their calculation as is the case for NDCI, NDVI, NDVI_B8A and MPHBI components. The describe() method of the dataframe API was used to provide with a rough statistical insight with respect to the distribution of features. Table below summarizes the average values per label for the engineered features NDCI, NDVI, NDVI_B8A, MPHBI, RD1, RD2 and RD3.

Label	NDCI	NDVI	NDVI_B8A	MPHBI	RD1	RD2	RD3
HAB	0.051	0.417	0.453	5.66	13.07	28.4	12.97
NO_HAB	0.054	0.210	0.212	1.225	2.47	3.75	3.58

Table 4: Average values for the engineered features per class label

The data ranges of these features per label were also noted to provide with the information of whether there is scope in truncating the dataset by detecting and removing datapoints that diverge significantly from the dataset distribution. This outlier detection and removal process can be accomplished by either setting a valid range for each feature or by employing a more standard technique such as "six-sigma". Table below shows the value ranges per label for each of the generated features. Using utilities of the sklearn library the dataset was shuffled and saved to a csv file for the subsequent modelling step using ANN with Tensorflow.

Label	NDCI	NDVI	NDVI_B8A	MPHBI	RD1	RD2	RD3
HAB	0 - 0.68	0.03 – 1.0	0.07 – 1.0	0.46 -56.2	1 – 22	13 – 52	3 – 64.0
NO_HAB	0 – 0.33	0 – 0.5	0 – 0.45	0 – 4.47	0 – 14.0	0 – 16	0 - 15

Table 5: Value ranges for the engineered features per class label

Last action of this processing step is to visualize the engineered features both in the whole dataset as well as per label. Since the features are purely numeric - and not categorical - in their nature, histograms are used for information display purposes. Figures below show these histogram plots that were designed using the pyplot utility of the matplotlib python library.

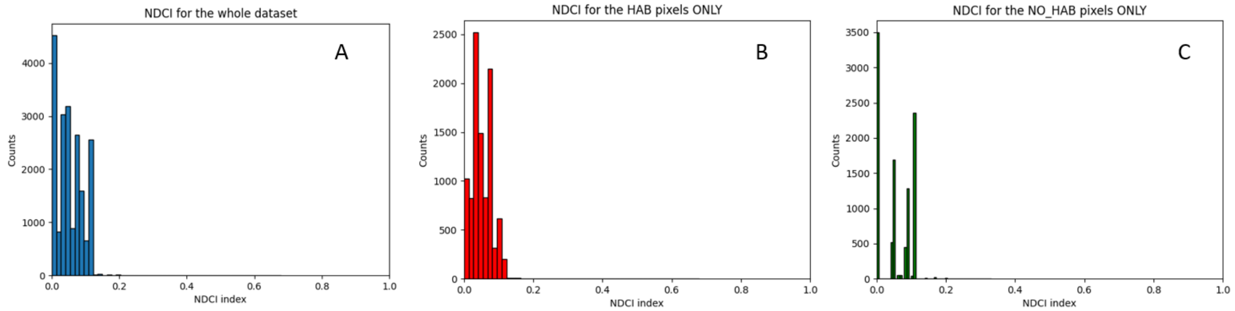


Figure 22: Generated feature NDCI – Comparison between feature distribution for the whole dataset (A), pixels labeled as HAB (B) and pixels for "clear water" – NO_HAB (C)

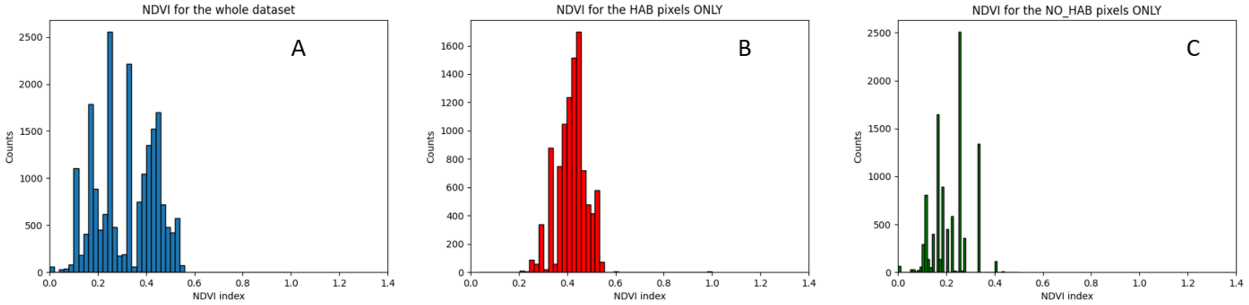


Figure 23: Generated feature NDVI – Comparison between feature distribution for the whole dataset (A), pixels labeled as HAB (B) and pixels for "clear water" – NO_HAB (C)

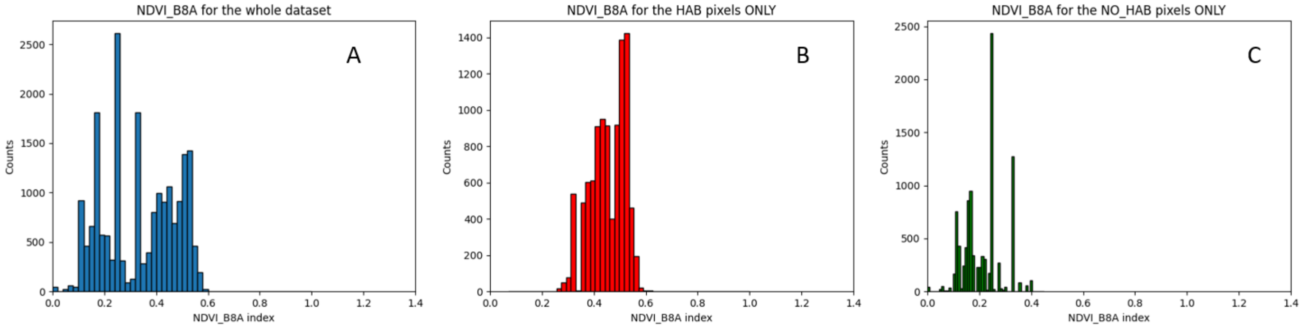


Figure 24: Generated feature NDVI_B8A – Comparison between feature distribution for the whole dataset (A), pixels labeled as HAB (B) and pixels for "clear water" – NO_HAB (C)

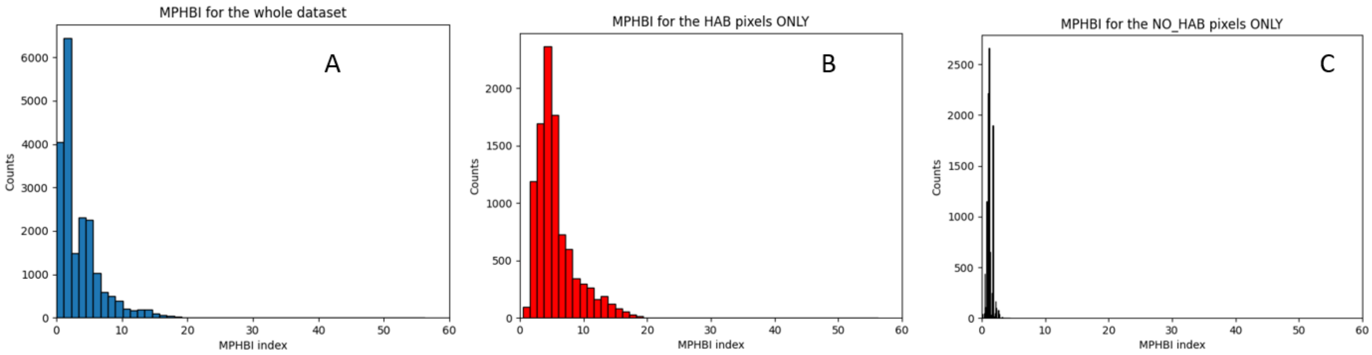


Figure 25: Generated feature MPHBI – Comparison between feature distribution for the whole dataset (A), pixels labeled as HAB (B) and pixels for "clear water" – NO_HAB (C)

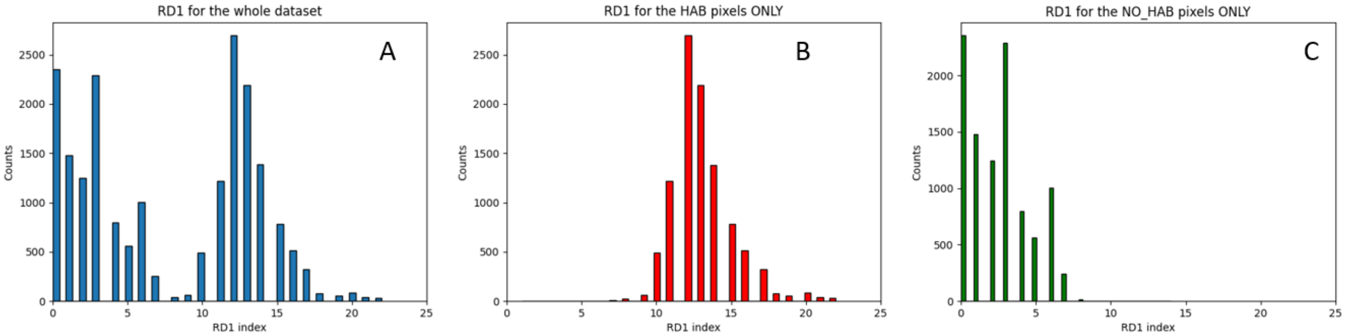


Figure 26: Generated feature RD1 – Comparison between feature distribution for the whole dataset (A), pixels labeled as HAB (B) and pixels for "clear water" – NO_HAB (C)

