Master in Computational Colour and Spectral Imaging (COSI)

Improving Ocean Chlorophyll Estimation in Satellite Hyperspectral Images Using Ensemble Machine Learning

Master Thesis

Presented by

Alvaro Flores-Romero

and defended at the

Norwegian University of Science and Technology

September 2023

Academic Supervisor(s): Dr. Sivert Bakken and Dr. Steven Yves Le Moan

Jury Committee:

1. Dr. Markku Keinänen
2. Dr. Jacob Bauer

Submission of the thesis: 9th August 2023
Day of the oral defense: 4th September 2023

Abstract

Chlorophyll can be used as a convenient way to study and monitor water quality in coastal regions. Traditional techniques to measure chlorophyll, such as fluorescence analysis, are expensive and time-consuming, as they require *in situ* long lasting campaigns. The use of high-frequency spectral images taken from space-borne imagers, such as the one in the HYPSO-1 mission, has become a more feasible alternative around the world. Although it is challenging, it is possible to retrieve surface reflectance from a hyperspectral imager after removing the atmospheric effects by aerosols and molecules.

This work addresses the problem of chlorophyll estimation with the surface reflectance of HYPSO-1 spectral images. First, a radiative transfer model of the 6SV1 algorithm was implemented for the first time on this small satellite so that the reflectance could be obtained from a spectral image. Second, pixels were classified to uniquely study those corresponding to ocean regions. Third, reflectance features that were highly correlated with measured chlorophyll concentrations were selected to improve the estimation process. Finally, fine-tuned features were combined with linear and polynomial regression, as well as ensemble machine learning techniques to identify the best performance of different approaches.

After atmospheric correction, it was possible to classify water pixels based on reflectance with a 98.3% accuracy, which is a useful method to overcome the problem of inconsistent telemetry data. Selecting the most appropriate features enhanced the performance of the model. The proposed "voting" ensemble machine learning approach performed better than traditional empirical methods, which was validated by using BOA reflectance measurements from the GLORIA dataset. Therefore, we can conclude that the concentration of ocean chlorophyll can be reliably estimated using hyperspectral images from the HYPSO-1 satellite.

Dedication

To my mom, dad and brother. They showed me that love can withstand the distance.

> Not to us, Oh Lord, not to us
> but to your name be the glory
>
> *Ps. 115:1 (Vulgate 113:9)*

Acknowledgment

I would like to express my deepest gratitude to Dr. Steven Le Moan and Dr. Sivert Bakken, as I consider myself very lucky for having their support and expertise.

I am deeply grateful to Dr. Joseph Garrett for his support in allowing me to investigate novel concepts and, most significantly, for his continued enthusiasm throughout the course of my thesis.

Lastly, I would like to thank the European Union and the coordinators and professors from the Norwegian University of Science and Technology, the University of Granada, and the University of Eastern Finland for supporting my professional growth in a rigorous scientific program.

Acronyms

6SV1 - Second Simulation of a Satellite Signal in the Solar Spectrum Vector
ANN - Artificial Neural Networks
AOD - Aerosol Optical Depth
AOT - Aerosol Optical Thickness
BOA - Bottom Of Atmosphere
CNN - Convolutional Neural Networks
CRS - Coordinate Reference System
CZCS - Coastal Zone Color Scanner
DEM - Digital Elevation Model
DL - Deep Learning
EM - Electromagnetic Radiation
ESA - European Space Agency
FOV - Field Of View
FWHM - Full Width Half-Maximun
FIFOV - Ground Instantaneous Field Of View
GCP - Ground Control Points
HAB - Harmful Algal Blooms
HSI - Hyperspectral Imaging
IFOV - Instantaneous Field Of View
IOP - Inherent Optical Properties
IR - Infrared
LR - Lower Right
MAE - Mean Absolute Error
MBR - Maximum Band Ratio
MIR - Mid Infrared
ML - Machine Learning
MSI - Multispectral Imaging
NASA - National Aeronautics and Space Administration
NDWI - Normalized Difference Water Index
NIR - Near Infrared
NN - Neural Networks
OCR - Ocean Color Radiometry
OLS - Ordinal Least Squares
OSM - Open Street Maps
RFE - Recursive Feature Elimination
RMSE - Root Mean Square Error
ROI - Region Of Interest

SFS - Sequential Feature Selector
SHAP - Shapley Additive Explanations
SPAD - Soil-Plant Analysis Development
SRF - Spectral Response Function
SWIR - Short Wave Infrared
TOA - Top Of Atmosphere
UAV - Unmanned Aerial Vehicles
UL - Upper Left
WVP - Precipitable Water Vapor

Contents

1 Introduction

The oceans cover approximately 71% of the Earth's surface (Spellman, 2019), which corresponds to 96. 5% of all global water reserves that perform important functions such as the regulation of the temperature of the planet (Barry, 2013; Gleick et al., 1993). Phytoplankton, also known as microalgae, is an important part of the marine food chain that is responsible for most of the oxygen produced by the ocean, which represents approximately half of the oxygen that humans breathe (Gualtieri and Barsanti, 2006; Barry, 2013).

Rapid growth of the algae population is known as a "bloom" and, although they are natural occurring phenomena, they can also have specific characteristics to cause negative impacts on economic activities and human health (UNESCO and Hallegraeff, Gustaaf M., 2003). When the bloom has conditions to harm animals and people, they are called "Harmful Algal Blooms (HABs)." These become dangerous when they produce toxins that can kill the consuming organisms and predators of such, having the potential to kill marine organisms and also affect human health. An early example of this impact occurred in May 1998 when a bloom severely damaged the sea farming industry on the western Norwegian coast; however, the most harmful bloom recorded in this region happened recently in June 2019 (Johannessen et al., 1989; John et al., 2022).

Phytoplankton has a photosynthetic pigment called chlorophyll that allows the absorption of sunlight that reaches the ocean through the atmosphere (Kirk, 2011). These conditions enable a biochemical process called photosynthesis in which the algae generate energy to exist while converting CO_2 and water into oxygen and carbohydrates (Hall and Rao, 1999). Due to the presence of this pigment in all algae, the chlorophyll-a concentration can be used as an index of the total biomass of algae in a specific area (Karlson et al., 2021; Pandey, 2014).

Being able to estimate the chlorophyll concentration in large ocean bodies quickly is key when constant monitoring is required for water quality. For these reasons, remote sensing (satellite, UAV, aircraft) solves accessibility, frequency, and coverage problems, making it a great alternative to monitoring blooms as they cover large areas in the water bodies, making them visible with satellite imaging sensors.

Multiple countries on every continent rely on seafood and water quality for the success of aquaculture (Trottet et al., 2022). In 2020, global production of this sector reached 122.6 million tonnes of live weight with an approximate value of 281.5 billion dollars (see Figure 1.1). According to United Nations (2022), the population is expected to grow to 8.5 billion in the next 7 years and, at the same time, UNDESA (2012) estimates that by 2050, the number of lakes with HAB will increase by 20%. These predictions show the importance of monitoring water bodies, as aquaculture is the fastest growing food production sector and fish mortalities can have a significant impact on the industry (Karlson et al., 2021; Tacon, 2020).

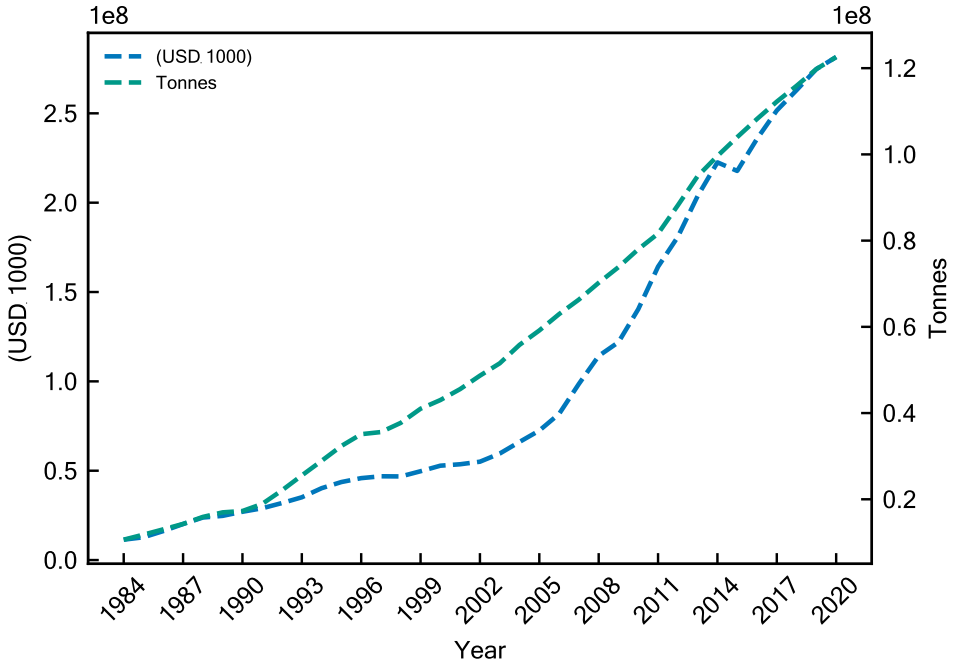


Figure 1.1: *Global aquaculture production (Value and Quantity) showing an increasing monetary importance in the last 40 years. (FAO, 2022)*

In recent years, there has been an increase in the number of publications that coincides with the growing trend of aquaculture revenue. This is observed primarily by the relative percentage of available publications related to the keywords "ocean color", "remote sensing", and "chlorophyll estimation". Figure 1.2 shows the number of publications only in the Google Scholar database classified with the keywords already mentioned, as well as the total number of yearly publications. Although the underlying assumption is that old material is either not included or not correctly labeled, there has been a clear increase over the past 15 years.

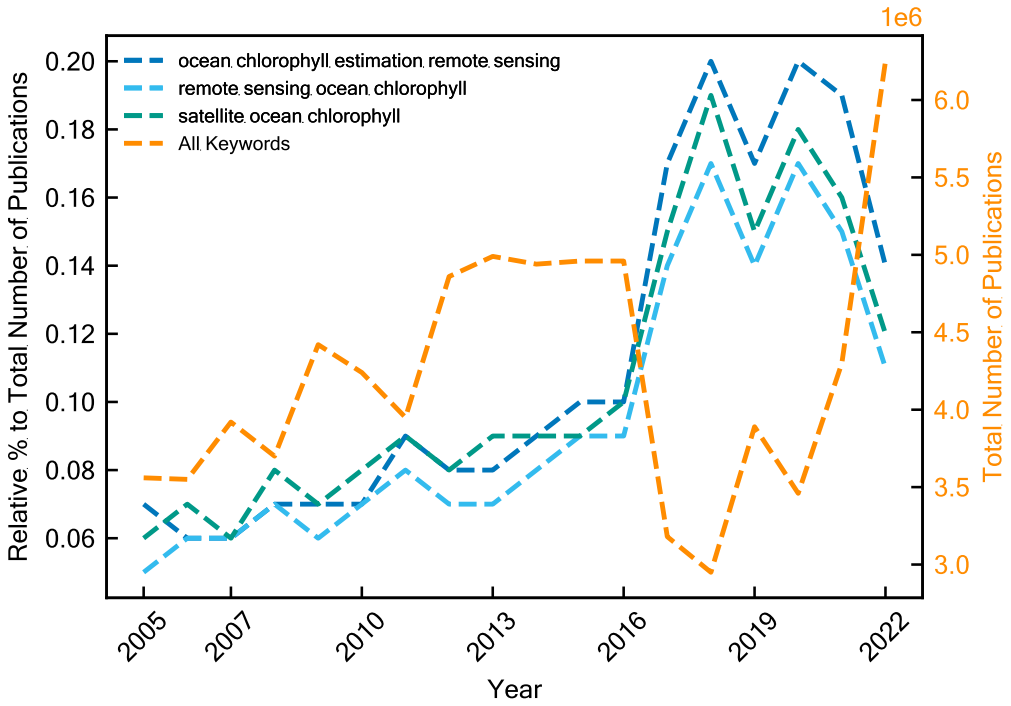


Figure 1.2: *Percentage of relative yearly publications tagged in the Google Scholar database with the indicated keyword for each curve.*

For the reasons described above, it is imperative to focus on improving and generalizing the methods that help obtain water quality ranking indicators. Remote sensing allows a convenient, non-invasive way to achieve this by retrieving chlorophyll concentration as an index of biomass as one of the study parameters. As a consequence, it is expected that the refinement through further study of this technique will provide enough information for multiple economic areas that rely on water bodies.

1.1 Motivation and Research Gap

HYPSO-1 is a SmallSat hyperspectral remote sensing mission operated by the Norwegian University of Science and Technology (NTNU) in which the current operational pipeline does not envisage a chlorophyll estimation stage (Grøtte et al., 2022). Due to the lack of a robust atmospheric correction process, recovery of BOA reflectance is not yet possible. As most of the reliable existing methods for chlorophyll inversion make use of reflectance, it is imperative to use a method that can recover the upward spectra from the surface of the Earth.

Chlorophyll estimation can be approached in different ways, but it requires

diverse and representative matches of reflectance (as seen by HYPSO-1) and chlorophyll ground truth. The quality of the estimation is dependent on the available information and the model used for the radiative transfer equation approximation.

This thesis will propose a solution to the following limitations:

1. Estimation of biomass concentration using surface reflectance R_{rs} obtained from the radiance observed by the HYPSO-1 satellite. Previous strategies have been designed for individual sensors using the information of custom spectral bands, but a solution for HYPSO-1 has not yet been implemented.

2. The atmospheric correction for HYPSO-1 has not been developed. Currently, only the radiometric correction by Henriksen et al. (2022) is used so that an L1B radiance spectral image can be obtained for all images.

3. At the moment of writing this work, there are no surface reflectance R_{rs} and ground truth chlorophyll matches that allow the study of estimation methods such as polynomial regression, ensemble machine learning, or artificial neural networks (ANN).

1.2 Scope

There are multiple parameters of water quality such as chlorophyll-a, harmful algae, turbidity, pollution sediment, submerged habitat, and temperature that can be derived from multispectral and hyperspectral satellite sensors (Lubac et al., 2008; Pahlevan et al., 2021; O'Shea et al., 2021). This work focuses on estimating chlorophyll, as it is a relevant parameter for the aquaculture industry (Muller-Karger, 1992, as cited in IOCCG, 2018).

The estimation of chlorophyll will be approached by using surface reflectance R_{rs} obtained from spectral images of HYPSO-1 *small-sat*; built and operated by the Norwegian University of Science and Technology (NTNU) (Prentice et al., 2021). The required reflectance R_{rs} can be obtained from the measured radiance detected by the hyperspectral camera after correcting for atmospheric influence using the 6SV1 model. To validate the generalization of the proposed method, *in situ* chlorophyll and reflectance measurements taken at sea level will be used; however, the main focus of this work continues to be the estimation of chlorophyll from hyperspectral images captured by the NTNU small satellite currently orbiting Earth.

To achieve a similar precision as that found in the ESA and NASA missions, the chlorophyll ground truth will be taken from the MODIS Aqua and Sentinel-3 OLCI satellites. For this study, the HYPSO-1 pixels will be matched by coordinates and time. The conditions for this process are described later in this work.

1.3 Contribution

In this work, both atmospheric correction and chlorophyll estimation are addressed for the HYPSO-1 satellite, introducing an atmospheric correction procedure, a satellite-matched dataset of hyperspectral measurements, and a prediction model. Chlorophyll values from NASA and ESA satellites were matched in the +/- 3-hour range to validate the estimation techniques proposed. The usage of multiple traditional ensemble machine learning methods allows to reduce external noise. The latter can compensate for the different characteristics of the satellites used, which, when paired, can generate in-process noise.

In summary the contributions are:

1. Chlorophyll inversion from BOA reflectance was implemented on HYPSO-1 hyperspectral images through the use of ensemble machine learning methods which result in a better prediction versus traditional polynomial approaches.

2. Implemented 6SV1 atmospheric correction for HYPSO-1 allowing to obtain surface reflectance from TOA radiance

3. Created the first HYPSO-1 dataset of more than 1×10^6 points with matching ESA chlorophyll values by coordinates, enabling for further study of chlorophyll estimation through the signal measured by the hyperspectral sensor onboard the NTNU satellite.

1.4 Structure of the Thesis

In Chapter 2 the technical history and background are given, providing an overview of the theoretic aspects of hyperspectral images and their use in chlorophyll inversion. Chapter 3 discusses the methodology chosen as the core of this research study, as well as the considerations regarding chlorophyll estimation, atmospheric correction, water detection, and model evaluation criteria. The results of all the chlorophyll estimation methods selected for this thesis are presented in Chapter 4, with the discussion of the results being presented in Chapter 5. Finally, the thesis is concluded in Chapter 6.

"Writefull" AI Grammatical Proof Reading Tools were used in the creation of this work to ensure the clarity of ideas.

2 Background

In the following sections, the technical background and the relevant historical context will be presented and discussed along with the best current approach for chlorophyll estimation from remote sensing hyperspectral images.

2.1 Technical Relevant History

In the early 1960s, the *TIROS* program was used to assess the potential for information extraction from orbiting sensors. Meteorological satellites (whose only purpose at the time was to track hurricanes) were able to take low-definition thermal IR images and identify differences in ocean temperature (Wilson et al., 2001). With these early approaches, it was clear that there was the possibility of extracting information from aerial devices; further attempts were made, such as in the work (Clarke et al., 1970, as cited in Gordon, 2010), where different chlorophyll concentrations were found to have different upward spectra. Although the experiment was done with a spectroradiometer mounted on a plane, the essence was later translated to satellite imagery missions.

Skylab was NASA's first space station, designed to be used by a team of astronauts to take different measurements that would allow studying various regions, including the ocean (Wilson et al., 2001). The availability of instruments such as multispectral, color, and grayscale cameras (S190A/B), multipsectral scanners (S192), IR spectrometer (S191), radiowave scatterometer (S193), and L/band radiometer (S194) was of great importance for collecting information to help us understand how the earth is perceived quantitatively from space (Eason and NASA, 1978). Figure 2.1 shows the spectral range in which the instruments of *Skylab* operated. Even though some damages occurred to some instruments during operation and that the duration of the mission was a little less than a year (from May 1973 to February 1974), the learnings set the foundations for the next generations of satellite sensing and imagery.

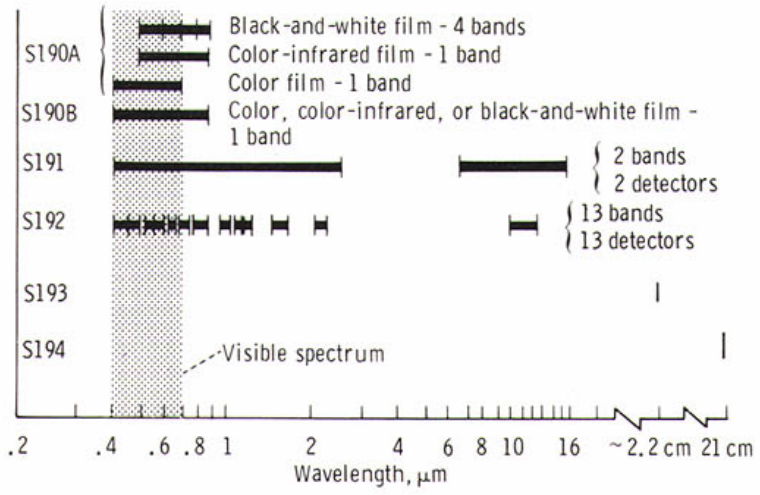


Figure 2.1: *Spectral range of instruments used in the Skylab mission. Image from (Eason and NASA, 1978).*

Nimbus-7, SeaSat, and Tiros-N are a triad of satellites focused on ocean monitoring (Wilson et al., 2001). OCR (Ocean Color Radiometry) is considered to start with NASA's Coastal Zone Color Scanner (CZCS), a sensor mounted on the Nimbus-7 as the parameters and design considerations were optimized for water sensing (NASA, nd). After the launch of this mission in 1978, it was possible for the first time to create maps of phytoplankton biomass (chlorophyll) using a satellite imaging sensor (National Research Council, 2011). This sensor can be considered as the beginning of passive remote sensing built on all previous experiments and missions to understand the Earth with space-borne and air-borne sensors.

The US led most of the early remote sensing attempts, but by the late 1980s and early 1990s other countries began satellite projects destined for earth monitoring on their own and in collaboration with other agencies. Some of the missions include the ERS-1 and ENVISAT (early 2000's EU satellite). (Wilson et al., 2001)

All of the previous experiments, studies, and satellite missions have allowed researchers to develop remote sensing to the current practices for sensor protection, measurement bias over time, and calibration and correction techniques which have led to more robust data processing. The increase in availability of hyperspectral measurements from space agencies (e.g. ESA, NASA) will continue to help researchers improve remote sensing observation.

Although the focus of this work has been on estimating chlorophyll using surface reflectance, other important information can also be retrieved from optical remote sensing, such as temperature profiles, altimetry, and meteorological profiles, which are useful for other disciplines. Regardless of the property to be estimated or studied, the complex interaction of the solar radiation that goes through the atmosphere of

the Earth and how it interacts with the surface needs to be understood.

2.2 Solar Radiation

When the electromagnetic radiation (EM) from the Sun reaches our atmosphere, a complex interaction begins between the emitted waves and the ionospheric constituents that are freely floating around the Earth (Elachi and Van Zyl, 2006). During this interaction, particles and aerosols in the atmosphere scatter the incoming sunlight in all directions, and some specific molecules absorb very efficiently at specific wavelengths (Randall B. Smith, 2012). The resulting radiation is transmitted through the atmosphere and reaches the surface where the scattering and absorption process may continue.

Areas with very low absorption and high transmission are usually referred to as "windows" and are used when designing measurement instruments, as sunlight can only reach the surface of the Earth through these areas throughout the spectrum (Wilson et al., 2001). According to Elachi and Van Zyl (2006), the molecules and particles that cause the different absorption regions are the ones in Table 2.1.

Table 2.1: *Absorption regions in the atmosphere for different wavelengths.*

Wavelength	Name	Property
$\lambda \geq 29m$	Radio Wave	All signals blocked by the ionosphere
$2.98cm \leq \lambda < 29m$		Transparent
$1.0mm \leq \lambda < 2.98cm$	Microwaves	Strong absorption from water vapour and oxygen
$15.0\mu m \leq \lambda < 1.0mm$	Far-IR	High absorption due to atmospheric constituents
$0.8\mu m \leq \lambda < 1.1\mu m$	NIR	High Absorption due to molecular vibration of water vapour and carbon dioxide
$0.4\mu m \leq \lambda < 0.8\mu m$	Visible	
$0.1\mu m \leq \lambda < 0.4\mu m$	UV	High absorption due to Ozone

This is visible in Figure 2.2, where the intensity per wavelength of solar radiation is shown at the top of the atmosphere (TOA) and at the bottom of the atmosphere (BOA). The simulated behavior generated by radiative transfer models perfectly shows the overall decrease in energy reaching the surface and specific regions that are filtered out by the atmosphere.

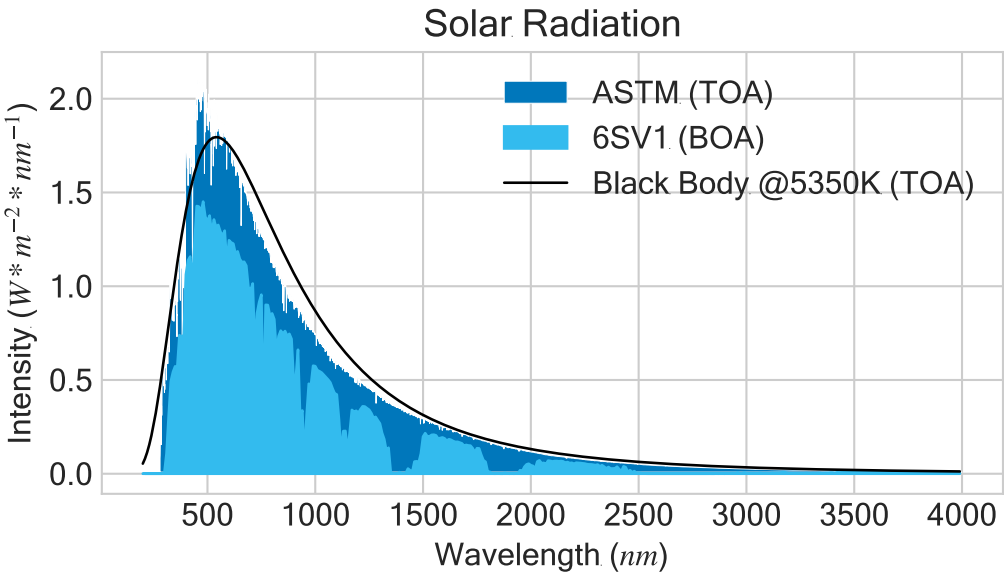


Figure 2.2: *Solar radiation at TOA comparing the sunlight without atmospheric absorption (ASTM) and the ideal black body radiation model from "Planck's Equation". The sunlight at BOA is given by the 6SV1 radiative transfer model showing the atmospheric absorption bands. The ASTM G173-03(2020) standard is derived from the SMARTS: Simple Model of the Atmospheric Radiative Transfer of Sunshine (ASTM, 2020).*

The black body intensity $I(\lambda, T)$ in terms of wavelength λ in meters can be expressed using Planck's equation 2.1. With the Boltzmann constant $k = 1.381 \times 10^{-23}\,\mathrm{J\,K^{-1}}$, the speed of light in vacuum $c = 2.998 \times 10^{8}\,\mathrm{m\,s^{-1}}$ and Planck's constant $h = 6.626 \times 10^{-34}\,\mathrm{J\,s}$, it is possible to obtain the distribution of a diffuse black body emitter at a given temperature T in Kelvin degrees, also known as thermal radiation (Incropera, 2007). Such an equation was used to display the "Black Body" energy signature at 5350K from the previous figure.

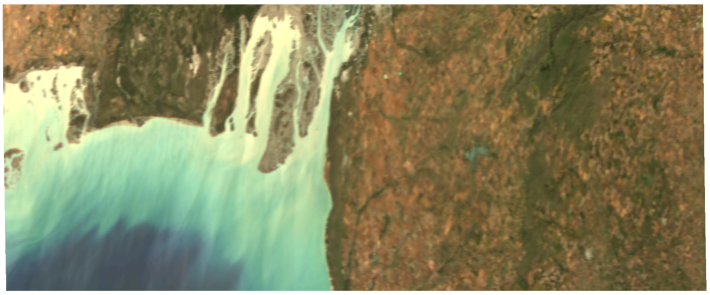
$$I(\lambda, T) = \frac{2hc^2}{\lambda^5 \left[exp(\frac{hc}{\lambda KT}) - 1\right]} \tag{2.1}$$

From the solar radiation that reaches the ocean (the one that is not scattered or absorbed by atmospheric particles), only $\approx 50\%$ can be used for photosynthesis as it falls in the photosynthetically active region between 400 and 700 nm, while the rest is weak energy that is not useful, as in the IR region (Hall and Rao, 1999). Similarly to the atmosphere, phytoplankton in the ocean (or any water body) will both absorb and scatter the incoming radiance from the Sun.

2.3 Ocean Color Radiometry

Ocean color radiometry (OCR) uses the intensity and measured upward flux through sensors mounted on UAVs, aircraft, or satellites to obtain information about the constituents of waterbodies and the corresponding Inherited Optical Properties (IOPs) (Mobley, 2021). The color of the ocean changes based on the materials or sediments in the water column, which can be defined as all the multiple depths of water at the geographical coordinate that is being monitored. In the case of a "standard" water body, photons in the red and green wavelengths are absorbed by water molecules, which results in the "blue" ocean color by the scattered photons from the blue wavelength region (Braun and Smirnov, 1993).

Organic matter and minerals can be transported from the land to the seas or lakes due to natural events and industrial activities. This results in changes in water clarity that are also measurable through reflectance, as variations in water constituents can be detected from a satellite (IOCCG, 2018). Figure 2.3 shows the difference of the same region with a 10-week difference between the captures. From the false images reconstructed from the HSI taken by HYPSO-1, the sediment distribution can be visualized and how it changes the color of the water.



(a) *2022-Dec-02 13:57*

(b) *2023-Feb-13 13:43*

Figure 2.3: *False color images from the "Bahia Blanca" region in Argentina (Lat: -38.8 Lon: -61.89) taken with HYPSO-1.*

According to Mobley (2021), the optical properties allow us to describe the water through which light passes. Due to the constant change in the waterbodies, a similar change can be observed in the vertical upward flux across the mean water surface covered by every pixel of an instrument sensor (IOCCG, 2008).

The improvement of ocean-color measuring instruments has allowed us to improve the study of IOPs and the relation with absorption and scattering (Dierssen and Randolph, 2012). It is now also possible to use passive remote sensing to monitor more ocean constituents other than chlorophyll, such as mineral particles, total suspended sediment, and color-dissolved organic matter (CDOM), type of water, and bottom depth (Dierssen and Randolph, 2012; Mobley, 2021). The influence of these water properties on the color of the water must still be studied, as well as the impact on the measured reflectance (IOCCG, 2018).

2.4 Chlorophyll Spectrum

One of the most important properties that can be monitored with OCR is the chlorophyll concentration, as this is an index of the biomass of phytoplankton (IOCCG, 2008). Any change in the biomass of the water column would modify the surface reflected upward flux that reaches the satellite, allowing us to correlate these two changes in a specific area. An increase in the total chlorophyll-a (chl-a) concentration can be used as an index of the total biomass of algae in a specific area (Karlson et al., 2021; Pandey, 2014; IOCCG, 2021).

Different materials with different properties on the surface can be identified due to the *spectral signature*, as the reflectance coming from a particular region (based on what is absorbed and what is not) will provide a wavelength-dependent characteristic curve (Randall B. Smith, 2012). In photosynthetic organisms that have chlorophyll-a and chlorophyll-b, the spectral signature is very characteristic with two peaks in the visible spectrum, as shown in Figure 2.4. These photosynthetic pigments have a considerable influence on the upward flux from the water surface because some of the incoming light is absorbed by biochemical processes (IOCCG, 2008). This process is not exclusively for aquatic organisms as it is also found in plants and trees on the ground, making the study of coastal regions particularly hard for low spatial resolution devices, as the influence of land and water areas may be combined in a single pixel.

