Estimating phytoplankton size classes from their inherent optical properties

Presented By:

David Stephen Berliner

BRLDAV003

Prepared For:

Dr. F. Nicolls

Dept. of Electrical Engineering

University of Cape Town

Declaration

I know the meaning of plagiarism and declare that all the work in the document, save for that which is properly acknowledged, is my own. This thesis/dissertation has been submitted to the Turnitin module and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor.

Signature:  D. S. Berliner

Acknowledgements

This thesis would not have been possible if it were not for a number of very special people in my life who provided support, guidance and mentorship through, what was at times, a rather gruelling process. First, and foremost, I have to pay a very special thanks to my wonderful wife, who had to make many sacrifices on my behalf, and who provided on-going support and encouragement throughout the process. I would like to thank my supervisor, Dr Fred Nicolls, for all of his guidance and involvement. His calm nature and pragmatic approach helped me steer clear of many potential catastrophes. It was through him that I got to meet all the wonderful people at the CSIR. Another very special thank you has to be given to Dr Sandy Thomalla and Dr Steward Bernard at SOCCO, both of whom provided me with all of the much-needed information and guidance on bio-optics and phytoplankton. Even though the SOCCO department had no obligation to help me, they sacrificed many hours for me, even integrating me into their team for the ACE expedition. Lastly, I would like to thank all my teammates of the ACE cruise – Dr William Moutier, Dr Nina Schuback, Dr Thomas Ryan-Keogh and Hazel Little; you guys played a key role in the success of this project.

Abstract

Phytoplankton plays a massive role in the regulation of greenhouse gases, with different functional types affecting the carbon cycle differently. The most practical way of synoptically mapping the ocean's phytoplankton communities is through remote sensing with the aid of ocean-optics algorithms.

This thesis is a study of the relationships between the Inherent Optical Properties (IOPs) of the ocean and the physical constituents within it, with a special focus on deriving phytoplankton size classes. Three separate models were developed, each focusing on a different relationship between absorption and phytoplankton size classes, before being combined into a final ensemble model.

It was shown that all of the developed models performed better than the baseline model, which only estimates the mean values per size class, and that the results of the final ensemble model is comparable to, and performs better than, most other published models on the NOMAD dataset.

Table of Contents

List of Figures

List of Tables

Glossary

Terms	Definition/Explanation
$a^*_{mean}(\lambda)$	Mean normalised absorption
$a^*_p(\lambda)$	Normalised phytoplankton absorption
$a_p(\lambda)$	Particulate absorption (m^{-1})
$a_{ph}(\lambda)$	Phytoplankton absorption (m^{-1})
$b_b(\lambda)$	Total backscattering (m^{-1})
$L_w(\lambda)$	Normalised spectral water-leaving radiance ($mW\ cm^2\ \mu m^{-1}\ sr^{-1}$)
$R_{rs}(\lambda)$	Spectral remote-sensing reflectance (sr^{-1})
T_{chla}	Total chlorophyll a
$< a(\lambda) >$	Mean absorption
$a(\lambda)$	Total absorption (m^{-1})
ACE	Antarctic Circumnavigation Expedition
AC-S	Spectral Absorption and Attenuation Sensor
ANN	Artificial Neural Network
AOP	Apparent Optical Properties
CDM	Colour Dissolved Matter
CDOM	Colour Dissolved Organic Matter
CSV	Comma Separated Value (file type)
DMS	Dimethyl Sulfide
DPA	Diagnostic Pigment Analysis
HPLC	High Performance Liquid Chromatography
IOP	Inherent Optical Properties
Micro (µm)	$\geq 20\mu m$
Nano (nm)	$2 - 20\mu m$
NAP	Non-Algal Particles
NMF	Non-Negative Matrix Factorisation
PAR	Photosynthetically Active Range
PCA	Principal Component Analysis
PFT	Phytoplankton Functional Type
Pico (pm)	$\leq 0.2\mu m$
PSD	Particle Size Distribution
RMSE	Root Mean Square Error
SQL	Structured Query Language
SVD	Singular-Value Decomposition
SVM	Support Vector Machine
SVR	Support Vector Regression

Chapter 1: Introduction

1.1. Background

This chapter serves to provide context to this research project before detailing the research aim and objectives. Some background information on phytoplankton and remote sensing is provided, with a special emphasis placed on their importance and why knowledge of their size distributions is of special interest in ocean ecology.

1.1.1 About Phytoplankton

Phytoplankton, also known as microalgae, is an extremely diverse set of photosynthetic microorganisms adapted to living in open water [1]. They are predominantly single-celled and are ubiquitous in almost all aquatic habitats, typically confined to the surface layer where there is an abundance of sunlight for photosynthesis [2]. Although the exact number of phytoplankton species that exist is unknown, with estimates varying wildly, it is currently estimated that between 30 000 and 100 000 species may exist [3], [4]. Figure 1 illustrates the extreme structural diversity found within phytoplankton species.

Figure 1: A selection of phytoplankton species illustrating extreme diversity (not to scale) [5].

In addition to their structural diversity, phytoplankton species span a massive range of sizes, from the smallest unicellular cyanobacteria at around ~1μm^3 to the largest microcystis having been recorded at ~10^9μm^3 [1]. Since phytoplankton cannot be seen with the naked eye, it's difficult to fully appreciate how large nine orders of magnitude really is. Finkel et al. [6] have managed to illustrate this by putting it into perspective, as can be seen in Figure 2.

Due to the sheer number of species and the large biochemical diversity that exists within them, it has been convenient and practical to group phytoplankton into functional

types (PFTs). PFTs are an arbitrary grouping of species, typically based on their biochemical functions (nitrogen fixers, calcifiers, silicifiers, dimethyl sulphide producers [DMS]) or size [7].

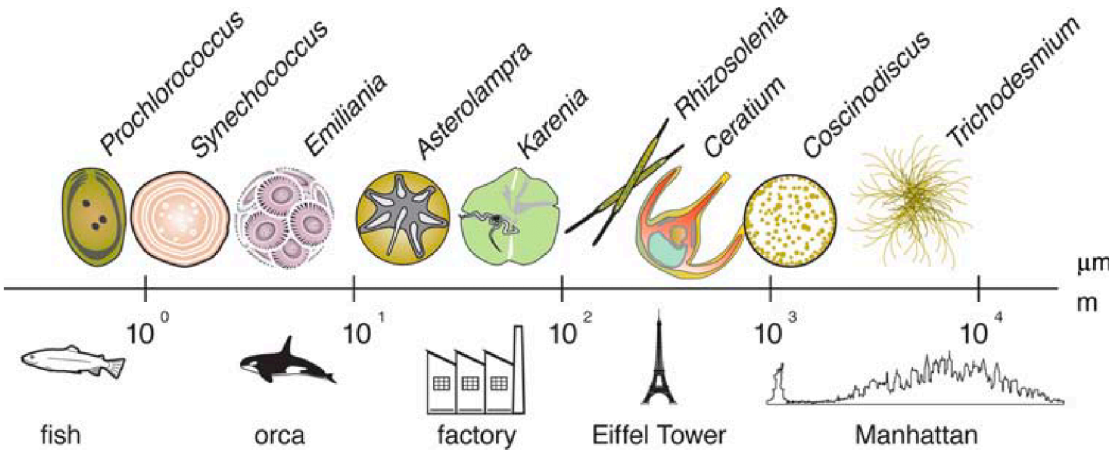


Figure 2: Phytoplankton size difference put into perspective [6].