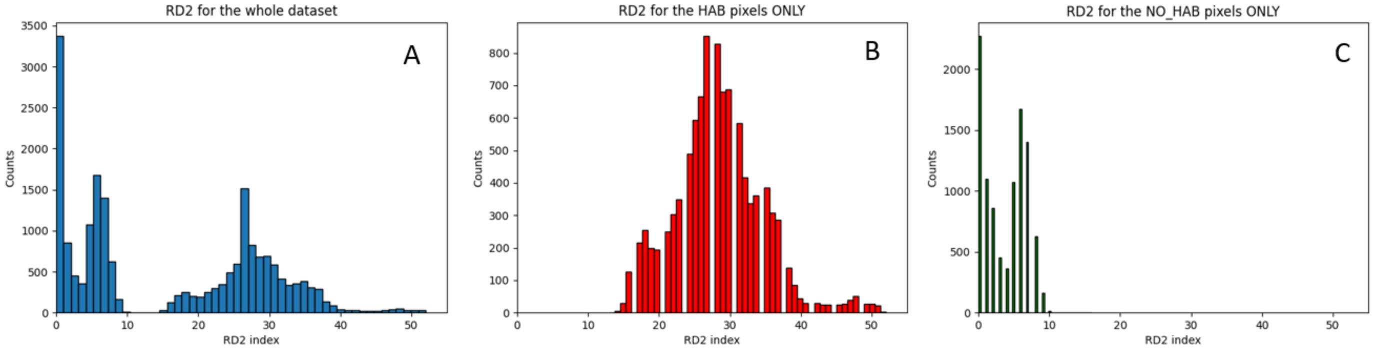


Figure 27: Generated feature RD2 – Comparison between feature distribution for the whole dataset (A), pixels labeled as HAB (B) and pixels for "clear water" – NO_HAB (C)

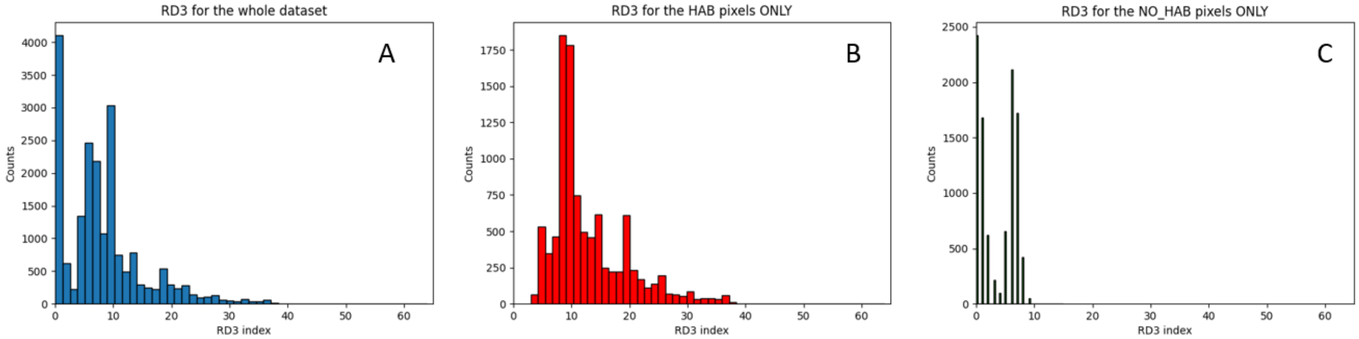


Figure 28: Generated feature RD3 – Comparison between feature distribution for the whole dataset (A), pixels labeled as HAB (B) and pixels for "clear water" – NO_HAB (C)

A preliminary inspection of the histograms shows that the selected features are quite informative for deciding the label per pixel. For instance, the NDVI feature shows a good separation between the class labels at around 0.25 unit. The most extreme observation is done when observing the RD1 feature family of histograms. There, and excluding the effect produced by outliers, there is a certain cutoff point at 10 units. It is obvious that all the pixels under this value are associated with clear water pixels whereas for RD1 values above this limit there are only HAB pixels observed.

The final step of the processing pipeline involves the creation of a model based on ANN that can predict the label of pixels when the appropriate feature vector is given to the classifier. To identify the optimum geometry, a range of ANN networks are explored. The other design parameter explored in these experiments is the activation function per layer. For reasons of simplicity, all layers in each experiment are made up by the same number of nodes. In any other case, the design space that would have to be explored becomes significant for the scope of the current research. The number of epochs and batch size are also fixed and equal to 100.

Initially, the dataset is read as a dataframe using pandas. The feature matrix X and the class vector Y are generated by dataframe manipulation. Subsequently, the dataset is split into a train and a test set with a ratio 70:30. Then, the model is formed using the sequential format that is appropriate for linear topologies. The input layer has a number of nodes that is equal to the number of features. Since the classification task is binary (i.e. HAB/NO_HAB) a single output node is placed as the exit layer of the ANN architecture. The model optimizer and loss type are fixed for all runs as "adam" and "binary_crossentropy" respectively. Metrics that are reported from the model are accuracy and "area under curve" (AUC). The optimizer, loss type and model metrics, are passed on the model using Tensorflow's compile() API as an object argument. F1 score is calculated using sklearn utility.

When these experiments arrive to a model with optimum performance, the model is re-trained. The only difference being that this time the dataset contains less records. Some of the data points are removed from the dataset to serve as a batch of "unseen" data. Once the final model is decided, the model is saved in a file of format h5. Using Tensorflow load_model() method, the saved model is retrieved and the "unseen" data are passed to it to perform the prediction.

The number of hidden layers used in these experiments were 2, 4 and 6. The number of nodes per layer were 10 and 20. The activation functions tried in the model were relu and sigmoid. The results for every experimental run in terms of accuracy ranged from 0.9997 to 1. The same holds true for the AUC and F1 scores. These observations initially led to the assumption that the model exhibited some form of overfitting. Thus, L1/L2 regularization and feature removal was tried to resolve the issue. It was observed though that the model had a 100% accuracy in predicting unseen datapoints. The effort of modifying the model to rectify this situation was abandoned after re-considering the histograms produced during data visualization. Evidently certain features are critical when deciding the label of a certain pixel (e.g. NDVI, RD1-3). Thus, for the scope of the given problem, the model constructed herein is general enough to predict efficiently the label of pixels. The final model that was promoted from this modelling activity possesses 4 hidden layers with 10 nodes each while each node uses a sigmoid activation function.

HAB Remote Sensing Service

The present chapter aims to offer a detailed description of the solution that is explored in the present work. Its main features are going to be outlined bringing the focus to the various areas of customer needs that it addresses. Each of the service features imposes limitations with respect to its technological implementation and add to the overall complexity of the proposed system. The end goal is to come up with a robust system capable of integrating a variety of data coming from various sources that are employed for water monitoring purposes.

A second topic covered in this chapter is the illustration of certain design solutions that can successfully accommodate the features of the proposed service. In order to achieve this, the types of data and their potential transformations are reconsidered from a systems' point of view. Moreover, the communication requirements between different components of the system are described. The potential protocols of communication regarding the sensor grid utilized for the solution as well as the service delivery to the end users fall also within the scope of this chapter.

Service Description and Features

The service aims to offer timely information with respect to water bodies in an affordable fashion. Certainly, the usage of satellite imagery obtained from a third party is extremely competitive cost as well as time wise. The output of the service can be utilized from the relevant stakeholders, both for fresh water as well as coastal regions. Each of these target areas bear an economic as well as a societal value and has to be protected from the adverse effects of harmful algal bloom according to the contamination mechanisms that have been outlined in the research literature.

The main benefits of integrating to an existing water management system data analysis that is collected from planetary observation artificial satellites is that the monitoring coverage is enlarged beyond the ordinary water sampling methods. Moreover, data can be acquired for areas that are remote and hard if not impossible to access. Due to the cost-effectiveness of the method, augmenting existing water management suites with satellite capabilities is a logical step for the protection of water resources. Nevertheless, relying alone on planetary observations for the detection and forecasting of harmful algal blooms is not sufficient. Water sampling and in-situ sensor arrays will always provide with strong quantitative data that provide with the appropriate validation and benchmarking for machine learning models based on satellite images that are charged with the task of measuring levels of chlorophyl and other HAB related toxins.

The data veracity provided by chemical sensing is of course unparalleled. The major drawback of these methods though is that sampling becomes extremely localized. By means of data fusion of in-situ with satellite data, color signatures acquired from multispectral space sensors can easily be correlated with the levels of measurands collected by sensor arrays deployed on the field. Besides constructing robust ML models, this data fusion technique offers the opportunity to gain insight with respect to the distribution of chlorophyl-a in areas located between in-situ sensors (i.e. via interpolation) or beyond the installed grid (i.e. extrapolation). This can easily yield contour-like plots that are far more informative and accurate when compared to plain field sampling. The data fusion is not only restricted to the combination of in-situ with satellite water characterization techniques. In order to remove all restrictions that are imposed by weather conditions, that limit the data quality of satellite imagery due to cloud coverage issues, drones that carry optical or chemical sensors can be used for a robust water management system. Able to operate/fly below the weather and without relying on the time resolution dependency of satellites that have scheduled orbital paths and cannot be located constantly/synchronously over an area, utilizing drones alongside the other sensing methods referred here, form a strong arsenal for water monitoring purposes.

Thus, in a nutshell, the proposed service consists of a cloud-based water monitoring system calibrated for HAB related pollution monitoring and able to combine and/or fuse data from three different types of datasources - namely, in-situ, drone and satellite. This conjunction of methods forms an intelligent system capable of data analysis and alertness for future HAB incidents that can be utilized for the timely deployment of mitigation platform and public as well as economic protection. The minimum viable product that can be realized with the combination of these monitoring resources as well as the data pipelines that are discussed in this work, can be augmented by data from weather sensors. Valuable macro-environmental input are water temperature as well as water current mapping to be able to allow predictions about the transport and evolution characteristics of HAB phenomena. This solution is able to fully digitalize and transform water quality monitoring tasks by providing high-resolution, three-dimensional observations.