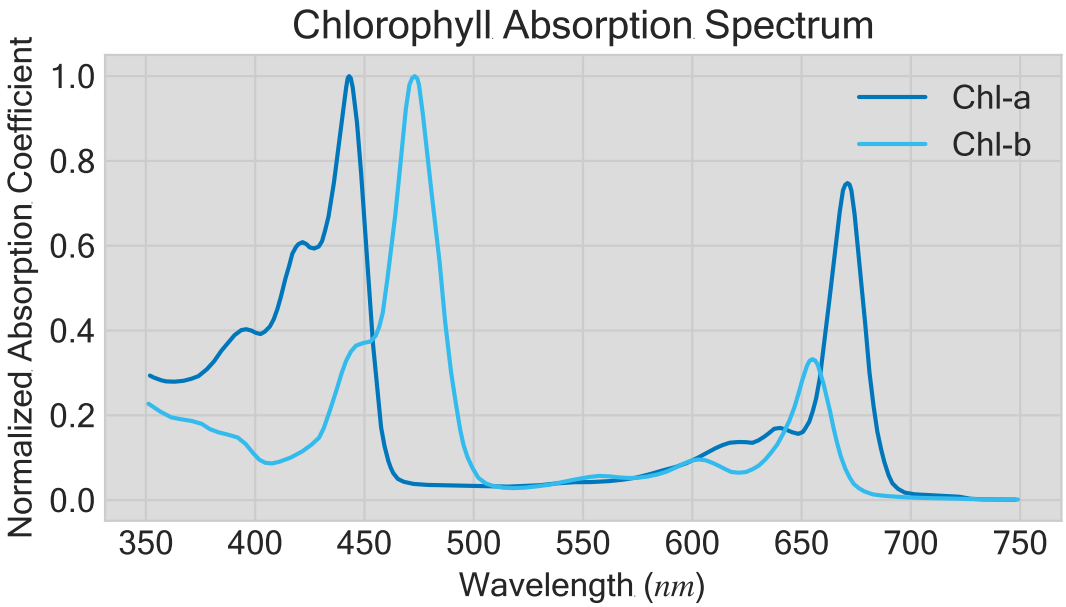


Figure 2.4: *Chlorophyll absorption spectrum for both Chlorophyll-a and Chlorophyll-b. Data from (Niedzwiedzki and Blankenship, 2010, as cited in Taniguchi and Lindsey, 2021).*

Knowing the characteristics of the type of surface or the material we want to study, we can monitor changes in the absorption and reflection of sunlight through remote sensing. In the case of Chl-*a*, the variation can be correlated with a higher phytoplankton count that will absorb more energy from the "blue" region and reflect photons in the "green". (O'Reilly et al., 1998, as cited in Dierssen and Randolph, 2012) This scattering effect creates a green color on the surface of the ocean.

2.5 Hyperspectral Imaging

According to Wilson et al. (2001) there are 4 ways in which ocean observations can be made in OCR but, for convenience, they have been grouped into 3 groups in this work: 1) radiometry (visible, IR, and microwave), 2) altimetry and 3) scatterometry. Passive sensors such as cameras can only provide radiometry data, since the others require an emitted signal, whose properties are known. Image acquisition from satellites captures the upwelling energy after it has interacted with the atmosphere and the surface of the earth collecting the photons reaching the sensor.

Traditional color RGB images are obtained with a single shot, and through interpolation, the 3 different bands (from the 3 different filters) are extracted in a process called "demosaicing" (see Figure 2.5). In modern cameras, the pixel density

of the sensors is very high, making it possible to diminish the potential artifacts caused by the *demosaicing* process. State-of-the-art methods have been designed to improve the band extraction, but they will not be covered in this work as they are not relevant for the objective of the thesis.

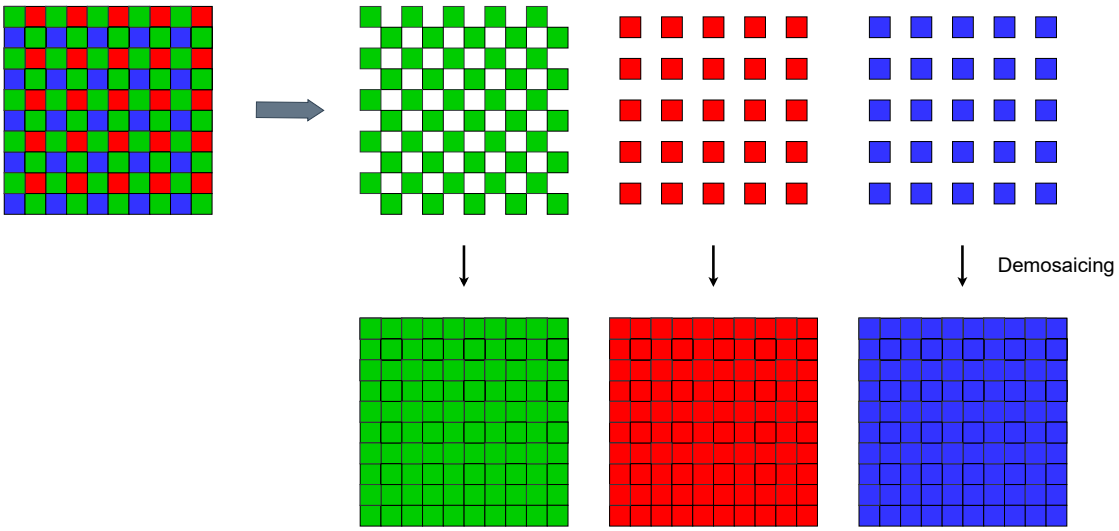


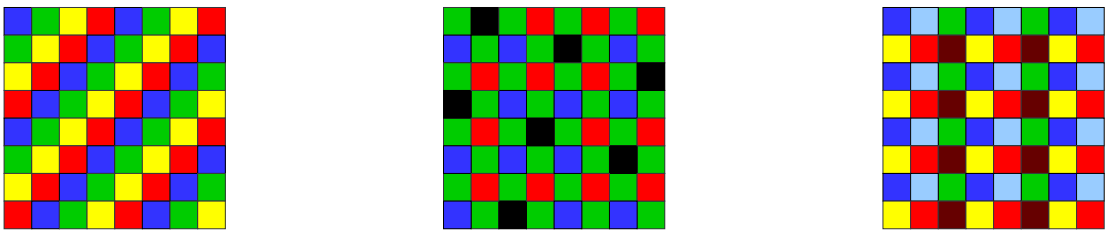
Figure 2.5: *"Demosaicing" process to go from a standard consumer camera to 3 bands through an interpolation process.*

When the photons from a scene are separated on the basis of their wavelength and captured (based on the property of the filter or grating used), different images of the same area can be obtained, which contain different information. Each of these images is called a band, and usually a higher number can reveal additional details about the captured scene. Most of the time, the term multispectral image (MSI) is attributed to images with ≈ 10 bands, while the term hyperspectral is used if there are hundreds of bands. This superlative convention is described by Polder and Gowen (2020) as inappropriate and thus suggests the use of "spectral imaging" for all cases. In any case, there is a fuzzy subjective distinction as to when an image is multispectral and hyperspectral, which is mostly based on the author's perception and field of expertise. Figure 2.6 shows an example composite of standard color RGB images versus a multispectral simulated composite to clarify the concept of multiple channels.

(a) *RGB* **(b)** *MSI*

Figure 2.6: *RGB vs MSI composites. Original image from (Flores-Romero, 2021) based on the "Indian Pines" dataset. (Baumgardner et al., 2015)*

Several attempts have been made to capture MSI with a single shot, an approach that is often referred to as "snapshot". This method proves rather difficult, as the more bands are added, the more artifacts are introduced during the demosaicing process due to the larger distance between pixels of the same wavelength. This can be seen in Figure 2.7, where each color represents a filter of the same spectral characteristics. Data inference in the "demosaicing" process is higher in the MSI snapshot filter array versus a classical Bayer filter pattern.



(a) *(Hemant Kumar Aggarwal and Majumdar, 2013)* **(b)** *(Kiku et al., 2014)* **(c)** *(Brauers and Aach, 2006)*

Figure 2.7: *MSI Snapshot Patterns using multispectral filter arrays (Lapray et al., 2014).*

To overcome the limitations of the filter array approach such as the reduction of spatial resolution leading to artifacts, the usage of a more complex setup using an optical element is preferred for remote sensing. Figure 2.8 shows the hyperspectral imager diagram used in the HYPSO-1 satellite. This setup allows us to decompose the incident light on the sensor that is reflected from the surface into different wavelengths (similar to Newton's prism experiment). This process is equivalent in purpose to having multiple filters on top of a camera sensor.

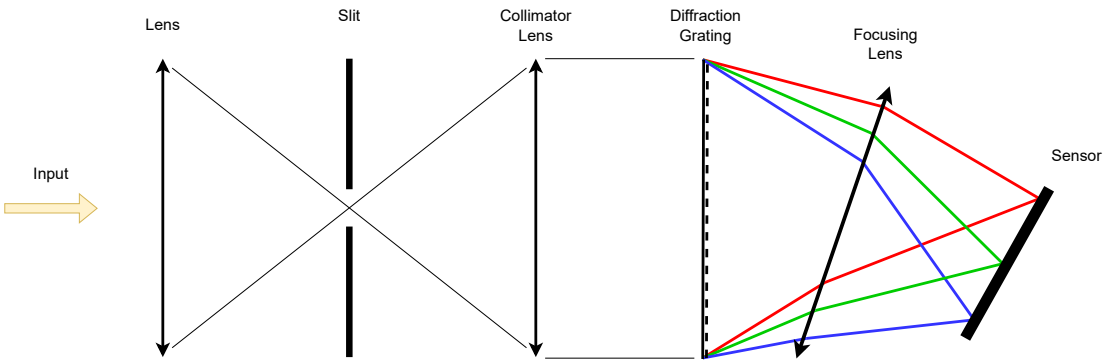


Figure 2.8: *Diagram of the HYPSO-1 optical pushbroom hyperspectral optical system based on (Prentice et al., 2021).*

The wavelength *"separation"* of such a system occurs in the diffraction grating element and is then focused on different parts of the sensor with a focusing lens. Photons from the physical EM signal can then be converted to digital values (counts) through the photodetectors on the sensor. The digital resolution in the NTNU satellite is 12 bits, which allows us to detect smaller signal variations in contrast to the traditional 8-bit depth of an RGB camera. (Henriksen et al., 2022)

The arrangement of optical elements such as the collimating lenses and gratings can be implemented in different ways so that the imager design can change depending on the requirements. There are two popular scanning styles, called *whiskbroom* and *pushbroom* (named after the way data is captured). The *snapshot* technique has been attempted using beam splitters, but there are still no viable techniques that do not sacrifice image quality.

HYPSO-1 uses the *"pushbroom"* capture method that captures the *swath* in the satellite FOV ω. The swath is a function of the design considerations of the optical system and the altitude at which the satellite orbits the Earth. The use of gratings requires an extra dimension of capture of what is being observed. This means that for a spatial point, a 1D array is needed; in the same way, measuring a 1D spatial area (observable swath) requires a 2D sensor. Each spatial pixel will be translated into a sensor row. Due to this limitation, *pushbroom* satellites must capture continuously while moving along its track. Figure 2.9 shows the hypercube as a function of wavelength based on the displacement of the satellite along the capture track.

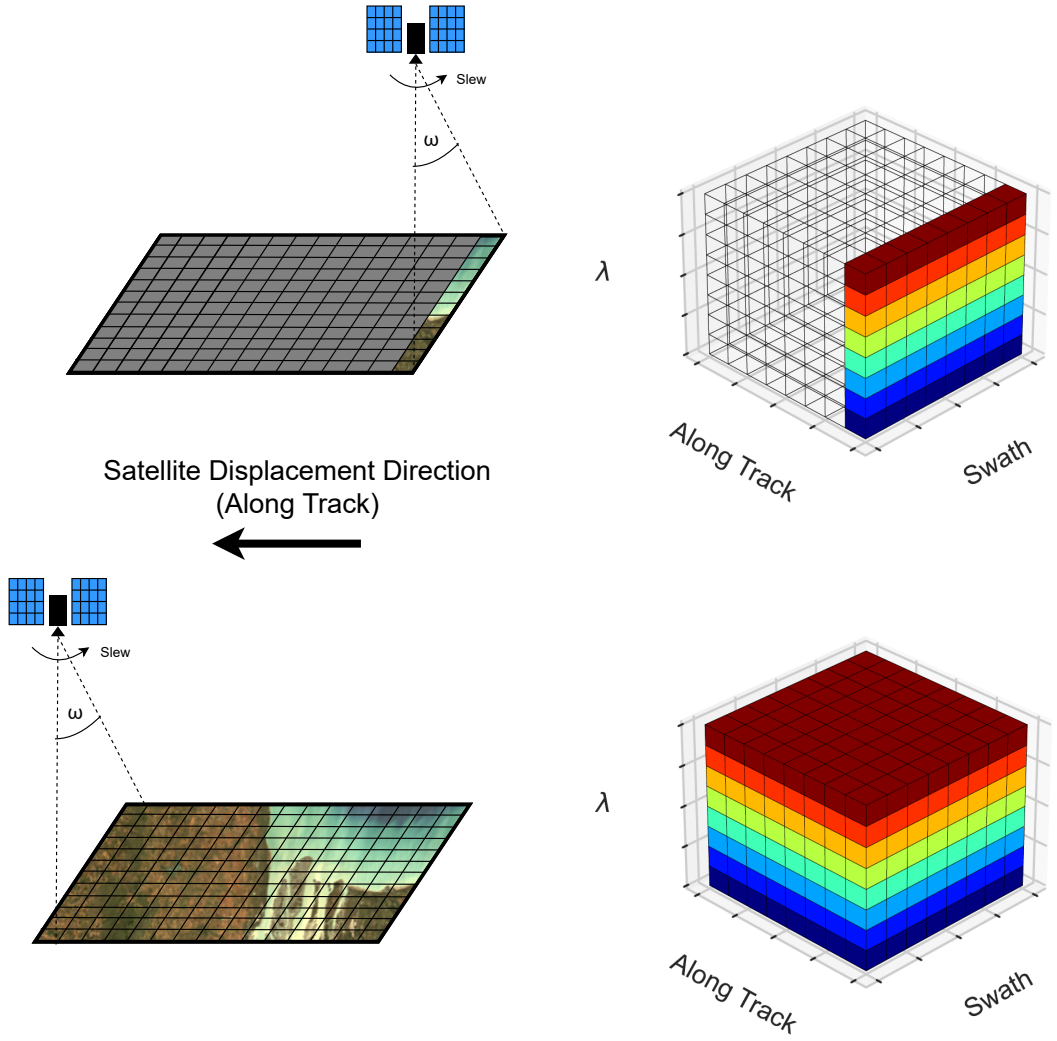


Figure 2.9: *Pushbroom*

The main benefit of extracting information from different wavelengths is that, contrary to the traditional RGB image where the wavelengths are chosen to mimic the color perception of a "standard" human being, spectral resolution is higher, allowing to extract more information of the same scene (see Figure 2.6). The selection of the spectral bands and ranges has to be defined in the design stage on the basis of the objectives, as different information can be obtained or derived depending on the region in the spectrum. Table 2.2 shows the potential use for each region based on previous satellites. From the perspective of using an HSI sensor for ocean studies (also applicable to multispectral devices), the important information to estimate chlorophyll is in the red-edge band (Lazzeri et al., 2021, as cited in Ruszczak et al., 2022).

Table 2.2: *Potential uses for specific regions of the spectrum. Wilson et al. (2001)*

Spectrum Region	Applications
Visible	Ocean Color
	Chlorophyll
IR	Surface Temperature
	Ice Detection
Microwave	Surface Temperature
	Wind Speed

2.6 Atmospheric Correction

An inherent problem of taking spectral measurements with satellites is that the EM radiation coming from the Sun goes through a complex interaction of solar irradiance, the optical path through the atmosphere (molecules and aerosols), and the ocean properties itself (IOCCG, 2021). A simplified version of these interactions is shown in Figure 2.10 for clarity. The ideal scenario would be to have no atmosphere so that the variation of the signal would only be caused by the properties of the surface from which it is reflected. In reality, not only is the atmosphere complex, but also is constantly changing.

Scattering, absorption, and emission occur at different stages, which can seriously affect the measured signal both at the surface and at the top of the atmosphere. Reliable quantitative analysis requires one to remove the effects caused by the atmosphere through a process called *"atmospheric correction"* (IOCCG, 2010).

In contrast, UAVs do not require a full atmospheric correction as the atmosphere has less influence on the measurements ("less atmosphere" between the surface and the sensor). The water vapor column and the aerosol scattering could be taken into account for a more robust correction using a physical model, but the use of calibration reference panels on the ground is a more common and simpler approach (Schläpfer et al., 2020; Guo et al., 2019).

As mentioned earlier in this work, Clarke et al. (1970) pioneered the early experiments on atmospheric measurement and understanding to acknowledge the high perturbation of EM signals entering and leaving Earth. One of the early atmospheric corrections on a satellite can be traced back to the CZCS sensor, which

used a simplified "single-scaterring" atmospheric correction. To have enough data to study the atmospheric effects on a measured signal at BOA and TOA, spectral images of Nimbus-7 and ground measurements had to be taken at the same time (Gordon, 2010).

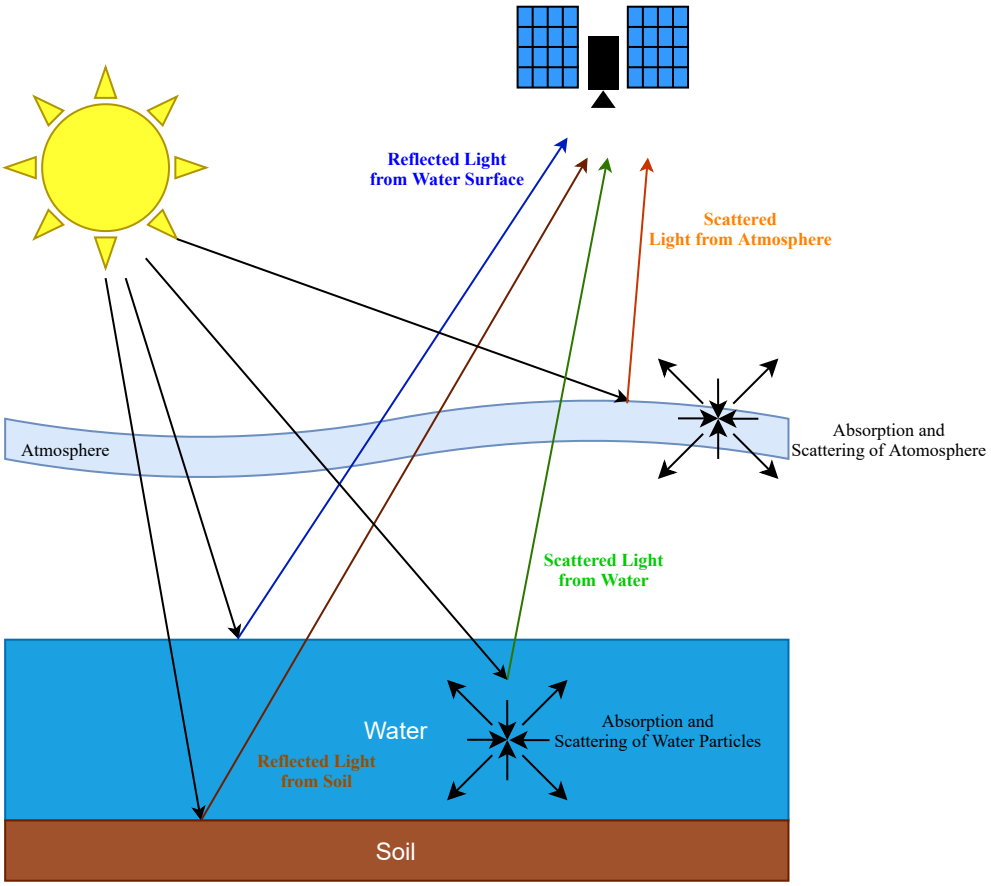


Figure 2.10: *Atmospheric scattering. Image from Flores-Romero (2021)*

The molecules corresponding to CO_2, H_2O, and O_3 dictate mainly the radiative characteristics of the atmosphere, as they are responsible for blocking sunlight from passing at specific wavelengths (Rao and Mahulikar, 2012). When considering the range from 400 to 800nm (see Figure 2.2), there is a loss of about 20% of the total energy before reaching the satellite (Gordon, 2019).

It is important to know the position of the Sun and the coordinates on the observed ground points since they describe the geometry of the incoming radiation. When utilizing a radiative transfer model, these parameters are used to describe the changing properties of the optical path, which is the area in which the sunlight goes through.

The optical path is generally described with two parameters: the water vapor column and the air mass. The former can be described as the integrated mass of gaseous water in the atmospheric column over a $1m^2$ area (Preusker, R. and El Kassar, R., 2022). The air mass can be approximated with equation 2.2, which defines the optical path length through the atmosphere with the angle between the zenith (vertical) and the solar beam z (Riordan and Hulstron, 1990). The higher the angle, the larger the AM index will be.

$$AM \approx \frac{1}{\cos z} \tag{2.2}$$

The optical path can change depending on these two parameters, resulting in a reduction in solar intensity. According to Riordan and Hulstron (1990), the longer the path due to a higher AM, the larger the scattering and solar absorption, causing an attenuated signal. The solar irradiance at BOA after it has been affected by the atmosphere is shown in Figure 2.11. Two different sources were used as comparisons, the standard ASTM G173-03 and the result of the 6SV1 radiative transfer model. It can be observed that most of the energy falls in the visible and NIR regions of the spectrum ($\approx 400 - 800nm$) where photosynthetic organisms can use it. Compared to TOA radiation in Figure 2.2, we can see attenuation of the signal in specific regions due to atmospheric effects.

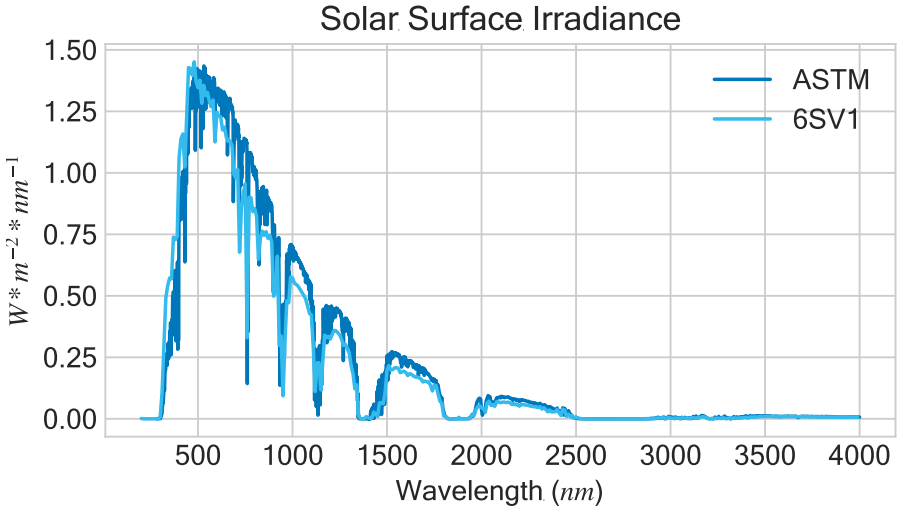


Figure 2.11: *Solar irradiance on the surface of the earth from ASTM G173-03(2020) direct and circumsolar measured standard and the output of the 6SV1 radiative transfer model used on this work (ASTM, 2020).*

The atmospheric correction problem can be simply expressed by the total radiance reaching the satellite (TOA) by the contribution of atmospheric scattering,

L_{atm}; the radiance reaching TOA reflected upward by the surface, L_{surf}^{TOA}; and the water-leaving radiance reaching TOA, L_w^{TOA} as described by the following equation (Mobley, 2021).

$$L_t = L_{atm} + L_{surf}^{TOA} + L_w^{TOA} \tag{2.3}$$

Gordon (2019) rewrites the radiance reaching TOA in an expanded way as seen in Equation 2.4. Table 2.3 contains the description of each parameter.

$$L_t(\lambda_i) = L_{path}(\lambda_i) + t_g(\lambda_i)L_g(\lambda_i) + t_{wc}(\lambda_i)L_{wc}(\lambda_i) + t_w(\lambda_i)L_w(\lambda_i) \tag{2.4}$$

Table 2.3: *Parameter notation from Equation 2.4 (Gordon, 2019; Mobley, 2021)*

Parameter	Definition
$L_{path}(\lambda_i)$	is the radiance in the optical path from the measured point to the sensor which accounts for scattering by air molecules, scattering by aerosols and the interaction between air molecules and aerosols (usually represented by $L_r(\lambda_i)$, $L_a(\lambda_i)$ and $L_{ra}(\lambda_i)$ respectively) (IOCCG, 2010). This term also accounts for skylight scattered by the atmosphere and reflected by the sea surface
$t_g(\lambda_i)L_g(\lambda_i)$	describes the direct specular reflection "Sun Glitter" from the surface to the sensor with the transmittance $t_g(\lambda_i)$ and the radiance $L_g(\lambda_i)$.
$t_{wc}(\lambda_i)L_{wc}(\lambda_i)$	describes the whitecap sea-surface regions with transmittance $t_{wc}(\lambda_i)$ and the radiance $L_{wc}(\lambda_i)$
$t_w(\lambda_i)L_w(\lambda_i)$	describes the whitecap free sea-surface regions with transmittance $t_w(\lambda_i)$ and the water-leaving radiance $L_w(\lambda_i)$

The problem of atmospheric correction can then be defined as obtaining approximations of the unknown parameters $L_{path}(\lambda_i)$, $t_g(\lambda_i)L_g(\lambda_i)$, $t_{wc}(\lambda_i)L_{wc}(\lambda_i)$ and $t_w(\lambda_i)$. All these terms that make up the radiative transfer equation are usually obtained through algorithms developed for specific instruments or by ignoring contributions such as the "sun glitter" (achieved by moving the sensor away from the glitter pattern) (Gordon, 2019). If the radiance $L_t(\lambda_i)$ measured on the satellite sensor is known, the water-leaving radiance (or reflectance) $L_w(\lambda_i)$ can be obtained using Equation 2.4.

There are multiple atmospheric correction algorithms that rely on specific sensors mounted on existing satellites and use data collected over the years. The Sen2Cor algorithm is the ESA approach that only works with Sentinel-2 satellites, as it is designed to use the SWIR and visible bands, taking into account the existence of a black pixel for the selection of aerosol optical depth (AOD) selection;

which accounts for the extinction effect of aerosols in the atmosphere (Gitahi and Hahn, 2020; Li et al., 2021). Another correction method designed to be used with the MODIS satellite is the Simplified High-Resolution MODIS Aerosol Retrieval Algorithm (SARA). This algorithm requires optical path measurements from AERONET stations to recover the reflected signal (Gitahi and Hahn, 2020).

ACOLITE is another method originally developed for Landsat and Sentinel-2 imagery, but has also been successfully implemented on other platforms like Sentinel-3 and hyperspectral satellites such as PRISMA (Braga et al., 2022; Vanhellemont and Ruddick, 2021, 2018). There are simpler implementations, such as the single scattering approximation (SSA), which was successfully used in the CZCS sensor (using bands in the NIR region). This method estimates a solution to the radiative transfer equation and obtains the BOA reflectance by assuming that the scattering between air molecules and aerosols $L_{ra}(\lambda_i)$ is zero (Gordon, 2019).

The 6SV1 algorithm proposed by Vermote et al. (2006) approaches the problem with the geometric considerations shown in Figure 2.12. The zenith θ and azimuth ϕ angles must be known for both the Sun and the satellite (subindices s and v respectively) in relation to the observed point.

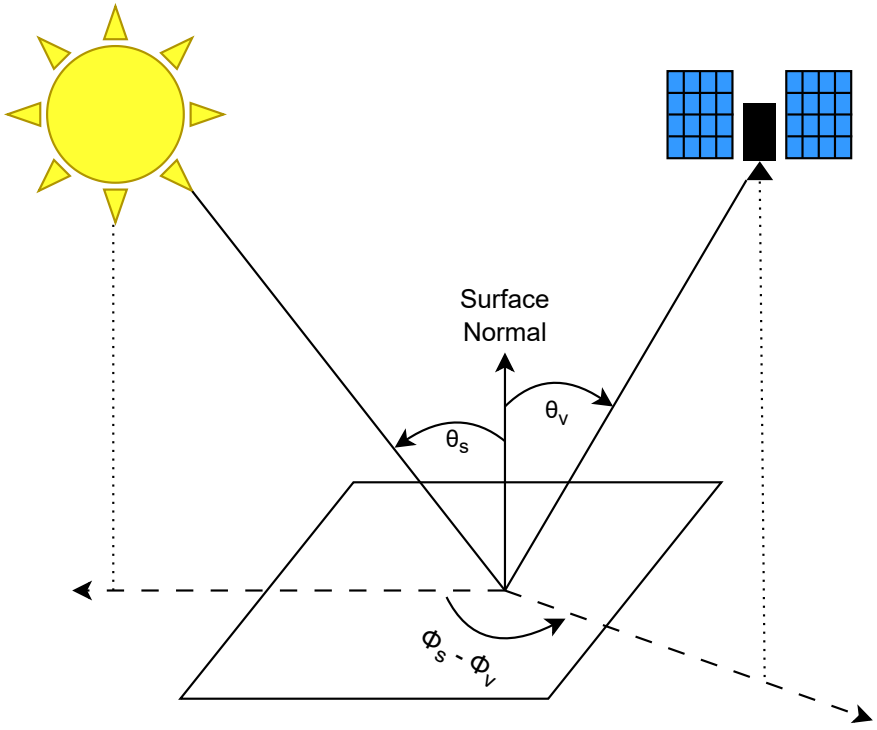
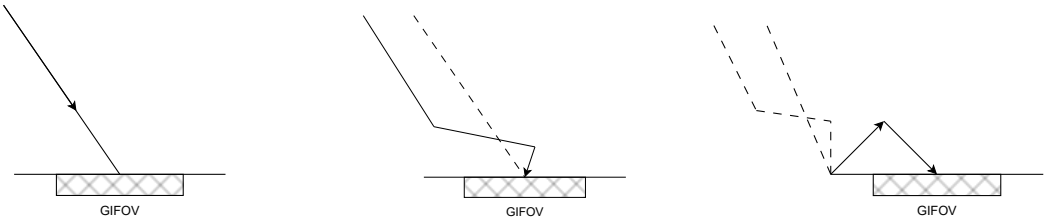


Figure 2.12: *Solar and viewing angles convention used in remote sensing (Vermote et al., 2006).*

The scattering on the surface (without considering any absorption)were described in the method by three main interactions shown in Figure 2.13.