1.1.2 Phytoplankton Importance

Interest in phytoplankton has gained momentum in the past two decades with the realisation of their role in the global carbon cycle. Despite only accounting for 1-2% of the total global biomass, phytoplankton are responsible for fixing some 50 gigatons of carbon into organic material through photosynthesis each year [8]. This process is known as primary production and the magnitude of this process is on the same order as the global total net production by terrestrial plants [9]. Figure 3 illustrates how the carbon is absorbed from the atmosphere, via photosynthesis, and transferred to higher trophic levels and sequestered in the deep ocean.

As atmospheric CO_2 levels increase so too does the amount of dissolved CO_2 in the ocean. In response to this, phytoplankton have been shown to increase their carbon uptake, thereby serving as a dampener of the global greenhouse effect [10]. Not only is phytoplankton responsible for producing roughly half of the oxygen on our planet, these microorganisms form the basis of the marine food chain as almost everything in the ocean, either directly or indirectly, feeds off of them.

In order to understand the driving forces behind phytoplankton, we need to better understand the communities or functional types (PFTs) that comprise phytoplankton,

each of which differ in a variety of ways, with distinct groups affecting the carbon cycle differently [11].

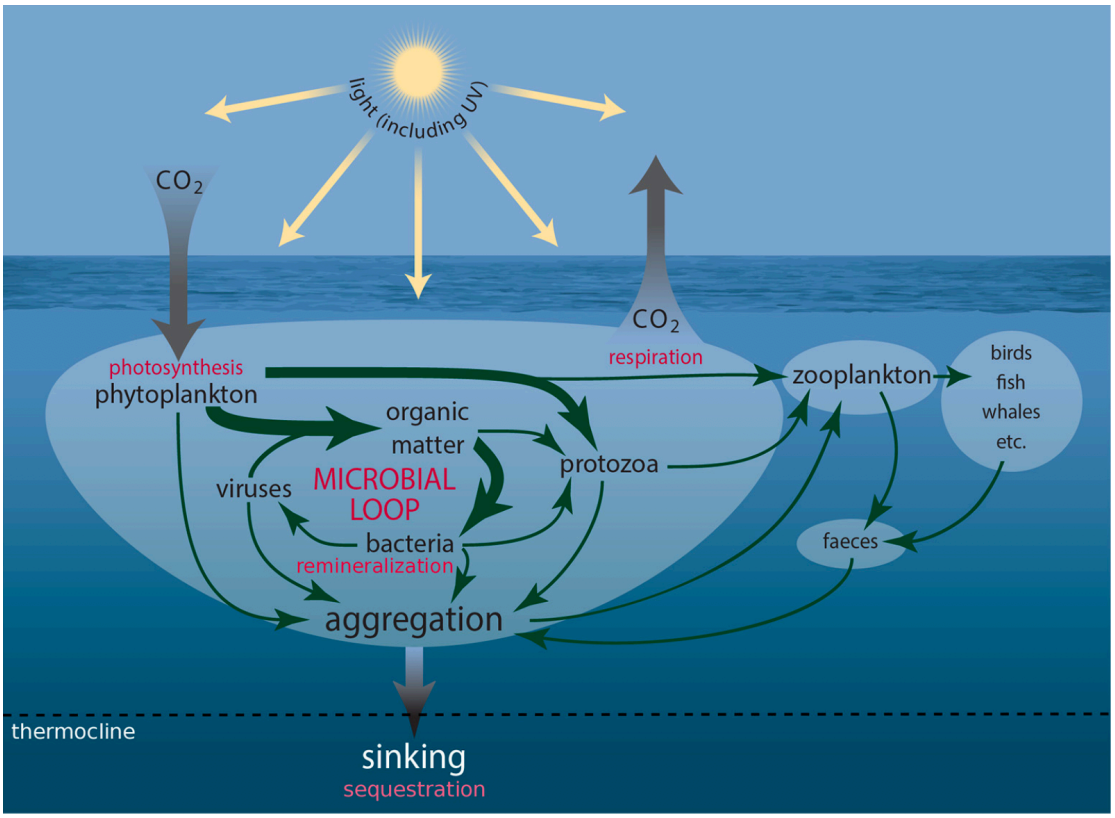


Figure 3: Microbial food web illustrating the transfer of carbon [12].

1.2. Research Motivation

Remote sensing of PFTs is an emerging field with high ecological importance. With an increase in the amount of publicly available data, possibilities have emerged for machine learning techniques to be applied for better modelling of these complex systems. It is my personal interest in machine learning, coupled with the global importance in understanding phytoplankton size structures, that has motivated me to focus this research into the investigation of the relationships between absorption and the size structures of phytoplankton.

1.2.1 Why knowledge of Size is Important

Given that phytoplankton are such an important role player in regulating greenhouse effects, and that they form the basis of the food chain, it is vital that we understand how they respond to changes in the environment. It is therefore unsurprising that there has been an increase in the number of ocean colour algorithms developed [7]. A large

number of these algorithms have been dedicated to working with PFTs due to the practicalities offered in working with these groupings [13].

The boundaries of these PFTs can be relatively arbitrary, but with increased academic attention the definitions of these functional types are starting to standardise. As the boundaries and definitions of functional types become clearer, inter-disciplinary research becomes more comparable.

One such PFT that emerged relates to the size of the phytoplankton, as knowledge of phytoplankton size is an aspect of particular importance in oceanographic science due to the sheer number of biochemical and ecological processes that it influences [14], [15]. These processes include, but are not limited to:

- **Photosynthesis Efficiency:**
  As a cell increases in volume, the surface area to volume ratio decreases and as a result the cell becomes less efficient at converting sunlight into energy. Numerous studies have shown that chlorophyll-specific productivity such as photosynthesis is inversely proportional to the cell size [16].

- **Sinking Rate:**
  Phytoplankton need to live in the epipelagic zone (the uppermost layer of the ocean) in order to convert sunlight and CO_2 into energy, but since they have no means of locomotion they will ultimately sink to the ocean floor, thereby sequestering significant amounts of carbon. Understanding phytoplankton sinking rates is therefore of vital importance for understanding carbon fluxes in the ocean [17]. Particle sinking rates have been shown to obey Stokes' law, where the sinking velocity is proportional to the radius squared [18].

- **Trophic Interactions and Food Web:**
  Phytoplankton is at the bottom of the food chain and as a result there are a number of zooplankton and fish species that rely on them as a source of food. Typically larger predators will eat larger phytoplankton prey [19], and it has also been found that these predators can be selective in their feeding preferences, preferring phytoplankton of certain sizes and types [20], [21].

1.2.2 Defining the Causal Relationships

In order to understand how knowledge of phytoplankton size can be gained, the relationships between the optical properties of the sample and the underlying particle size information need to be determined. Figure 4 defines some of the relationships between the constituents of a water sample and ultimately their optical properties.