To accomplish the tasks described before the service has to have certain core internal technical features. Building on top of these central service facets ensures the continuous development of the HAB detection platform to match the ever-changing status of the market and the needs of the customers. These core features can be summarized as:

1. Remote Data gathering: Pipelines from Sentinel and other providers that are able to provide planetary observation satellite imagery available at regular intervals to satisfy the needs of customers.
2. On site Data Collection: Pipelines for data generation, collection and storage from drones, in-situ sensors and laboratories charged with the task of conducting measurements and laboratory analysis
3. Pre-processing: Pipelines and algorithms that are compatible to process raw data and transform them to feature vectors that are produced from the fusion of the various datasources
4. Adaptive Machine Learning Models: Enforced re-learning in cases where temporarily the effectiveness of the service falls short when compared to the specified parameters. These criteria could be defined within the service-level-agreement that is activated upon customer subscription and outlines the quality related specifications for the service in question
5. Integration/Fusion of Data: Integration of data gathered from in-situ and drone units to incorporate these in the models for labeling purposes. Scope for enhancing those models with other macro-environmental types of data (e.g. temperature, water currents) as well as hydrological models
6. Models Correlation: Correlate satellite imagery with the presence of the HAB phenomenon and its severity
7. Reporting: A service that is able to provide the customers with meaningful and self-explanatory reports that they can use to issue their next actions
8. Archiving: A service for the customers as well as the general public. This feature allows all stakeholders to build an understanding and raise awareness about the phenomenon and use this knowledge for their ends - regardless of the fact that this might be economic, societal or educational in its nature – by gaining access to historical data from the service that allows them to view HAB incidents at larger time scales
9. Alertness and notification: A service feature that is able to inform customers via the appropriate and predefined communication channels about an abrupt evolution of the phenomenon that requires immediate action to prevent some environmental catastrophe

Service Design Issues

In this section, the requirements in terms of subsystem interoperability are discussed. The starting point is the feature list of the HAB service outlined before and some preliminary assumptions that mostly deal with the collection of data from the various datasources that do not operate from the earths' orbit. Each feature is accompanied with its own hardware and software requirements for its operation. Moreover, the communication and integration between these subsystems is of paramount importance and issues like fault tolerance and synchronicity among the different modules of the service need to be addressed in order to be able to deliver to customers the features that are needed and have subscribed for.

As an initial overview of the system, it has been hypothesized that the HAB detection will be offered via the web and in a software-as-a-service (SaaS) format. The usable outcome is provided to the customers using a cloud infrastructure. This diminishes the need for drawing strategically the server network for the provision of the service. In any other case, the promoted choice would be to have regional servers for customer access. Probably, these servers could be distributed geographically to address the needs for customers and other stakeholders in a per-continent basis (i.e. 5 servers for access and provisioning purposes). Finally, and because the learning models have to be adhoc for specific regions due to variations of the color signatures of native algae species across the globe, the ML server can be located in a specific central node of the HAB service network. The task of this server is to gather all data required for learning and re-distribute the trained models as executables back to the service provisioning nodes in the cloud to be utilized by the end-users.

For the purpose of gathering satellite data, the API of the provider is used. In this manner, data access can be automated and also scheduled according to the service-level-agreement with the customers and the flight plan of the orbital sensor array. For the purpose of on-site data collection from drones and in-situ chemical sensors, the acquired data must be funneled to the data collection/fusion software unit. The operation and control of these sensor arrays can be regulated with networks that are suitable for internet of things (IoT) applications. The data can be transferred via satellite. A more cost effective option is to use the recently spreading of 5G network nodes and infrastructure for that purpose that allows the creation of mobile adhoc networks of extremely high speed and data bandwidth. Data can be published in a message passing software integration unit (e.g. Apache Kafka or UDP multicast networks for information exchange amongst components) and collected by the recipient applications that have subscribed to the appropriate message type. Lastly, for the case of collecting data by regular water sampling by laboratories, a custom web service is appropriate. Through this, researchers and laboratory technicians can upload their data in a predefined format (e.g. csv). After these uploaded files are collected and parsed, the contained data be added to the integrated HAB dataset.

The data pre-processing entitles the transformation of raw image and numerical data to feature vectors that are suitable for subsequent learning. Some of these transformations are explained in the previous chapter with the main focus on feature selection and generation. The 19 features presented there, constitute a robust initial basis for the formation of machine learning models. It is important to note that the collection of data from different datasources can create different models for the same region. One solution to make the most out of this, is to implement ensemble models and finalize the weights per model-participant with respect to the synthesized score that predicts the output label. These models can be informed further by meteorological and hydrological data that can be acquired from third-party web services.

Another important requirement is for the models to be adaptive. To accomplish this, frequent re-train cycles are required. To appreciate the scale of this task it needs to be reminded that since no HAB model can have global validity, normally the model population is to be comparable to the actual customer basis for adhoc service provisioning. The re-learning cycles can either happen at regular intervals or by setting thresholds in the accuracy during operation. Should the deviations between remote sensing data and other types of collected data become significant, an alert system can notify the service operator to commence a new learning cycle. Using this methodology, re-training can also be initiated automatically should the deviations in model performance exceed a certain level. The decision for what type of re-learning pattern to be followed can be coined by the service-level-agreement itself.

The models developed by the HAB service perform the task of correlating HAB phenomenon with images. Although in the present work a binary classifier is investigated (i.e. HAB/NO_HAB), this is not sufficient for the end-customers. The needs of the customers are focused on getting information about the severity of the phenomenon. This is mainly due to the fact that they abide to the directives of regulatory agents charged with the task of monitoring the usage of water resources and enforcing penalties in the cases where the quality standards are not met. For that purpose, the classification engines of the HAB service must be able to predict a discrete list of classes (i.e. multiclass) to provide with the necessary resolution for high service quality.

In order to provide customers with human readable results, a reporting service has to be in place. The customer should be able to choose from a variety of report templates that can range from a concise up to a very detailed document. This can be offered for free or prescribed by the service-level-agreement between the binding parties. For this service feature, it is critical to decide information-rich visualization and tabulation formats for the output data. The generation of the reports can happen at regular intervals (e.g. once a month) and can be forwarded to mail or file servers of the client without being requested by him/her.

The archiving and alertness/notification features of the service can be realized as different applications that can operate over HTTP and probably by using the REpresentational State Transfer (REST) protocol as RESTful APIs. The alertness service in particular should be able to push notifications to the service clients probably via email or SMS to authorized employees of the customer.

This description of the solution proposed herein facilitates any further architectural activities to combine the subsystems of the proposed service. The current trend is to utilize a micro-service based architecture where each application participates in the overall service ecosystem as a discrete unit while exposing its API to the rest of the members. Although the discussion about this architectural pattern is vast and outside the scope of the present work, it is worth mentioning that major benefits of following this architectural pattern is that the services within the ecosystem can be realized in different computer languages and software frameworks as long as their API are of well-defined scope and exposed to the dependent parties. This pattern offers opportunity for simplicity and fast developments cycles as well as re-usability. Candidate application servers for the ecosystem in question based on the present discussion are:

1. Data Collection application server
2. Data Integration/Fusion server
3. ML server
4. Reporting and alertness mail server
5. Archiving web server

HAB Service: Business Analysis

This chapter aims to explore the HAB service presented in here from a business perspective. Firstly, any information regarding the current state of the water management and monitoring market are presented. For that purpose, an interview with a subject matter expert was conducted. The next step is to formulate hypotheses with respect to the needs of clients by forming the customers' profile. The business features of the HAB service must, at least partially, overlap and satisfy the clients' needs. Having conducted these tasks, the activity that follows is to draw core business concepts for the HAB service that will support subsequent planning endeavors. Finally, the pricing model for the HAB service is explored with the scope of capturing the cost scale for its operation and by adding on top of this an appropriate mark up to allow for profitability that ensures service continuity.

State of Market

The total market size for water quality monitoring systems globally, is expected to reach $6.7 billion by 2025, from a $3.8 billion in 2017, growing at a compound annual growth rate (CAGR) of 7.3% from 2018 to 2025 [68]. The markets' subdomains are the utilities and industrial sector as well as the usage of water quality monitoring technologies for commercial and residential customers (e.g. pH measuring kits for swimming pools). Core products of this market are pH and conductivity sensors and dissolved oxygen analyzers. For the case of HAB detection, it is assumed that only the utilities domain as well as some governmental agents form a set of potential buyers.

The global market size for HAB related products and services, was not found in the literature. Nevertheless, a measure for it can be obtained indirectly by measuring its economic impact as this dictates the marginal expenditure required for the aversion of the phenomenon. As it was discussed in the literature (see Table 2), the annual cost for the EU and the US amounted to $895 million in the year 2005. Thus, a safe estimate would be that currently that figure - and extrapolating globally - is of the order of $6 billion on an annual basis. This includes costs related to health, aquaculture, leisure industry as well as for quality monitoring and mitigation purposes. For the future task of building a cost model for the service proposed here, it is assumed that the overall market size (business scope) for HAB is around 2-4% of the overall water management market size (i.e. \$120 - \$240 million p.a) to achieve a significant cost saving effect on commerce and communities that are affected by the phenomenon.

In order to gain a much better insight in the market and related technologies, an open interview was conducted with an individual who is a subject matter expert in the domain of environmental remote sensing applications for water management. The interview was recorded for future reference and a transcript was generated and shared between the participants. The interviewee was Professor Stefan Simis who is a project coordinator of MONOCLE-H2020 that is aiming to develop earth observation-based solutions for inland and coastal waters. He also holds a position in Plymouth Marine Laboratory as an earth observation scientist and is involved in the Copernicus Land Monitoring Service - Lake water quality project [69]. During the interview several topics were discussed such as the state of current technologies used, the variety of potential services that could be spun out from the work done on MONOCLE-H2020 [70] project as well as the system architectures that could be involved in realizing water monitoring service where orbital planetary observations play a key role. The main focus areas of the interview regarding the business side of the issue were: (1) to identify the potential stakeholders and agents that would be involved in a buy-in and/or subscribe to a HAB monitoring service as well as (2) speculate on future prospects for the market in question.

Although in the present work ImgLab was used for labeling purposes, during the interview with Professor Simis it became evident that for a quality service based on remote sensing data, in-situ validation is of paramount importance to be able to make confident use of satellite imagery - that is of course the most cost-effective method of water quality data acquisition. In his words and during describing the work done in MONOCLE-H2020: "**... the project is focused on developing the technology that will make in situ validation of satellite products cheaper and more widespread as a result of being more affordable**". From a purely technical point of view, a question that arose was whether a system that relies on remote sensing data has also to have some knowledge of the flight plans of satellites in order to know in advance when a certain area of interest is being observed from orbit and satellite products become available. From the discussion it was found that this is a minor concern since nowadays satellite overpasses are very frequent and therefore there is no need/problem to make in-situ measurements coincide with an overpass. The single eventuality where satellite scheduling is mandatory is when requesting high resolution data and want these readings to coincide with in-situ measurements.