(a) **(b)** **(c)**

Figure 2.13: *Surface Contribution. Reconstructed images from (Vermote et al., 2006)*

The direct solar flux in Figure 2.13a that becomes attenuated by AOT τ, a solar flux at TOA E_s and the solar zenith angle contribution $\mu_s = cos(\theta_s)$ can be expressed as:

$$E_{Sol}^{Dir} = \mu_s E_s e^{-\tau/\mu_s} \tag{2.5}$$

$t_d(\theta_s)$, a diffuse transmittance factor can be used to express the diffuse solar irradiance E_{Sol}^{Diff} that is completely independent of the surface properties as seen in Figure 2.13b.

$$t_d(\theta_s) = \frac{E_{Sol}^{Diff}(\theta_s)}{\mu_s E_s} \tag{2.6}$$

Lastly, the scattering that occurs due to *"trapping"* of the irradiance is generated by successive scatterings and reflections in both the surface and the immediate atmosphere of the studied area (see Figure 2.13c). This successive behavior can expressed in a series such that for a spherical albedo of the atmosphere S and a reflection ρ_t the series would be defined as:

$$\sum_{n=1}^{\infty} \rho_t^n S^n = [\rho_t S + \rho_t^2 S^2 + \cdots \rho_t^n S^n] = \frac{1}{1 - \rho_t S} \tag{2.7}$$

With the series solution the total normalized surface level illumination can be expressed in terms of total the total transmittance $T(\theta_s)$ as:

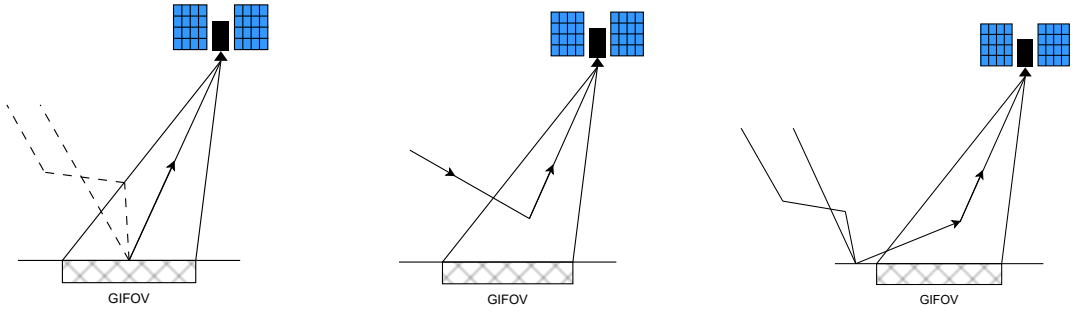
$$E_{Surf}^{Norm} = \frac{T(\theta_s)}{1 - \rho_t S} \quad \text{where} \quad T(\theta_s) = e^{-\tau/\mu_s} + t_d(\theta_s) \tag{2.8}$$

Radiance, as perceived by the satellite from the scattering alone, can be described as seen in Figure 2.14. Vermote et al. (2006) The total contribution of direct and

diffuse components reflected by the surface is shown in Figure 2.14a and can be expressed by the following equation:

$$e^{-\tau/\mu_v} \quad \text{where} \quad \mu_v = cos(\theta_v) \tag{2.9}$$

Figure 2.14b shows the intrinsic radiance of the atmosphere that Vermote et al. (2006) named $\rho_a(\theta_s, \theta_v, \phi_s, \phi_v)$. On the other hand, the contribution of the environment from direct and diffuse sources on external surfaces (see Figure 2.14c) is denoted as the diffuse transmittance $t_d(\theta_v)$.



(a) **(b)** **(c)**

Figure 2.14: *Satellite Contribution. Reconstructed images from (Vermote et al., 2006)*

Considering that $t_d(\theta_s) = t_d(\theta_v)$ due to the reciprocity principle and by considering the previous scattering interactions based on the incident angles, the apparent reflectance ρ^* can be expressed as:

$$\rho^*(\theta_s, \theta_v, \phi_s - \phi_v) = \rho_a(\theta_s, \theta_v, \phi_s - \phi_v) + \frac{\rho_t}{1 - \rho_t S} T(\theta_s) T(\theta_v) \tag{2.10}$$

where,

$$T(\theta_v) = e^{-\tau/\mu_v} + t_d(\theta_v) \tag{2.11}$$

The approach taken by Vermote et al. (2006) to the scattering problem has been described above, however, there are multiple interactions and conditions for which the 6SV1 radiative transfer model finds approximate solutions. Some of these are listed below:

- Non-uniform surface reflectance is considered based on the weighted spatial average taking into account the efficiency of different points based on the geometry.

- Non-uniform altitude can be solved using variations of the Rayleigh Optical Thickness

- Intrinsic atmosphere reflectance over a black target (dark pixel) to account for $\rho_r + \rho_a$ (Rayleigh + aerosol contributions)

- Scattering and absorption interaction is computed through the different molecules absorption for each scattering path. The approximations are made on the basis of the previous knowledge of the affected bands by each molecule.

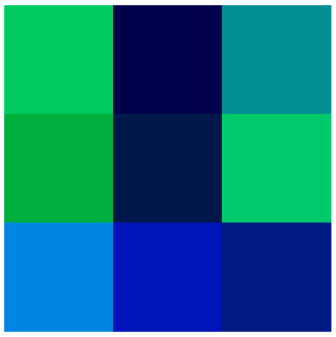
Multiple correction methods exist, which main purpose is to estimate the reflectance of the surface by eliminating atmospheric effects on a measured radiance signal by a space-borne or air-borne sensor. As the reflectance changes based on the properties of the surface, a high-quality correction capable of retrieving this signal becomes of high importance for the study of the Earth's surface.

2.7 Water Detection

Being able to create a water mask to focus on specific regions can result in a beneficial reduction of the computational complexity of the analysis, as well as avoiding outliers with different characteristics (e.g., the signature of a forest is different from an ocean). This is particularly important when dealing with hyperspectral images with millions of pixels and tenths or even hundreds of spectral bands.

For water detection to be possible, the contrast between water and non-water pixels needs to be increased. Zhai et al. (2015) found that the differences for vegetation, water, land and shadow can be greatly improved based on the characteristics of the instrument. During the design phase of a satellite, the selection of bands plays an important role to facilitate the detection of specific regions based on the mission objectives.

To create a water mask, usually the difference between pixels is increased by extracting the main features that can describe each of them. There is no predefined number of features to obtain per pixel, depending entirely on the technique used. Figure 2.15a contains an example of a RGB image (3 channels), from which two features corresponding to the green and blue values are extracted (see Figure 2.15b).



(a) *3x3 Image to Classify*

G: 202 B: 97	G: 0 B: 75	G: 143 B: 146
G: 175 B: 61	G: 23 B: 74	G: 203 B: 108
G: 133 B: 225	G: 20 B: 185	G: 27 B: 132

(b) *Green and Blue Values*

Figure 2.15: *3x3 image and the corresponding Green and Blue values (randomly selected).*

For the particular image used, there are only two features, green and blue values, making the 2D plot easier to visualize in Figure 2.16.

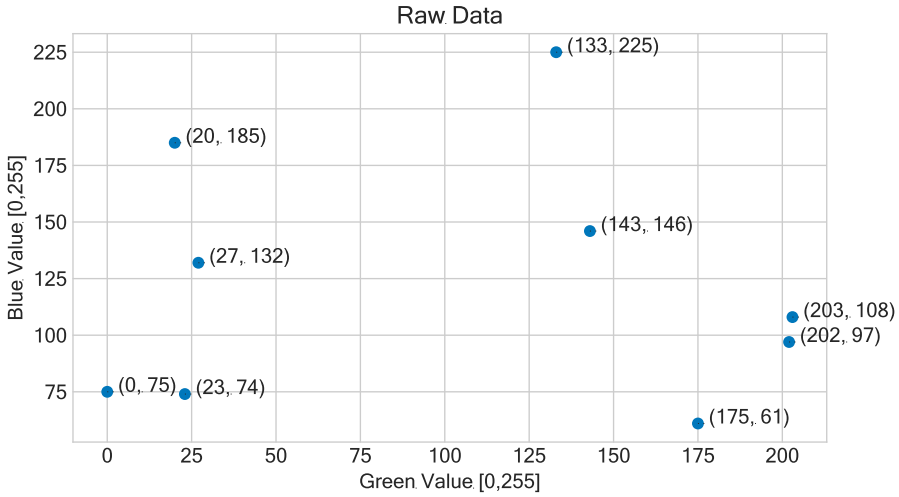


Figure 2.16: *Scatter plot of the Green-Blue values from Figure 2.15*

To enhance the pixel difference in a spectral image by means of features, techniques such as single-band analysis or water detection indices have been developed. These methods are popular as they enhance the difference by applying band subtraction or the ratio of the bands from a spectral image. Simple methods were used in early remote sensing studies, where the detection of regions of interest using a single band was preferred. To detect wet and non-wet regions Wang et al. (2002) used the SWIR spectrum of Landsat 7 Thematic Mapper (TM) from 2.08-2.35 μm. As stated by Ji et al. (2015), hyperspectral mixing due to low spatial resolution can have an impact on the detection of water bodies correctly;

therefore, a single band method makes the classification process more sensitive and error-prone.

Using water detection indices has the advantage of considering information from multiple bands and taking into account the particular variations that characterize specific substances (Xu, 2006). Since indices are easy to compute and explain, they have become quite popular. The most common indices used in the literature can be found in Table 2.4, and although this list is not exhaustive, many of the new approaches are variations of those presented.

Table 2.4: *Classic indices used for water detection*

Name	Formulation	Reference
NDWI	$\frac{Green - NIR}{Green + NIR}$	(McFEETERS, 1996)
MNDWI	$\frac{Green - MIR}{Green + MIR}$	(Xu, 2006)
NDVI	$\frac{NIR - Red}{NIR + Red}$	(JUSTICE et al., 1985; Jiang et al., 2006)

To improve pixel classification, multiple attempts have been made to increase the pixel difference, but the proposals are most of the time satellite-dependent, as the spectral response functions (SRFs) for each one are different. The works of Jiang et al. (2020) and Milczarek et al. (2017) are approaches that focus mainly on Sentinel-2 data due to the available bands. Similarly, the "Automatic Water Extraction Index" of Feyisa et al. (2014) aims to improve the classification of Landsat water pixels by empirical determination of the polynomial coefficients for each satellite band.

Newer missions such as Sentinel-3 with instruments such as the Ocean and Land Color Instrument (OLCI) have been used to describe workflows that use recovered surface reflectance (Santer, 2010). For a pixel to be classified as a water body, it must pass all the tests specified in Table 2.5 for the NIR band (band 12 at 753.75nm), the red band (band 10 at 681.25nm), and the blue band (band 2 at 412.5nm) (L. Bourg et al., 2023). Even though these tests have been determined empirically and applied only to this specific satellite, they have been proven reliable.

Table 2.5: *Water Body detection spectral tests for Sentinel-3 OLCI (L. Bourg et al., 2023).*

$0 < \rho_{blue} < 0.3$	$0 < \rho_{red} < 0.5$	$0 < \rho_{nir} < 0.7$	$\rho_{blue} > \rho_{nir}$

Surprisingly, vegetation indices can be used to detect water surfaces, such is the case with the NDVI index for negative values (Zhai et al., 2015). A drawback is that

for water pixels with a high phytoplankton concentration (similar to that observed in algal blooms), the chlorophyll can create variations on the spectrum, making this particular index unreliable in some situations. Every method may have its drawbacks, emphasizing the need of using a combination of information-highlighting techniques to handle different areas in a scene such as forests, urban, and water. In different studies, the use of NDWI (alone or in combination with other indices such as NDVI) has been found to enhance the contrast between water and non-water regions, making the classification task easier (Acharya et al., 2018). Depending on the existing elements, some indices might work better as different spectrum regions emphasize particular features.

Applying the same indices to different satellites can yield different results. Regions in the spectrum are not clearly defined, making their limits fuzzy. This variation mixed with different satellite bands properties like center wavelength and bandwidth makes not every single band the same, even if, for example, they are both considered NIR.

Once the difference has been increased, the water pixels can be selected on the basis of an optimal threshold, which can be found with different statistical and mathematical formulations that account for local, global, and sometimes time variations (Sekertekin, 2019). Thresholds can be defined empirically for specific sensors and remote sensing scenes, so that the classification of specific scenes is reliable and repeatable. To generalize this process, several methods have been defined to automatically find the threshold using information from a grayscale "feature map".

A very popular option for automatic threshold selection is the Otsu algorithm of Otsu (1979) that uses the histogram variance of the feature map or entropy-based methods such as the one proposed by Kapur et al. (1985). Geometric approaches such as the one from (Zack et al., 1977, as cited in Sekertekin, 2019) have been widely used, while other techniques try to account for fuzziness in the threshold selection process (Huang and Wang, 1995). Regardless of the method used, pixels can be grouped using clustering techniques based on the similarity of features. If the number of features n is plotted in the n-dimensional space, we can visualize it similarly to Figure 2.16.

With iterative methods such as agglomerative clustering (which initially assume that each point is a subgroup), it is possible to create groups from individual points based on how far they are from each other in the feature space, reducing them until the desired number of clusters k is reached. Euclidean distance is a common approach to measuring the distance between two points a and b based on the number of features n using equation 2.12. The "Average Linkage" distance between two clusters A and B is given by the distance between the points of both groups and the total number of points $|A|$ and $|B|$ within the cluster A and B,

respectively (see Equation 2.13).

$$d_{ab} = \sqrt{\sum_{i=1}^{n}(a_i - b_1)^2} \tag{2.12}$$

$$AL = \frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d_{ab} \tag{2.13}$$

At the end of this iterative optimization approach, a desired number of clusters are found that have the smallest distance between the points. Each cluster can be assigned a "centroid", which would represent the point that has the smallest distance to all the points of the group. Figure 2.17 visually shows the distance from each point in each cluster to the center.

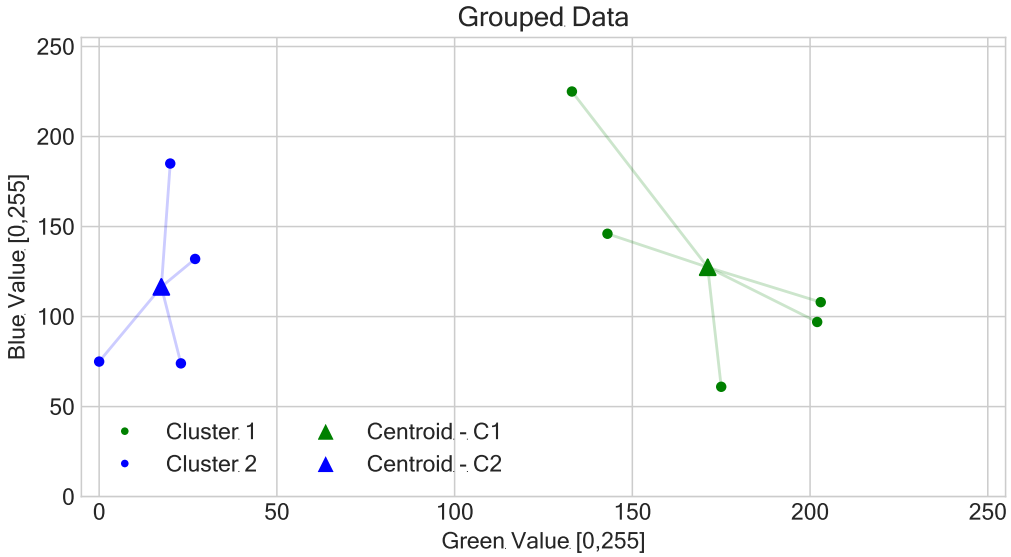


Figure 2.17: *Distance from each of the points in each cluster to the centroid of each group.*

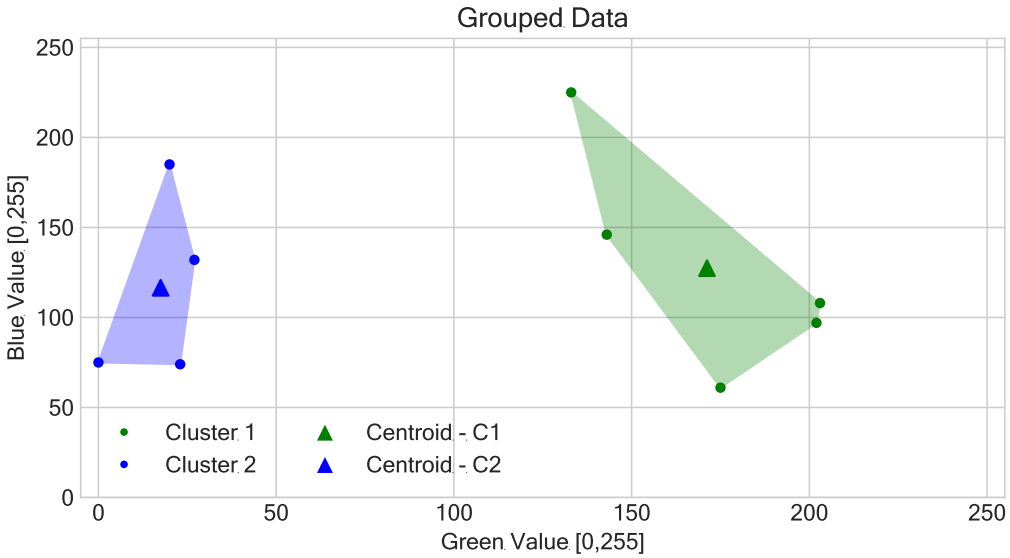


Figure 2.18: *Scatter plot of the groups made with agglomerative clustering using the Green-Blue as input parameters. The centroid of each of the convex-hulls is shown for visualisation.*

Geometric boundaries can be defined with methods like "convex Hull" to define the are at which the boundary conditions are met, as shown in Figure 2.18. It can be said that inside these boundaries all points belong to the same group. If more points are to be added, the best approach would be to repeat the iterative process to determine new centroids and new boundaries.

It has been proven useful to use a combination of water indices with feature clustering through machine learning to enhance the water pixels classification. Cordeiro et al. (2021) proposed the extraction of spectral features through the use of NIR and SWIR bands using the indices NDWI and MNDWI on random pixel sampling. Agglomerative clustering was implemented, defining the number of target groups k with the Calinsk-Harabasz index (based on the spread and density of the data). The resulting groups could then be classified as water or non-water using a custom implementation of the MBWI index (originally developed for Landsat), generalizing classification to the remaining pixels using Naive Bayes. Machine learning approaches such as the maximum likelihood classifier (MLC) are another useful alternative to classify water pixels in an image as done by Gong et al. (2013). To avoid overfitting a method to a particular set of points, we must provide good generalization, which is generally achieved by using data that cover different types of ocean, seasons, and scenes.

Currently, there are multiple data sets on water masks, such as the Global Lakes and Wetland Database (GLWD) (Lehner and Doell, 2004), to more modern ones like MODLAND (for use in MODIS raster images) (Carroll et al., 2009) and

the FROM-GLC which was generated from four machine learning classifiers (Gong et al., 2013, as cited in Ji et al., 2015). Probably the one that stands out the most is GLOBELAND30, which has masks with a resolution of 30m obtained from data from Landat TM and China's HJ-1 satellite with revision in 2000, 2010 and 2020 (Jun et al., 2014).

Satellites such as AQUA-MODIS utilize data sets publicly available from (JPL, 2013), such as SRTM-SWBD (Shuttle Radar Topography Mission Water Body Data). This is a 30m coordinate grid that has flagged which positions belong to a waterbody. Although interpolation can allow us to obtain information in-between pixels, telemetry variations and inconsistent data (sometimes caused by sensor aging) can make the use of geolocation water masks unreliable.

Different descriptors can be used to improve water detection; however, the task becomes more complicated when more classes are introduced, such as urban regions and different types of water. As the SRF of each sensor is different, different approaches that involve multiple features may be used to enhance the contrast of the pixels and improve the classification and generalization of the pixels in a spectral image.

2.8 Chlorophyll Estimation

Estimating chlorophyll concentration from satellite imagery is a complex problem that has yet to be reduced to a simple answer that fits all approaches. At the laboratory level, a fluorometer is commonly used to get accurate results of the chlorophyll concentration, and the spectrum is often measured with a spectrometer. To estimate chlorophyll from a satellite, physical formulations based on the ocean and atmosphere are deemed to be the most accurate; the difficulty of acquiring input parameters and the inherit complexity have also led to semiphysical methods as a less complex alternative.

A well-known empirical approach to chlorophyll estimation is the OCx band ratio algorithm developed by O'Reilly and Werdell (2019) in combination with the Color Index (CI) of Hu et al. (2019). The former method was developed on the basis of empirical data regression, and it has been fine-tuned to account for different water types. Using this polynomial approach, sensor-specific implementations have been designed so that the SRFs are taken into account. Equation 2.14 shows the 4^{th} degree polynomial formulation used for all implementations of this method.

$$log_{10}(chl_a) = a_0 + \sum_{i=1}^{4} a_i \left(log_{10} \left(\frac{R_{rs}(\lambda_{blue}}{R_{rs}(\lambda_{green}} \right) \right)^i \tag{2.14}$$

The second version of the CI algorithm works well in tandem with OCx as it

can improve chlorophyll detection (smaller error and higher R^2) in the range of 0.25-0.40 $\mathrm{mg/m^3}$. Equations 2.15 and 2.16 show the formulation of this method.

$$CI = R_{rs}(\lambda_{green}) - \left[R_{rs}(\lambda_{blue}) + \frac{\lambda_{green} - \lambda_{blue}}{\lambda_{red} - \lambda_{blue}} \left(R_{rs}(\lambda_{red}) - R_{rs}(\lambda_{blue})\right)\right] \quad (2.15)$$

$$log_{10}(chl_a) = a + CI * b \quad (2.16)$$

Each satellite has different spectral bands with specific SRF, so the application of this empirical formulation needs to be recalculated for every sensor to obtain the applicable coefficients.

To aid in the creation of more general methods, multiple publicly available datasets have been created for ocean color measurements that combine *in situ* chlorophyll analysis with BOA (Bottom of Atmosphere) reflectance. Perhaps the most popular one is the dataset by Valente et al. (2022) which has two older versions released in 2019 and 2016 and contains BOA reflectance matches from a spectrometer with satellite estimated R_{rs} and the corresponding chlorophyll measured in the lab. Ruszczak et al. (2022) went as far as introducing a new dataset that includes more complex parameters such as the SPAD (Soil-Plant Analysis Development) parameter and the maximum quantum yield of PSII photochemistry (Fv/Fm).

Most recently Lehmann et al. (2023) released a dataset that aims to have multiple water-type representations with the proper methodology documented for the more than 7,000 measurements. Total suspended solids and absorption are also documented, making it a very good alternative to the other datasets mentioned to analyze the water quality of ocean and inland water bodies.

As the quality of chlorophyll estimation depends highly on the quality of the atmospheric correction (better correction translates to better recovered R_{rs}), it is valuable to retrieve the surface reflectance from the ground as the problem of atmospheric correction has not yet had a definitive solution.

Although the polynomial approach of O'Reilly and Werdell (2019) is one of the most widely used approaches due to its simplicity (currently used by NASA in the AQUA-MODIS mission), there are other approaches that have yielded good results with different techniques. In the following sections, two strategies to estimate chlorophyll from surface reflectance R_{rs} will be presented. The spectral indices method will be covered first, leaving the machine learning approach for last. Although there are many other methods, they will not be discussed in this work to maintain the scope of the thesis.

2.8.1 Spectral Indices

As with water detection indices, the spectral indices for chlorophyll estimation are calculated with band subtraction and ratio, making this a quick and easy method to obtain the chlorophyll concentration based on the type of characteristics of a hyperspectral image. Since both ground and aquatic photosynthetic organisms contain chlorophyll, most of the spectral indices can be reused, as there is absorption in the blue and red spectral regions (see the chlorophyll signature in Figure 2.4).

Indices based on two bands, such as band ratio and band subtraction, may be heavily affected if the atmospheric correction process fails to accurately estimate R_{rs}. Viewing conditions and moisture may be the main contributors to a low-quality correction process, and consequently, the index would not accurately represent the chlorophyll concentration (Abderrazak et al., 1996).

Jiang Hai-ling et al. (2014) compared the precision of multiple indices to estimate the chemical parameters of vegetation, finding that precision is a trade-off with the complexity of the method. Depending on the photosynthetic organism, different chlorophyll indices have shown to have a better correlation between the spectrum change and biomass concentration, such as the Optimized Soil Adjusted Vegetation Index (OSAVI) or the Vegetation Index Based on Universal Pattern Decomposition Method (VIUPD).

Spectral indices have also been mixed with polynomial models to improve biomass estimation using multispectral imaging (Che et al., 2021). Although this approach is pretty simple, it works better when fitted to a particular type of algae or particle. Although this approach is as straightforward, it is usually affected by many parameters and, as has been seen before, can have questionable results (Bannari et al., 2007, as cited in Ruszczak et al., 2022).

2.8.2 Machine Learning

Deep learning supervised techniques have been used with remote sensing hyperspectral images to extract low-level features and perform a pixel-wise data analysis (Zhang et al., 2016). Although deep learning approaches generally perform better than other methods, they lack explainability and require very large datasets to avoid overfitting and underrepresentation (Barbedo, 2018).

Ye et al. (2021) retrieved chlorophyll-a concentration values using MODIS-AQUA multispectral images (particularly focused in the visible bands) to train a two-stage CNN but faced the dataset size restrictions, thus using synthetic oversampled *in situ* data to compensate. Representation of different types of water bodies at different concentrations also presented a challenge for reproducibility, as samples need to be coordinated worldwide.

Multiple branch CNNs for satellite imagery have also been used by Tulczyjew et al. (2022), where spectral, spatial, and spectral-spatial characteristics were separated to aid in the unmixing of hyperspectral signals. By separating the signal composition, it is thought that the estimation of chlorophyll could be enhanced, but the problem of complex absorption and scattering on the surface remains.

Pyo et al. (2019) implemented a point-centered regression CNN (PRCNN) in which windows of varying size were defined in which the center pixel would contain the chlorophyll measurement. By this the surrounding pixels were considered in the estimation of chlorophyll by using physical measurements as total flux and path radiance as part of the input spectral cube. Although the approach to this complex problem used CNN, the solution can be considered an analytical approximation due to the physical measurements included in the process. The results presented in this work seemed to have a better performance than the chlorophyll estimation methods based on Inherit Optical Properties from Gons et al. (2002); Li et al. (2015). The results were presented for a specific area and rely on physical measurements for each spectral image, making it not a viable option for wide area coverage.

Bakken et al. (2021) utilized linear models such as the "Partial Least Squares Regression" (PLSR) and the "Least Absolute Shrinkage and Selection Operator" (LASSO) to estimate chlorophyll-a based on the TOA reflectance from HICO satellite spectral images. While this helps in reducing the complexity with good accuracy, it does not account for scattering and absorption on the optical path.

Many strategies have been proposed to estimate chlorophyll from R_{rs}, ranging from empirical methods to artificial neural networks (ANN). Each of the implementations has been useful in overcoming specific challenges such as the sensitivity to non-optimal atmospheric corrections as well as small datasets. Most of the approaches seem to focus on a specific sensor; using the available bands to extract features with spectral indices and principal component techniques. Generalization is still important to consider, as environmental and atmospheric conditions change constantly.

3 Methodology

3.1 Overview

To begin to address the problem of chlorophyll inversion, HYPSO-1 hyperspectral images had to be manually georeferenced to match a standard coordinate reference system using known ground points and their equivalents in the hypercube. This became an essential step as the coordinate grid of some captures did not accurately represent the scene.

After completion of this process, the radiance cube was atmospherically corrected to eliminate the effects of the atmosphere as perceived by the sensor. This process allows one to obtain the BOA reflectance, which describes the properties of the surface. To achieve this, in this work, the radiative transfer model of the 6SV1 algorithm was implemented for the first time for HYPSO-1 spectral images.

Knowing the reflectance values of scene pixels, it was possible to classify each one of them as water or non-water. The former is of interest, as the purpose of this work is to estimate ocean chlorophyll. To create a working dataset, the spectral images of Sentinel-3 and MODIS-AQUA were matched with the coordinates of the HYPSO-1 captures in a $\pm$ 3 hour window (Seegers et al., 2018). This was done to extract chlorophyll estimates from ESA / NASA due to the lack of *in situ* ground truth. Linear interpolation was used to retrieve the Chl-a value at the HYPSO-1 coordinates from the matching satellite images.

Having both the HYPSO-1 and the GLORIA dataset, a data analysis stage was introduced to create and select relevant features to improve the estimation process. The problem of estimating chlorophyll with the surface reflectance of spectral images was then approached as follows:

1. Multiple features were created from the reflectance spectrum using existing chlorophyll descriptor models. Simpler descriptors, such as band ratios and band difference, were also considered, along with the corresponding logarithmic transformations. For each descriptor X that requires r different wavelengths, all permutations ${}_nP_r$ were tested using all n bands of HYPSO-1/GLORIA, selecting the five with the highest correlation with chlorophyll

prediction. The selected r bands in each of the five selected permutations were ensured to correspond to a different region (see Table 3.4) and that set was unique with respect to the remaining four (irrespective of order).

2. Polynomial combinations for the models with the highest correlation were created up to 3^{rd} degree. The process was to account for more complex relationships than the contribution of a single feature to chlorophyll regression.

3. Five features were selected by cross-validation using recursive feature elimination (RFE), sequential feature selector (SFS), and stepwise forward-backward regression. Both RFE and SFS were implemented with random forest regression as used by Adam et al. (2014) to reduce overfitting in the subset selection process.

4. The features in the optimal subset were combined with both linear regression and ensemble machine learning voting and weighting techniques. The multivariate regression approach is similar to the one used by Cao et al. (2020), but the technique is improved by using multiple models.

The workflow of the entire process is shown in Figure 3.1, where each of the colored elements of this diagram will be detailed later in this chapter.

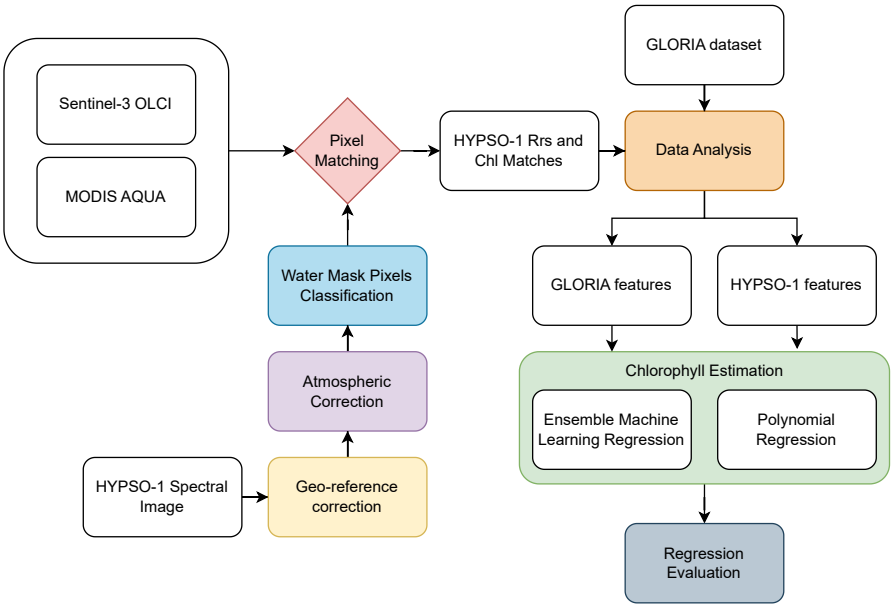


Figure 3.1: *Methodology overview workflow diagram of the current work. Colored regions are carefully described in the sections of this chapter.*

3.2 Geo-reference Correction

Due to the inconsistency in the telemetry data coming from the HYPSO-1 captures, manual correction of the geolocation coordinates had to be performed. An example of the coordinate mismatch to ground points can be seen in Figure 3.2 where the coast of Florida does not correspond to the Open Street Maps (OSM) reference used.