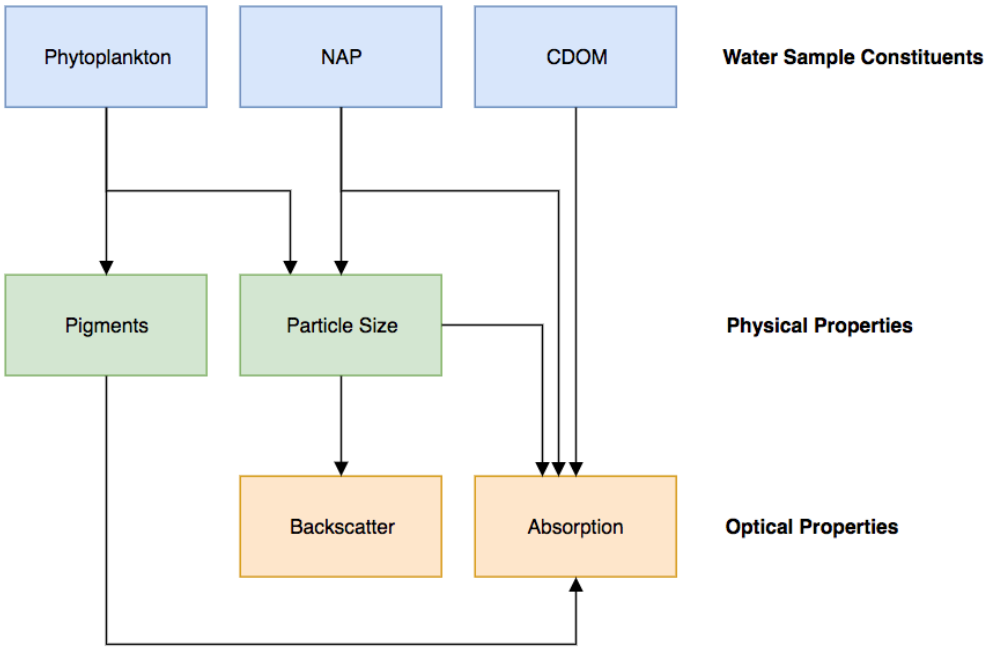


Figure 4: Causal relationships describing the optical properties of a water sample.

It can be seen that multiple factors influence the absorption signal, and that both direct and indirect relationships between the absorption signal and the phytoplankton cell size exist. Both Non-Algal Particles (NAP) and Colour Dissolved Organic Matter (CDOM) directly influence the absorption signal and therefore influence any model trying to estimate particle size. If they are not accounted for directly, they will contribute to the overall noise and error of the model. Other than absorption, the other optical property that is influenced by particle size is backscatter. Models that incorporate both backscatter and absorption should perform better at estimating particle size (both phytoplankton and other particles), as compared to a model only utilising absorption data.

1.3. Hypothesis

Given the optical absorption signal of a water sample containing phytoplankton cells, the size distribution of those cells can be estimated with reasonable accuracy.

1.4. Aim

To estimate the percentage of phytoplankton present, per size class (Pico, Nano and Micro), from their optical absorption spectra through the use of multiple statistical models.

1.5. Objectives

1. Develop a model to estimate phytoplankton size class ratios through the use of absorption basis vectors and matrix factorisation.
2. Develop a model to estimate phytoplankton size class ratios through the use of estimated pigment concentrations.
3. Develop a model to estimate phytoplankton size class ratios through the use of an empirical parameterisation technique.
4. Develop an ensemble model, as a composition of the other models, to estimate phytoplankton size class ratios.

1.6. Overview of Methods

Four separate models were developed, each exploiting a different relationship between the phytoplankton cell size and their optical properties. All of the models are developed against data that is labelled through the use of Diagnostic Pigment Analysis. The first model estimates the cell size concentrations by factorising the absorption signal into three basis vectors, one per size class, and then calculates the relative abundances using these basis vectors.

The second model decomposes the absorption signal into a set of 12 Gaussian bands, which are then used to estimate pigments and pigment compositions through non-linear regression. The cell size concentrations are then estimated from the derived pigment concentrations through the use of Support Vector Regression.

The third model is a parameterised approximation of the absorption signal, where the parameters are regressed against the known cell size concentrations and the final model is an ensemble model, combining the first three models via Support Vector Regression.

1.7. Apparatus

All of the models were developed in Python 2.7, with the use of a Postgres 9 database, running on a 2015 MacBook Pro 2,2 GHz Intel Core i7 with 16 GB of 1600 MHz DDR3 RAM. The following python libraries were used:

1. Scikit-learn
2. Numpy
3. Matplotlib
4. iPython Notebook.

1.8. Thesis Structure

Chapter 2 provides a review of the relevant literature needed to provide context for this work. A broad overview of the ocean-optics research landscape, including a history of remote sensing and what is currently achievable is provided. This includes an introduction to both ocean colour algorithms as well as more general machine learning techniques. The final section in this chapter discusses dimensionality reduction and parameterisation techniques, why they are important when working with high-dimensional data, and how they have previously been applied to oceanographic datasets.

Chapter 3 deals with the collection and preparation of the datasets used in this research. A breakdown of the various types of data (optical, size, pigment) is provided along with details on how this data was collected or acquired. Details are supplied on how the datasets were prepared, along with the final tally of usable data. This chapter also provides a cursory analysis of the underlying structure of the data. Finally, the size class labels are calculated, through the use of Diagnostic Pigment Analysis (DPA), so that they can later be used in model development.

Chapter 4 represents the main methodology of this paper, where four separate models are developed for the estimation of phytoplankton size class (Pico, Nano and Micro), from absorption coefficients. Each model has its own aim, method, data and results. At the end of the chapter, all of the results across the different models are tabulated, along with a table of other published results for comparison.

Chapter 5 provides a summary and an overview of the results obtained in this thesis along with suggestions and recommendations for possible improvement.

Chapter 2: Literature Review

This chapter provides a background to ocean colour algorithms as well as more general machine learning techniques that have previously been used in the modelling of phytoplankton size classes. A brief introduction to the field of ocean-colour research is provided before some of the main techniques are discussed. These techniques focus more on the biological and physical relationships between phytoplankton and their optical properties. The chapter then moves on to discussing machine learning, both in general and how it has been used in ocean colour algorithms, along with some technical underpinnings of the more common techniques. The final section in this chapter discusses dimensionality reduction and parameterisation techniques, why they are important when working with high-dimensional data, and how they have previously been applied to oceanographic datasets.

2.1. Ocean Colour Algorithms

The colour of seawater is directly attributed to the relative concentrations of the optically active constituents present in its uppermost layers [22]. This knowledge allowed researchers to investigate whether these optical properties could be exploited such that information on the underlying constituents could be extracted. The first proof-of-concept device attempting to perform such tasks was developed in the early 1980s and was called the Coastal Zone Colour Scanner (CZCS) [23]. The creation of this, and other such devices, marked the start of the ever-growing field of ocean-colour research. These initial approaches focused on calculating the concentration of Chlorophyll-a (used as a proxy for biomass) present in the sample.

It was soon discovered that one of the most practical methods for monitoring and studying ocean colour was through the use of satellite imagery, as this approach provided the highest spatial and temporal resolution for measuring the optical properties of the ocean's surface layer [7]. Since then, a number of approaches have been developed for determining phytoplankton community composition from satellite imagery. Remote sensing algorithms, however, cannot be developed and validated without in situ measurements for comparison and validation.

Before any models can be developed the manner in which light behaves, and is influenced by both water and the atmosphere, needs to be understood and accounted

for. The atmosphere alone can account for more than 90% of the radiance signal measured by satellite-borne radiometers [24].

Once the data from the radiometer has been adjusted to account for atmospheric effects, the primary output is either remote sensing reflectance $R_{rs}(\lambda)$ or normalised water-leaving radiance $L_w(\lambda)$. These are known as Apparent Optical Properties (AOPs), which can be related to the physical constituents of the water through a series of inversions and parameterisations as shown in Figure 5. Where AOPs describe the way that light moves through the atmosphere, Inherent Optical Properties (IOPs) describe how light interacts with particles inside the water and, as such, cannot be measured directly from the satellite. The AOPs are typically converted into IOPs, such as absorption and backscatter, so that they can more easily be compared and modelled against in situ measurements.