During the interview with Professor Simis, a range of types of stakeholders were discussed as potential end-buyers. These include: (1) national Authoroties, (2) municipal authoroties, (3) water utility agents, (4) leisure/recreational, (5) aquaculture as well as (5) insurance companies. The opinion of the interviewee is that from this list, probably water utilities and aquaculture businesses would more likely be interested in upgrading the available resource monitoring systems as these types of stakeholders suffer the direst effects of HAB and other environmentally adverse incidents. In the exact words of the interviewee during the conducted dialog:

- George Skoufias: **"...Who would buy in such a product first? I mean, who has the bigger stakes over there, do you think?"**

- Prof. Simis: **"...I think between water utilities and aquaculture, there is...water utilities companies have to pay fines for any sewage bills, for example, those are really, really big factors to determine their investment"**

Nevertheless, one type of stakeholder that was not anticipated prior to the time of the interview, was national authorities. In the words of the interviewee: **"...if you were to have a national player with lots of local interests, I think that's probably where you can see more, rapid uptake because they would know which sides to focus efforts on..."**. From a purely business perspective, all the interested parties - nation-wide - can request from a public national agent to provide this type of service. Therefore, a company that specializes in orbital earth observations and satellite image analysis can provide these products to national agencies (to some domain of environmental ministry for example) and these can afterwards be distributed to municipal and local agents (private or public) to diversify their access to the available infrastructure for environmental protection. At some point this is already being done within the European Union as Copernicus is a product of collaboration between the nations of the continent and that comes in some form of public investment on building more robust infrastructures for the future. Business-wise what became evident from the interview is that the next step, after the multi-national collaboration described here, are public-private joint investments.

In conclusion, this interview was very informative on many aspects that are tangent to the HAB service proposed in this work. Nevertheless, the main outcome for further business analysis was that from all the potential stakeholders and end-customers hypothesized in the beginning of this research, national authorities are going to be profiled further since they are probably the first stakeholder category that could express an interest. Although the current trend of the market is to focus on direct sampling methods and in-situ sensor arrays that are more expensive to deploy, operate and maintain but come with the benefit of providing with very accurate measurements, satellite-based solutions and data-fusion techniques are expected to gain ground bearing in mind the aggravation and/or increased frequency of occurrence of phenomena that are closely related to climate change. Harmful algal blooms, being recognized as co-stressors of climate change, are expected to be in the center-point of these endeavors in the future.

Customer Profile and Business Modelling

This section covers the topic of correlating the client profile of the selected stakeholder type (i.e. water domains of environmental national authorities) with the business features of the HAB service. Subsequently, the outcomes of this correlation help to identify the service features that qualify for the final deliverable to the market. Based on these, key aspects of the business can be coined and provide the appropriate information for drawing a more detailed business strategy and plan. The chosen tools herein to perform these tasks is the Strategyzer framework and more specifically the value proposition [71] (VPC) and business model [72] (BMC) canvases.

In particular, the VPC aims to create a match between the product/service features and the needs of customers. Namely, the product-market fit comprises of two segments - the customer profile and the value map. The market/target-customer is profiled by forming hypotheses regarding the functions the agent performs (i.e. "customer jobs"), the needs and expectations from using a service (i.e. "customer gains") as well as any situations of some adversity that the customer experiences (i.e. "customer pains") during task delivery. The value map segment of the VPC is concerned with the proposed value that the service/product has to offer with respect to the customer profile. It lists the product/service marketable features (business & services segment), the aspects that the customer gains value by using the service (gain creators), as well as the characteristics that relieve or completely alleviate customer pains (pain relievers). For the proposed HAB service, the different areas of the VPC are filled as:

1. **Customer Tasks**:

- Maintain early warning systems for HAB incidents
- Continuous education of stakeholders
- Build archive of field data for monitoring
- Abide to environmental laws and regulations

2. **Customer Pains**

- Accessibility to remote areas for monitoring activities
- Bear costs for deployment and maintenance of environmental monitoring equipment
- Monitoring tasks are labor intensive
- Monitoring tasks require IT infrastructure
- Coordination between interested stakeholders to disseminate results
- Informed about amendments in regulations
- Bear cost of penalties due to mis-compliance

3. **Customer Gains**

- Control economic impact of HAB incidents
- Maintain environmental quality that allows all relevant stakeholder business entities to continue their activities
- Encourage stakeholder and citizen participation, raise awareness and engagement
- Better specificity of monitoring techniques to phenomena like HAB
- Contribute to the general knowledge about the phenomenon
- Reduce penalties

4. **Product/services**

- Gathering satellite data
- Perform data fusion/integration with providers of in-situ and drone monitoring rigs
- Correlate satellite imagery to HAB status/severity
- Reporting and notification system

5. **Gain Creators**

- Early warning tool for HAB incidents
- Frequent reporting according to SLA

- Adaptive HAB models with varying environmental factors

6. Pain relievers

- Remote sensing at large spatial scales and at remote areas
- Tapping on ESA/NASA infrastructure for earth observation to minimize costs
- No labor required
- Service deployed as a SaaS
- Early warning that allows rapid action; hence potential to avoid paying fines for transgressing environmental laws

The BMC on the other hand is segmented in the following 9 areas that answer certain business questions:

1. Value proposition: Which are the product/service features that satisfy customers?
2. Customer Segments: Who the product/service is targeting?
3. Channels: How the business communicates with its customers and delivers the product/service?
4. Customer Relationships: How relationships with customers are maintained?
5. Revenue Streams: How the business generates revenue?
6. Key Resources: What resources are required for product/service delivery?
7. Key Activities: What activities are required for product/service delivery?
8. Key Partnerships: What collaborations will benefit the business?
9. Cost Structure: What are the main cost factors for running the business?

For the HAB service proposed here the respective BMC segments are filled as:

1. Value proposition

- A low-cost HAB satellite-based detection system
- A high accuracy system as a result of data fusion
- Early warning and notification for future HAB incidents

2. Customer Segments

- Governmental environmental agencies

3. Channels

- SaaS

- Website

4. **Customer Relationships**

- Informative customer on-boarding process
- Reporting and notification system
- Online help

5. **Revenue Streams**

- Subscription-based service
- B2B and B2G partnerships for research and development

6. **Key Resources**

- API for collection of orbital data
- In-situ data
- Drone data
- Verification pipeline with laboratory data
- HAB detection algorithms and pipelines

7. **Key Activities**

- Gather satellite data
- Perform data fusion with drone, in-situ, laboratory and weather data
- Maintain accurate models
- Maintain API and other services (e.g. notification)

8. **Key Partnerships**

- Environmental drone operators
- In-situ field data gathering operator
- Satellite imagery provider
- Laboratories that perform gathering and analysis of water samples

9. **Cost Structure**

- Satellite API usage

- Cost of deploying and running cloud service

The respective VPC, BMC canvases in a diagram format can be seen in appendices B and C.

HAB Service: Pricing and Cost Issues

This section concerns the estimation of a cost figure for the HAB service. The scope is to capture the correct cost scale rather than arriving to an accurate estimate. Subsequently, a reasonable markup is going to be added to that cost to justify the viability of the business. For this task, the memory usage of the models developed as well as the system design considerations are going to provide with the necessary input for this analysis.

The cost model is partially based on a set of techno-economic assumptions that were coined in the previous business modelling section. These can be summarized as:

1. The potential market size per annum is $120-240 million. The final figure to form a hypothesis is taken as $200 million.
2. The company can aim for a net 2% of that market at its onset (i.e. $4 million)
3. The operational cost components as seen in the BMC are: (a) satellite API usage and (b) cost of running cloud service
4. Key partnerships that enable data fusion and validation activities are not considered cost-wise
5. Income is purely due to customer subscriptions
6. The initial customer base for the first year is 10 subscription-based accounts that require HAB remote monitoring services for an area of 10,000 km^2 each

The cost model is partially based on a set of techno-economic assumptions that were coined in the previous business modelling

section as well as the service design section where all the micro-services of the ecosystem were listed. These assumptions can be summarized as:

1. The potential market size per annum is $120-240 million. The final figure to form a hypothesis is taken as $200 million.
2. The company can aim for a net 2% of that market at its onset (i.e. $4 million)

3. The operational cost components as seen in the BMC are: (a) satellite API usage and (b) cost of running cloud service
4. Key partnerships that enable data fusion and validation activities are not considered cost-wise
5. Income is purely due to customer subscriptions
6. Before the customer can start using the service an initial pre-train cycle must take place as part of client on-boarding
7. No model can serve more than one customer (i.e. one model per customer)
8. A new model is required every month to capture any seasonality and abrupt meteorological and hydrological alteration effects
9. The machine learning server needs to be in its "ON" state only when modelling takes place
10. No adaptive HAB modeling is considered during this cost evaluation exercise
11. For the generation of each model at least 30 sets of satellite products are used
12. For each set, 100,000 pixels are picked for model generation. Thus, **1,000,000 pixels** are required per model
13. To serve the clients with zero interruptions and unlimited requests performed to Sentinel hub API, a premium account of the satellite web service is required

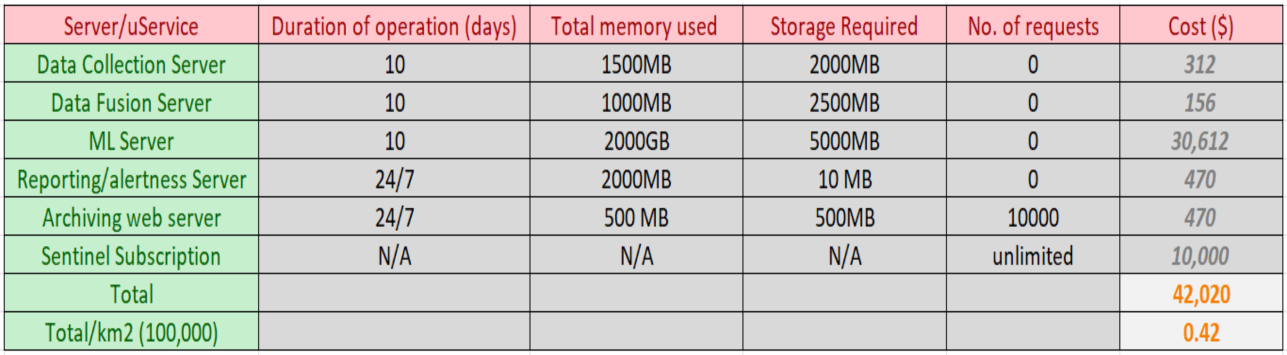
Other findings of the current work that are useful for this analysis deal with the memory usage and data manipulation involved in the services' operation. Based on the work done during HAB modelling the following outcomes are of value for the task at hand:

1. For a total of 20,000 pixels (randomly sampled by 2 set of satellite products), the memory required for the learning process alone was **358.5MB of RAM**. This figure was found by using the memory_profiler python utility that was installed via pip
2. The learning for an ANN of 6 hidden layers having 20 nodes each took approximately **5 minutes on a 8GB RAM station**
3. The pipeline itself and all the processing prior to training the model was found to consume around **20MB** of RAM for approximately 3 minutes with a 70:30 train/test split
4. The size of the downloaded data products from Sentinel hub API amounted to approximately **0.25 MB**. This is important when bearing in mind the storage requirements of the file server where the raw downloaded are saved.
5. For the given model the size of the executable model (.h5 file type) is **0.1MB**
6. The size of the libraries used in the modelling (e.g. matplotlib, Tensorflow etc.) is **1,7GB**
7. Data from Sentinel hub were requested at a resolution of 60 meters

Other assumptions that facilitate the drawing of a complete cost profile involve the following:

1. The archiving server is supposed to sustain a load of **10,000** requests per year. This includes simple access to the website of the service where general info for the HAB phenomenon is listed and of course history data can be accessed in CSV or image formats. From these 10,000 request it is assumed that only **1,000 involve data retrieval and downloads**
2. Compared to the image data gathered from the satellite provider, the volume of data from other datasources is insignificant
3. For the learning process we assume 4 data acquisitions per month
4. The resolution for the acquired data is 10m
5. The memory specs for the virtual machines in the cloud are 16GB RAM

For the cost calculation task at hand, each micro-service is considered according to its operational time memory and storage requirements as well as the amount of user requests issued to its API. To produce cost figure per micro-service ecosystem participant, the Amazon Web Services cost calculator utility is used =[73, 74]=[73, 74]. The cost component related to a premium subscription to Sentinel hub is found the providers' website. The acquired access is considered to match the characteristics of "Enterprise L" service offering. The results of this activity are shown in the table below on a yearly basis based on all the assumptions outlined before.



Server/uService	Duration of operation (days)	Total memory used	Storage Required	No. of requests	Cost ($)
Data Collection Server	10	1500MB	2000MB	0	*312*
Data Fusion Server	10	1000MB	2500MB	0	*156*
ML Server	10	2000GB	5000MB	0	*30,612*
Reporting/alertness Server	24/7	2000MB	10 MB	0	*470*
Archiving web server	24/7	500 MB	500MB	10000	*470*
Sentinel Subscription	N/A	N/A	N/A	unlimited	*10,000*
Total					42,020
Total/km2 (100,000)					0.42

Table 6: Breakdown of operating costs per application/web/ML server in the proposed HAB micro-service cluster – assumed initial clientele size of 10 accounts requesting remote monitoring services for 10,000 km^2 each

The table above informs about the cost side of the HAB service delivery to the customer. For the actual price calculation further research needs to be conducted. Of course, and from a purely entrepreneurial standpoint, the business development strategy that is to be followed dictates what would the capital needs be in the near future and this information allows the appropriate markup to operate on the marginal cost to define a price tag for service usage that guarantees business viability. Nevertheless, the service is offered as a SaaS and customer subscriptions can be arranged in tiers to assist the go-to-market and selling strategies. It is important to note that the service can be marketed as FREEMIUM; thus, some features can be offered for free whereas if customers require access to further functionalities, they can select the subscription package that fits their needs. This marketing strategy has been shown to be successful in generating customer interest. For the HAB service explored herein the possible subscription tiers could be:

1. Simple data acquisition of satellite data in image or CSV format (**FREE service tier**)
2. Acquisition of data that it has been amplified by mean of data fusion from other datasources (drone, in-situ, meteorological, hydrological). This tier allows marketing the HAB service as a data platform able to provide other businesses with the appropriate datasets that will support their models and ultimately their service delivery activities
3. The usage of the service for HAB detection and severity prediction with the procurement of reports and receival of notifications at regular intervals. Alerts for imminent HAB incidents are also generated and relayed via customer relations channels that have been specified during customer onboarding
4. Service pack is same as above but with the added feature of adaptive modeling whenever prediction efficiency falls below a certain threshold. This event triggers an immediate re-train cycle and the customer/region-specific model gets updated

Discussion

This section offers the opportunity to conduct critique on the work done thus far in order to accomplish the aims and objectives defined in the beginning of the research. Major talking points deal with the assessment of the activity outcomes regarding the data modelling and the formation of a viable business model based on the HAB service features as well as the profile of the targeted audience.

As far as creating a HAB prediction model is concerned, it should be noted that all work dealt with data processing manipulations on a pixel level. This choice comes with the drawback of having to handle a big number of data points. An alternative to this, revising the research literature on the subject, is to manipulate and group pixels in tiles and afterwards perform the construction of feature vectors. This is lightweight in terms of model training requirements and allows for more satellite data products to be included for ANN modeling purposes without adding extra memory overhead during model training.

The other issue with the proposed ANN model arouses when looking at classifier performance. For a wide range of ANN architectures, the efficacy of the model in terms of F1 score, accuracy and area under curve performance merits implied overfitting issues. At first glance, this issue can be tackled by using a larger dataset either by selecting more pixels from the labeled regions of the acquired satellite products or simply by passing more satellite product through the pre-processing pipeline. Nevertheless, a careful observation of the histograms for the generated dataset yields the conclusion that certain engineered features, such as RD1, are extremely performant in segregating the two class labels (i.e. HAB/NO_HAB) at well-defined cutoff pixel intensity values. Although this comes as a surprise while rendering the implementation of any statistical learning technique redundant - since the model could be constructed using a finite set of simple rules - it also uncovers the central limitation of the current work. Having done the labeling only by means of color judgement using ImgLab web utility, the resulting models are based solely on color recognition. In literature, most developed models were informed by data collected in-situ that provides with accurate labeling. Should such data become available, it is possible also that the selected ANN architecture may need to be revisited.

Another matter of importance is that the model was tested against "unseen" data that were gathered from the same pre-processing pipeline as the one used for the training data and not from raw image input. The latter presents the complication of having available a method to discern between land and water pixels, especially when considering coastal and lake or freshwater reservoir regions. This fact calls for the formulation of a new processing pipeline able to selectively group pixels that comprise the region(s) of interest (ROI) corresponding to water volumes. There are several approaches to achieve this, as:

1. Use image processing techniques such as texturing - to identify directly land pixels - or edge-detection to locate the coastline
2. Use of a radar satellite product from an earth observation provider (e.g. Sentinel-1). The advantage of this technique relies on the apparent "blackness" of water surfaces in radar images due to its minimal roughness. It is well known that increased surface roughness gives strong radar signatures. The other advantages of radar acquired data is that it needs to be downloaded only once per monitoring region, supposing that in the near future no significant land erosion phenomena will occur. Also, radar imagery is not affected by cloudiness.
3. Use the classifier itself to select land pixels by running predictions on them and setting a threshold in the probability outcome

Evidently, selecting the correct ROI for pelagic satellite products requires no additional pre-processing as the whole image corresponds to water volume.

The cost analysis activities conducted in the present work aimed to capture the correct scale for the operating expenses that was also calculated in a per area format to reveal the underlying marginal cost for the HAB service. It is important to note that this analysis was based strongly on the results of memory monitoring for model training. Nevertheless, the model developed herein is binary (i.e. HAB/NO_HAB) whereas as it outlined in the business section of the thesis, customers strive for answers regarding the severity of the HAB phenomenon. This business requirement calls for a multiclass classification engine that should have increased complexity when compared to its binary counterpart. This fact might add to the memory requirements for the model and in turn increase the marginal cost for service delivery.

As far as the pricing strategy is concerned, simply applying a markup rate on the calculated cost figure is one method of producing a reasonable figure for service usage fee but this is informed solely by internal factors for the business and market input is completely absent. An alternative method is to query potential customers about their willingness-to-pay for the provisioned service. It is important to note that since multiple tiers in service delivery are possible, observing the reaction of end-buyers towards each offered service pack is paramount. Concludingly, the estimated cost figure allows for a very competitive service offering and offers potential for economies of scale. To understand this better, should the marginal cost figure calculated herein be trusted, the monitoring fee for forecasting HAB phenomena in an area a large as the Aegean sea, could be achieved on a yearly charge as low as 85,000 dollars.

Conclusion and Future Work

The final assessment of the work conducted here is made by correlating the aims and objectives set in the beginning of the study with its research outcomes. The HAB phenomenon was studied, and literature review findings helped to understand its root causes from an environmental science perspective. The major methods for prevention, mitigation and detection were researched as well as some algorithms applicable to remote sensing implementations.

The modelling activities conducted yielded a robust pipeline for feature extraction from raw satellite image products. The evaluation of the developed ANN model revealed a fundamental research limitation. Namely, the absence of in-situ data for data labelling purposes. The HAB service envisioned herein was explored from a systems' point of view bearing in mind the technical requirement of gathering and analysis of data from various sources such as drones, chemical in-situ sensors and direct water sampling.

The outline of the technical capabilities of the proposed service were translated into required business features that can be communicated with the interested market stakeholders. Exploration of the potential market segments that the proposed HAB service is addressing was conducted both via customer profiling techniques as well as by gathering market data from an interview with a subject expert on the matter. Subsequently, a business wireframe was drawn using the business model canvas framework. A cost analysis activity based on the findings of the HAB modelling work carried out here, yielded the opportunity for an aggressive go-to-market strategy based on a low marginal cost for service provisioning.