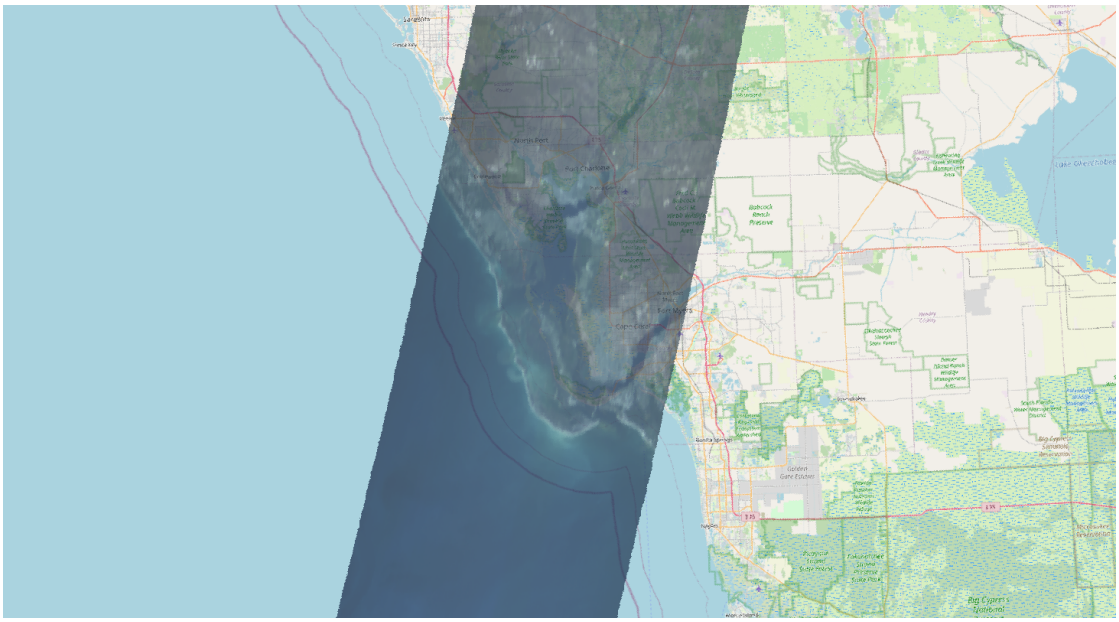


Figure 3.2: *Overlap of the HYPSO-1 capture wrong coordinates and the OSM reference. Through visual inspection and the change in opacity of the spectral image, it is possible to see a mismatch along the coast line.*

To correct each image, characteristic areas have to be selected in both the map and the HYPSO-1 capture. At least 8 points (4 in the spectral image, plus 4 on the map) are required to perform this correction. To make the process easier, the matches were put together by loading the GeoTIFF of each spectral image into the software "QGIS Desktop" version 3.28.3 and using the "Georeferencer" tool. Figure 3.3 shows the manually selected points so that a geometric transformation can be implemented.If the nature of the image allowed, the selected points were as far apart as possible from each other.

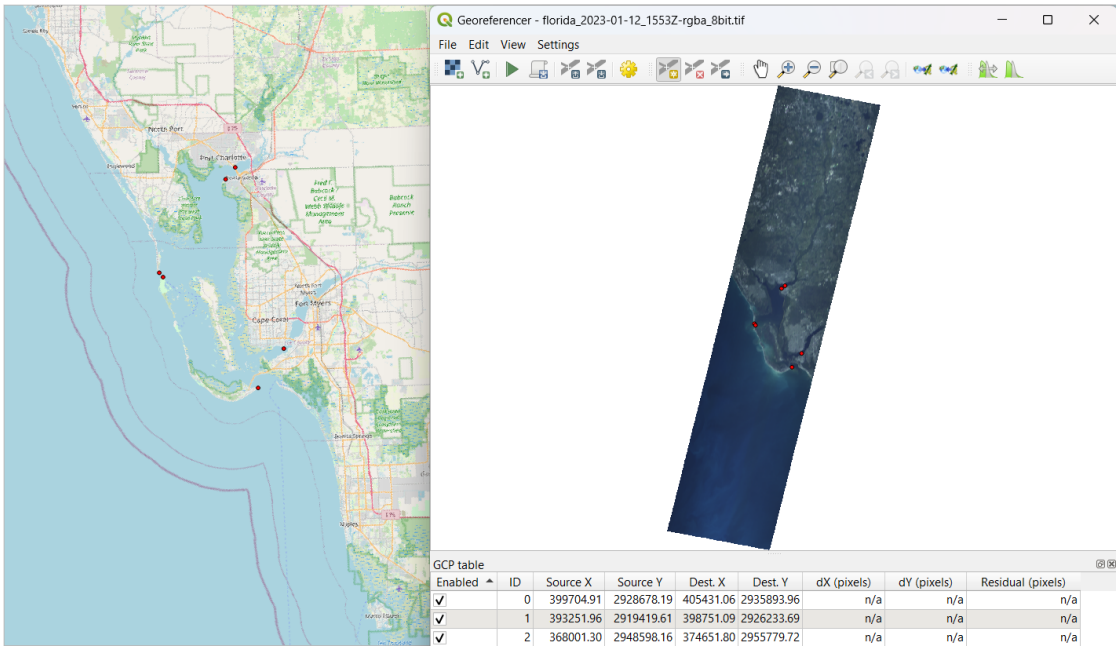


Figure 3.3: *Ground control points (GCP) are shown in red which were manually matched between the source points of the spectral image and the desired destination on the reference map using the same reference system. The X and Y coordinates from source to destination can be seen in the Coordinate Reference System (CRS) units for both latitude and longitude.*

Initially, the exported coordinates from the "Georeferencer" tool were used to estimate a transformation from the source (HYPSO-1) to the destination (OSM map). The Python OpenCV library was used to approximate the homography matrix 3×3 of Equation 3.1 as well as a second-degree polynomial, but the results were far from optimal for both of these methods.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.1}$$

Using the "linalg.lstsq" function in the *numpy* library, the approximation to x in equation 3.2 was derived. This linear transform created the most visually accurate correction of all the alternatives tested. Source coordinates were placed in matrix A and destination points in matrix b. This process was carried out independently for both latitude and longitude, since each parameter has its own source-destination pairs (see the bottom right area of Figure 3.3). The estimation had to be repeated for each image as the mismatch was not constant between captures.

$$Ax = b \tag{3.2}$$

Figure 3.4 shows the result of the HYPSO-1 image after a proper transformation was implemented.

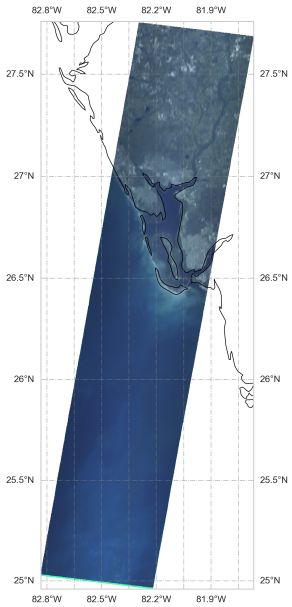


Figure 3.4: *RGB image of the HYPSO-1 capture after correcting the coordinates using a the least-squares solution from Equation 3.2. Image was fitted to another CRS using "Cartopy".*

3.3 Atmospheric Correction

The 6SV1 algorithm for atmospheric correction was selected for this work, being the first implementation available on the HYPSO-1 satellite (Vermote et al., 2006). All modifications and considerations made to the algorithm are discussed in this section. Parameters from the HYPSO-1 spectral image metadata were used to define the geometry of the scene, which is fundamental to implement a correct atmospheric correction. To avoid confusion with other authors, the naming conventions of Jensen (2014) for the atmospheric parameters were used in this section.

- **INPUT PARAMETERS:**

  1. Radiance: From the HYPSO-1 process pipeline, the L1B radiance hypercube can be obtained after it has been radiometrically corrected using the coefficients defined in previous calibration studies (Henriksen et al., 2022). Radiance per band was the input of the model, which means that for n bands, atmospheric correction had to be performed n times. The

correction performed for each band is considered to apply to every pixel it contains.

2. Month and day: The ISO time in the HYPSO-1 metadata was parsed in a Python-friendly format so that the month and day could be extracted and fed to the model.
3. Latitude and Longitude: The boundaries surrounding the capture were obtained with the 2D latitude and longitude grid of the spectral image. At the same time, the approximate center of the captured region was estimated with Equation 3.3.

$$lon_c = \frac{max(lon) + min(lon)}{2} \qquad lat_c = \frac{max(lat) + min(lat)}{2} \tag{3.3}$$

4. Surface altitude: A Digital Elevation Model (DEM) with a resolution of 2 km was used; the mean altitude value in m was obtained from all values within the region described by the upper left (UL) and lower right (LR) points (see Table 3.1).

Table 3.1: *UL and LR considerations for obtaining the mean altitude through a DEM.*

Upper Left Corner	Lat	max(lat)
	Lon	min(lon)
Lower Right Corner	Lat	min(lat)
	Lon	max(lon)

5. Ground Reflectance: An uniform parameterized BRDF (lambertian) was used to describe aquatic regions as only water pixels will be used and continental areas will be ignored. From all the built-in options in the 6SV1 model, the "Lake Water" reflectance profile was selected as it better describes the oceanic regions studied in this work compared to the other profiles. To avoid unnecessary assumptions, it was decided to select a standard profile rather than choosing an arbitrary constant reflectance value for the BRDF model.
6. Atmospheric Profile: The latitude at the center of the spectral image was used to select the atmospheric profile for the correction. The region, in combination with the month, allows one to follow the criteria of Table 3.2 to define the atmospheric profile between tropical (T), mid-latitude summer (MLS), mid-latitude winter (MLW), subarctic summer (SAS),

and subarctic winter (SAW). A simplified criterion as defined early by NASA (1966) can also be used (see Table A.1 in the Appendix).

Table 3.2: *Atmospheric profile conditions based on the month of the capture and the center latitude of the spectral image. Values defined for the 6SV1 algorithm and adapted from the FLAASH atmospheric correction documentation (Felde et al., 2003). A more accurate selection would require vapor information in the optical path or surface air temperature .*

Lat (°N)	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
80	SAW	SAW	SAW	SAW	SAW	SAW	MLW	MLW	MLW	MLW	SAW	SAW
70	SAW	SAW	SAW	SAW	MLW	MLW	MLW	MLW	MLW	MLW	SAW	SAW
60	MLW	MLW	MLW	MLW	MLW	MLW	SAS	SAS	SAS	SAS	MLW	MLW
50	MLW	MLW	MLW	MLW	SAS	SAS	SAS	SAS	SAS	SAS	SAS	SAS
40	SAS	SAS	SAS	SAS	SAS	SAS	MLS	MLS	MLS	MLS	SAS	SAS
30	MLS	MLS	MLS	MLS	MLS	MLS	T	T	T	T	MLS	MLS
20	T	T	T	T	T	T	T	T	T	T	T	T
10	T	T	T	T	T	T	T	T	T	T	T	T
0	T	T	T	T	T	T	T	T	T	T	T	T
-10	T	T	T	T	T	T	T	T	T	T	T	T
-20	T	T	T	T	T	T	MLS	MLS	MLS	MLS	T	T
-30	MLS	MLS	MLS	MLS	MLS	MLS	MLS	MLS	MLS	MLS	MLS	MLS
-40	SAS	SAS	SAS	SAS	SAS	SAS	SAS	SAS	SAS	SAS	SAS	SAS
-50	SAS	SAS	SAS	SAS	SAS	SAS	MLW	MLW	MLW	MLW	SAS	SAS
-60	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW
-70	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW	MLW
-80	MLW	MLW	MLW	MLW	MLW	MLW	SAW	SAW	MLW	MLW	MLW	MLW

7. Aerosol Optical Thickness at 550nm: The value of the parameter was manually retrieved using information provided by the NASA Earth Observation (NEO) website. The aerosol optical thickness (AOT) can be recovered by defining a geographic region of interest (ROI). The AOT value closest to the center of the HYPSO-1 spectral image was chosen because the spatial resolution of MODIS-AQUA is lower than that of the NTNU satellite. As an alternative, it is possible to use AERONET profiles from an available station, but since no images were captured on top of these areas, this option was not used.
8. SRF and wavelength: A Gaussian distribution for the spectral sensitivity of each band of the hyperspectral imager was assumed to describe the sensitivity in a defined wavelength range. The hyperspectral nature of the device allows this to be an educated approach, although the actual SRF would be desired. Using the calculation of FWHM by Henriksen et al. (2022), Equation 3.4 can be established using $w = 50\mu m$ as the width of the slit, $a = 3.33\mu m$ for the spacing of the groove of the grating,

the incident angle of light as $\alpha = 0\,\text{deg}$, the order of the grating $k = 0$ and the focal length fixed at $f = 50mm$.

$$FWHM = \frac{wa\cos(\alpha)}{kf} = 3.33nm \tag{3.4}$$

With a constant FWHM value defined, σ can be calculated with Equation 3.5.

$$\sigma = \frac{FWHM}{2\sqrt{2 \cdot log(2)}} = 1.40138nm \tag{3.5}$$

Equation 3.6 is used to establish the SRF that will be used for every band of HYPSO-1. The set g_x is given by Equation 3.7, and is defined according to the 6SV1 algorithm requirement of a fixed step of $2.5nm$ in the device response function.

$$SRF = exp\left(-\frac{1}{2}\frac{g_x^2}{\sigma^2}\right) \tag{3.6}$$

$$g_x = \{r : \exists n \in \mathbb{N}, \text{ such that } r = -3\sigma + 2.5n, \text{ and } r \in [-3\sigma, 3\sigma)\} \tag{3.7}$$

For the central wavelength λ_o of each band, the spectral range in which the SRF corresponds can be easily obtained using the set g_x from $[-3\sigma, 3\sigma)$. Equation 3.8 is used to center the SRF around the central wavelength λ_o. For each band, the first and last values of wl_x are inputs to the 6SV1 model along with the SRF.

$$wl_x = g_x + \lambda_o \tag{3.8}$$

9. Solar and viewing angles: As previously shown in Figure 2.12, the zenith angles θ and azimuth angles ϕ are used to describe the locations of the capture surface with respect to the satellite and the Sun. Both solar angles θ_s and ϕ_s are floating point values obtained from the metadata in the HYPSO-1 GeoTIFF file. For the satellite (viewing) angles θ_v and ϕ_v, the 6SV1 implementation requires a zenith-azimuth pair for each spectral band. Ideally, the mean value per band is desired, but because the HYPSO-1 metadata contains only two values per capture, the same value was repeated for each of the 120 bands.
10. Aerosol Profile: The aerosol profile is used to describe the behavior of the particles and aerosols in the capture area. The maritime aerosol model was selected based on the description given by F.X. Kneizys et al. (1996)

which corresponds to areas with mainly sea salt particles and conditions of high relative humidity. All the spectral images chosen for this study were from coastal regions, which perfectly fits the characteristics of a maritime model.

- **OUTPUT:**
  1. E_{DIR}: Direct Solar Irradiance
  2. E_{DIFF}: Diffuse Solar Irradiance
  3. L_p: Path radiance from scattering
  4. $T_{abs\uparrow}$: Transmittance coefficient of upward absorption
  5. $T_{scat\uparrow}$: Transmittance coefficient of upward scattering

The output had to be recalculated per band, updating the zenith and azimuth angles, as well as the SRF with its corresponding spectral range. To retrieve the L2A spectral image (also known as reflectance hypercube), additional steps had to be taken.

The total radiance L_S reaching the sensor can be defined in terms of L_T and L_P, which represent the total radiance coming from the target on the surface and the path radiance from scattering, respectively (Mobley, 1999). From Equation 3.9 only L_S is available with the spectral image, the rest is unknown.

$$L_S = L_T + L_P \qquad [\mathrm{W\,m^{-2}\,sr^{-1}}] \quad (3.9)$$

The global spectral irradiance on the surface $E_{g\lambda}$ is given by the integration of two components. The first component is the solar spectral irradiance at TOA $E_{s\lambda}$, which is affected when entering the atmosphere by the downward transmittance T_{θ_s} in the direction of the solar zenith angle with the cosine of the same angle. The second component is the diffuse spectral sky irradiance $E_{d\lambda}$) that comes from all directions. The integral in a defined wavelength range results in irradiance that reaches the surface. The sum rule of integration allows one to separate an integral of the sum of functions into the sum or their integrals, as shown in Equation 3.10.

$$\begin{aligned} E_{g\lambda} &= \int_{\lambda_1}^{\lambda_2} (E_{s\lambda} T_{\theta_s} cos(\theta_s) + E_{d\lambda}) d\lambda \\ &= \int_{\lambda_1}^{\lambda_2} E_{s\lambda} T_{\theta_s} cos(\theta_s) d\lambda + \int_{\lambda_1}^{\lambda_2} E_{d\lambda} d\lambda \end{aligned} \quad (3.10)$$

For simplicity, the previous equation can be separated into two simplified values:

$$E_{DIR} = \int_{\lambda_1}^{\lambda_2} E_{o\lambda} T_{\theta_s} cos(\theta_s) d\lambda \qquad E_{DIFF} = \int_{\lambda_1}^{\lambda_2} E_{d\lambda} d\lambda \quad (3.11)$$

A simplified version of equation 3.10 can be written as follows:

$$E_{g\lambda} = E_{DIR} + E_{DIFF} \tag{3.12}$$

L_T, the radiance directly from the target can be defined by knowing $E_{g\lambda}$. Equation 3.13 considers the reflectance properties of the surface as ρ_λ in a specific wavelength range and also the total transmittance T_{θ_v} in the viewing direction. Based on the Lambertian assumption for the surface reflectance selected for the 6SV1 parameters, a factor of $1/\pi$ is added in equation 3.13 to take into account that radiation gets reflected evenly in the upper hemisphere.

$$\begin{aligned} L_T &= \frac{1}{\pi}\int_{\lambda_1}^{\lambda_2} \rho_\lambda T_{\theta_v}(E_{s\lambda}T_{\theta_s}cos(\theta_s) + E_{d\lambda})d\lambda \\ &= \frac{1}{\pi}\rho_\lambda T_{\theta_v} E_{g\lambda} \\ &= \frac{1}{\pi}\rho_\lambda T_{\theta_v}(E_{DIR} + E_{DIFF}) \end{aligned} \tag{3.13}$$

Using Equation 3.9 it can be established that $L_T = L_S - L_P$, which when combined with Equation 3.13 produces the following equality.

$$L_S - L_P = \frac{1}{\pi}\rho_\lambda T_{\theta_v}(E_{DIR} + E_{DIFF}) \tag{3.14}$$

Rearranging Equation 3.14 into Equation 3.15 leads to the same model that Moran et al. (1992) has used before to retrieve surface reflectance from satellites. This final equation makes it possible to use all the outputs of the 6SV1 model, which are floating-point scalars. The parameter T_{θ_v} accounts for both absorption and upward scattering transmittance such that $T_{\theta_v} = T_{abs\uparrow} \cdot T_{scat\uparrow}$. With this model, ρ_λ per satellite band can be obtained, which represents the surface target reflectance at the central wavelength λ_o for the integration range defined by the first and last element of wl_x as λ_1 and λ_2. The parameter L_S is the only input, which corresponds to the corrected radiance per channel that comes directly from the HYPSO-1 GeoTIFF spectral image.

$$\rho_\lambda = \frac{\pi(L_S - L_P)}{T_{\theta_v}(E_{DIR} + E_{DIFF})} \tag{3.15}$$

The Python implementation of the 6SV1 algorithm offers the option of using the TOA reflectance instead of the radiance L_S perceived by HYPSO-1. To do the conversion, the formulation included in the documentation by USGS (2019) can be used (see Equation 3.16). Where $L_{S\lambda}$ is the spectral radiance at the sensor after radiometric correction, θ_s is the solar zenith angle and ESUN is the mean solar exo-atmospheric irradiance.

$$\rho_\lambda = \frac{\pi \cdot L_{S\lambda} \cdot d^2}{ESUN_\lambda \cdot cos(\theta_s)} \tag{3.16}$$

For each spectral band b the ESUN can be calculated by numeric integration using the extraterrestrial solar spectral irradiance at the top of the atmosphere $E_{s\lambda}$. In the case of HYPSO-1, the SRF for every band is the same as per the Gaussian assumption stated earlier in this work. Initially the SRF needs to be normalized by using equation 3.17 and then the integration in the entire spectrum of equation 3.18 will return the ESUN value for band b.

$$SRF'_b = \frac{SRF_b}{\int^\lambda SRF_b} \tag{3.17}$$

$$ESUN_b = \int^\lambda SRF'_b \cdot E_{s\lambda} \qquad [\mathrm{mW\,m^{-2}\,nm^{-1}}] \tag{3.18}$$

The distance to the Sun d in astronomical units (AU) can be approximated by Kepler's first law to describe the orbital ellipses, knowing the eccentricity $e = 0.01672$ that describes the orbit of the Earth. Equation 3.19 shows this law where for simplification $x = e \cdot cos(\theta)$.

$$r = a\frac{1-e^2}{1+e\cdot cos(\theta)} \Rightarrow a\frac{1-e^2}{1+x} \tag{3.19}$$

The Taylor series expansion around very small values of x (mainly due to e), allows one to approximate the function $\dfrac{1}{1+x}$ to $(1-x)$, resulting in equation 3.20.

$$r = a(1-e^2)(1-x) \Rightarrow a(1-e^2)(1-e\cdot cos(\theta)) \tag{3.20}$$

$360°$ can be divided by 365.2422, the number of true days in a year to get the angle equivalence per day $360°/365.2422 = 0.9856°/day$. Four days need to be subtracted from the used Julian day for this equation to be valid, as the periphelion (closest point to the Sun) occurs approximately on January 4th.

$$a(1-e^2)(1-e\cdot cos(0.9856(jday-4))) \tag{3.21}$$

If instead of meters, the calculation is performed with AU, the term $a(1-e^2) \approx 1$. The distance in AU to the Sun can be approximated with Equation 3.22. This approximation starts from the periphelion and thus has higher variations the more we diverge from it. An alternative is to use the lookup tables by USGS (2019), which change annually, and interpolate the AU based on the month and day.

$$d = 1 - 0.0167 \cdot cos(0.9856(jday-4)) \tag{3.22}$$

3.4 Water Mask Pixels Classification

Having a spectral range in HYPSO-1 of approximately 390 to 804nm limits the options of water detection indices that use SWIR and NIR ranges. Based on this limitation, the approach by Cordeiro et al. (2021) was selected because it builds a robust strategy for water detection regardless of the index used. Figure 3.5 shows the overall implementation of HYPSO.

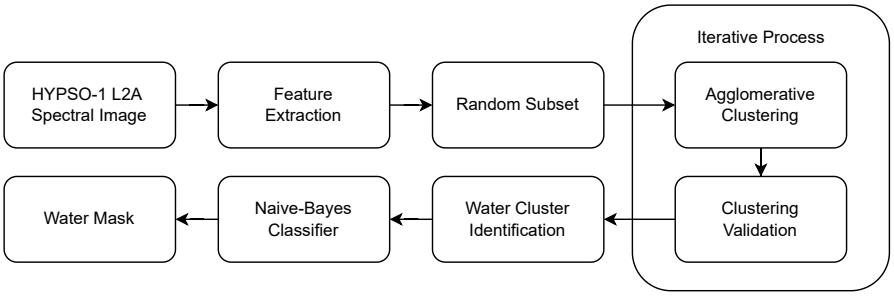


Figure 3.5: *Water detection process overflow based on Cordeiro et al. (2021) as implemented on HYPSO-1.*

The Python implementation of this method was used on the HYPSO-1 spectral image after atmospheric correction was performed. Some parameters were tuned to improve the water detection as explained below.

1. Feature extraction: Although this method is capable of using complex indices such as the Modified Normalized Difference Water Index (MNDWI) and the Multi-Band Water Index (MBWI), because of the lack of SWIR bands, only the NDWI index and the NIR bands were used. The HYPSO-1 bands closest to those used in the original work were selected. Table 3.3 shows the equivalent Sentinel-2 bands for each HYPSO-1 band used. In addition to these two features, the Otsu thresholding algorithm was also applied on the NDWI index to create a third characteristic for the classification.

Table 3.3: *Equivalent HYPSO-1 bands to the ones of Sentinel-2 used in the original work of Cordeiro et al. (2021). Band 3 in Sentinel-2 has an FWHM of* $\approx 34.798nm$ *while band 8 is* $\approx 104.784nm$*. Although the band 120 on HYPSO-1 is closer in wavelength to the band 8 of Sentinel-2 at* $\approx 803nm$*, it was not used as the last band tends to be noisy on hyperspectral systems.*

Band	Sentinel-2	HYPSO-1
Green	B3 (560nm)	B49 (560nm)
NIR	B8 (835nm)	B119 (801nm)

2. Random Subset: Having calculated the three selected features for each pixel in the spectral image, 20% of them were randomly selected as the training subset.

3. Iterative Process:

    - Agglomerative Clustering: Although agglomerative clustering occupies more memory than K-means on a high pixel count ($O(n^3)$ vs. $O(n^2)$), it was used because it was the selected choice of the original work.
    - Clustering Validation: The Calisnk-Harabasz index is used to identify the best number of clusters K through the intercluster and intracluster variance (Cordeiro et al., 2021). This process is repeated testing multiple values for the K number of clusters from 2 to 7 (default values in the Python implementation). A higher Calisnk-Harabasz index is better as it denotes a higher density which is desired for data clustering, selecting it as the target number of clusters.

4. Water Cluster Identification: In the original work, the MBWI index was calculated for the centroid of each of the K clusters found to classify as a water cluster the one with the maximum MBWI value. For HYPSO-1 images, the water cluster was selected using the max NDWI value.

5. Naive-Bayes Classifier: The generalization of the previous training is implemented for the rest of the pixels using the Naive-Bayes classifier, although SVM also gives good results at the cost of a longer computation time.

The water mask was used in the later stages to identify with pixels to work with, discarding nonwater regions for this work.

3.5 Pixel Matching

Two conditions were established to select matches of HYPSO-1 with Sentinel-3 and MODIS-AQUA spectral images, one for the entire capture and the second for individual pixels. A HYPSO-1 pixel is considered "matched" if both conditions are met, and although matching with one satellite is enough, having both matches is the desired case.

1. Time match: The condition was met if the HYPSO-1 capture time was within a $\pm$ 3 hours window with respect to the time at which Sentinel-3 or MODIS-AQUA captured the same region with at least a 20% overlap (Seegers et al., 2018).

2. Coordinate match: The condition was met if the HYPSO-1 pixel is within the capture region of the Sentinel-3 satellite or the MODIS-AQUA satellite.

For each of the HYPSO-1 pixels marked as "matched", the chlorophyll value was obtained via 2D linear interpolation from the coordinate and chlorophyll grid of Sentinel-3 and MODIS-AQUA independently (one or both). The coordinates for all three satellites are given at the center of the pixel, so no additional considerations were implemented regarding this.

Once the two-dimensional chlorophyll map was obtained for each HYPSO spectral image, a 2D normalized Gaussian convolution was applied so that the values would be smoothed. The 3x3 discrete Gaussian kernel in Equation 3.23 was used considering a symmetrical padding for the edge pixels.

$$\begin{bmatrix} 1/16 & 1/8 & 1/16 \\ 1/8 & 1/4 & 1/8 \\ 1/16 & 1/8 & 1/16 \end{bmatrix} \tag{3.23}$$

The HYPSO-1 data set is then formed by combining reflectance and interpolated chlorophyll values on a single data frame.

3.6 Data Analysis

In addition to the HYPSO dataset, the *in situ* surface reflectances and laboratory-measured chlorophyll of the GLORIA dataset by Lehmann et al. (2023) was also considered. This was done to confirm the generalization of the chlorophyll inversion process on other high-spectral-resolution measurements as a means to verify the used methodology.

The GLORIA and HYPSO data points were divided into independent training, testing and validation splits by 33.5%, 50% and 16.5%, respectively. Only the training and validation splits were used during this stage, leaving the test split for the estimation and evaluation at the end.

Currently, there is no consensus on the limits of each region in the spectrum. Different authors may use different ranges mostly based on the available bands in the satellite they are studying. Because HYPSO-1 and the GLORIA dataset have reflectances with high spectral resolution, boundary selection is not obvious. To eliminate the risk of potential confusion when comparing the approach of this work with others, the ranges in Table 3.4 will be used for the remainder of this study.

Table 3.4: *Spectral ranges used for this work. The values may change in name and values between different literature sources.*

Spectral Region	Range (nm)
Blue	$[400, 500)$
Green	$[500, 600)$
Red	$[600, 700)$
Far-Red	$[700, 750)$
NIR	$[750, 1000)$

From the Quantile-Quantile plots in Figure 3.6, it can be seen that the interpolated chlorophyll does not fall on the red line, which means that it does not follow a Gaussian distribution. Many developed chlorophyll models use the log transformation of chlorophyll because it has been shown before that a normal distribution is followed by this correction (Campbell, 1995, as cited in Seegers et al., 2018).

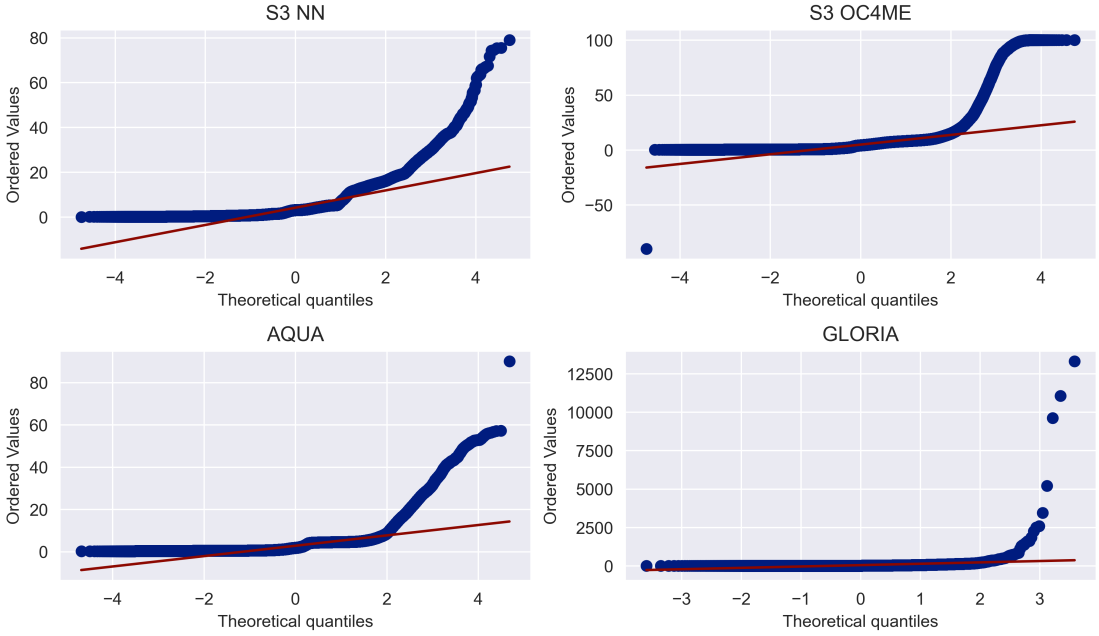


Figure 3.6: *Quantile-Quantile Plot for matching chlorophyll values on HYPSO-1 pixels. Normal distribution can be verified by following the straight line pattern.*

Considering the non-normality of the values, a log transformation was implemented on the training and validation splits attempting to achieve a more normal-distributed behavior. This will, of course, have to be considered for the final estimates, where *unlogging* would be required to compare with the test split. Figure 3.7 shows the distributions after implementing this change. After transformation, a tendency towards normality can be seen for all the distributions in the Q-Q plot, as they have moved closer to the red line. Due to these results, the log transformation will be used as a regression target in this work.