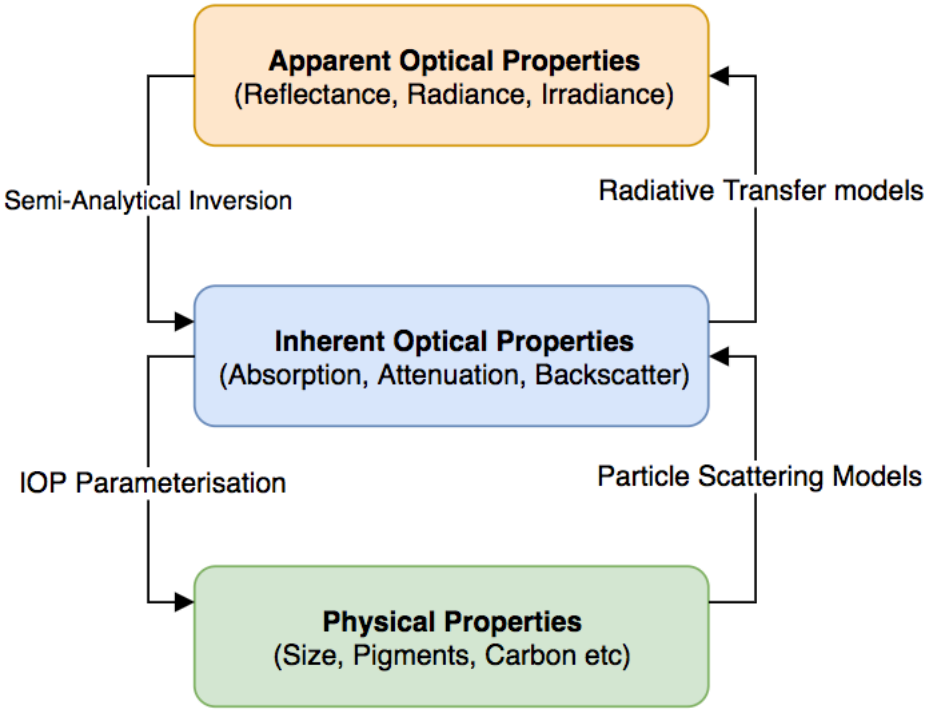


Figure 5: Relationship between the physical constituents in the water and how they are related to their optical properties, both in the water (IOPs) and in the atmosphere (AOPs).

All of these models, which rely on satellite ocean-colour sensors, are limited by the spectral resolution of these sensors, and the current satellites (SeaWiFS, MODIS and MERIS) are limited to a relatively small number of wavelengths [25]. This can make it particularly difficult to detect the small changes in the spectral slope that are required to differentiate between PFTs.

The sensors available for measuring the absorption of in situ samples provide far higher resolution than the satellite-based sensors. One such sensor, used in two of the datasets of this research, known as the spectral absorption and attenuation sensor (AC-S), developed by Sea-bird Scientific, measures absorption and attenuations from 400nm – 730nm at a resolution of 4nm [26]. The AC-S has a flow through system for water to be continuously pumped through the instrument, making it particularly well suited to collecting samples on a ship.

2.1.1. Overview of Relevant Techniques

Table 1 represents a selection of techniques that are used to estimate accessory pigment concentrations as well as phytoplankton size classes and distributions. It is important to note that these case studies only represent a small fraction of all the bio-optical models that have been developed, and that many combinations of IOPs and AOPs have been used to estimate the physical and chemical properties of the ocean.

Table 1: Relevant case studies grouped by model output for a number of input types.

Input	Output	Case Study
Absorption	Accessory Pigments	Chase et al., 2013 [27]
Radiance		Bracher et al., 2015 [28]
Radiance	Phytoplankton Size Classes (Pico, Nano, Micro)	Varunan and Shanmugam, 2015 [29]
Absorption + Radiance		Wang et al., 2015 [15]
Absorption + Pigments		Brito et al., 2015 [30]
Absorption		Zhang et al., 2015 [31]
Chl-a		Uitz et al., 2006 [32]
		Brewin and Hirata, 2011 [33]
		Hirata et al., 2011 [34]
		Brewin et al., 2010 [25]
Absorption + Chl-a		Wang et al., 2013 [35]
Backscatter	Phytoplankton Size Distribution	Slade and Boss, 2015 [36]
Backscatter		Kostadinov, Siegel and Maritorena, 2009 [37]
Radiance		Kostadinov, Siegel and Maritorena, 2010 [38]
Carbon		Kostadinov et al., 2016 [13]

2.2. Estimating Phytoplankton Size Classes

Phytoplankton range nine orders of magnitude in size and any given ocean sample could contain a mixture of many species. Trying to determine the exact size distribution of all phytoplankton present in the sample is a difficult and impractical task. It is therefore common practice to estimate the relative contributions per size class. These classes are, however, not very well defined and slight variations exist [39]. One of the more common categorisations is the three-size class model (Pico, Nano, Micro), as used by Uitz et al. [40], with the boundaries defined as Pico ≤ 0.2μm, Nano 2 – 20μm and Micro ≥ 20μm.

2.2.1. Chlorophyll-a Based Methods

Chlorophyll-a is the dominant pigment found inside phytoplankton and has distinct absorption peaks at roughly 440nm and 675nm. Due to its influence on the absorption signal it has been widely used as a means of deriving phytoplankton size structure [25], [32], [33]. These methods typically work on the assumption that Chlorophyll-a concentration is proportional to cell size, i.e. larger cells are typically dominant in waters with high Chlorophyll-a concentration. This assumption, however, is not always valid in local regions such as the East China Sea [15].

Methods based on Chlorophyll-a alone are very rudimentary and while their simplicity gains them some merit, especially in estimating the total biomass present, their ability to differentiate between size classes is very limited. These methods are still important since many oceanic cruise datasets contain Chlorophyll-a without any other incidental data which could be used to estimate phytoplankton size information.

2.2.2. Accessory Pigments Based Methods

Phytoplankton contain many different accessory pigments that they use in photosynthesis. These pigments are typically measured through a process known as High Performance Liquid Chromatography (HPLC). HPLC is the process of passing a liquid sample through a solid adsorption material, which interacts with each of the constituent particles (pigments in this case) differently, thereby causing different flow rates through the adsorption material [41]. By knowing how long each pigment takes to pass through the adsorption material allows the absorption reading to be associated with the corresponding pigment. Most of the pigments can be detected at 450nm, while Chlorophyll-a and its derivatives can be detected at 667nm and Bacteriochlorophyll-a at 770nm [42]. Figure 6 shows an example chromatogram from one of the samples

taken by [42], showing how the absorbance at 450nm changes over time as a result of different pigments taking longer to pass through the adsorption material.

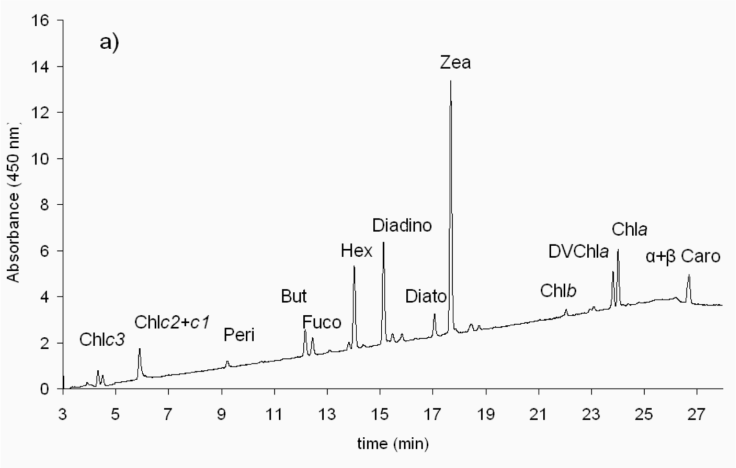


Figure 6: HPLC chromatograms from a surface water sample measured at 450nm, taken from [42].