In conclusion, the proposed service offers its customers the opportunity to cut down drastically their HAB monitoring expenses by tapping on a low-cost solution that is based on satellite image acquisition, data-fusion and artificial intelligence. Subscribing to such a service allows customers to decrease the required frequency for regular water sampling and analysis as well as to cut-down any expenses related to deploying and servicing in-situ sensor grids. The HAB service proposed here can become operational only with a minimal infrastructure from the clients' side that is necessary for labeling and calibration purposes.

Although the present work is holistic with respect to the facets of the HAB service that have been explored - science, data modelling, system architecture and business modelling - some shortcomings and opportunities were revealed. The issues that are appropriate to be addressed in future work include:

1. Acquisition of in-situ, drone and direct sampling HAB data for coastal, fresh water as well as pelagic cases for model labelling purposes
2. Development of a multiclass classifier able to handle HAB severity predictions
3. Implementation of a processing pipeline for "unseen" data capable of discerning confidently land from sea/freshwater image areas
4. Further investigation of the design aspects that are required for data fusion of satellite, drone and in-situ data
5. Amplification/augmentation of the selected feature vectors to perform data fusion from the various datasources
6. Perform market research regarding asking prices and features offered currently by other HAB prediction platforms (research the competition to further inform pricing models)
7. Gathering of new interview data from the relevant stakeholders to detect their willingness-to-pay and subscribe to any of the service tiers described in this work

References

[1] N. Oceanic and A. Administration. (2016, Apr.) What is a harmful algal bloom? [Online]. Available: https://www.noaa.gov/what-is-harmful-algal-bloom

[2] P. Manivasagan and S.-K. Kim, “Chapter 34 - an overview of harmful algal blooms on marine organisms,” in *Handbook of Marine Microalgae*, S.-K. Kim, Ed. Boston: Academic Press, 2015, pp. 517–526.

[3] C. J. Band-Schmidt, J. J. Bustillos-Guzman, D. J. Lapez-Cortes, I. Garate-Lizarraga, E. J. Nunez-Vazquez, and F. E. Hernandez-Sandoval, “Ecological and physiological studies of gymnodinium catenatum in the mexican pacific: a review,” *Marine drugs*, vol. 8, no. 6, pp. 1935 – 1961, 2010.

[4] C. H. Mortimer, “The exchange of dissolved substances between mud and water in lakes,” *Journal of Ecology*, vol. 30, no. 1, pp. 147–201, 1942.

[5] J. Heisler, P. Glibert, J. Burkholder, D. Anderson, W. Cochlan, W. Dennison, Q. Dortch, C. Gobler, C. Heil, E. Humphries, A. Lewitus, R. Magnien, H. Marshall, K. Sellner, D. Stockwell, D. Stoecker, and M. Suddleson, “Eutrophication and harmful algal blooms: A scientific consensus,” *Harmful Algae*, vol. 8, no. 1, pp. 3–13, 2008.

[6] C. J. Gobler, “Climate change and harmful algal blooms: Insights and perspective,” *Harmful Algae*, vol. 91, p. 101731, 2020.

[7] [Online]. Available: https://hab.whoi.edu/maps/regions-world-distribution/

[8] [Online]. Available: https://www.youtube.com/watch?v=lO_2p9nHApw

[9] [Online]. Available: https://www.youtube.com/watch?v=ak84op0uQ9A&t=42s

[10] S. Isabella, C. Diana, P. Luca, D. Srdan, and L. Teresa, “Algal bloom and its economic impact,” European Commission, Joint Research Centre, Tech. Rep., 2016.

[11] P. Hoagland and S. Scatasta, *The Economic Effects of Harmful Algal Blooms*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 391–402.

[12] [Online]. Available: https://storymaps.arcgis.com/stories/

[13] LitRev/2_in. [Online]. Available: https://www.s3eurohab.eu/portal/

[14] W. K. Dodds, W. W. Bouska, J. L. Eitzmann, T. J. Pilger, K. L. Pitts, A. J. Riley, J. T. Schloesser, and D. J. Thornbrugh, "Eutrophication of U.S. freshwaters: Analysis of potential economic damages," *Environmental Science & Technology*, vol. 43, no. 1, pp. 12–19, 2009.

[15] J. T. Turner and E. Granéli, *"Top-Down" Predation Control on Marine Harmful Algae*. Springer Berlin Heidelberg, 2006.

[16] N. S. G. College, "Prevention, control and mitigation of harmful algal blooms - a research plan," Office of Oceanic and Atmospheric ResearchNational Oceanic and Atmospheric Administration Department of Commerce (United States Congress), Tech. Rep., September 2001.

[17] T. Kim, "Prevention of harmful algal blooms by control of growth parameters," *Advances in Bioscience and Biotechnology*, vol. 9, pp. 613–648, 2018.

[18] M. Pal, P. J. Yesankar, A. Dwivedi, and A. Qureshi, "Biotic control of harmful algal blooms (HABs): A brief review," *Journal of Environmental Management*, vol. 268, 2020.

[19] O. T., "Sustainable development in the seto inland sea, japan : from the viewpoint of fisheries," in *Red tides in the Seto Inland Sea*. Tokyo, Japan: Terra Scientific Publishing Company, 1997.

[20] D. M. Anderson, "Approaches to monitoring, control and management of harmful algal blooms (HABs)," *Ocean & Coastal Management*, vol. 52, no. 7, pp. 342–347.

[21] M. R. Sengco, A. Li, K. Tugend, D. Kulis, and D. M. Anderson, "Removal of red- and brown-tide cells using clayflocculation. i. laboratory culture experiments with gymnodinium breve and aureococcus anophagefferens," *Marine Ecology Progress Series*, vol. 210, 2001.

[22] U.S. National Office for Harmful Algal Blooms. [Online]. Available: https://hab.whoi.edu/-response/control-and-treatment/

[23] [Online]. Available: https://www2.whoi.edu/site/andersonlab/current-projects/florida-clay-mitigation/

[24] M.-C. Archambault, V. Bricelj, J. Grant, and D. Anderson, "Effects of suspended and sedimented clays on juvenile hard clams, mercenaria mercenaria, within the context of harmful algal bloom mitigation," *Marine Biology*, vol. 144, pp. 553–565, 01 2004.

[25] Z. Yu, X. Song, X. Cao, and Y. Liu, “Mitigation of harmful algal blooms using modified clays: Theory, mechanisms, and applications,” *Harmful Algae*, vol. 69, pp. 48–64, 2017.

[26] W. Song, T. Teshiba, K. Rein, and K. E. O’Shea, “Ultrasonically induced degradation and detoxification of microcystin-lr (cyanobacterial toxin),” *Environmental Science & Technology*, vol. 39, no. 16, pp. 6300–6305, 2005.

[27] H. Huang, G. Wu, C. Sheng, J. Wu, D. Li, and H. Wang, “Improved cyanobacteria removal from harmful algae blooms by two-cycle, low-frequency, low-density, and short-duration ultrasonic radiation,” *Water*, vol. 12, no. 9, 2020.

[28] M. Rome, R. E. Beighley, and T. Faber, “Sensor-based detection of algal blooms for public health advisories and long-term monitoring,” *Science of The Total Environment*, vol. 767, 2021.

[29] H. JW and K. BJ, “Associations between chlorophyll a and various microcystin health advisory concentrations,” *F1000Res*, vol. 5, no. 151, 2016.

[30] L. Lawton, J. Metcalf, B. Zegura, R. Junek, M. Welker, A. Torokne, and L. Blaha, *Laboratory analysis of cyanobacterial toxins and bioassays*, 2021, pp. 745–800.

[31] M. Laycock, J. Jellett, E. Belland, P. Bishop, B. ThÃ©riault, A. Russell-Tattrie, M. Quilliam, A. Cembella, and R. Richards, *Mist Alert(TM): A Rapid Assay for Paralytic Shellfish Poisoning Toxins*, 2001, pp. 254–256.

[32] K. Spilling and J. Seppälä, *Measurement of Fluorescence for Monitoring Algal Growth and Health*. Springer New York, 2020.

[33] Y. Zhang, L. Liu, Y. He, P.-f. Zhang, and Z.-H. Cai, “A fiber-based fluorometric system for in situ algal classification,” *Optics & Laser Technology*, vol. 76, 2016.

[34] [Online]. Available: https://ldi.ee/products/algal-bloom-detector/

[35] F. Lefevre, A. Chalifour, L. Yu, V. Chodavarapu, P. Juneau, and R. Izquierdo, “Algal fluorescence sensor integrated into a microfluidic chip for water pollutant detection,” *Lab on a chip*, vol. 12, 2011.

[36] K. Sellner, G. Doucette, and G. Kirkpatrick, “Harmful algal blooms: Causes, impacts and detection,” *Journal of industrial microbiology & biotechnology*, vol. 30, 2003.

[37] [Online]. Available: https://www.fluidimaging.com/

[38] K. Flynn and S. Chapra, “Remote sensing of submerged aquatic vegetation in a shallow non-turbid river using an unmanned aerial vehicle,” *Remote Sensing*, vol. 6, 2014.

[39] T.-C. Su and H.-T. Chou, “Application of multispectral sensors carried on unmanned aerial vehicle (uav) to trophic state mapping of small reservoirs: A case study of tain-pu reservoir in kinmen, taiwan,” *Remote Sensing*, vol. 7, no. 8, 2015.

[40] D. Van der Merwe and K. P. Price, “Harmful algal bloom characterization at ultra-high spatial and temporal resolution using small unmanned aircraft systems,” *Toxins*, vol. 7, no. 4, 2015.

[41] A. S. K. Ngo, M. I. Cordel, R. L. Uy, and J. Ilao, “Determining the correlation between concentration levels and the visual features of algae in water surfaces,” 2015.

[42] S. Shang, Z. Lee, G. Lin, C. Hu, L. Shi, Y. Zhang, X. Li, J. Wu, and J. Yan, “Sensing an intense phytoplankton bloom in the western taiwan strait from radiometric measurements on a uav,” *Remote Sensing of Environment*, vol. 198, 2017.