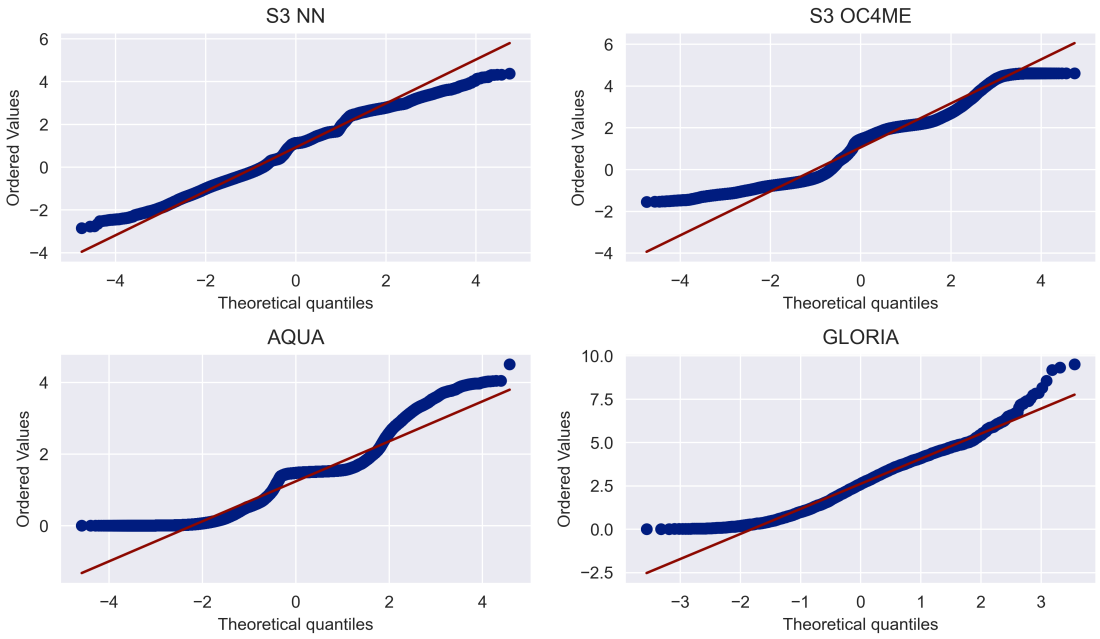


Figure 3.7: *Q-Q Plot matching chlorophyll values on HYPSO-1 pixels after applying log transformation.*

3.6.1 Feature Creation

The feature creation process is shown in the diagram in Figure 3.8, which will be described in this section.

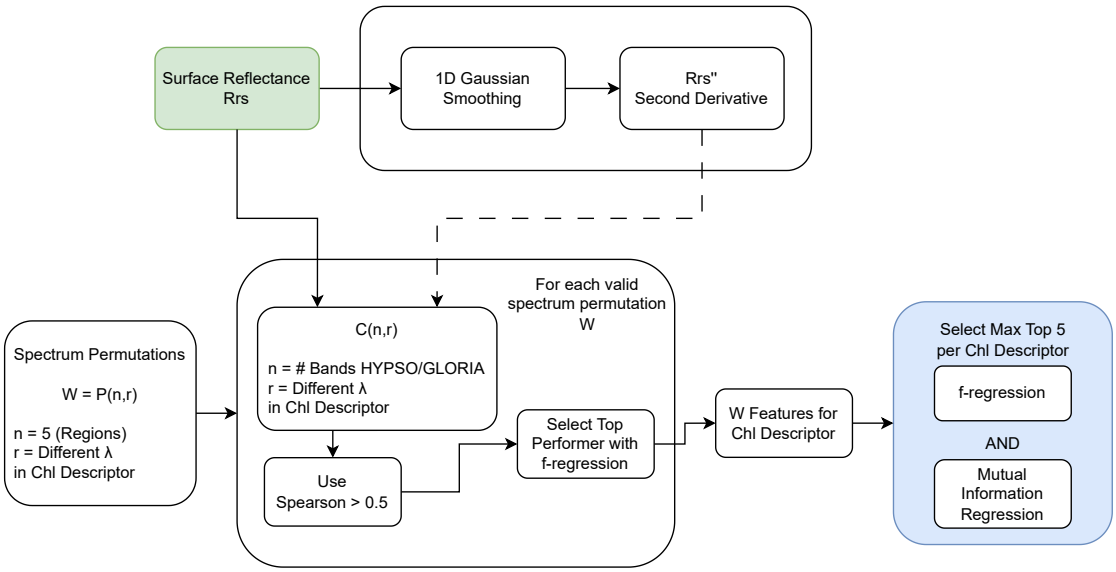


Figure 3.8: *Workflow of the feature creation process that was followed for the training split of the HYPSO and GLORIA dataset independently.*

Lubac et al. (2008) found that the chlorophyll variation can be studied by the maximum variation in the second derivative of the reflectance, being particularly sensible in the blue and green regions of the spectrum. Taking this into account, the second derivative of R_{rs} used in this work was obtained by the method of central difference on the interior points. On the edges, a one-sided difference is applied to not reduce the size of each set. Before calculating the derivative, a normalized Gaussian filter was convoluted on every reflectance. The Gaussian filter was based on the sigma value previously defined in Equation 3.5 of $\sigma = 1.4013nm$. The Gaussian 1D filter is defined by Equation 3.24 where d_x is given by Equation 3.25 where z is the defined size of the filter (5 chosen for this case) . The normalization of this filter is given as $|w| = w/sum(w)$ which was used to convolute each reflectance before the second derivative calculation.

$$w = \exp\left(-\frac{1}{2}\frac{d_x^2}{\sigma^2}\right) \tag{3.24}$$

$$d_x = \{r : r = -3\sigma + n \cdot (3\sigma - \min 3\sigma)/(z-1),\ n \in \{0, 1, ...4\}\} \tag{3.25}$$

The second order derivative was included in this work as it has been shown to be less sensitive to sunlight and skylight contributions, making it easier to detect small variations and to find patterns through the surface reflectance (Tsai and Philpot, 1996, as cited in Lubac et al., 2008).

To be able to estimate chlorophyll, it is imperative to have sufficient and high-quality descriptors derived from the surface reflectance. Using R_{rs} and the second derivative $\frac{d^2 R_{rs}}{d\lambda}$, the chlorophyll descriptors in Table 3.5 were selected based on their importance in the literature. Each one of these was fine-tuned, which, in the scope of this work, refers to testing all permutations so that the correlation with $log(chl)$ increases. Due to the hyperspectral nature of HYPSO, there is more freedom in selecting which band to use for different indices; therefore, this approach was taken.

The relevant indices chosen for this study are the following (see Table 3.5 for the mathematical formulation):

- Two-Band Vegetation Index (TBVI): This index (also known as the normalized difference vegetation index NDVI) in 3.5-A has been used in different works, with many combinations used to achieve different degrees of precision (Adam et al., 2014; Marshall and Thenkabail, 2015). In the work of Pérez et al. (2000) the wavelengths for λ_1 and λ_2 were most likely 550nm and 600nm for the green and red bands, respectively, since consumer grade RGB cameras were used. This index is usually combined with a NIR band but is mostly replaced by a far-red one. In the case of Wang et al. (2019), λ_1 was established in the NIR range from 848 to 881 nm while λ_2 was fixed in the red region from 646 to 684 nm. In general, the ranges used with this index change relative to the sensor used.

- Three-wavelength model: The equation shown in Table 3.5-B proposed by Gitelson et al. (2008) will be used, as it has had good results when combined in multivariate approaches (Matthews, 2011). Different MSI sensors have been used for this method, but a hyperspectral approach is possible.

- OCVI: The canopy-optimized OCVI index by Vincini et al. (2008) increases the relationship between the red and green bands with an empirical parameter c that has been found to have values between 0.64 and 1.31. In the original work, λ_1 was placed in the NIR range, while λ_2 was located in green and λ_3 in the red region.

Indices that go beyond the spectral range of the sensors in this work were not considered, such as the Green-Brown Vegetation Index GBVI, which reaches 2,000 nm (Cui and Kerekes, 2018).

Table 3.5: *Chlorophyll descriptors from the literature used to build the dataset. The subindex on each parameter shows the wavelength in nm used based on the original methods. Due to hyperspectral information from HYPSO-1 and the GLORIA dataset, the closest band in wavelength value was used for the marked sensor without interpolation.*

ID	Descriptor	Source
A	$TBVI = \frac{(\lambda_1 - \lambda_2)}{\lambda_1 + \lambda_2}$	(Rouse et al., 1974)
B	$TBM = \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)\lambda_3$	(Gitelson et al., 2008)
C	$OCVI = \frac{\lambda_1}{\lambda_2}\left(\frac{\lambda_3}{\lambda_2}\right)^c$	(Vincini et al., 2008)

Chlorophyll descriptor approaches will be considered in their simplest form, as shown in Table 3.6 where the band difference and the ratio are used to emphasize reflectance features.

Table 3.6: *Fundamental features computed as chlorophyll descriptors.*

$BD = (\lambda_1 - \lambda_2)$	$LBD = log\,(\lambda_1 - \lambda_2)$
$BR = \left(\frac{\lambda_1}{\lambda_2}\right)$	$LBR = log\left(\frac{\lambda_1}{\lambda_2}\right)$

Each descriptor in Tables 3.5 and 3.6 requires r different wavelengths λ identified by their unique subindex. For example, in the case of the TBM index in Table 3.5-B, $\lambda_1 \neq \lambda_2 \neq \lambda_3$ making $r = 3$. All valid spectrum subgroups are obtained with ${}_nP_r$, where $n = 5$ for the spectrum regions (see Table 3.4) and r is the number of bands on a given descriptor. In Table 3.7 the valid subgroups $w = 20$ for when $r = 2$ are observed, showing that the number of valid spectrum combinations can then be established as 20, 60, 120 for 2, 3 and 4 different numbers of bands in a descriptor.

Table 3.7: *Valid spectrum combinations based on the 5 band ranges from Table 3.4. For a descriptor that uses two distinct* λ

blue-farred	green-blue	red-blue	farred-blue	nir-blue
blue-green	green-farred	red-farred	farred-green	nir-farred
blue-nir	green-nir	red-green	farred-nir	nir-green
blue-red	green-red	red-nir	farred-red	nir-red

For each descriptor, the permutations of all R_{rs} bands were tested and separated into its corresponding subgroup. A permutation for band ratio (BR) using 535-643 would fall in the "green-red" subgroup, while the same descriptor with bands 643-535 would be part of the "red-green" one. An invalid permutation which would not be part of any group could be one with bands 643-643 as both fall in the red region and "red-red" is not valid. For a λ selection to be valid, each must belong to a different region of the spectrum.

For each valid spectrum combination of each descriptor, the Spearman correlation coefficient between the descriptor with the valid wavelengths and $log(chl)$ was calculated. Only those tested cases with coefficients greater than 0.5 were kept. Of all the selected descriptors in each subgroup, the one with the highest f-regression score was selected as a means to use F statistics to select the best combination. A total of W fine-tuned combinations are selected.

To reduce the colinearity of the combinations, a final refinement is implemented to choose a maximum of five features from the fine-tuned W per descriptor. Mutual Information Regression and F-Regression were implemented, and only the features selected by both were designated as the optimal combinations for that descriptor. A maximum of five combinations per chlorophyll descriptor can be selected with this process that was repeated independently for HYPSO, HYPSO", GLORIA, and GLORIA", where the double quote indicates the second derivative.

To the extent of the research done, polynomial features have not been used in combination with chlorophyll descriptors as a means to account for nonlinearity and improve the performance of a model. The number of new features can be described as a combination with replacement ${}^{n}C_{r} = (n + r - 1)!/(r!(n - 1))!$, where n is the total number of fine-tuned features and r is the degree of selected polynomial. A self-imposed limit of second-degree polynomial features has been defined since, as stated by Sohil et al. (2022), high degrees create overly complicated shapes at the boundary of the variables.

Standardization was performed for each of the created features such that the median and standard deviation were subtracted from the values using the equation 3.26. Parameters μ and σ were obtained from the training split to avoid data leakage. According to Zheng and Casari (2018), this occurs when the information

from the test split reaches the training stage. This can be achieved indirectly by applying normalization or any other data-preparation technique with statistical features from the entire dataset to each of the individual splits.

To improve the performance of models that rely on the weight and the n-dimensional distance, input parameters have to be scaled. Input variables were standardized, as not a single one follows a normal distribution. Normalization between 0 and 1 was not selected as the band ratios may change depending on the sensor SRF, thus making the minimum and maximum values potentially capable of fluctuating more.

$$x' = \frac{x - \mu}{\sigma} \tag{3.26}$$

3.7 Chlorophyll Estimation

Chlorophyll regression was performed using fine-tuned descriptors found in the previous section, including those created by polynomial combination. The three approaches used are described in the following subsections.

3.7.1 Multivariate Linear Regression

Multivariate linear regression was implemented to study the relationship of all the features created and their impact on chlorophyll prediction. In favor of explainability, only optimized combinations and polynomial characteristics of Table 3.6 were used for this type of regression due to the simple nature of the descriptors.

The desired number of selected features for this linear regression was set at four. With the Lasso algorithm, the optimal features were evaluated to optimize a regularized L1 multivariate linear regression (Muthukrishnan and Rohini, 2016). Further elimination or features occurred as this method can define the coefficient of specific features as 0. With the remaining features, the "ElasticNet" method, which uses L1 and L2 as regularizers, is computed to obtain the coefficients a for the features x such that Equation 3.27 is satisfied for $n = 4$.

$$log(chl) = a_0 + \sum_{i=0}^{n} a_i x_i \tag{3.27}$$

3.7.2 OCx Polynomial

Without further modification, the MBR algorithm of O'Reilly and Werdell (2019) for the HICO sensor in Equation 3.28 was used with both the HYPSO-1 and the

GLORIA datasets, as it shares a similar spectral resolution. The coefficients for the 4th degree polynomial are $[0.26869, 0.96178, -3.43787, 2.80047, -1.59267]$.

$$MBR_{OC6} = max\left[\frac{(416, 444, 490, 513)}{mean(553, 668)}\right] \tag{3.28}$$

3.7.3 Ensemble Machine Learning

To avoid relying on a single estimator, the prediction was combined by regression of different independent methods. This ensemble technique allowed for refinement estimation based on the best performance of different approaches.

The ensemble techniques used with their respective regressors are the following:

1. Weighted Average Voting: For each regressor, an MAE is calculated to rank them in terms of performance, allowing their individual predictions to be weighted according to their score. After ordering them, their individual prediction can be weighted against their score. If the output is independent, the voting mechanism can work as an optimal combination model (Zhang and Ma, 2012).
    - K-Neighbors Regressor
    - Decision Tree Regressor
    - Histogram Gradient Regressor
    - XGBoost Regressor
2. Stacking: A regressor is used to calculate chlorophyll concentrations, learning from the estimations of multiple different regressors that are used as input features (Witten and Witten, 2017).
    - K-Neighbors Regressor
    - Decision Tree Regressor
    - Histogram Gradient Regressor
    - Final Estimator: Gradient Boosting Regressor
3. Extreme Trees: Random samples are chosen when spawning new internal trees, which makes it different from the random forest method as it does not follow a greedy algorithm (which causes fewer trees correlated) (Geurts et al., 2006). 30 decision trees with a depth from 1 to 30 where the prediction of all is averaged.
4. Gradient Boost: 100 estimators for a combined gradient prediction

5. Bagging:
    - K-Neighbors Regressor
    - Decision Tree Regressor
    - Histogram Gradient Regressor
    - XGBoost Regressor
6. Data Transformation Voting: Data is transformed with linear and non-linear approaches to the equivalent chlorophyll value in the training set. A voting approach is selected to merge the different mappings used.
    - Min-Max Scaler
    - Standard Scaler
    - Robust Scaler
    - Power Transformer
    - Quantile Transformer (100 quantiles)
    - K-bins Discretizer (20 bins)

To avoid data leakage on all the methods mentioned above, during training cross-validation, normalization was performed for every smaller split of the k-fold subset independently of the large training set.

3.8 Evaluation

3.8.1 Atmospheric Correction

To compare the atmospheric correction surface reflectance R_{rs} of the method implemented in this work to that of the ACOLITE implementation, the root mean square difference (RMSD) and the average unbiased absolute relative difference ϵ were used as suggested by Li et al. (2022). Equations 3.8.1 and 3.8.1 corresponding to the symmetric signed percentage bias (β) and the median symmetric accuracy (MdSA), respectively, were also considered as proposed by Pahlevan et al. (2021) due to their resistance to outliers when used as percentages.

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R_{rrs} - \hat{R}_{rrs}\right)^2} \tag{3.29}$$

$$\epsilon = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{|R_{rrs} - \hat{R}_{rrs}|}{R_{rrs} + \hat{R}_{rrs}} \right] \times 200 \tag{3.30}$$

$$SSPB = \beta = 100 \times sign(z) \cdot (10^{|Z|} - 1)[\%]$$
$$\text{where } Z = Median \left(log_{10} \left(\frac{\hat{R}_{rs}(\lambda_i)}{R_{rs}(\lambda_i)} \right) \right)$$

$$MdSA = 100 \times (10^{Y} - 1)[\%]$$
$$\text{where } Y = Median \left| log_{10} \left(\frac{\hat{R}_{rs}(\lambda_i)}{R_{rs}(\lambda_i)} \right) \right|$$

3.8.2 Chlorophyll Regression Evaluation

Common evaluation metrics for regression are the coefficient of determination R^2 and the root mean square error (RMSE). As stated by Seegers et al. (2018), these alternatives are sensitive to outliers and therefore must be interpreted adequately.

O'Shea et al. (2021) considers that for biomass estimation, linear metrics such as room mean squared difference (RMSD) and median absolute percentage difference (MAPD) should be avoided, favoring alternatives such as root mean square logarithmic difference (RMSLD) and mean absolute difference (MAD). Similarly, Pahlevan et al. (2021) cautions against using the mean average percentage error (MAPE) and the root mean square error (RMSE) as the description of logarithmic models may not be reliable. Finally, Seegers et al. (2018) advises us to focus on the mean average error (MAE) and bias as a means to compare chlorophyll estimates between works.

Under these considerations, the following metrics will be used, including R^2 and RMSE for traceability with the rest of the literature.

$$BIAS = 10^{\left(\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \right)} \tag{3.31}$$

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (log(\hat{y}_i + 1) - log(y_i + 1))^2} \tag{3.32}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | \hat{y}_i - y_i | \tag{3.33}$$

where:

- $\hat{y}$ - Predicted value of y
- y_i - Ground truth value of y
- n - number of samples

4 Results

4.1 Pre-Processing

To visually compare the performance of the atmospheric correction of the 6SV1 method (implemented in this work) with the ACOLITE algorithm by Vanhellemont and Ruddick (2021), the reflectance of four random pixels was included in Figure 4.1. This plot shows the similarity of both reconstructions having a slight divergence in the green region of the spectrum.

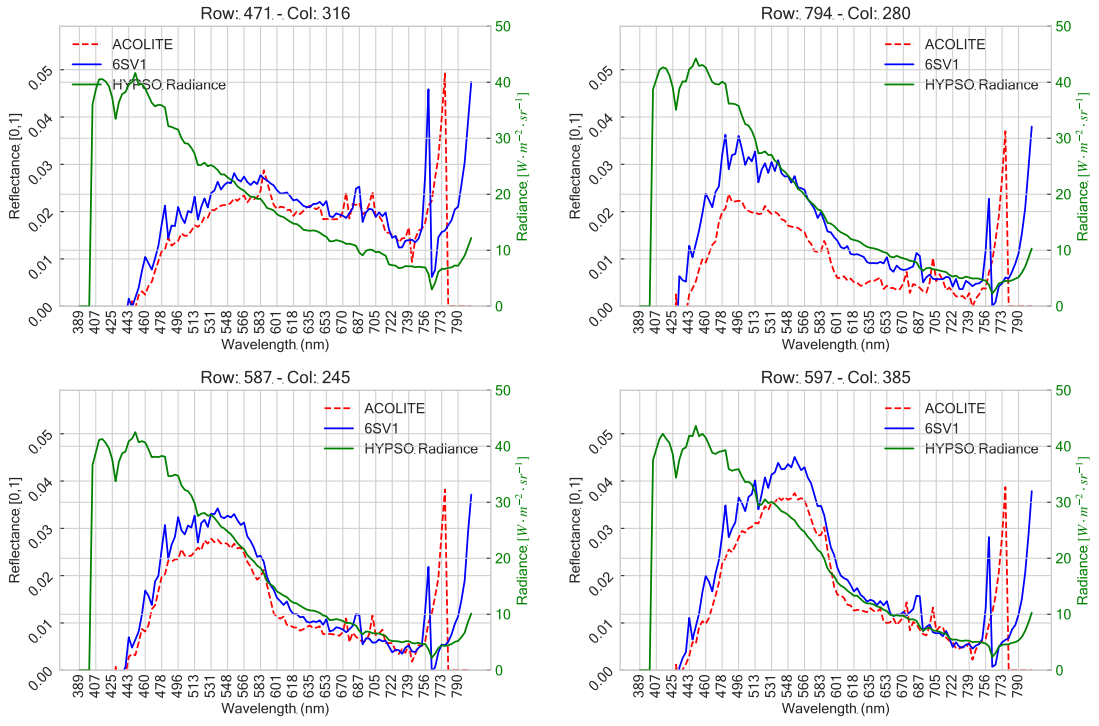


Figure 4.1: *Comparison of R_{rs} recovered from the "florida_2023-01-12_1553Z" HYPSO capture using the 6SV1 algorithm and the ACOLITE method by Vanhellemont and Ruddick (2021).*

The regression plot for the same pixels is plotted in Figure 4.2. The difference per region is visible when comparing the one-to-one equivalence. A high correlation of both estimations can be seen in the blue regions, the same one that decreases towards the NIR area.

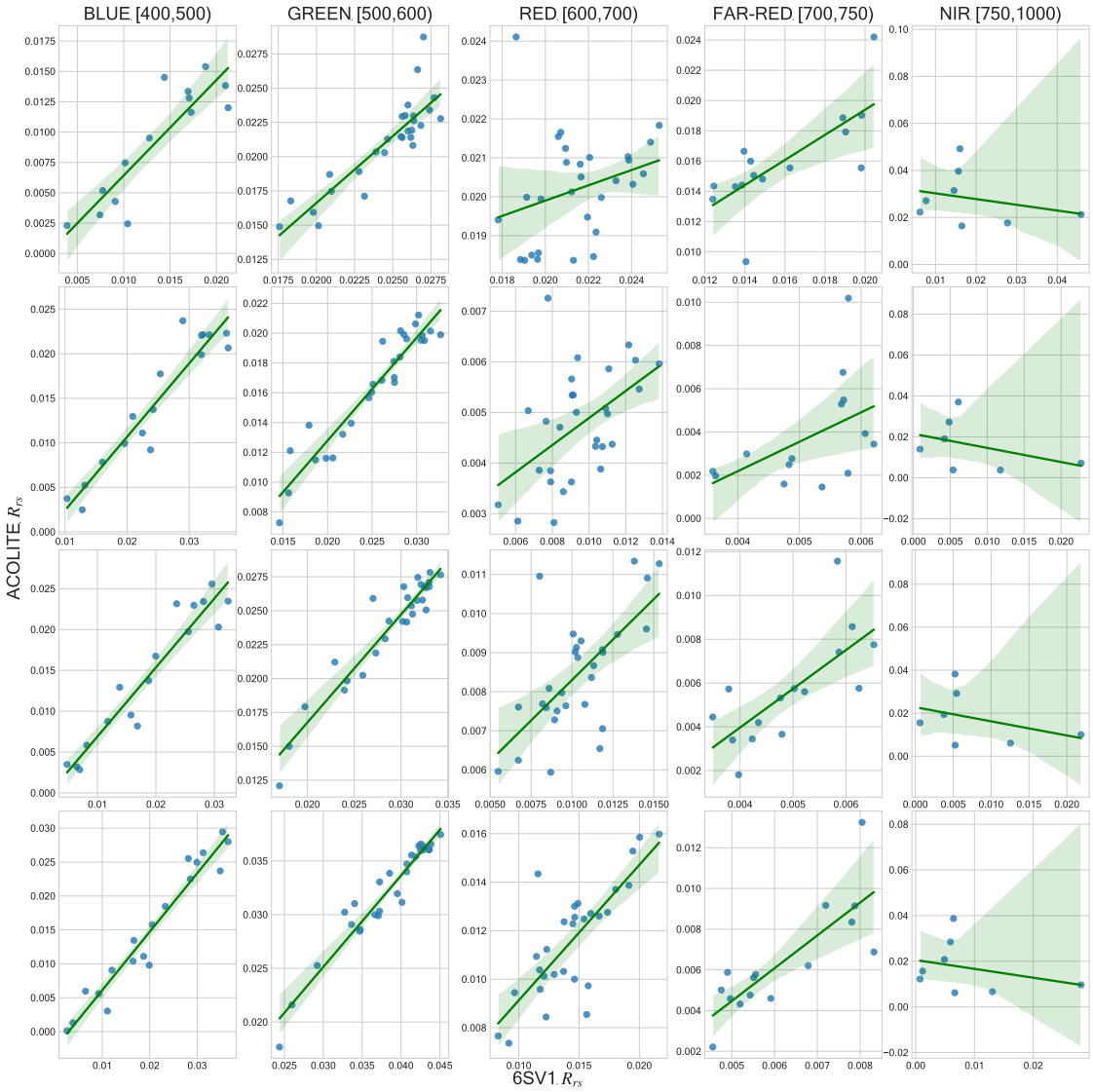


Figure 4.2: *Regression plots per region of the spectrum where each row represents a pixel. The X-axis is for the 6SV1 reflectance and the Y-axis for the ACOLITE reflectance for the same pixel.*

The performance of all R_{rs} in the training subset is shown in Table 4.1. Contrary to different works in which the estimated reflectance is compared with the one measured at the surface, these metrics help to note how different the ACOLITE

and the 6SV1 approach are. An increase in variance σ is evident in the MdSA for the further regions of the spectrum, as well as for the maximum differences from the RMSD. Based on the data for the SSPB (bias), the reflectance of HYPSO seems to be overestimated compared to the ACOLITE method (which is visually shown in Figure 4.1).

Table 4.1: *Results for the spectrum comparison metrics on the points of the entire training split. m stands for the minimum value and M for the maximum value of the set tested.*

		RMSD ↓	ϵ ↓	SSPB	MdSA
Blue	**m**	0.0034	10.7577	9.0952	14.4356
	M	0.0176	101.5534	719.9941	719.9941
	$\bar{X}$	0.0098	56.1188	205.3332	206.9220
	σ	0.0037	17.7446	111.6825	111.2835
Green	**m**	0.0036	7.0548	18.0763	18.0763
	M	0.0148	68.7920	414.7315	414.7315
	$\bar{X}$	0.0088	35.8355	137.8106	137.8106
	σ	0.0029	13.4550	73.3239	73.3239
Red	**m**	0.0022	7.3028	12.3838	17.8375
	M	0.0157	119.9313	2342.9951	2342.9951
	$\bar{X}$	0.0050	55.3348	395.7801	396.8008
	σ	0.0018	27.1878	377.5938	376.6646
Far-RED	**m**	0.0016	10.3459	0.0617	15.6342
	M	0.0241	101.5451	2005.4562	2005.4562
	$\bar{X}$	0.0028	48.0204	269.6985	287.9747
	σ	0.0014	23.9327	342.7256	331.4295
NIR	**m**	0.0156	42.7438	30.8596	113.7177
	M	0.0899	139.6188	98.4187	6223.8089
	$\bar{X}$	0.0180	115.2221	93.7457	2701.2345
	σ	0.0039	14.5483	7.3528	926.7569

Visually, the mean of the positive metrics is shown in Figure 4.3 where the average variations are more obvious.

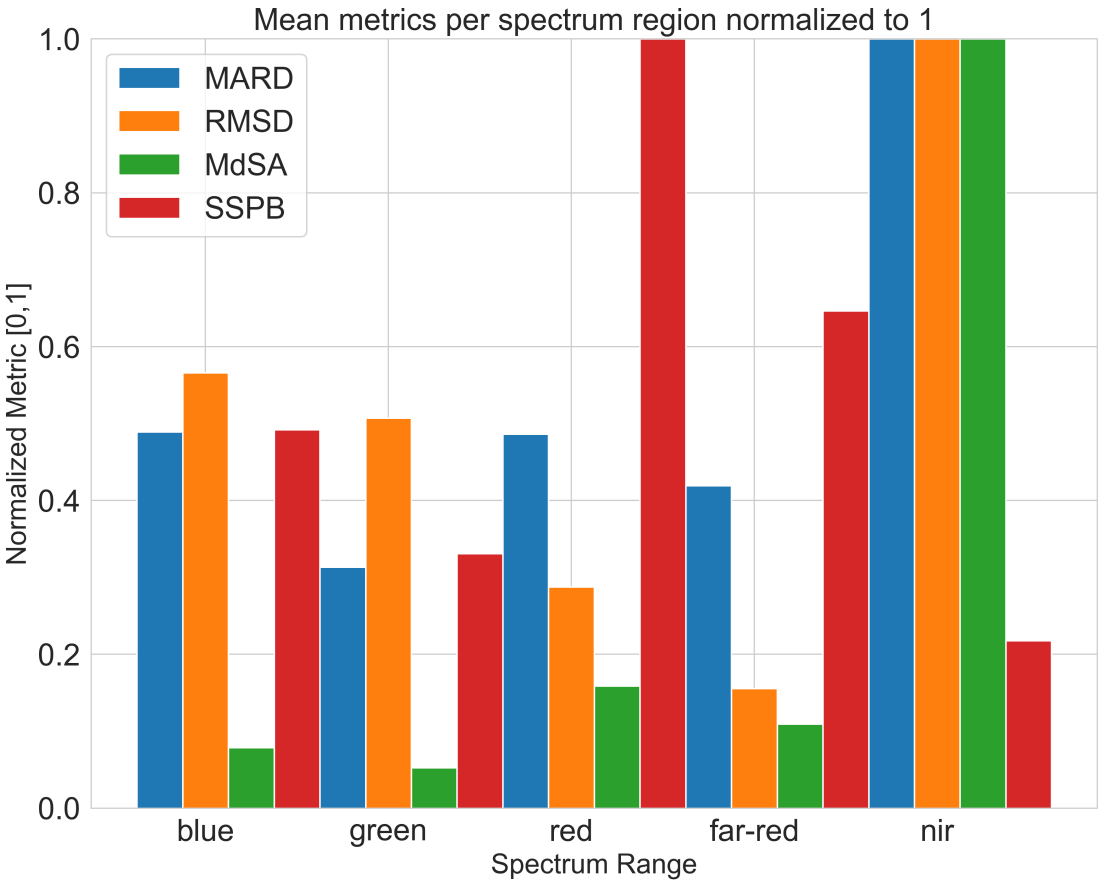


Figure 4.3: *Mean for each metric of Table 4.1 per spectrum range.*

To validate the extraction of the water mask, the OSM Water Layer, a high resolution surface mapping by Yamazaki et al. (2017) was used in its latest release in July 2021. The surface of the Earth is divided into 2,160 GeoTiff files of $\approx 1GB$ in size. The implementation of Cordeiro et al. (2021) to detect the mask is about 2,000 times lighter by not relying on lookup tables. In satellite processing, the size difference can be a very important factor depending on where the process is implemented. Telemetry data must be correct for use of mappings such as the OSM Water Layer. Every spectral image of this work was verified and, if needed, manually registered, making it possible to use the lookup tables for verification.