Figure 7 shows the absorption spectra of some of the pigments found within phytoplankton, and it can be seen why 450nm in HPLC would be able to detect most of the pigments due to the large absorption overlap. The absorption of phytoplankton is normally only measured over the photosynthetically active range (PAR) between ~400nm – 700nm, as light waves longer than 700nm are mostly absorbed by the water itself and do not penetrate the water column, and wavelengths in ultra violet (UV) spectrum do not play a large role in photosynthesis [43]. Phytoplankton do still absorb light within the UV spectrum but excessive exposure can lead to photoinhibition and cell damage [12].

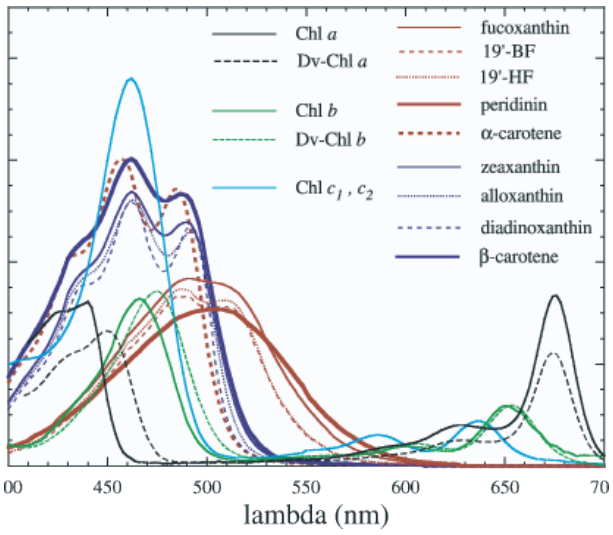


Figure 7: Absorption spectra of the different accessory pigments in solution, taken from [44].

Certain diagnostic pigments only occur in certain species and these species may fall into specific size ranges, therefore knowledge of the relative concentrations of these pigments can be used to infer size information. It was Vidussi et al. [45] who first estimated phytoplankton size fractions from diagnostic pigments, a technique now known as Diagnostic Pigment Analysis (DPA). Table 2 shows the seven diagnostic pigments used and to which taxonomic group and size class they belong.

Table 2: Taxonomic pigments used by [45] to estimate size classes.

Pigments	Taxonomic Significance	Size (µm)
Zeaxanthin	Cyanobacteria and Prochlorophytes	< 2
Divinyl Chlorophyll-a	Prochlorophytes	< 2
Chlorophyll-b + Divinyl Chlorophyll-b	Green flagellates and Prochlorophyll	< 2
19'-Hexanoyloxyfucoxanthin	Chromophytes Nanoflagellates	2 – 20
19'-Butanoyloxyfucoxanthin	Chromophytes Nanoflagellates	2 – 20
Alloxanthin	Cryptophytes	2 – 20
Fucoxanthin	Diatoms	> 20
Peridinin	Dinoflagellates	> 20

DPA was further improved by Uitz et al. [32], where specific weights were associated with each of the diagnostic pigments. This technique has undergone minor adjustments since then and has been widely adopted by the ocean science community [25], [31].

Even though DPA is faster than measuring phytoplankton sizes via microscopy, it is still a relatively slow process, as samples need to be collected and pigments need to be extracted and measured in a lab via HPLC. DPA also suffers from both temporal and scale coverage limitations, as the results are only applicable to the area where the samples were taken. It is because of these shortcomings that researchers have developed absorption-based alternatives, where the optical properties of the water can be measured either by continuous flow through systems like the AC-S, or through the use of satellite imagery. From these absorption measurements the individual pigment concentrations cannot be accurately calculated, as diagnostic pigments have overlapping absorption spectra. Without being able to reliably estimate diagnostic pigment concentrations from absorption, DPA cannot be used directly in absorption-

based models. These absorption-based methods, discussed further in section 2.2.3, typically exploit some other optical property in order to estimate cell size, but will still use data that is labelled via DPA, to be trained against.

2.2.3. Absorption Based Methods

Phytoplankton are full of pigments, each absorbing light at some range of wavelengths, as shown earlier in Figure 7. When adding up the absorption spectra from each of these pigments, the combined and complete pigment spectrum is obtained. This spectrum is then distorted, in what is known as the packaging effect, by the fact that pigments are found within packets called chloroplasts and that the chloroplasts are packaged inside the cell. This packaging effect is highly dependent on cell size and results in a flattening of the spectrum [46]. Ciotti et al. [47] found that 80% of the variability in the spectral shape could be attributed to the size of the phytoplankton in the sample. It is therefore this spectral flattening that most absorption-based models revolve around. Phytoplankton cells are not the only particles in the water. Sediment and other detrital matter referred to as “colour dissolved material” (CDM) also absorb light within the water. Figure 8 show the typical absorption spectra for the various size classes and CDM, where the solid lines represent the hyperspectral results estimated by Uitz et al. [40] and the dashed lines represent optically weighted [48] specific absorption at Sea-viewing Wide Field-of-view Sensor (SeaWiFS) bands, within the first optical depth.

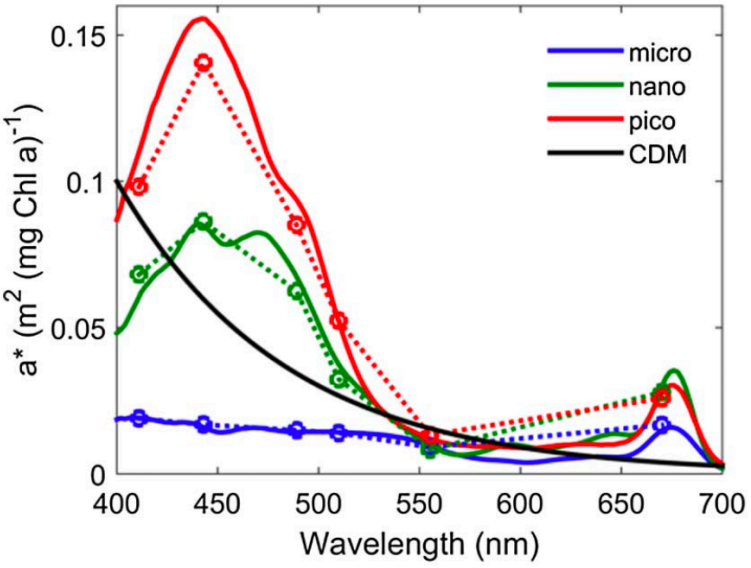


Figure 8: Absorption basis vectors for three size classes (Pico, Nano, Micro), where the solid lines represent the hyperspectral basis vectors calculated by Uitz *et al.* [48] and the dashed lines connect points that represent optically weighted specific absorption at SeaWiFS bands [31].

The approach taken by Zhang et al. [31] was to make use of pre-calculated basis vectors, as shown by the solid lines in Figure 8, to estimate the total contributions of these basis vectors in the given absorption signal. The contributions or weightings of these basis vectors would indicate the total contributions per size class present. Singular Value Decomposition (SVD) was used as the method for factorizing the signal matrix into its constituent weightings, and good results were obtained. This technique was used in this thesis in Model 1 of Chapter 4, both to calculate basis vectors and to estimate size contributions from an absorption signal.