[43] F. Xu, Z. Gao, X. Jiang, J. Ning, X. Zheng, D. Song, J. Ai, and M. Chen, “Mapping of green tide using true color aerial photographs taken from a unmanned aerial vehicle,” in *Remote Sensing and Modeling of Ecosystems for Sustainability XIV*, vol. 10405, International Society for Optics and Photonics. SPIE, 2017.

[44] [Online]. Available: https://en.wikipedia.org/wiki/Nimbus_7

[45] J. L. Mueller, “Prospects for wuring phytoplankton bloom extent and patchiness using remotely sensed ocean color images: An example,” *Toxic Dinoflagellate Blooms*, pp. 303–308, 1979.

[46] C. Rodriguez Benito, C. Haag, M. Fea, and H. Gutierrez, “Monitoring marine life from space envisat experience in chile,” *Esa Bulletin-european Space Agency - ESA BULL-EUR SPACE AGENCY*, vol. 126, 2006.

[47] J. Gower, S. King, and P. Goncalves, “Global monitoring of plankton blooms using meris mci,” *International Journal of Remote Sensing*, vol. 29, no. 21, 2008.

[48] C. Binding, T. Greenberg, and R. Bukata, “The meris maximum chlorophyll index; its merits and limitations for inland water algal bloom monitoring,” *Journal of Great Lakes Research*, vol. 39, 2013.

[49] I. Ogashawara, “The use of sentinel-3 imagery to monitor cyanobacterial blooms,” *Environments*, vol. 6, no. 6, 2019.

[50] V. Klemas, “Remote sensing of algal blooms: An overview with case studies,” *Journal of Coastal Research*, vol. 28, 2012.

[51] D. Wu, R. Li, F. Zhang, and J. Liu, “A review on drone-based harmful algae blooms monitoring,” *Environmental Monitoring and Assessment*, vol. 191, no. 4, 2019.

[52] L. Shen, H. Xu, and X. Guo, “Satellite remote sensing of harmful algal blooms (HABs) and a potential synthesized framework,” *Sensors (Basel)*, vol. 12, no. 6, 2012.

[53] S. Groom and P. Holligan, “Remote sensing of coccolithophore blooms,” *Advances in Space Research*, vol. 7, no. 2, 1987.

[54] [Online]. Available: https://en.wikipedia.org/wiki/Rayleigh_scattering

[55] M. H. Khalili and M. Hasanlou, “Harmful algal blooms monitoring using sentinel-2 satellite images,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-4/W18, 2019.

[56] [Online]. Available: https://step.esa.int/main/snap-supported-plugins/sen2cor/sen2cor-v2-10/

[57] M. E. Smith and S. Bernard, “Satellite ocean color based harmful algal bloom indicators for aquaculture decision support in the southern benguela,” *Frontiers in Marine Science*, vol. 7, 2020.

[58] [Online]. Available: https://docs.sentinel-hub.com/api/latest/

[59] [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/resolutions/radiometric

[60] [Online]. Available: https://sentinelhub-py.readthedocs.io/en/latest/

[61] S. Mishra and D. R. Mishra, “Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters,” *Remote Sensing of Environment*, vol. 117, 2012.

[62] [Online]. Available: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/-ndvi/

[63] [Online]. Available: https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/-maximum_peak_height_bloom_index/

[64] [Online]. Available: https://coastalscience.noaa.gov/

[65] [Online]. Available: https://coastalscience.noaa.gov/news/2021-lake-erie-algal-bloom-was-more-severe-than-predicted-by-seasonal-forecast/

[66] [Online]. Available: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/-Sentinel-2

[67] [Online]. Available: https://imglab.in/#

[68] [Online]. Available: https://www.alliedmarketresearch.com/water-quality-monitoring-systems-market

[69] [Online]. Available: https://land.copernicus.eu/global/

[70] [Online]. Available: https://www.monocle-h2020.eu/About

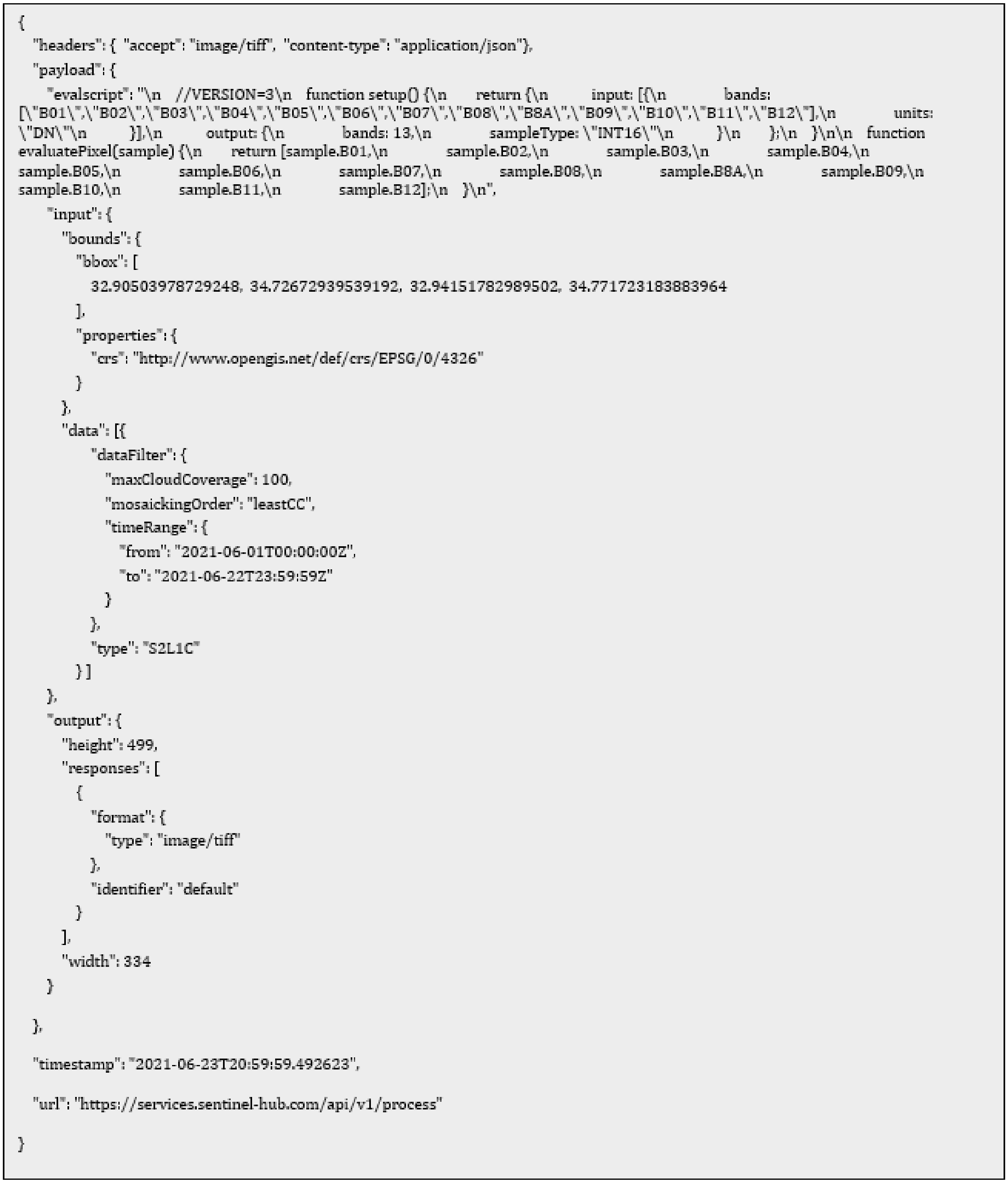
[71] A. Osterwalder, Y. Pigneur, P. Papadakos, G. Bernarda, T. Papadakos, and A. Smith, *Value proposition design*. John Wiley & Sons, Oct. 2014.

[72] A. Osterwalder and Y. Pigneur, *Business model generation*. John Wiley & Sons, Jun. 2010.

[73] [Online]. Available: https://docs.aws.amazon.com/machine-learning/latest/dg/-pricing.html

[74] [Online]. Available: https://docs.aws.amazon.com/apigateway/latest/developerguide/-simple-calc-lambda-api.html