An accuracy of 96.6% was obtained from the matching process, while the precision, recall, and F1 score were found to be 98.3%, 91.3%, and 94.7%, respectively. The confusion matrix in Figure 4.4 shows the classification results versus what is considered in this work the ground truth.

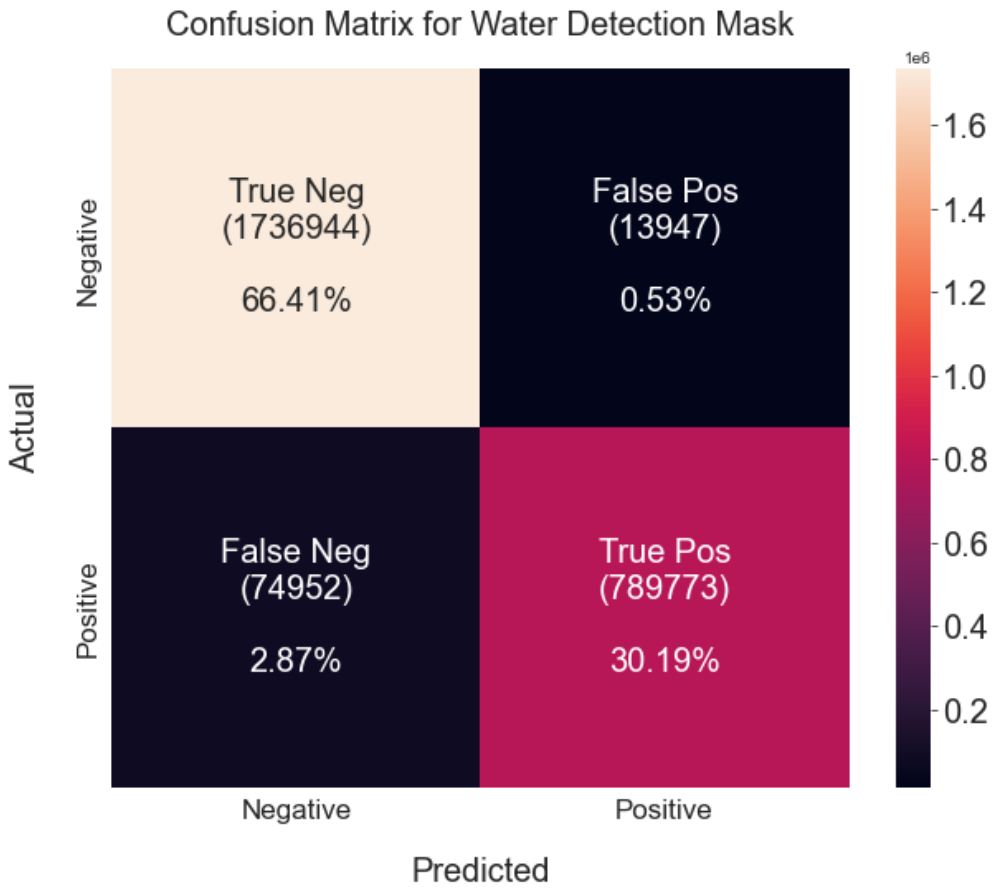


Figure 4.4: *Confusion matrix for pixel classification predicted using the methodology by Cordeiro et al. (2021) and using the high-resolution surface mapping by Yamazaki et al. (2017) as the ground truth.*

After estimating the water masks and matching with the ESA and NASA satellites, a total of 1,362,318 water pixels were found to build the HYPSO-1 dataset, which is small considering that per image there are $956 \times 684 = 653,904$ pixels. For the points of GLORIA 4,145 measurements were obtained after trimming the spectral range to match that of HYPSO and by discarding all inputs with missing values. This process was implemented to have the same spectral coverage for both devices and to be able to use the same indices. The geographical locations of the HYPSO-1 images used are shown in Figure 4.5, while in Figure 4.6 is the location of the GLORIA *in situ* measurements after selecting the usable ones. In general, there is agreement on the regions covered by both datasets except for South Argentina, Japan, Philippines, India, and the Middle East.

Figure 4.5: *Location of the HYPSO-1 points used based on the matching conditions established previously.*

Figure 4.6: *Location of the GLORIA dataset points.*

4.2 Feature Creation

Following the workflow described previously in Figure 3.8, five of the most important fine-tuned chlorophyll descriptors were extracted when selected by mutual agreement of the f-regression and the mutual information regression method. The results for each of the descriptors used can be found in Tables 4.2 to 4.7 where the bands λ_1, λ_2, and λ_3 were selected according to the form of each descriptor shown in Tables

3.5 and 3.6, using the ranges from Table 3.4 so that not two wavelengths would be in the same range.

Table 4.2: *TBVI optimal band selection results.*

	λ_1	λ_2	Combination	Spearman	Score f-reg	Score Mutual Reg.
HYPSO	701	460	farred-blue	0.708	218425.619	0.744
	569	489	green-blue	0.745	297164.534	0.920
HYPSO"	715	478	farred-blue	0.658	167441.840	0.462
	687	481	red-blue	0.673	163633.485	0.483
GLORIA	699	496	red-blue	0.794	2415.769	0.710
	699	599	red-green	0.801	2407.804	0.707
GLORIA"	798	575	nir-green	0.617	216.833	0.324

Table 4.3: *TBM optimal band selection results.*

	λ_1	λ_2	λ_3	Combination	Spearman	f-reg	Score Mutual Reg.
HYPSO	503	601	801	green-red-nir	0.728	3860.992	0.897
HYPSO"	481	687	804	blue-red-nir	0.626	2100.371	0.418
GLORIA	599	698	701	green-red-farred	0.773	755.869	0.601
	497	533	602	blue-green-red	0.775	409.995	0.605
	497	698	704	blue-red-farred	0.782	267.685	0.664
GLORIA"	599	698	701	green-red-farred	0.773	755.869	0.601
	497	698	704	blue-red-farred	0.782	267.685	0.664
	497	542	698	blue-green-red	0.814	253.047	0.669

Table 4.4: *OCVI optimal band selection results.*

	λ_1	λ_2	λ_3	Combination	Spearman	f-reg	Mutual Reg.
HYPSO	499	503	701	blue-green-farred	0.676	1898.946	0.756
	499	503	615	blue-green-red	0.723	2713.703	0.850
HYPSO"	-	-	-	-	-	-	-
GLORIA	482	503	761	blue-green-nir	0.715	453.979	0.519
	464	500	698	blue-green-red	0.806	1602.926	0.649
	461	500	701	blue-green-farred	0.814	1573.953	0.673
	599	605	701	green-red-farred	0.822	1217.736	0.613
GLORIA"	-	-	-	-	-	-	-

Table 4.5: *Band-Ratio optimal band selection results.*

	λ_1	λ_2	Combination	Spearman	f-reg	Mutual Reg.
HYPSO	701	503	farred-green	0.703	125451.700	0.686
	555	489	green-blue	0.735	262587.155	0.855
HYPSO"	517	684	green-red	0.670	181451.307	0.439
GLORIA	516	499	green-blue	0.769	1947.723	0.560
	700	591	farred-green	0.791	1264.605	0.666
GLORIA"	-	-	-	-	-	-

Table 4.6: *log(Band-Ratio) optimal band selection results.*

	λ_1	λ_2	Combination	Spearman	f-reg	Mutual Reg.
HYPSO	701	478	farred-blue	0.720	197796.251	0.811
	569	489	green-blue	0.745	239131.294	0.926
HYPSO"	-	-	-	-	-	-
GLORIA	699	500	red-green	0.791	1653.306	0.645
	709	497	farred-blue	0.802	1826.660	0.687
	700	670	farred-red	0.902	2894.991	0.917
GLORIA"	698	750	red-nir	0.517	51.760	0.175
	500	600	green-red	0.574	58.115	0.186

Table 4.7: *Band Difference optimal band selection results.*

	λ_1	λ_2	Combination	Spearman	f-reg	Mutual Reg.
HYPSO	615	766	red-nir	0.617	99225.650	0.716
	517	489	green-blue	0.729	245888.184	0.819
HYPSO"	506	687	green-red	0.640	121290.534	0.501
	506	715	green-farred	0.664	125275.108	0.508
	722	517	farred-green	0.685	149059.233	0.480
GLORIA	705	489	farred-blue	0.704	502.499	0.466
GLORIA"	669	705	red-farred	0.746	743.151	0.478
	504	701	green-farred	0.746	660.551	0.540

4.3 Chlorophyll Estimation

4.3.1 Multivariate Linear Regression

With the optimal features in Tables 4.2 to 4.7, the linear regression was obtained with a maximum of four terms. The proposed selection method for optimal features failed to find optimal combinations based on the second derivative of the reflectance, therefore they were ignored for the rest of the work. The selection of four parameters through the Lasso model discarded some features, thus the blank spaces in Table 4.8, where the best results from the multivariate regressor "ElasticNet" are shown.

Table 4.8: *Linear regression results from the best features generated out of the fine tuning process (best results marked in gray and with an "*"). Coefficients correspond to Equation 3.27. G stands for GLORIA and H for HYPSO.*

ID	Data	X_1 $[a_1]$	X_2 $[a_2]$	X_3 $[a_3]$	X_4 $[a_4]$	a_0	BIAS	RMSLE ↓	MAE ↓	RMSE ↓	R^2
A	G	TBVI(699,496) [0.7003]	TBM(497,533,602) [−0.3417]	LBR(700,670) [0.8462]	BR(656,562) [−0.3158]	2.296	0.531	2.439	0.978	1.209	0.517
BR		BR(516,499) [0.961]	BR(700,591) [0.545]	-	-	2.296	34.798	4.366	1.702	2.127	-0.492
BR		BR(443,562) [−1.399]	BR(450,670) [1.575]	-	-	2.296	0.992	4.879	1.587	2.379	-0.866
SO*		BR(700,591) [−0.275]	LBR(699,500) [0.547]	LBR(700,670) [1.232]	-	2.296	0.492	2.092	0.780	1.004	0.667
CO		TBVI(699,496) [0.6562]	TBVI(699,599) [0.6062]	-	-	2.296	0.650	2.8289	1.132	1.402	0.351
PSO		LBR(700,670) [1.4738]	LBR(699,500) * LBR(700,670) [1.0365]	LBR(700,670) [2] [−0.7933]	-	2.296	1.272	2.303	0.931	1.197	0.527
A	H	TBVI(569,489) [1.2136]	BD(517,489) [−0.2907]	-	-	0.576	1.055	1.363	0.723	0.809	0.348
SO*		BR(701,503) [−0.2703]	LBR(701,478) [0.2912]	LBR(569,489) [1.2498]	BD(517,489) [−0.2586]	0.576	0.850	1.120	0.536	0.616	0.622
BRO		BR(555,489) [1.0941]	-	-	-	0.576	14.263	2.280	1.191	1.353	-0.821
LBRO		LBR(701,478) [−0.2887]	LBR(569,489) [1.2032]	-	-	0.576	2.530	1.190	0.673	0.832	0.310
BDO		BD(615,766) [−0.3959]	BD(517,489) [1.1819]	-	-	0.576	0.976	1.707	0.903	0.999	0.006
CO		TBVI(701,460) [−0.2505]	TBVI(569,489) [1.3515]	-	-	0.576	1.315	1.403	0.771	0.866	0.253
PSO		BR(555,489) [0.5261]	LBR(569,489) [0.3758]	LBR(701,478)* LBR(569,489) [−0.7022]	BR(701,503)* BR(555,489)* LBR(701,478) [0.4001]	0.576	1.542	1.061	0.599	0.695	0.519

The best linear regression for HYPSO (marked with an "*" in Table 4.8), based on a lower error and bias closer to 1.0, is shown in Figure 4.7 to demonstrate the prediction results in an actual image. The model for GLORIA cannot be used as the fine-tuned features and regressions are specific for each sensor, and thus for each dataset.

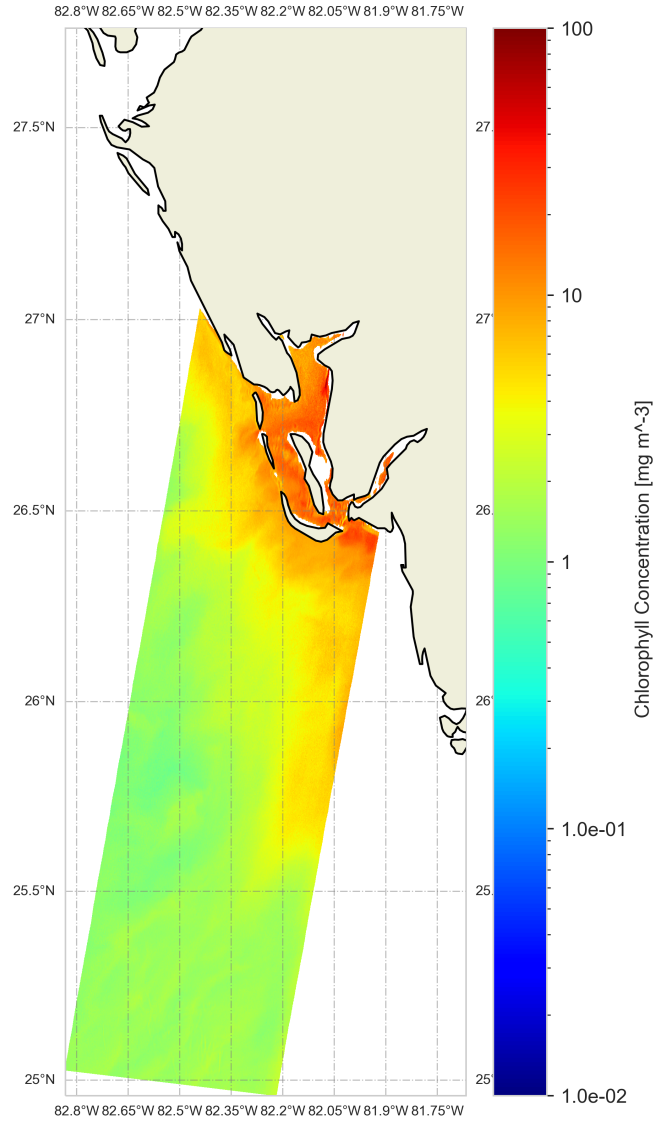
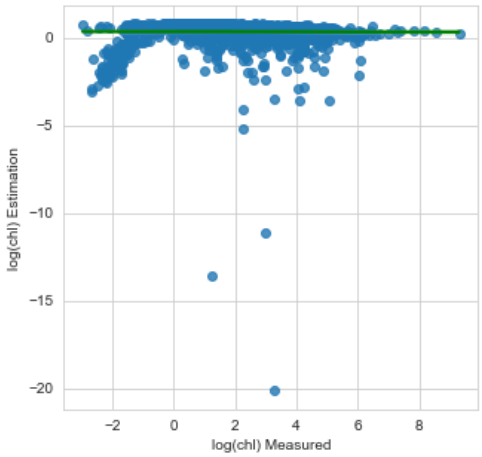


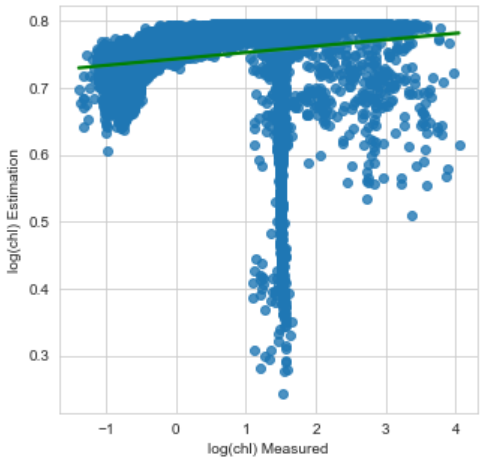
Figure 4.7: *Best LR Plot for HYPSO using log(BR), BR and BD.*

4.3.2 OCx MBR Algorithm

The implementation of the OCx algorithm on the data did not return the expected results. The regression plots in Figure 4.8 and the corresponding metrics in Table 4.9 show that the model does not translate well to other spectral sensors with similar resolution. The data distribution seemed similar between the HYPSO and GLORIA plots, which could be generated by the 4th degree polynomial behavior applied incorrectly to the data used.



(a) *GLORIA*



(b) *HYPSO*

Figure 4.8: *Regression plot of implementing the HICO OC6 MBR algorithm on the GLORIA and HYPSO dataset.*

Table 4.9: *HICO MBR 4th degree polynomial regression results on the HYPSO (H) and GLORIA (G) datasets.*

Data	BIAS	RMSLE ↓	MAE ↓	RMSE ↓	R^2
G	0.015	5.464	2.116	2.689	-1.255
H	1.560	1.618	0.896	1.000	-0.022

Due to visually inconsistent results in the regression plot, the fourth-degree polynomial calculation as described by O'Reilly and Werdell (2019) was estimated using the data from both datasets. The result is shown in Figures 4.9 and 4.10

using the same MBR bands from Equation 3.28 but defining the coefficients for our specific data.

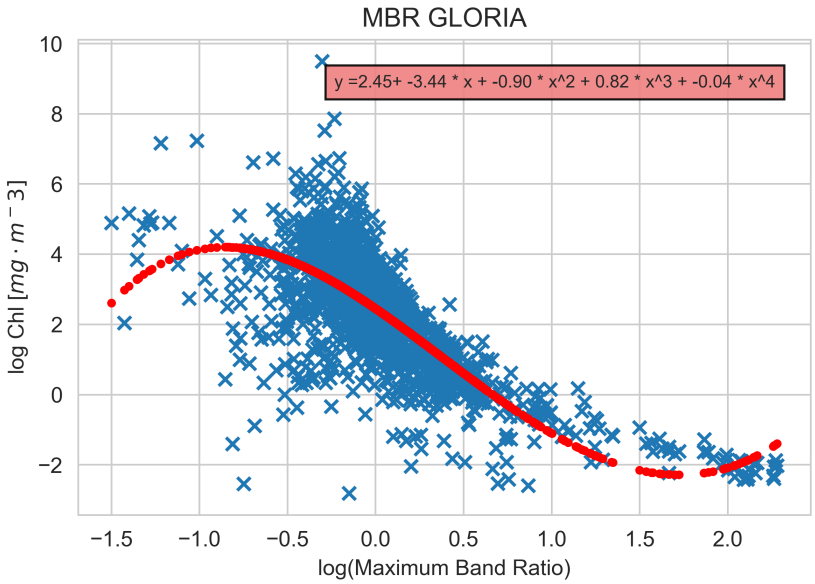


Figure 4.9: *MBR 4th Polynomial calculated for the GLORIA dataset.*

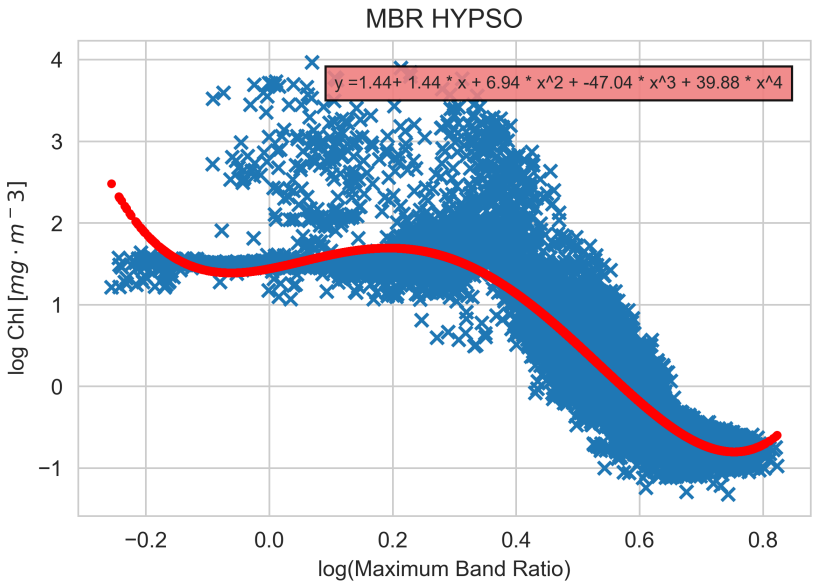


Figure 4.10: *MBR 4th Polynomial calculated for the HYPSO dataset.*

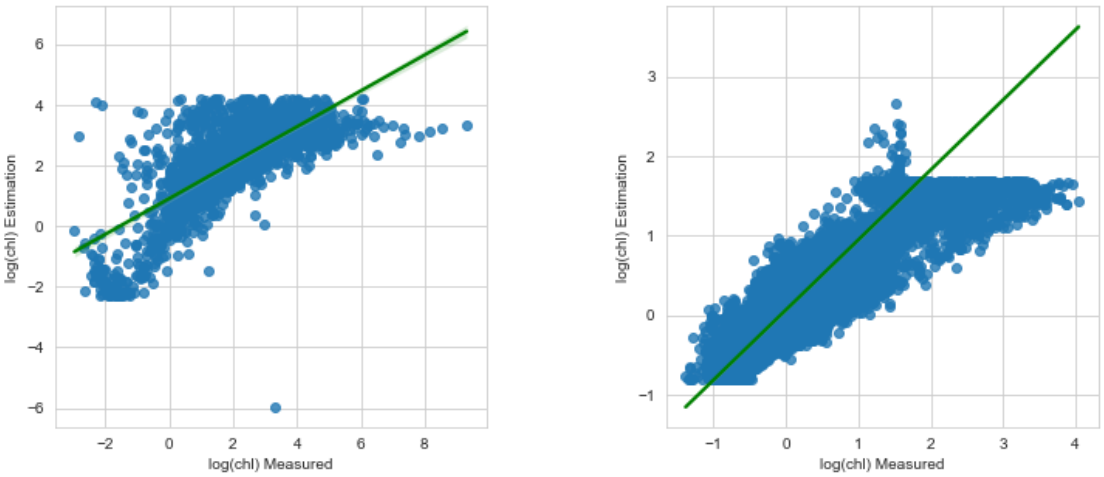
The metrics to evaluate this regression are included in Table 4.10, including

the coefficients for the custom OCX polynomial. When comparing the results of Table 4.9 and Table 4.10, it can be seen that the custom approach improves bias and reduces error metrics.

Table 4.10: *MBR 4th degree polynomial regression results for HYPSO (H) and GLORIA (G). The coefficients are included in the table where subindex is the feature exponent.*

Data	a0	a1	a2	a3	a4	BIAS	RMSLE	MAE	RMSE	R^2
G	2.44939	-3.44061	-0.89645	0.82480	-0.04171	1.060	2.496	0.870	1.187	0.560
H	1.44003	1.44035	6.93572	-47.03517	39.88372	0.999	1.054	0.420	0.335	0.884

The equivalent regression plots are shown in Figure 4.11. Both datasets show a hard estimation limit when predicting values using the polynomial, which could be improved by a curated selection of points, as done by O'Reilly and Werdell (2019).



(a) *GLORIA* **(b)** *HYPSO*

Figure 4.11: *Regression plot of estimation vs. ground truth after implementing the custom MBR 4th degree polynomial for prediction, giving the results of Table 4.10.*

The custom HYPSO MBR polynomial with the coefficients of Table 4.10 was used to predict chlorophyll on a sample spectral image, as shown in Figure 4.12.

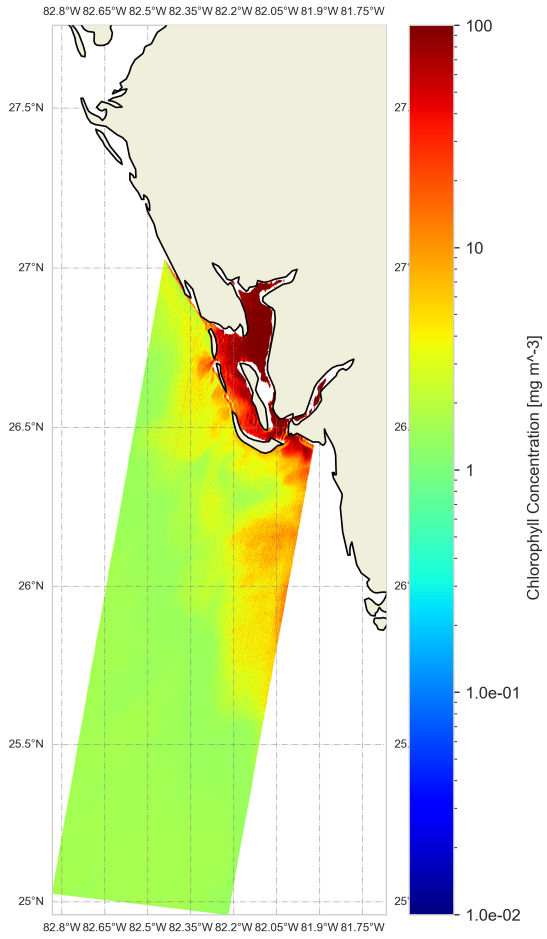


Figure 4.12: *Chlorophyll prediction on a HYPSO spectral image using the MBR 4th degree polynomial with the coefficients of Table 4.10 and the bands of Equation 3.28.*

4.3.3 Ensemble Machine Learning

To improve the results obtained so far, the ensemble machine learning approach was used considering a 10 K-Fold validation and presenting the mean results in this section. Only the features related to the band ratio, log band ratio, band difference, and TBVI index were used for the ensemble model, as based on empirical testing they were the best performing features. Table 4.11 shows the results where, based on a lower error (MAE and RMSLE) and a bias closer to 1.0, it was determined that the voting strategy outperformed the rest (see row marked with "*").

Table 4.11: *Ensemble machine learning results both datasets. The best result is shown in gray and marked with an "*".*

Data	Type	BIAS	RMSLE ↓	MAE ↓	RMSE ↓	R^2
H	Voting*	1.302	0.972	0.320	0.568	0.696
	Stacking	1.317	0.975	0.323	0.571	0.692
	n-Trees	1.403	1.018	0.347	0.591	0.670
	G-Boost	1.420	1.060	0.374	0.614	0.645
	Transform	1.141	1.155	0.369	0.638	0.617
	Bagging	1.254	0.997	0.323	0.573	0.690
G	Voting*	1.460	1.510	0.531	0.752	0.785
	Stacking	2.836	1.783	0.701	1.022	0.655
	n-Trees	3.096	1.987	0.744	1.083	0.613
	G-Boost	-	-	-	-	-
	Transform	1.263	1.790	0.585	0.871	0.746
	Bagging	1.566	1.677	0.538	0.817	0.778

As the voting ensemble performed the best based on an overall lower error, the SHAP values (Shapley Additive Explanations) were calculated to study the contribution of each feature to the prediction of a model. The farther a SHAP value from zero is, the greater the contribution of a characteristic to the estimation of chlorophyll. For HYPSO-1, the mean SHAP values are shown in Figure 4.13 while the individual values for the test set are shown in Figure 4.14. For the GLORIA dataset, the equivalent graphs are shown in Figures 4.15 and 4.16, respectively. The inter-feature contribution for the features on HYPSO-1 is not as big as that on the see in the GLORIA SHAP values. While on average the contribution decreases by approximately 0.20 on the HYPSO features, GLORIA sees a higher drop from the top feature of approximately 8.4 points. On the individual plots, it can also be seen that the features selected for the GLORIA dataset are responsible for higher SHAP values when compared to those in the HYPSO results.