Wang et al. [15] employed a different technique for estimating size classes from the absorption signal. Instead of using the absorption basis vectors representing Pico, Nano and Micro, they made use of the first three principal components, calculated via Principal Component Analysis (PCA), and related these to size classes. This model was further evaluated by deriving the phytoplankton specific absorption $a_{ph}(\lambda)$ from reflectance $R_{rs}(\lambda)$, with good results. Even though this technique worked well, it will probably not be as widely adopted as the basis vector approach performed by Zhang et al. [31] due to the fact that principal components are not as meaningful as basis vectors, which relate directly to the absorption spectra of that size class. Zhang's technique is something that is more tangible for biologists to understand, and something that can be refined independently.

These techniques describe how pigment or absorption information has previously been used to provide information on the size distribution of phytoplankton. In the following section a more general background on machine learning techniques is provided, along with some of the technical details that underpin the more common techniques.

2.3. Machine Learning Techniques

Running statistical techniques on a computer, with the aim of making predictions on data, is known as Machine Learning [49]. The rise in computational abilities coupled with the high availability of data has led to machine learning being found in many fields and industries [50].

Typically, machine learning techniques are employed to estimate some continuous value ("regression") or some categorical value ("classification"). Many algorithms exist for making these estimations / classifications and choosing the correct algorithm is dependent on a number of factors such as the availability of data (both labelled and

unlabelled), whether the number of categories is known, and the number of dimensions in the dataset [51].

Put simply, machine learning techniques try to quantify the relationship between a number of input features (independent variables) and some desired output (dependent variables). The relationship between the independent and dependent variables may be linear or non-linear, and the correct algorithm needs to be selected accordingly. In certain scenarios it is very difficult to know upfront what the relationship is, and different algorithms will need to be evaluated in order to see what works best in the given context.

In order to determine how well a particular model is performing, a contextually-appropriate error metric needs to be defined. The measures and consequences of a model's success are highly dependent on the context. For example, making an incorrect medical diagnosis is far worse than serving an irrelevant advert on a website. Accuracy, however, is not the only characteristic evaluated when choosing the correct model. There are other factors, such as the time it takes to train the model, the rate at which the model will need to be retrained, and whether the model is capable of being retrained incrementally with the introduction of new data.

Many machine learning techniques and models have been used extensively within oceanography to estimate the synoptic distribution of phytoplankton cell size from their measured optical properties. For example, Brewin et al. made use of Artificial Neural Networks (ANN) [52], Li et al. made use of a SVR [53], and Belgiu and Drăguţ made use of Random Forests (RF) [54]. Hu et al. [55] do an in-depth analysis of all of these techniques and show that RF based methods perform the best, closely followed by SVR and then ANNs, given their datasets and model configurations. It is for this reason that a more technical explanation of the internal mechanics of ANNs, SVR and Random Forests will be provided.

2.3.1 Artificial Neural Networks

ANNs encompass a large class of learning models and algorithms that were originally inspired by biological neural networks [56]. An ANN is comprised of layers of interconnected artificial neurons, where each layer is connected to the next through a weight matrix. These weights, coupled with some non-linear activation function are what define the threshold for a given artificial neuron output [51], as shown in Figure 9.

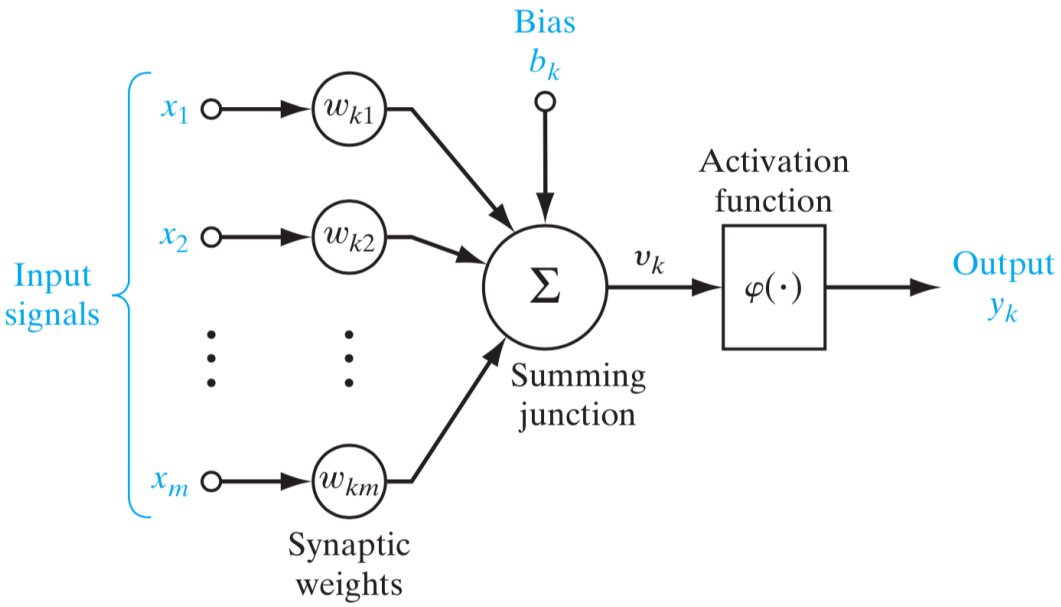


Figure 9: Non-linear model of a neuron labelled *k* [56].

Without the non-linear activation function the artificial neuron would be nothing more than a linear regression model, and the entire network would be reducible to a simple linear model. This activation function, often the sigmoid or relu functions, allow the ANN to capture complex non-linear relationships within the data. Training is the process of finding synaptic weights that improve the accuracy of the ANN. One such approach is known as back-propagation, where the synaptic weights are iteratively updated for each of the layers.

The ANN in Figure 9 represents a simplistic feed-forward neural network, otherwise known as a perceptron [57]. One of the challenging tasks in the design of an ANN is determining the number of hidden layers as well as the number of nodes in each layer. The greater the number of layers, the greater the complexity the network is capable of representing, but so too is the amount of data required and the amount of time it takes to train.

Many architectures of ANN exist, for example, Convolution Neural Networks (CNNs) are a class of ANN that are made up of feature extractors and classifiers, and are well suited to the task of 2D and 3D image recognition [58].

Another popular class of ANN is the Recurrent Neural Network (RNN), which allows the output to be fed back into the input of the model. A fundamental difference between

RNNs and classical ANNs is the ability of an RNN to process arbitrary length inputs, and to generate arbitrary length outputs [58]. This feedback mechanism gives the model a form of memory making it well adept to tasks such as language translation, where the meaning of a word in a sentence may be contextual.

The Generative Adversarial Network (GAN) is also a widely used architecture that consists of a generator model and a discriminator model. The generator model captures the distribution of the dataset and generates an output that is representative of the input distribution. The discriminator model estimates the probability that the input came from the original dataset or from the generator. The training function of the generator is such that it is trying to maximize the probability of the discriminator making a mistake [59]. GANs are used in a wide variety of applications including dataset generation, image upscaling and image generation.

The ANN implemented by Hu et al. for estimating phytoplankton size classes, only had 1 hidden layer with 10 artificial neurons and performed reasonably well [55].

2.3.2 Support Vector Regression

SVR is a subset of the models known as Support Vector Machines (SVMs), which themselves are a set of supervised learning techniques that can be employed for both classification and for regression [60]. SVMs attempt to find a hyperplane that best separates the data into two classes. The support vectors are the points that fall closest to the hyperplane in each class and if removed would alter the position of the hyperplane. They are thus important elements in the dataset. The goal of an SVM is to maximise the margin between the hyperplane and the nearest point from each class, as shown in Figure 10.