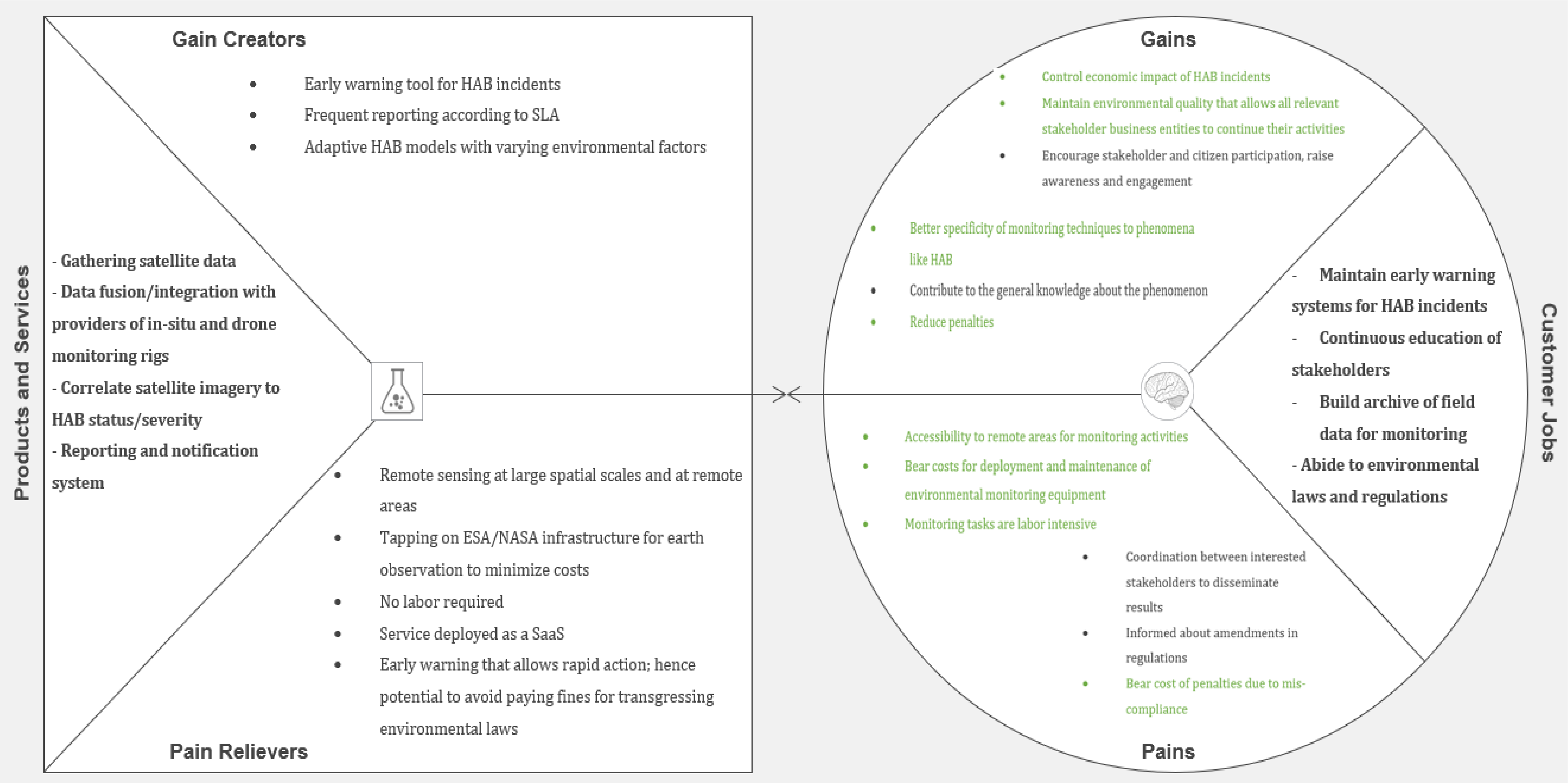
Appendix A – Sentinel hub API Example Request Payload



```
{
  "headers": {  "accept": "image/tiff",  "content-type": "application/json"},
  "payload": {
    "evalscript": "\n    //VERSION=3\n    function setup() {\n        return {\n            input: [{\n                bands:
[\"B01\",\"B02\",\"B03\",\"B04\",\"B05\",\"B06\",\"B07\",\"B08\",\"B8A\",\"B09\",\"B10\",\"B11\",\"B12\"],\n                units:
\"DN\"\n            }],\n            output: {\n                bands: 13,\n                sampleType: \"INT16\"\n            }\n        };\n    }\n\n    function
evaluatePixel(sample) {\n        return [sample.B01,\n                sample.B02,\n                sample.B03,\n                sample.B04,\n
sample.B05,\n                sample.B06,\n                sample.B07,\n                sample.B08,\n                sample.B8A,\n                sample.B09,\n
sample.B10,\n                sample.B11,\n                sample.B12];\n    }\n",
    "input": {
      "bounds": {
        "bbox": [
          32.90503978729248, 34.72672939539192, 32.94151782989502, 34.771723183883964
        ],
        "properties": {
          "crs": "http://www.opengis.net/def/crs/EPSG/0/4326"
        }
      },
      "data": [{
          "dataFilter": {
            "maxCloudCoverage": 100,
            "mosaickingOrder": "leastCC",
            "timeRange": {
              "from": "2021-06-01T00:00:00Z",
              "to": "2021-06-22T23:59:59Z"
            }
          },
          "type": "S2L1C"
        } ]
    },
    "output": {
      "height": 499,
      "responses": [
        {
          "format": {
            "type": "image/tiff"
          },
          "identifier": "default"
        }
      ],
      "width": 334
    }

  },

  "timestamp": "2021-06-23T20:59:59.492623",

  "url": "https://services.sentinel-hub.com/api/v1/process"

}
```

Appendix B: Value Proposition Canvas (VPC)

Customer pains and gains that are hypothesized to overlap with features of the proposed HAB service, are color-coded in green



Products and Services

Gain Creators

- Early warning tool for HAB incidents
- Frequent reporting according to SLA
- Adaptive HAB models with varying environmental factors

- Gathering satellite data
- Data fusion/integration with providers of in-situ and drone monitoring rigs
- Correlate satellite imagery to HAB status/severity
- Reporting and notification system

Pain Relievers

- Remote sensing at large spatial scales and at remote areas
- Tapping on ESA/NASA infrastructure for earth observation to minimize costs
- No labor required
- Service deployed as a SaaS
- Early warning that allows rapid action; hence potential to avoid paying fines for transgressing environmental laws

Gains

- Control economic impact of HAB incidents
- Maintain environmental quality that allows all relevant stakeholder business entities to continue their activities
- Encourage stakeholder and citizen participation, raise awareness and engagement

- Better specificity of monitoring techniques to phenomena like HAB
- Contribute to the general knowledge about the phenomenon
- Reduce penalties

Customer Jobs

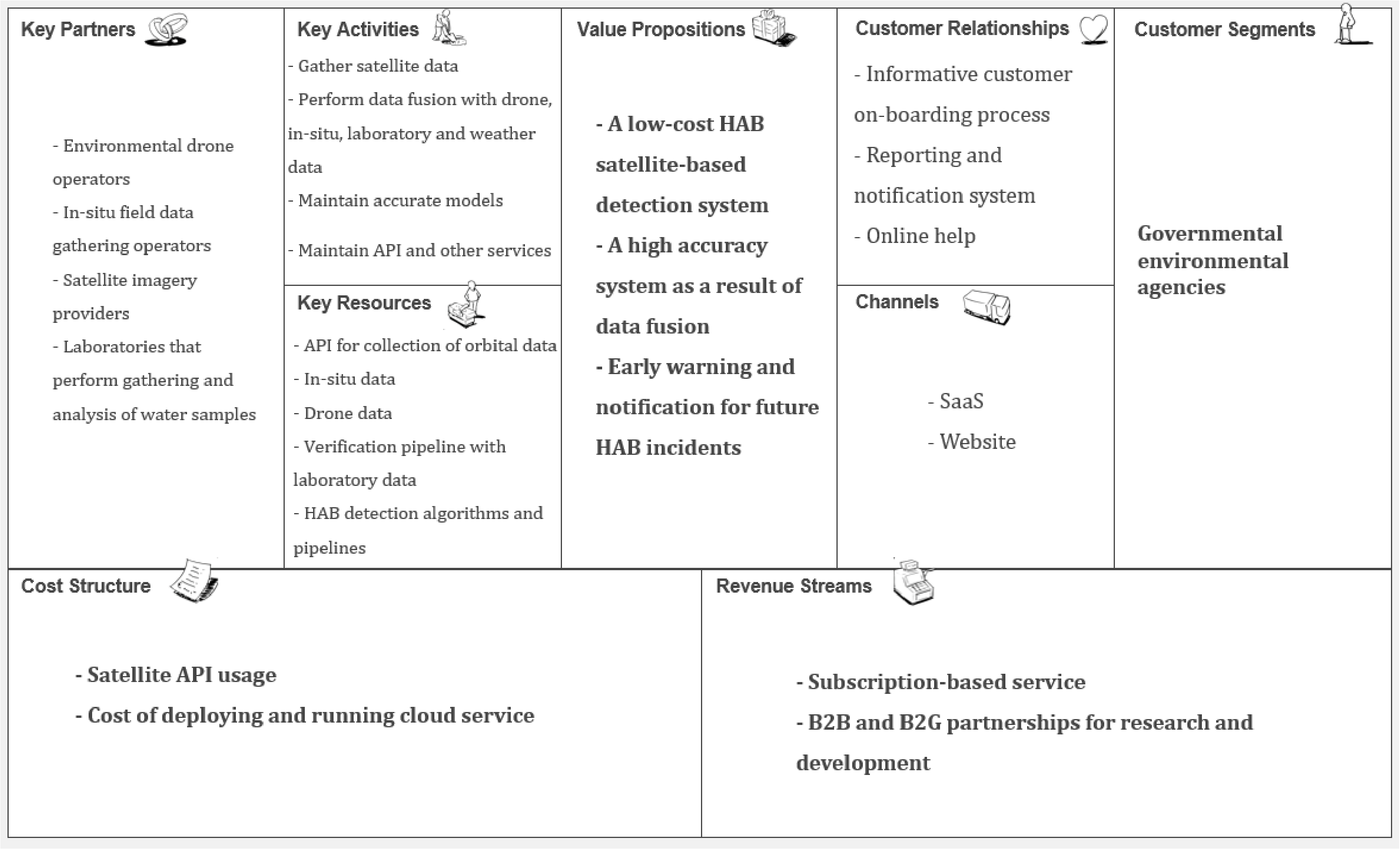
- Maintain early warning systems for HAB incidents
- Continuous education of stakeholders
- Build archive of field data for monitoring
- Abide to environmental laws and regulations

Pains

- Accessibility to remote areas for monitoring activities
- Bear costs for deployment and maintenance of environmental monitoring equipment
- Monitoring tasks are labor intensive

- Coordination between interested stakeholders to disseminate results
- Informed about amendments in regulations
- Bear cost of penalties due to mis-compliance

Appendix C: Business Model Canvas (BMC)



Key Partners

- Environmental drone operators
- In-situ field data gathering operators
- Satellite imagery providers
- Laboratories that perform gathering and analysis of water samples

Key Activities

- Gather satellite data
- Perform data fusion with drone, in-situ, laboratory and weather data
- Maintain accurate models
- Maintain API and other services

Key Resources

- API for collection of orbital data
- In-situ data
- Drone data
- Verification pipeline with laboratory data
- HAB detection algorithms and pipelines

Value Propositions

- **A low-cost HAB satellite-based detection system**
- **A high accuracy system as a result of data fusion**
- **Early warning and notification for future HAB incidents**

Customer Relationships

- Informative customer on-boarding process
- Reporting and notification system
- Online help

Channels

- SaaS
- Website

Customer Segments

Governmental environmental agencies

Cost Structure

- **Satellite API usage**
- **Cost of deploying and running cloud service**

Revenue Streams

- **Subscription-based service**
- **B2B and B2G partnerships for research and development**