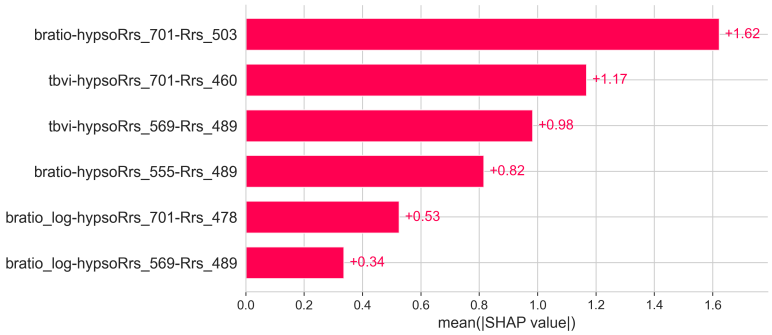


Figure 4.13: *Mean SHAP Values for HYPSO*

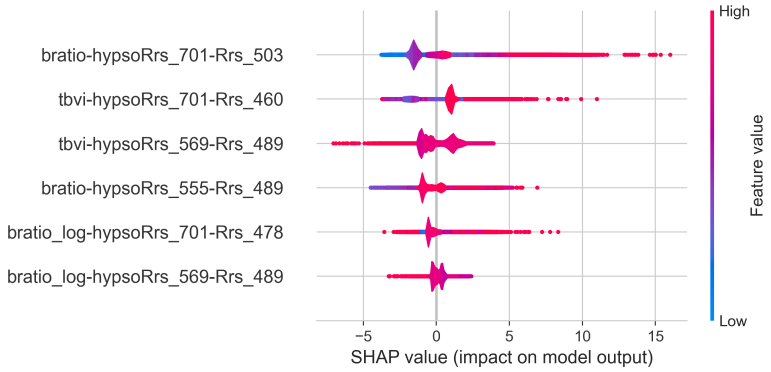


Figure 4.14: *SHAP Values for HYPSO*

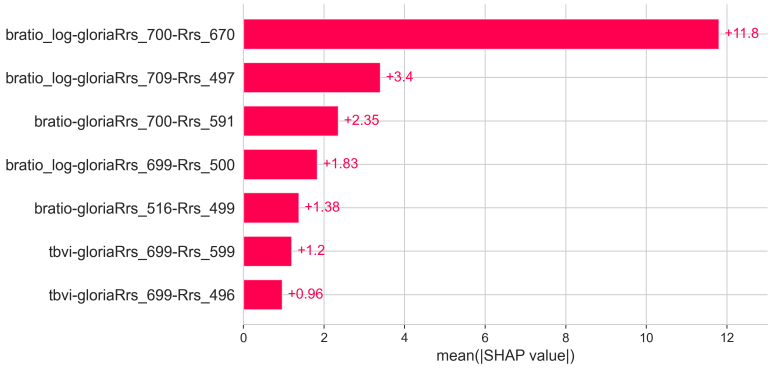


Figure 4.15: *Mean SHAP Values for GLORIA*

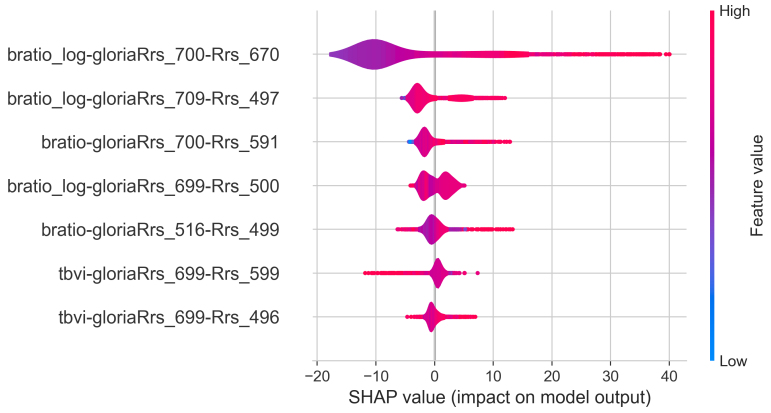


Figure 4.16: *SHAP Values for GLORIA*

Because the best method is the voting ensemble, an estimation of chlorophyll on a HYPSO-1 spectral image was done and shown in Figure 4.17.

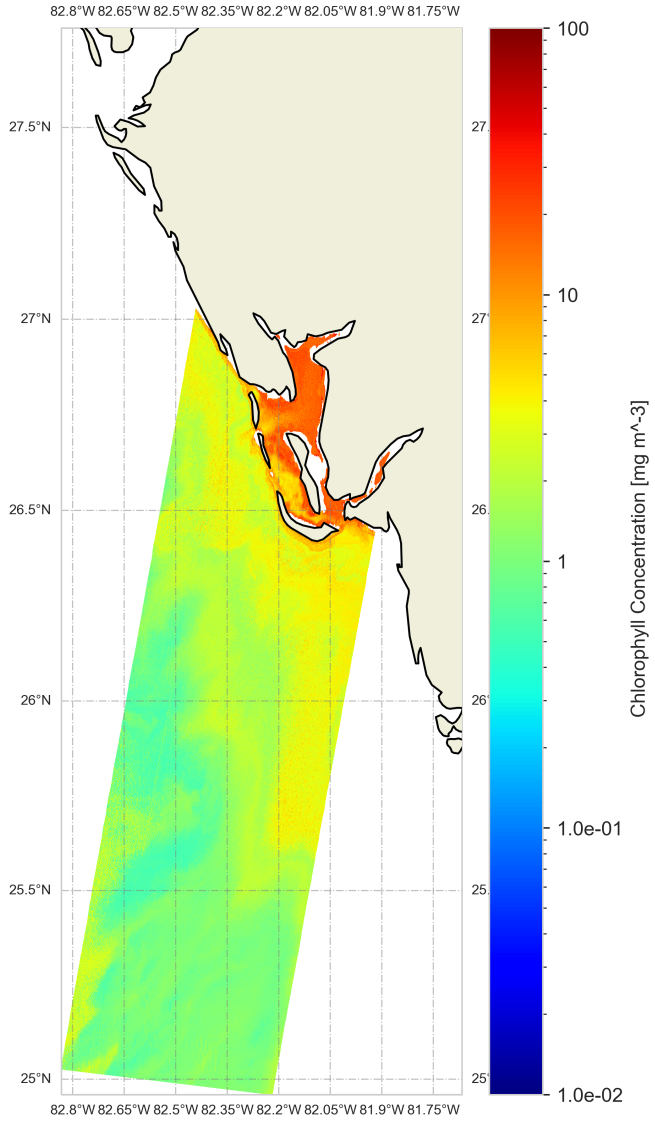


Figure 4.17: *Chlorophyll prediction on a HYPSO spectral image using the voting ensemble method.*

5 Discussion

5.1 Atmospheric Correction

The performance of the atmospheric correction method to obtain surface reflectance R_{rs} was evaluated taking as reference the ACOLITE method, which was developed almost in parallel for HYPSO-1. The average reflectance difference is higher in the far red and NIR regions, as shown previously in Table 4.1. Both methods struggle to estimate the reflectance of $\lambda < 443nm$ as the values become negative, which in this context is not a valid response.

Regarding the atmospheric profile, which was selected based on the criteria of Table 3.2, it should be noted that over the years the relative center of the regions changes (Chen and Chen, 2013). A more accurate and robust approach would be to define the atmospheric profile based on surface measurements such as water vapor, pressure, and temperature in combination with the lookup tables defined by F.X. Kneizys et al. (1996). The approach used in this work was selected due to the current inability of HYPSO-1 to measure any of these properties.

For each spectral image from HYPSO-1 it was assumed that the correction per band was applied equally for all pixels. Although this helps to speed up the correction process, a more accurate approach would require pixel-level information such as the solar azimuth and zenith angles as well as the altitude. Due to information availability, the mean of the previous two parameters was used for every pixel.

At the time of writing this work, the spectral response function has not been measured for the HYPSO-1 hyperspectral camera. To account for this, a Gaussian distribution peak normalized to 1 was used for the 6SV1 atmospheric correction algorithm. This assumption is an educated approach, as the device spectral resolution is high at approximately 4nm; however, the measurement of the SRF is recommendable for future work as the lack of it adds another layer of uncertainty to the correction process.

It must be recalled that the atmospheric correction process is still a problem without a definitive answer, but the approximations through radiative transfer

models are constantly being developed. The quality of the corrections is based on the characteristics of the on-board sensors and the knowledge of the optical path from the satellite to the surface.

Enhancing the spectral range of the satellite can be a good way to approach a better atmospheric correction, although the effect of not having them can be studied by matching HYPSO-1 passes with stations that measure the optical path variables. AERONET stations could serve this purpose in understanding the implications of this process; however, Gordon (2019) established that AERONET aerosol measurements are sparse and are not really applicable for satellites with low spatial resolution (>1km). If they were available in HYPSO-1, the spectral bands in the SWIR region could be used to remove the contribution of water vapor, as previously done for the HYPERION sensor of the EO-1 mission (Goetz et al., 2002). NIR bands in $\approx 788nm$ and $\approx 885nm$ have also been used for atmospheric correction in the MERIS sensor, but due to the lack of the latter in HYPSO-1, a similar approach could not be replicated (Schroeder et al., 2007). The 6SV1 atmospheric correction method used had to be limited to the limited spectral range of HYPSO-1, but it is evident that to achieve further improvements, an extension of this range would have to be evaluated for future satellite missions.

Li et al. (2022) found that when analyzing atmospheric correction against ground truth reflectance, the RMSD and ϵ oscillate in the range of 0.0013 to $0.0049sr^{-1}$ and 8% to 47%, respectively. The metrics in Table 4.1 show that the reflectance signals estimated by the 6SV1 and ACOLITE methods have a smaller RMSD in the red region, while the average unbiased absolute difference ϵ is smaller in the green. The metrics of each region fall outside the empirical ranges found by Li et al. (2022), so further analysis is needed to evaluate both methods against a known ground truth.

Different types of atmospheric environmental conditions can make the estimation process of R_{rs} extremely difficult, as the high number of particles and aerosols may be challenging for current atmospheric correction models (Gokul et al., 2019, as cited in IOCCG, 2021). This poses the need to make corrections with the information available from the onboard sensors to facilitate the study of different surface conditions and water types.

The quality of the atmospheric correction process can have a direct impact on the accuracy of chlorophyll estimation, as the features used to predict chlorophyll are based on the relationship between different bands (see Table 3.5). All conditions that may result in noise or variations in atmospheric correction should be reduced so that their influence is minimized in the estimation of surface reflectance R_{rs}, resulting in more accurate chlorophyll estimations.

5.2 Water Mask Pixels Classification

High-resolution surface maps, such as the one of Yamazaki et al. (2017) are excellent options for classifying pixels on satellite images, as different water bodies are already identified based on their coordinates. To uniquely rely on this type of surface mapping, HYPSO-1 would need to have accurate coordinates; which was not the case for most of the used images. The implementation of a method that could classify water pixels based on reflectance was necessary to overcome inconsistent telemetry data.

Regardless of the quality of the atmospheric correction, the clustering technique of the method by Cordeiro et al. (2021) was shown to be robust enough to achieve a precision of 98.3% when classifying water pixels of a single type. Different water types (i.e. lakes and oceans) on the same image could cause higher classification errors, but it was not tested in this work as the ocean water type was the main focus. For spectral images with multiple water types, SWIR bands could prove useful for improving classification by including pixel features from the MNDWI and MBWI indices in the pipeline described in Figure 3.5.

5.3 Pixel Matching

Synthetic chlorophyll data for HYPSO-1 was generated by interpolation using matching spectral images from Sentinel-3. To account for the differences in spatial resolution that could lead to deficient chlorophyll values, a Gaussian filter was applied to the scene to smooth out the chlorophyll map. Even after this process, it is possible to find variations due to the time difference of the captures, even after limiting it with a window of $\pm$ 3 hours. The results of the QQ-Plot of Figures 3.6 and 3.7 show that even after applying a logarithmic transformation, normality is not achieved, which can be attributed to the nature of the defined interpolation process. The GLORIA dataset was not affected by the previously mentioned variations as all elements were matched with *in situ* chlorophyll measurements.

To improve the quality of the HYPSO dataset, spectral images would have to be paired with *in situ* samples from permanent ocean stations, as done by Vanhellemont and Ruddick (2021) on the Belgian coast. Temporal campaigns can also be deployed to collect local measurements, but it can take more time Binh et al. (2022); Li et al. (2022).

5.4 Chlorophyll Estimation

In this section, the contribution of feature tuning and feature selection to the estimation of chlorophyll is discussed. A summary of the best results for each model is found in Table 5.1 independently for the HYPSO (H) and GLORIA (G) datasets, where bold letters indicate the best result per metric.

Table 5.1: *Summary of best performing models.*

		BIAS	**RMSLE** ↓	**MAE** ↓	**RMSE** ↓	R^2
Multivariate Linear Regression	H	0.850	1.120	0.536	0.616	0.622
	G	0.492	2.092	0.780	1.004	0.667
OCx Polynomial	H	**0.999**	1.054	0.420	**0.335**	**0.884**
	G	1.060	2.496	0.870	1.187	0.560
Ensemble Voting	H	1.302	**0.972**	**0.320**	0.568	0.696
	G	1.460	1.510	0.531	0.752	0.785

Feature selection was used to reduce the complexity of the machine learning models used. Statistical methods were preferred to describe the relationship between each of the features and the estimation of chlorophyll. F-regression and mutual information regression were preferred over alternative recursive feature elimination methods that grow cubic to the number of feature points.

The GLORIA data points were reduced 40% because the spectral range of all the records was not the same. This caused issues during the analysis stage, as band ratios could not be calculated in some cases due to missing data. Constant expansion and maintenance of this dataset as done by others like Valente et al. (2022) is suggested. Although this process may take multiple years, GLORIA has potential to aid the study of hyperspectral remote sensing sensors due to its main focus on high spectral resolution.

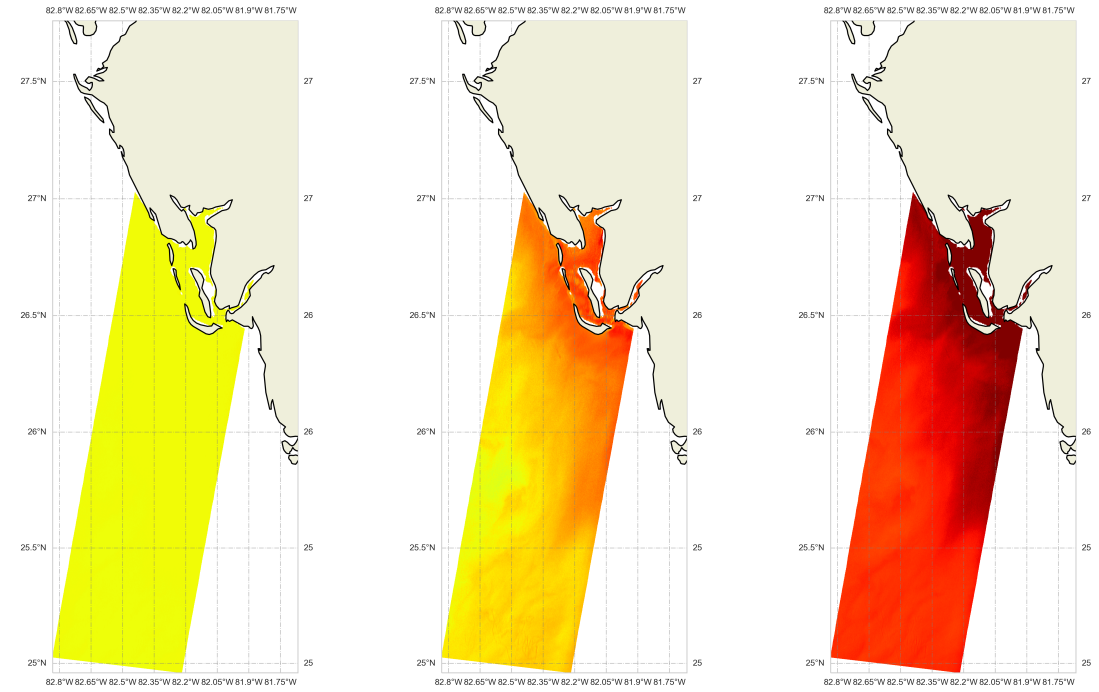
Regarding chlorophyll descriptors, the LBD feature was not used because when the band difference was very small, leading to infinite asymptotically values. Similarly, the second derivative was discarded as it was shown to not give consistent results when implementing the feature elimination process (see Tables 4.4 to 4.6).

Previous studies like the one performed by Tan et al. (2017) showed that multivariate regression through stepwise methods can help to show the relevant characteristics to improve predictions. From the multivariate linear regression analysis performed in this work, it was clear that the only "complex" chlorophyll descriptor selected as the highest contributor was TBVI (see Table 4.8). The Lasso approach discarded TBM and OCVI as the least contributing features.

According to Matthews (2011), the 700nm and 670nm band ratio has been used as it is highly correlated with chlorophyll estimation. Different studies have

explored possible variations close to this range, such as in the work of Sun et al. (2012). With the feature optimization method of this work, the same ranges were found for the GLORIA dataset where the logarithm of the same band ratio is consistently selected, being also used as a feature of the best multivariate model (see Table 4.8) and the feature with a highest mean SHAP value (see Table 4.15). In the case of HYPSO, the same ratio was not found. The closest ratio to LBR(700,670) is BR(701,503), supporting the observations that the atmospheric correction performs differently on different regions of the spectrum, and thus finding 503nm instead of 670nm.

An interesting behavior was observed for the multivariate linear regressions when exclusively using the features BR, BD, and LBR features independently (see rows in Table 4.8 with IDs "BRO", "LBRO" and "BDO" where "O" stands for "Only"). Figure 5.1 shows how the chlorophyll in a HYPSO image is estimated with these three regression models. Visually, BD seems to estimate relatively lower values, LBR estimates values in the medium regions of chlorophyll, and BR goes to the high end. This might explain why the best regression for HYPSO in Table 4.8 (marked with a *) has the contribution of all three.



(a) *BD [0.05, 1.0)* $\mathrm{mg\,m^{-3}}$ **(b)** *LBR [1.0, 5.0]* $\mathrm{mg\,m^{-3}}$ **(c)** *BR >10* $\mathrm{mg\,m^{-3}}$

Figure 5.1: *Individual contributions of independent feature groups on a HYPSO spectral image (same color scale).*

From the summary Table 5.1, it can be seen that the best results of the multivariate linear regression are similar to those of the OCx polynomial. Bias is better in the polynomial regression of OCx by being closer to 1.0, establishing that, on average, the prediction is closer to the ground truth. The MAE is smaller for the multivariate linear regression on the GLORIA dataset while the same metric is higher for the HYPSO data.

Visually, the OCx polynomials from Figures 4.9 and 4.10 show very little similarity to the model defined by O'Reilly and Werdell (2019) which is replicated in Figure 5.2 for the SeaWiFS sensor. The observed difference may be due to a curated selection of data points that allowed a good polynomial fit. After comparing the results, there is place to suspect that the good bias results from the OCX polynomial in this study are caused by an over-fitted regression.

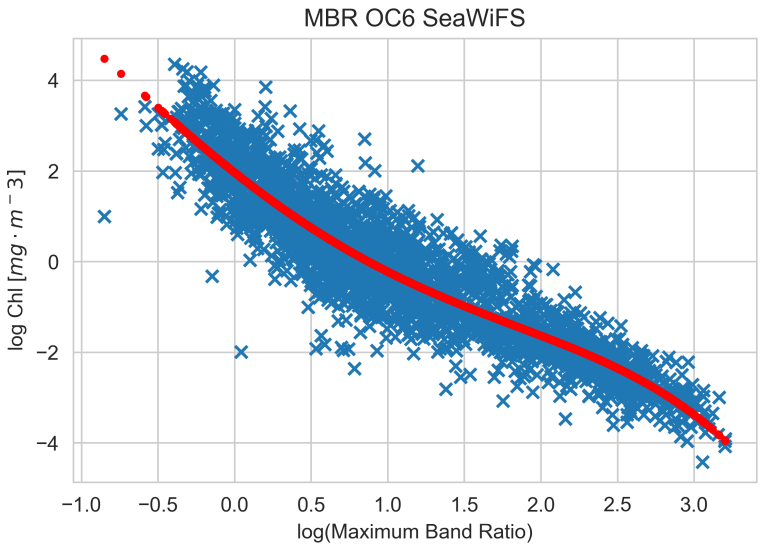


Figure 5.2: *OC6 MBR for SeaWiFS*

Of all the ensemble machine learning approaches, the voting strategy performed the best. K-fold splits and data normalization were performed separately in training and test splits to avoid data leaks, which could lead to optimistic conclusions based on the results of Nalepa et al. (2019). The ensemble machine learning voting regressor outperformed every other strategy in this work (see Table 5.1), making it the best alternative for chlorophyll prediction using R_{rs}.

The features selected for the ensemble model of each dataset are shown in Figures 4.13 and 4.15 where a higher SHAP value represents a greater contribution of the specific feature. The model for the GLORIA dataset shows that the feature LBR(700,670) is the most important while for HYPSO it is BR(701,503), which are the same features observed in the best performing multivariate linear regression.

The difference between features is greater on the model trained with the GLORIA dataset, finding a more homogeneous feature contribution when training the model with the HYPSO data.

Although a super-learner approach was tested using nearly twice the number of regressors, the performance did not increase with the model complexity. In a visual comparison of the chlorophyll map shown in Figures 4.7, 4.12, and 4.17, it can be observed that the voting approach estimates lower chlorophyll concentrations, which are not visible in the other images (lower values are shown in light blue). The performance of each machine learning regression used for the "voting" approach can be further enhanced by tuning individual hyperparameters. A grid or random search can be used so that multiple combinations of parameters can be tested to minimize the error. This process can lead to an exhaustive and long search, which was not implemented in this work.

Comparing the results of this thesis with the current literature is not straightforward as most of them use the coefficient of determination R^2 as the evaluation metric. The ensemble model outperforms the empirical OCX regression of this work and that of Binh et al. (2022); O'Reilly and Werdell (2019) by having a lower error. A similar bias and a lower MAE are found when comparing the voting ensemble technique with simpler traditional machine learning models such as those used by Cao et al. (2020).

More complex models can be used instead of simpler ones to reduce the error in chlorophyll estimation. The presented models achieve that compromising the explainability, which is possible with a multivariate linear regression, as the features and coefficients are clearly defined.

Toxic HAB can have an impact on the environment without being detected from space due to low biomass concentration (IOCCG, 2021). This characteristic may require additional properties other than chlorophyll to study and monitor water quality in coastal regions. Seasonal factors such as temperature, wind, and currents can change the conditions under which phytoplankton proliferate. This was not a factor considered in this work, but separating the data used for regression based on season may be of interest for future work based on the results of this thesis.

6 Conclusion

The problem of chlorophyll estimation with surface reflectance is approached by presenting an ensemble machine learning approach using fine-tuned chlorophyll descriptors. Permutations of bands were tested according to the requirement of different features to find the most correlated to chlorophyll estimation. Feature selection was implemented to reduce the number of descriptors without affecting the estimation, ensuring a low correlation between the selected ones. The performance of multivariate linear regression (including polynomial features), OCX maximum band ration polynomial, and ensemble machine learning was compared. The best performing model was quantitatively evaluated using prior work using relevant metrics. As a means to test generalization, the HYPSO approach was replicated using the GLORIA *in situ* reflectance dataset to eliminate the contribution of atmospheric correction to chlorophyll estimation. The proposed voting ensemble machine learning model is quantitatively evaluated with multiple metrics to better describe the estimation of the chlorophyll logarithmic pattern. The voting method achieves better results than its counterparts by using 6 fine-tuned features for HYPSO and 7 for GLORIA.

6.1 Future Work

Improved atmospheric correction:
In this work the atmospheric correction process was not evaluated against known ground surface reflectances, because of this, it is not possible to accurately assess the performance of the method introduced. Differences in surface reflectance between atmospheric correction and *in situ* R_{rs} are visible in the selected bands of the fine-tuned descriptors of HYPSO and GLORIA, respectively. More studies are needed for the HYPSO satellite to evaluate this correction process based on known ground truths.

ANN with Fine Tuned Features:
Ye et al. (2021) used artificial neural networks to reduce the chlorophyll estimation

error. With the existing radiance dataset of $\approx 1e^6$ points, it is possible to use the relevant features of this work to further develop this area of knowledge.

Effective Chlorophyll Ranges:
Hu et al. (2012, 2019) showed that lower concentrations follow a pseudolinear pattern, thus proposing a chlorophyll index (CI) for values less than 0.25 $\mathrm{mg\,m^{-3}}$. In later work, the performance of the models presented in this thesis should be evaluated in different chlorophyll ranges to potentially improve the chlorophyll estimation performance.

A Appendix

Table A.1: *Atmospheric model used in NASA (1966) based on the latitude and month of the year. The altitude in the region has not been considered, as only the surface model is of interest.*

Center Latitude	**Month (1 to 12)**	**Atmospheric Profile**
-15>Lat<15	Any	Tropical
15<Lat<=45	5 to 9	MidLatitudeSummer
-45<=Lat<-15	1 to 4 & 10 to 12	MidLatitudeWinter
45<Lat<=60	5 to 9	SubArticSummer
-60<=Lat<-45	1 to 4 & 10 to 12	SubArticWinter

Bibliography

Abderrazak, B., Morin, D., Bonn, F., and Huete, A. (1996). A review of vegetation indices. *Remote Sensing Reviews*, 13:95–120. (cited on page 33)

Acharya, T. D., Subedi, A., and Lee, D. H. (2018). Evaluation of Water Indices for Surface Water Extraction in a Landsat 8 Scene of Nepal. *Sensors*, 18(8):2580. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute. (cited on page 28)

Adam, E., Mutanga, O., Abdel-Rahman, E. M., and Ismail, R. (2014). Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: exploratory of in situ hyperspectral indices and random forest regression. *International Journal of Remote Sensing*, 35(2):693–714. Publisher: Taylor & Francis. (cited on pages 36 and 52)

ASTM (2020). Standard Tables for Reference Solar Spectral Irradiances: Direct Normal and Hemispherical on 37° Tilted Surface. (cited on pages 10, 20, 107, and 108)

Bakken, S., Johnsen, G., and Johansen, T. A. (2021). Analysis and Model Development of Direct Hyperspectral Chlorophyll-A Estimation for Remote Sensing Satellites. In *2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–5. ISSN: 2158-6276. (cited on page 34)

Bannari, A., Khurshid, K. S., Staenz, K., and Schwarz, J. W. (2007). A Comparison of Hyperspectral Chlorophyll Indices for Wheat Crop Chlorophyll Content Estimation Using Laboratory Reflectance Measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3063–3074. Conference Name: IEEE Transactions on Geoscience and Remote Sensing. (cited on page 33)

Barbedo, J. G. A. (2018). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and Electronics in Agriculture*, 153:46–53. (cited on page 33)

Barry, J. (2013). The Sustainability of Ocean Resources. In Madhavan, G., Oakley, B., Green, D., Koon, D., and Low, P., editors, *Practicing Sustainability*, pages 201–205. Springer, New York, NY. (cited on page 1)

Baumgardner, M., Biehl, L., and Landgrebe, D. (2015). 220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3. Version Number: 1.0 Type: dataset. Accesed: 08-11-2021. (cited on pages 15 and 107)

Binh, N. A., Hoa, P. V., Thao, G. T. P., Duan, H. D., and Thu, P. M. (2022). Evaluation of Chlorophyll-a estimation using Sentinel 3 based on various algorithms in southern coastal Vietnam. *International Journal of Applied Earth Observation and Geoinformation*, 112:102951. (cited on pages 81 and 85)

Braga, F., Fabbretto, A., Vanhellemont, Q., Bresciani, M., Giardino, C., Scarpa, G. M., Manfè, G., Concha, J. A., and Brando, V. E. (2022). Assessment of PRISMA water reflectance using autonomous hyperspectral radiometry. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192:99–114. (cited on page 22)

Brauers, J. and Aach, T. (2006). A color filter array based multispectral camera. In Group, G. C., editor, *12. Workshop Farbbildverarbeitung*, Ilmenau. (cited on page 15)

Braun, C. L. and Smirnov, S. N. (1993). Why is water blue? *Journal of Chemical Education*, 70(8):612. Publisher: American Chemical Society. (cited on page 11)

Campbell, J. W. (1995). The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research: Oceans*, 100(C7):13237–13254. (cited on page 49)

Cao, Z., Ma, R., Duan, H., Pahlevan, N., Melack, J., Shen, M., and Xue, K. (2020). A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes. *Remote Sensing of Environment*, 248:111974. (cited on pages 36 and 85)

Carroll, M., Townshend, J., DiMiceli, C., Noojipady, P., and Sohlberg, R. (2009). A new global raster water mask at 250 m resolution. *International Journal of Digital Earth*, 2(4):291–308. Publisher: Taylor & Francis. (cited on page 30)

Che, S., Du, G., Wang, N., He, K., Mo, Z., Sun, B., Chen, Y., Cao, Y., Wang, J., and Mao, Y. (2021). Biomass estimation of cultivated red algae Pyropia using unmanned aerial platform based multispectral imaging. *Plant Methods*, 17(1):12. (cited on page 33)

Chen, D. and Chen, H. W. (2013). Using the Köppen classification to quantify climate variation and change: An example for 1901–2010. *Environmental Development*, 6:69–79. (cited on page 79)

Clarke, G. L., Ewing, G. C., and Lorenzen, C. J. (1970). Spectra of Backscattered Light from the Sea Obtained from Aircraft as a Measure of Chlorophyll Concentration. *Science*, 167(3921):1119–1121. Publisher: American Association for the Advancement of Science. (cited on pages 7 and 18)

Cordeiro, M. C. R., Martinez, J.-M., and Peña-Luque, S. (2021). Automatic water detection from multidimensional hierarchical clustering for Sentinel-2 images and a comparison with Level 2A processors. *Remote Sensing of Environment*, 253:112209. (cited on pages 30, 46, 47, 64, 65, 81, 108, 109, and 111)

Cui, Z. and Kerekes, J. P. (2018). Potential of Red Edge Spectral Bands in Future Landsat Satellites on Agroecosystem Canopy Green Leaf Area Index Retrieval. *Remote Sensing*, 10(9):1458. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute. (cited on page 52)

Dierssen, H. M. and Randolph, K. (2012). Remote Sensing of Ocean Color. In Meyers, R. A., editor, *Encyclopedia of Sustainability Science and Technology*, pages 8952–8975. Springer, New York, NY. (cited on pages 12 and 13)

Eason, R. L. and NASA (1978). SP-399, Skylab EREP Investigations Summary, Appendix-A. Accesed: Jun 5th, 2023. URL: https://history.nasa.gov/SP-399/app-a.htm. (cited on pages 7, 8, and 107)

Elachi, C. and Van Zyl, J. (2006). *Introduction to the physics and techniques of remote sensing*. Wiley series in remote sensing. Wiley-Interscience, Hoboken, N.J, 2nd ed edition. OCLC: ocm61309501. (cited on page 9)

FAO (2022). FAO Fisheries & Aquaculture - Global aquaculture production. https://www.fao.org/fishery/statistics-query/en/aquaculture. (cited on pages 2 and 107)

Felde, G., Anderson, G., Cooley, T., Matthew, M., adler Golden, S., Berk, A., and Lee, J. (2003). Analysis of Hyperion data with the FLAASH atmospheric correction algorithm. In *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)*, volume 1, pages 90–92 vol.1. URL: https://www.l3harrisgeospatial.com/docs/FLAASH.html. (cited on pages 41 and 111)

Feyisa, G. L., Meilby, H., Fensholt, R., and Proud, S. R. (2014). Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sensing of Environment*, 140:23–35. (cited on page 27)

Flores-Romero, A. (2021). Remote Sensing Processing Methods: A comprehensive Review. Technical report, NTNU, Gjøvik, Norway. (cited on pages 15, 19, and 107)

F.X. Kneizys, L.W. Abreu, G.P. Anderson, J.H. Chetwynd, E.P. Shettle, A. Berk, L.S. Bernstein, D.C. Robertson, P. Acharya, L.S. Rothman, J.E.A. Selby, W.O. Gallery, and S.A. Clough (1996). The MODTRAN 2/3 Report and LOWTRAN 7 MODEL. Technical report, Phillips Laboratory - Geophysics Directorate, North Andover, MA. (cited on pages 42 and 79)

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42. (cited on page 56)

Gitahi, J. and Hahn, M. (2020). High-resolution urban air quality monitoring using sentinel satellite images and low-cost ground-based sensor networks. *E3S Web of Conferences*, 171:02002. (cited on page 22)

Gitelson, A. A., Dall'Olmo, G., Moses, W., Rundquist, D. C., Barrow, T., Fisher, T. R., Gurlin, D., and Holz, J. (2008). A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: Validation. *Remote Sensing of Environment*, 112(9):3582–3593. (cited on pages 52 and 53)