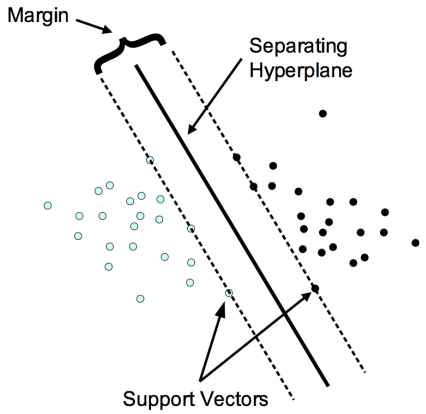


Figure 10: SVM maximizing the margin between support vectors of two classes [60].

In order to solve the regression case where a continuous output is required, SVM is extended. Given a dataset $\{X = [x_i, y_i] \in \Re^m \times \Re, i = 1 \ldots n\}$, the linear function can be represented as:

$$f(x) = \langle w, x \rangle + b, \qquad w \in X, b \in \Re, \tag{2.1}$$

where w represents a weight vector, b represents a bias value and $\langle \cdot, \cdot \rangle$ denotes the dot product in X. Certain constraints are then added to the linear function, where it is expressed as the following convex optimisation problem:

$$\text{minimise} \qquad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{2.2}$$

$$\text{subject to} \qquad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}, \tag{2.3}$$

where $\|w\|^2$ represents the Euclidean norm, a term added to minimise overfitting, $C > 0$ is a constant which helps define the ratio between the "flatness" of $f(x)$, where ε defines the radius of the tube within which the regression function must lie, and where tolerances of no more than ξ are allowed. The error in approximation is then measured using Vapnik's so-called ε-intensity loss function, described by:

$$|\xi|_\varepsilon := \begin{cases} 0, & if\ |\xi| \leq \varepsilon \\ |\xi| - \varepsilon, & otherwise. \end{cases} \tag{2.4}$$

This is visually depicted in Figure 11 where only values outside of the shaded area contribute to the loss.

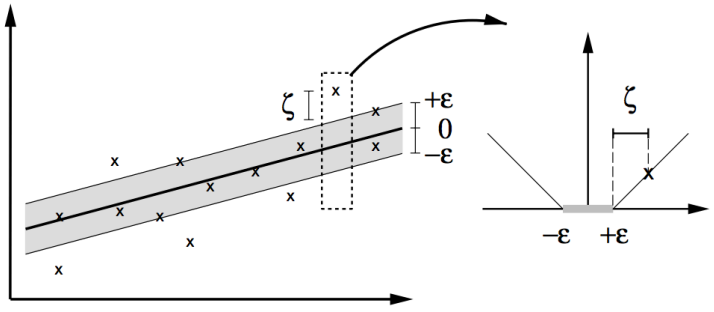


Figure 11: Soft margin loss for a linear SVR [60].

This quadratic optimisation problem can be solved through the use of Lagrange multipliers such that the linear function can be shown to be:

$$f(x) = \langle w, x \rangle + b = \sum_{i}^{n} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b. \quad (2.5)$$

This is known as the Support Vector expansion, where α_i, α_i^* are introduced parameters which need to be learned, and the number of support vectors is equal to the number of non-zero α_i and α_i^*.

In order to model non-linear relationships, the linear function in Equation 2.5 can be modified through what is known as "the kernel trick", where a non-linear kernel is substituted into the equation. The following represents the generalised kernel equivalent:

$$f(x) = \sum_{i}^{n} (\alpha_i - \alpha_i^*) k(x, x') + b, \quad (2.6)$$

where $k(x, x')$ represents the kernel of choice. One such example of a suitable kernel is the polynomial mapping:

$$k(x, x') = (\gamma \langle x, x' \rangle + r)^d. \quad (2.7)$$

Another popular kernel, and the one that is used in this research, is the Radial Bias Function (RBF):

$$k(x, x') = e^{-\gamma \|x, x'\|^2} \quad (2.8)$$

This ability to add in non-linear kernels makes SVM and SVR very powerful. Their flexibility through kernel choice coupled with the fact that they have very few hyperparameters and are guaranteed to reach a global optimum [61], make them a very popular choice in many machine learning contexts.

2.3.3 Random Forests

Random Forests are another popular machine learning technique, which combine bagging and decision trees and can be used for both regression and classification-

based problems. Multiple decision trees, each of which have been given some subset of the data, are combined and averaged in order to get a more stable and accurate prediction. This accuracy and robustness, however, does comes at the expense of interpretability [62].

A decision tree has a number of nodes, each of which perform a test on an attribute and dictates which branch needs to be followed. There is no requirement for the tree to be balanced, although balance typically leads to fewer decisions that need to be made, with the most significant tests performed first. The leaf node of the tree represents the outcome of the decision tree.

The Random Forest algorithm works by creating *N* decision trees, where each tree is given some subset of the available training data, typically around 80% (although this is tuneable). This sampling is "with replacement", so each decision tree will have some overlapping training data. When constructing a tree, instead of searching for the most important feature while splitting each node, the best feature is selected from a random subset of the features. This results in far greater diversity between each tree, and generally results in better overall performance [62].

2.4. Dimensionality Reduction

Working with high dimensional data can introduce a number of problems when trying to develop statistical models. This phenomenon was first described by Richard Bellman, who called it "the curse of dimensionality" [63]. As the number of parameters describing the input space increases, the volume of the possible output space can grow exponentially, requiring an ever-increasing amount of data for a model to have any statistical significance. In other words, the sparsity of the data increases with the number of dimensions. Another problem that arises from high dimensionality is that it becomes increasingly difficult to find a meaningful distance metric. When using something like Euclidean distance, for example, the distance between pairs of samples becomes increasingly similar as you increase the dimensionality. Even if you managed to find enough data to make your model statistically significant and you were able to define a meaningful distance metric for training, the computational overhead required to train your model might render it intractable. It is therefore common practice to perform some sort of dimensionality reduction on the dataset, converting it from a high dimensional space to a lower one while trying to maintain as much of the relevant information as possible.

All dimensionality reduction techniques that are effective on a given dataset rely on an underlying assumption that there is redundancy in the given dimensions. That is to say that the dataset is expressed in more dimensions than is necessary to fully represent it, implying that some mutual information exists between the various dimensions. One way of measuring this joint variability, for linearly-related dimensions, is by calculating the covariance between each dimension:

$$COV(x,y) = \sum_{i=0}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \ , \tag{2.8}$$

where x_i and y_i are the two variables (dimensions) being compared, with $\bar{x}$ and $\bar{y}$ their respective means and n the total number of sample points. This can be done for all sets of points and put into what is known as a covariance matrix. This matrix is symmetrical around the principal diagonal, along which the variance of each variable is contained. Even if there is a low covariance between variables, there might still exist some non-linear relationship between them that can be exploited for dimensionality reduction.

2.4.1 Junge Slope

The most commonly used technique for parameterizing marine particle size distributions for optical modelling is the power law or Junge distribution [14], [64] as defined by the following equation:

$$N(D) = N_0 \left(\frac{D}{D_0}\right)^{-\xi}, \tag{2.9}$$

where $N(D)$ (m^{-4}) is the concentration of particles at diameter D (m), and N_0 (m^{-4}) is a reference concentration of particles at the reference diameter D_0 (typically 2μm) and ξ is the differential slope parameter.

Unlike PCA, these results are a little bit more comparable between different datasets. This, coupled with the simplicity of the equation, makes it rather appealing to marine scientists. It was only later discovered that phytoplankton have complex shapes with variable size distributions, and that the inherent bias in the Junge slope severely limits its ability to represent anything more than a monospecific phytoplankton population [64].