Gleick, P. H., Pacific Institute for Studies in Development, Environment, a. S., and Institute, S. E., editors (1993). *Water in crisis: a guide to the world's fresh water resources*. Oxford University Press, New York. (cited on page 1)

Goetz, A., Ferri, M., Kindel, B., and Qu, Z. (2002). Atmospheric correction of Hyperion data and techniques for dynamic scene correction. In *IEEE International Geoscience and Remote Sensing Symposium*, volume 3, pages 1408–1410 vol.3. (cited on page 80)

Gokul, E. A., Raitsos, D. E., Gittings, J. A., Alkawri, A., and Hoteit, I. (2019). Remotely sensing harmful algal blooms in the Red Sea. *PLOS ONE*, 14(4):e0215463. Publisher: Public Library of Science. (cited on page 80)

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., Niu, Z., Huang, X., Fu, H., Liu, S., Li, C., Li, X., Fu, W., Liu, C., Xu, Y., Wang, X., Cheng, Q., Hu, L., Yao, W., Zhang, H., Zhu, P., Zhao, Z., Zhang, H., Zheng, Y., Ji, L., Zhang, Y., Chen, H., Yan, A., Guo, J., Yu, L., Wang, L., Liu, X., Shi, T., Zhu, M., Chen, Y., Yang, G., Tang, P., Xu, B., Giri, C., Clinton, N., Zhu, Z., Chen, J., and Chen, J. (2013). Finer resolution observation and monitoring of global land

cover: first mapping results with Landsat TM and ETM+ data. *International Journal of Remote Sensing*, 34(7):2607–2654. Publisher: Taylor & Francis. (cited on pages 30 and 31)

Gons, H., Rijkeboer, M., and Ruddick, K. (2002). A chlorophyll-retrieval algorithm for satellite imagery (Medium Resolution Imaging Spectrometer) of inland and coastal waters. *Journal of Plankton Research*, 24(9):947–951. (cited on page 34)

Gordon, H. R. (2010). Some Reflections on ThirtyFive Years of Ocean Color Remote Sensing. In Barale, V., Gower, J., and Alberotanza, L., editors, *Oceanography from Space: Revisited*, pages 289–306. Springer Netherlands, Dordrecht. (cited on pages 7 and 19)

Gordon, H. R. (2019). *Physical Principles of Ocean Color Remote Sensing*. University of Miami. (cited on pages 19, 21, 22, 80, and 111)

Grøtte, M. E., Birkeland, R., Honoré-Livermore, E., Bakken, S., Garrett, J. L., Prentice, E. F., Sigernes, F., Orlandić, M., Gravdahl, J. T., and Johansen, T. A. (2022). Ocean Color Hyperspectral Remote Sensing With High Resolution and Low Latency—The HYPSO-1 CubeSat Mission. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19. Conference Name: IEEE Transactions on Geoscience and Remote Sensing. (cited on page 3)

Gualtieri, P. and Barsanti, L. (2006). *Algae: anatomy, biochemistry, and biotechnology*. Taylor & Francis, Boca Raton. (cited on page 1)

Guo, Y., Senthilnath, J., Wu, W., Zhang, X., Zeng, Z., and Huang, H. (2019). Radiometric Calibration for Multispectral Camera of Different Imaging Conditions Mounted on a UAV Platform. *Sustainability*, 11(4):978. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. (cited on page 18)

Hall, D. O. and Rao, K. K. (1999). *Photosynthesis*. Studies in biology. Cambridge University Press, Cambridge, UK ; New York, 6th ed edition. (cited on pages 1 and 10)

Hemant Kumar Aggarwal and Majumdar, A. (2013). Multi-spectral demosaicing technique for single-sensor imaging. In *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pages 1–4, Jodhpur, India. IEEE. (cited on page 15)

Henriksen, M. B., Prentice, E. F., Johansen, T. A., and Sigernes, F. (2022). Pre-Launch Calibration of the HYPSO-1 Cubesat Hyperspectral Imager. In *2022 IEEE Aerospace Conference (AERO)*, pages 1–9. ISSN: 1095-323X. (cited on pages 4, 16, 39, and 41)

Hu, C., Feng, L., Lee, Z., Franz, B., Bailey, S., Werdell, J., and Proctor, C. (2019). Improving Satellite Global Chlorophyll a Data Products Through Algorithm Refinement and Data Recovery. *Journal of Geophysical Research: Oceans*, 124. (cited on pages 31 and 88)

Hu, C., Lee, Z., and Franz, B. (2012). Chlorophyll aalgorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *Journal of Geophysical Research: Oceans*, 117(C1). (cited on page 88)

Huang, L.-K. and Wang, M.-J. J. (1995). Image thresholding by minimizing the measures of fuzziness. *Pattern Recognition*, 28(1):41–51. (cited on page 28)

Incropera, F. P., editor (2007). *Fundamentals of heat and mass transfer*. John Wiley, Hoboken, NJ, 6th ed edition. OCLC: ocm62532755. (cited on page 10)

IOCCG (2008). Why ocean colour? the societal benefits of ocean-colour technology. Technical report, IOCCG. Platt, T., Hoepffner, N., Stuart, V. and Brown, C. (eds.), Reports of the International Ocean-Colour Coordinating Group, No. 7, IOCCG, Dartmouth, Canada. (cited on page 12)

IOCCG (2010). Atmospheric Correction for Remotely-Sensed Ocean-Colour Products. Technical report, IOCCG. Wang, M. (ed.), Reports of the International Ocean-Colour Coordinating Group, No. 10, IOCCG, Dartmouth, Canada. (cited on pages 18 and 21)

IOCCG (2018). Earth Observations in Support of Global Water Quality Monitoring. Technical report, IOCCG. Greb, S., Dekker, A. and Binding, C. (eds.), IOCCG Report Series, No. 17, International Ocean Colour Coordinating Group, Dartmouth, Canada. (cited on pages 4, 11, and 12)

IOCCG (2021). Observation of Harmful Algal Blooms with Ocean Colour Radiometry. Technical report, IOCCG. Bernard, S., Kudela, R., Robertson Lain, L. and Pitcher, G.C. (eds.), IOCCG Report Series, No. 20, International Ocean Colour Coordinating Group, Dartmouth, Canada. http:// dx.doi.org/ 10.25607/ OBP-1042. (cited on pages 12, 18, 80, and 85)

Jensen, J. R. (2014). *Remote sensing of the environment: an earth resource perspective*. Pearson, Harlow, second edition [exclusive edition only for the benefit of students outside the united states and canada] edition. (cited on page 39)

Ji, L., Gong, P., Geng, X., and Zhao, Y. (2015). Improving the Accuracy of the Water Surface Cover Type in the 30 m FROM-GLC Product. *Remote Sensing*, 7(10):13507–13527. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute. (cited on pages 26 and 31)

Jiang, W., Ni, Y., Pang, Z., He, G., Fu, J., Lu, J., Yang, K., Long, T., and Lei, T. (2020). A NEW INDEX FOR IDENTIFYING WATER BODY FROM SENTINEL-2 SATELLITE REMOTE SENSING IMAGERY. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-3-2020:33–38. (cited on page 27)

Jiang, Z., Huete, A. R., Chen, J., Chen, Y., Li, J., Yan, G., and Zhang, X. (2006). Analysis of NDVI and scaled difference vegetation index retrievals of vegetation fraction. *Remote Sensing of Environment*, 101(3):366–378. (cited on page 27)

Jiang Hai-ling, Zhang Li-fu, Yang Hang, Chen Xiao-ping, Wang Shu-dong, Li Xue-ke, and Liu Kai (2014). Comparison of accuracy and stability of estimating winter wheat chlorophyll content based on spectral indices. In *2014 IEEE Geoscience and Remote Sensing Symposium*, pages 2985–2988, Quebec City, QC. IEEE. (cited on page 33)

Johannessen, J. A., Johannessen, O. M., and Haugan, P. M. (1989). Remote sensing and model simulation studies of the Norwegian coastal current during the algal bloom in May 1988. *International Journal of Remote Sensing*, 10(12):1893–1906. (cited on page 1)

John, U., Šupraha, L., Gran-Stadniczeñko, S., Bunse, C., Cembella, A., Eikrem, W., Janouškovec, J., Klemm, K., Kühne, N., Naustvoll, L., Voss, D., Wohlrab, S., and Edvardsen, B. (2022). Spatial and biological oceanographic insights into the massive fish-killing bloom of the haptophyte Chrysochromulina leadbeateri in northern Norway. *Harmful Algae*, 118:102287. (cited on page 1)

JPL, N. (2013). NASA Shuttle Radar Topography Mission Water Body Data Shapefiles. (cited on page 31)

Jun, C., Ban, Y., and Li, S. (2014). Open access to Earth land-cover map. *Nature*, 514(7523):434–434. Number: 7523 Publisher: Nature Publishing Group. (cited on page 31)

JUSTICE, C. O., TOWNSHEND, J. R. G., HOLBEN, B. N., and TUCKER, C. J. (1985). Analysis of the phenology of global vegetation using meteorological satellite data. *International Journal of Remote Sensing*, 6(8):1271–1318. Publisher: Taylor & Francis. (cited on page 27)

Kapur, J. N., Sahoo, P. K., and Wong, A. K. C. (1985). A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*, 29(3):273–285. (cited on page 28)

Karlson, B., Andersen, P., Arneborg, L., Cembella, A., Eikrem, W., John, U., West, J. J., Klemm, K., Kobos, J., Lehtinen, S., Lundholm, N., Mazur-Marzec, H., Naustvoll, L., Poelman, M., Provoost, P., De Rijcke, M., and Suikkanen, S. (2021). Harmful algal blooms and their effects in coastal seas of Northern Europe. *Harmful Algae*, 102:101989. (cited on pages 1, 2, and 12)

Kiku, D., Monno, Y., Tanaka, M., and Okutomi, M. (2014). Simultaneous capturing of RGB and additional band images using hybrid color filter array. In Sampat, N., Tezaur, R., Battiato, S., and Fowler, B. A., editors, *Proceedings Volume 9023, Digital Photography X*, page 90230V, San Francisco, California, USA. (cited on page 15)

Kirk, J. T. O. (2011). *Light and photosynthesis in aquatic ecosystems*. Cambridge University Press, Cambridge, UK ; New York, 3rd ed edition. OCLC: ocn650821926. (cited on page 1)

L. Bourg, J. Bruniquel, C. Henocq, H. Morris, J. Dash, R. Preusker, and S. Dransfeld (2023). Copernicus Sentinel-3 OLCI Land User Handbook (v1.2). Technical Report OMPC.ACR.HBK.001, ESA. Ref: OMPC.ACR.HBK.001. URL: https://sentinel.esa.int/documents/247904/4598066/Sentinel-3-OLCI-Land-Handbook.pdf. (cited on pages 27 and 111)

Lapray, P.-J., Wang, X., Thomas, J.-B., and Gouton, P. (2014). Multispectral Filter Arrays: Recent Advances and Practical Implementation. *Sensors*, 14(11):21626–21659. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute. (cited on pages 15 and 107)

Lazzeri, G., Frodella, W., Rossi, G., and Moretti, S. (2021). Multitemporal Mapping of Post-Fire Land Cover Using Multiplatform PRISMA Hyperspectral and Sentinel-UAV Multispectral Data: Insights from Case Studies in Portugal and Italy. *Sensors*, 21(12):3982. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute. (cited on page 17)

Lehmann, M. K., Gurlin, D., Pahlevan, N., Alikas, K., Conroy, T., Anstee, J., Balasubramanian, S. V., Barbosa, C. C. F., Binding, C., Bracher, A., Bresciani, M., Burtner, A., Cao, Z., Dekker, A. G., Di Vittorio, C., Drayson, N., Errera, R. M., Fernandez, V., Ficek, D., Fichot, C. G., Gege, P., Giardino, C., Gitelson, A. A., Greb, S. R., Henderson, H., Higa, H., Rahaghi, A. I., Jamet, C., Jiang, D., Jordan, T., Kangro, K., Kravitz, J. A., Kristoffersen, A. S., Kudela, R., Li, L., Ligi, M., Loisel, H., Lohrenz, S., Ma, R., Maciel, D. A., Malthus, T. J., Matsushita, B., Matthews, M., Minaudo, C., Mishra, D. R., Mishra, S., Moore, T., Moses, H., Novo, E. M. L. M., Novoa, S., Odermatt, D., O'Donnell, D. M., Olmanson, L. G., Ondrusek, M., Oppelt, N., Ouillon, S., Pereira Filho, W.,

Plattner, S., Verdú, A. R., Salem, S. I., Schalles, J. F., Simis, S. G. H., Siswanto, E., Smith, B., Somlai-Schweiger, I., Soppa, M. A., Spyrakos, E., Tessin, E., van der Woerd, H. J., Vander Woude, A., Vandermeulen, R. A., Vantrepotte, V., Wernand, M. R., Werther, M., Young, K., and Yue, L. (2023). GLORIA - A globally representative hyperspectral in situ dataset for optical sensing of water quality. *Scientific Data*, 10(1):100. Number: 1 Publisher: Nature Publishing Group. (cited on pages 32 and 48)

Lehner, B. and Doell, P. (2004). Development and Validation of a Global Database of Lakes, Reservoirs and Wetlands. *Journal of Hydrology*, 296:1–22. (cited on page 30)

Li, J., Ge, X., He, Q., and Abbas, A. (2021). Aerosol optical depth (AOD): spatial and temporal variations and association with meteorological covariates in Taklimakan desert, China. *PeerJ*, 9:e10542. (cited on page 22)

Li, L., Li, L., and Song, K. (2015). Remote sensing of freshwater cyanobacteria: An extended IOP Inversion Model of Inland Waters (IIMIW) for partitioning absorption coefficient and estimating phycocyanin. *Remote Sensing of Environment*, 157:9–23. (cited on page 34)

Li, X., Shang, S., Lee, Z., Lin, G., Zhang, Y., Wu, J., Kang, Z., Liu, X., Yin, C., and Gao, Y. (2022). Detection and Biomass Estimation of Phaeocystis globosa Blooms off Southern China From UAV-Based Hyperspectral Measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13. Conference Name: IEEE Transactions on Geoscience and Remote Sensing. (cited on pages 57, 80, and 81)

Lubac, B., Loisel, H., Guiselin, N., Astoreca, R., Felipe Artigas, L., and Mériaux, X. (2008). Hyperspectral and multispectral ocean color inversions to detect *Phaeocystis globosa* blooms in coastal waters. *Journal of Geophysical Research*, 113(C6):C06026. (cited on pages 4 and 51)

Marshall, M. and Thenkabail, P. (2015). Developing in situ Non-Destructive Estimates of Crop Biomass to Address Issues of Scale in Remote Sensing. *Remote Sensing*, 7(1):808–835. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. (cited on page 52)

Matthews, M. W. (2011). A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters. *International Journal of Remote Sensing*, 32(21):6855–6899. Publisher: Taylor & Francis. (cited on pages 52 and 82)

McFEETERS, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7):1425–1432. Publisher: Taylor & Francis. (cited on page 27)

Milczarek, M., Robak, A., and Gadawska, A. (2017). Sentinel Water Mask (SWM) - new index for water detection on Sentinel-2 images. Technical report, Szent István University. (cited on page 27)

Mobley, C. (2021). Remote Sensing - Ocean Color. https://www.oceanopticsbook.info/view/remote-sensing/ocean-color. (cited on pages 11, 12, 21, and 111)

Mobley, C. D. (1999). Estimation of the remote-sensing reflectance from above-surface measurements. *Applied Optics*, 38(36):7442. (cited on page 43)

Moran, M. S., Jackson, R. D., Slater, P. N., and Teillet, P. M. (1992). Evaluation of simplified procedures for retrieval of land surface reflectance factors from satellite sensor output. *Remote Sensing of Environment*, 41(2):169–184. (cited on page 44)

Muller-Karger, F. E. (1992). Remote sensing of marine pollution: A challenge for the 1990s. *Marine Pollution Bulletin*, 25(1):54–60. (cited on page 4)

Muthukrishnan, R. and Rohini, R. (2016). LASSO: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. (cited on page 55)

Nalepa, J., Myller, M., and Kawulok, M. (2019). Validating Hyperspectral Image Segmentation. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1264–1268. arXiv:1811.03707 [cs]. (cited on page 84)

NASA (1966). U.S. Standard Atmosphere Supplements, 1966. Technical Report NASA-CR-88870, NASA. NTRS Document ID: 19670028571 NTRS Research Center: Legacy CDMS (CDMS). (cited on pages 41, 89, and 112)

NASA (n.d.). CZCS. Ocean Color Web. https://oceancolor.gsfc.nasa.gov/data/czcs/2. (cited on page 8)

National Research Council (2011). *Assessing the Requirements for Sustained Ocean Color Research and Operations*. National Academies Press, Washington, D.C. (cited on page 8)

Niedzwiedzki, D. M. and Blankenship, R. E. (2010). Singlet and triplet excited state properties of natural chlorophylls and bacteriochlorophylls. *Photosynthesis Research*, 106(3):227–238. (cited on pages 13 and 107)

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegel, D. A., Carder, K. L., Garver, S. A., Kahru, M., and McClain, C. (1998). Ocean color chlorophyll algorithms for SeaWiFS. *Journal of Geophysical Research: Oceans*, 103(C11):24937–24953. (cited on page 13)

O'Reilly, J. E. and Werdell, P. J. (2019). Chlorophyll algorithms for ocean color sensors - OC4, OC5 & OC6. *Remote Sensing of Environment*, 229:32–47. (cited on pages 31, 32, 55, 71, 73, 84, and 85)

O'Shea, R. E., Pahlevan, N., Smith, B., Bresciani, M., Egerton, T., Giardino, C., Li, L., Moore, T., Ruiz-Verdu, A., Ruberg, S., Simis, S. G., Stumpf, R., and Vaičiūtė, D. (2021). Advancing cyanobacteria biomass estimation from hyperspectral observations: Demonstrations with HICO and PRISMA imagery. *Remote Sensing of Environment*, 266:112693. (cited on pages 4 and 58)

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics. (cited on page 28)

Pahlevan, N., Mangin, A., Balasubramanian, S. V., Smith, B., Alikas, K., Arai, K., Barbosa, C., Bélanger, S., Binding, C., Bresciani, M., Giardino, C., Gurlin, D., Fan, Y., Harmel, T., Hunter, P., Ishikaza, J., Kratzer, S., Lehmann, M. K., Ligi, M., Ma, R., Martin-Lauzer, F.-R., Olmanson, L., Oppelt, N., Pan, Y., Peters, S., Reynaud, N., Sander de Carvalho, L. A., Simis, S., Spyrakos, E., Steinmetz, F., Stelzer, K., Sterckx, S., Tormos, T., Tyler, A., Vanhellemont, Q., and Warren, M. (2021). ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters. *Remote Sensing of Environment*, 258:112366. (cited on pages 4, 57, and 58)

Pandey, A., editor (2014). *Biofuels from algae*. Elsevier, Amsterdam, first edition edition. (cited on pages 1 and 12)

Polder, G. and Gowen, A. (2020). The hype in spectral imaging. *Journal of Spectral Imaging*, page a4. (cited on page 14)

Prentice, E. F., Grøtte, M. E., Sigernes, F., and Johansen, T. A. (2021). Design of a hyperspectral imager using COTS optics for small satellite applications. In Sodnik, Z., Cugny, B., and Karafolas, N., editors, *International Conference on Space Optics — ICSO 2020*, page 187, Online Only, France. SPIE. (cited on pages 4, 16, and 107)

Preusker, R. and El Kassar, R. (2022). Monthly total column water vapour over land and ocean from 2002 to 2012 derived from satellite observations.

Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.8e0e4724 (Accessed on 07-JUN-2023). (cited on page 20)

Pyo, J., Duan, H., Baek, S., Kim, M. S., Jeon, T., Kwon, Y. S., Lee, H., and Cho, K. H. (2019). A convolutional neural network regression for quantifying cyanobacteria using hyperspectral imagery. *Remote Sensing of Environment*, 233:111350. (cited on page 34)

Pérez, A. J., López, F., Benlloch, J. V., and Christensen, S. (2000). Colour and shape analysis techniques for weed detection in cereal fields. *Computers and Electronics in Agriculture*, 25(3):197–212. (cited on page 52)

Randall B. Smith (2012). *Introduction to Remote Sensing of Environment (RSE)*. MicroImages. URL: https://www.microimages.com/documentation/Tutorials/introrse.pdf. (cited on pages 9 and 12)

Rao, G. A. and Mahulikar, S. P. (2012). Effect of Atmospheric Transmission and Radiance on Aircraft Infared Signatures. *Journal of Aircraft*. (cited on page 19)

Riordan, C. and Hulstron, R. (1990). What is an air mass 1.5 spectrum? (solar cell performance calculations). In *IEEE Conference on Photovoltaic Specialists May 1990*, pages 1085–1088 vol.2. (cited on page 20)

Rouse, J. W., Haas, R. H., Schell, J. A., and Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. In *NASA*. NTRS Author Affiliations: Texas A&M Univ. NTRS Report/Patent Number: PAPER-A20 NTRS Document ID: 19740022614 NTRS Research Center: Legacy CDMS (CDMS). (cited on page 53)

Ruszczak, B., Wijata, A. M., and Nalepa, J. (2022). Unbiasing the Estimation of Chlorophyll from Hyperspectral Images: A Benchmark Dataset, Validation Procedure and Baseline Results. *Remote Sensing*, 14(21):5526. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute. (cited on pages 17, 32, and 33)

Santer, R. (2010). OLCI Level 2 Algorithm Theoretical Basis Document - Pixel Classification (v2.3). Technical report, ESA. Ref: S3-L2-SD-03-C01-LISE-ATBD. URL: https://www-cdn.eumetsat.int/files/2020-04/pdf_s3_l2_atbd_pixel_class.pdf. (cited on page 27)

Schläpfer, D., Popp, C., and Richter, R. (2020). DRONE DATA ATMOSPHERIC CORRECTION CONCEPT FOR MULTI- ANDHYPERSPECTRAL IMAGERY – THE DROACOR MODEL. *The International Archives of the Photogrammetry,*

Remote Sensing and Spatial Information Sciences, XLIII-B3-2020:473–478. (cited on page 18)

Schroeder, T., Behnert, I., Schaale, M., Fischer, J., and Doerffer, R. (2007). Atmospheric correction algorithm for MERIS above case-2 waters. *International Journal of Remote Sensing*, 28(7):1469–1486. (cited on page 80)

Seegers, B. N., Stumpf, R. P., Schaeffer, B. A., Loftin, K. A., and Werdell, P. J. (2018). Performance metrics for the assessment of satellite data products: an ocean color case study. *Optics Express*, 26(6):7404. (cited on pages 35, 47, 49, and 58)

Sekertekin, A. (2019). Potential of global thresholding methods for the identification of surface water resources using Sentinel-2 satellite imagery and normalized difference water index. *Journal of Applied Remote Sensing*, 13(4):044507. Publisher: SPIE. (cited on page 28)

Sohil, F., Sohali, M. U., and Shabbir, J. (2022). An introduction to statistical learning with applications in R: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7. *Statistical Theory and Related Fields*, 6(1):87–87. (cited on page 54)

Spellman, F. R. (2019). *The Handbook of Nature*. Rowman & Littlefield. Google-Books-ID: JTbDDwAAQBAJ. (cited on page 1)

Sun, D., Li, Y., Wang, Q., Le, C., Lv, H., Huang, C., and Gong, S. (2012). A novel support vector regression model to estimate the phycocyanin concentration in turbid inland waters from hyperspectral reflectance. *Hydrobiologia*, 680(1):199–217. (cited on page 83)

Tacon, A. G. J. (2020). Trends in Global Aquaculture and Aquafeed Production: 2000–2017. *Reviews in Fisheries Science & Aquaculture*, 28(1):43–56. Publisher: Taylor & Francis. (cited on page 2)

Tan, W., Liu, P., Liu, Y., Yang, S., and Feng, S. (2017). A 30-Year Assessment of Phytoplankton Blooms in Erhai Lake Using Landsat Imagery: 1987 to 2016. *Remote Sensing*, 9(12):1265. (cited on page 82)

Taniguchi, M. and Lindsey, J. S. (2021). Absorption and Fluorescence Spectral Database of Chlorophylls and Analogues. *Photochemistry and Photobiology*, 97(1):136–165. (cited on pages 13 and 107)

Trottet, A., George, C., Drillet, G., and Lauro, F. M. (2022). Aquaculture in coastal urbanized areas: A comparative review of the challenges posed by Harmful Algal

Blooms. *Critical Reviews in Environmental Science and Technology*, 52(16):2888–2929. Publisher: Taylor & Francis. (cited on page 2)

Tsai, F. and Philpot, W. (1996). Derivative Analysis of Hyperspectral Data. *Remote Sensing of Environment*, 66:41–51. (cited on page 51)

Tulczyjew, L., Kawulok, M., Longépé, N., Saux, B. L., and Nalepa, J. (2022). A Multibranch Convolutional Neural Network for Hyperspectral Unmixing. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5. arXiv:2208.02361 [cs, eess]. (cited on page 34)

UNDESA (2012). Back to our Common Future: Sustainable Development in the 21 st Century (SD21) project. Technical report, United Nations Department of Economic and Social Affairs, New York. (cited on page 2)

UNESCO and Hallegraeff, Gustaaf M. (2003). *Manual on Harmful Marine Microalgae*, chapter 1, pages 25–26. Number 11 in Monographs on oceanographic methodology. UNESCO Publishing, France, 2 edition. (cited on page 1)

United Nations (2022). World Population Prospects 2022: Summary of Results. Technical report, Department of Economic and Social Affairs - Population Division. UN DESA/POP/2022/TR/NO. 3. (cited on page 2)

USGS (2019). Landsat 7 (L7) Data Users Handbook. Technical Report LSDS-1927 v2.0, Department of the Interior U.S. Geological Survey, Sioux Falls, South Dakota. (cited on pages 44 and 45)

Valente, A., Sathyendranath, S., Brotas, V., Groom, S., Grant, M., Jackson, T., Chuprin, A., Taberner, M., Airs, R., Antoine, D., Arnone, R., Balch, W. M., Barker, K., Barlow, R., Bélanger, S., Berthon, Y., Bracher, A., Brando, V., Brewin, R. J. W., Canuti, E., Chavez, F. P., Cianca, A., Claustre, H., Clementson, L., Crout, R., Ferreira, A., Freeman, S., Frouin, R., García-Soto, C., Gibb, S. W., Goericke, R., Gould, R., Guillocheau, N., Hooker, S. B., Hu, C., Kahru, M., Kampel, M., Klein, H., Kratzer, S., Kudela, R., Ledesma, J., Lohrenz, S., Loisel, H., Mannino, A., Martinez-Vicente, V., Matrai, P., McKee, D., Mitchell, B. G., Moisan, T., Montes, E., Muller-Karger, F., Neeley, A., Novak, M., O'Dowd, L., Ondrusek, M., Platt, T., Poulton, A. J., Repecaud, M., Röttgers, R., Schroeder, T., Smyth, T., Smythe-Wright, D., Sosik, H. M., Thomas, C., Thomas, R., Tilstone, G., Tracana, A., Twardowski, M., Vellucci, V., Voss, K., Werdell, J., Wernand, M., Wojtasiewicz, B., Wright, S., and Zibordi, G. (2022). A compilation of global bio-optical in situ data for ocean colour satellite applications – version three. *Earth System Science Data*, 14(12):5737–5770. Publisher: Copernicus GmbH. (cited on pages 32 and 82)

Vanhellemont, Q. and Ruddick, K. (2018). Atmospheric correction of metre-scale optical satellite data for inland and coastal water applications. *Remote Sensing of Environment*, 216:586–597. (cited on page 22)

Vanhellemont, Q. and Ruddick, K. (2021). Atmospheric correction of Sentinel-3/OLCI data for mapping of suspended particulate matter and chlorophyll-a concentration in Belgian turbid coastal waters. *Remote Sensing of Environment*, 256:112284. (cited on pages 22, 61, 81, and 109)

Vermote, E., Tanre, D., Deuze, J., Herman, M., and Morcrette, J.-J. (2006). Second simulation of a satellite signal in the solar spectrum-vector (6SV). Technical report, Department of Geography - University of Maryland. (cited on pages 22, 23, 24, 39, and 108)

Vincini, M., Frazzi, E., and D'Alessio, P. (2008). A broad-band leaf chlorophyll vegetation index at the canopy scale. *Precision Agriculture*, 9(5):303–319. (cited on pages 52 and 53)

Wang, J., Xiao, X., Bajgain, R., Starks, P., Steiner, J., Doughty, R. B., and Chang, Q. (2019). Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 154:189–201. (cited on page 52)

Wang, Y., Colby, J. D., and Mulcahy, K. A. (2002). An efficient method for mapping flood extent in a coastal floodplain using Landsat TM and DEM data. *International Journal of Remote Sensing*, 23(18):3681–3696. Publisher: Taylor & Francis. (cited on page 26)

Wilson, S., Apel, J., and Lindstrom, E. (2001). Satellite Oceanography, History and Introductory Concepts. In *Encyclopedia of Ocean Sciences*, pages 2517–2530. Elsevier. (cited on pages 7, 8, 9, 13, 18, and 111)

Witten, I. H. and Witten, I. H., editors (2017). *Data mining: practical machine learning tools and techniques*. Elsevier, Amsterdam, fourth edition edition. (cited on page 56)

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14):3025–3033. Publisher: Taylor & Francis. (cited on page 27)

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and Bates, P. D. (2017). A high-accuracy

map of global terrain elevations. *Geophysical Research Letters*, 44(11):5844–5853. (cited on pages 64, 65, 81, and 109)

Ye, H., Tang, S., and Yang, C. (2021). Deep Learning for Chlorophyll-a Concentration Retrieval: A Case Study for the Pearl River Estuary. *Remote Sensing*, 13(18):3717. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute. (cited on pages 33 and 87)

Zack, G. W., Rogers, W. E., and Latt, S. A. (1977). Automatic measurement of sister chromatid exchange frequency. *The Journal of Histochemistry and Cytochemistry: Official Journal of the Histochemistry Society*, 25(7):741–753. (cited on page 28)

Zhai, K., Wu, X., Qin, Y., and Du, P. (2015). Comparison of surface water extraction performances of different classic water indices using OLI and TM imageries in different situations. *Geo-spatial Information Science*, 18(1):32–42. Publisher: Taylor & Francis. (cited on pages 25 and 27)

Zhang, C. and Ma, Y., editors (2012). *Ensemble Machine Learning: Methods and Applications*. Springer, New York, NY. (cited on page 56)

Zhang, L., Zhang, L., and Du, B. (2016). Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40. Conference Name: IEEE Geoscience and Remote Sensing Magazine. (cited on page 33)

Zheng, A. and Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. O'Reilly, Beijing : Boston, first edition edition. OCLC: ocn957747646. (cited on page 54)

List of Figures

List of Tables