I applied a Junge slope fit to the available particle size data from the ACE cruise so that its performance could be evaluated. The coefficient of determination (also known as R^2) was used as a metric for defining how well the approximated particle size distribution (PSD) matched the original. Here R^2 is defined by the following:

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.10}$$

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 \tag{2.11}$$

$$SS_{tot} = \sum_i (y_i - \bar{y}_i)^2, \tag{2.12}$$

where y_i represents the original signal, $\hat{y}_i$ the estimated signal and $\bar{y}_i$ the mean of the original signal.

The final R^2 value obtained when applying the Junge slope to the particle size data was 0.64. This implies that 64% of the variation from the mean is explained by this model. Figure 12 shows an example where the Junge slope fits the data reasonably well and another example where the approximation is poor.

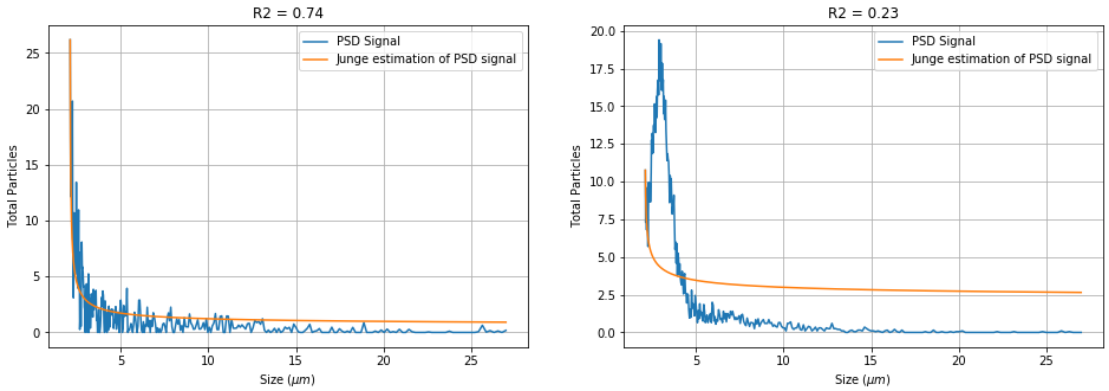


Figure 12: A comparison of a reasonably good (left) and bad (right) Junge slope fit from random size samples from the ACE cruise.

2.4.2 Effective Diameter

Another technique used for parameterising particle size is to take the ratio of particle volume to particle area. This metric is known as the Effective Diameter (D_{eff}) and is a

technique that was originally used by atmospheric physicists [65]. The quantity D_{eff} can be calculated as follows:

$$D_{eff} = \frac{\int_2^{50} \frac{\pi}{6} d^3 F(d)}{\int_2^{50} \frac{\pi}{4} d^2 F(d)}, \tag{2.13}$$

where d is the particle diameter and $F(d)$ is the number of particles per unit of volume. The integral is taken between the diameters of 2μm and 50μm.

This technique is not meant to be an approximation of the PSD, but rather a metric that is useful in describing the optical properties of the particle mixture. This technique has been successfully used to infer relationships between the optical properties of phytoplankton and a range of size distributions [64]. It is therefore not possible to go back from D_{eff} to a size distribution.

2.4.3 Principal Component Analysis

Principal Component Analysis (PCA) is a popular factorisation technique that has been used in many oceanographic studies [40], [15], [66]. PCA works by projecting the data onto a set of orthogonal basis vectors that point in the direction of largest variance [67]. This has been illustrated in Figure 13, where the long arrow represents the first principal component and the short arrow the second.

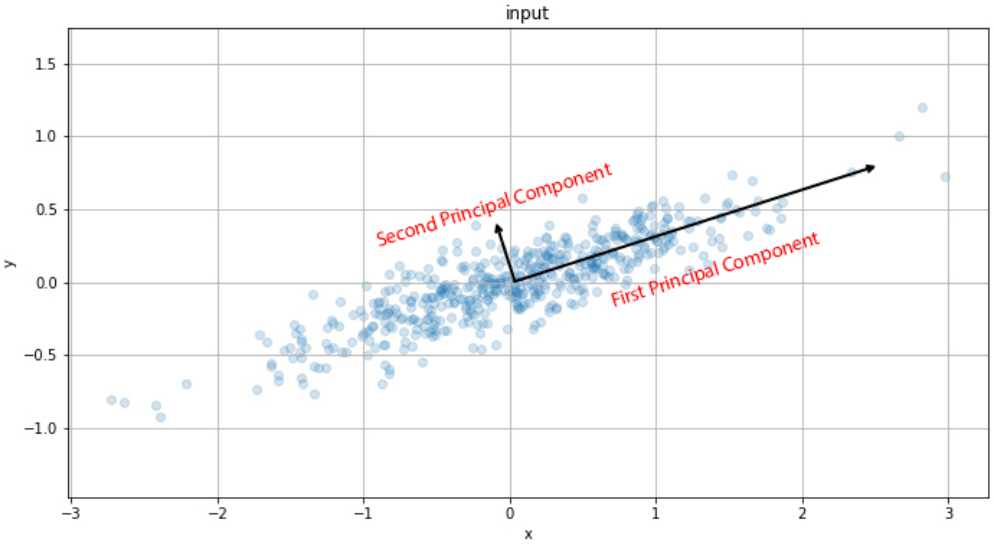


Figure 13: An illustration of the two principal components of a random, normally distributed dataset, where the length of the arrows represents the unit eigenvector scaled by its corresponding eigenvalue.

The principal components in PCA are determined by calculating the eigenvectors of the covariance matrix, and then selecting those that have the largest corresponding eigenvalues [51]. The arrows in Figure 13 represent the eigenvectors, where every sample is represented as some magnitude (eigenvalue) of these vectors. PCA can be seen as a factorisation where U and V are found, such that

$$X \equiv U \cdot V^T , \tag{2.14}$$

where X represents the original samples, in matrix form, and U and V represent the eigenvalues and eigenvectors respectively. This can be seen as an optimisation problem where U and V are found by solving the following minimisation [62]:

$$argmin \sum_{i,j} (x_{ij} - u_i \cdot v_j)^2. \tag{2.15}$$

2.5. Conclusion

This chapter has discussed how the optical properties of phytoplankton relate to their size as well as the statistical techniques employed to exploit this relationship. The methods discussed have successfully shown that relationships between optical properties and phytoplankton size do exist and that both apparent and inherent optical properties can be used. A number of machine learning techniques were then presented, with references to existing research in oceanography, where relevant. Finally, a number of parametrisation techniques were discussed as a means of dimensionality reduction. The following chapter describes the data used in this research as well as how it was collected and prepared.

Chapter 3: Data Acquisition and Analysis

This chapter describes what data is required in order to successfully model particle size, along with how such data was acquired. All of the data preparation and pre-processing is described for each of the data types. This chapter also includes some basic data analysis, in the form of covariance analysis and PCA, in order to get a better understanding of the structure of the data. These linear relationships serve as the basis for the development of Model 3, presented in Chapter 4.6. Finally, the size class labels are calculated, through the use of DPA, so that they can later be used in model development.

3.1 Data Requirements

Both optical and pigment information need to be measured in order to analyse the relationships between phytoplankton optical properties and the community size structure. Absorption was the only optical property used in this research, with wavelengths between 350nm and 750nm used, depending on the measuring apparatus. The specific wavelengths measured and used are described in this chapter under the relevant dataset sub-heading. In the absence of actual particle size information, pigment concentrations were used as a proxy for inferring size classes through DPA, covered later. The collected samples needed to be both spatially and temporarily aligned across the various sensors, for them to be compared to each other and analysed.

The data used in this research was gathered from a number of cruises, namely the Tara expedition, the Antarctic Circumnavigation Expedition (ACE) and a combined dataset from the National Space Agency (NASA), called the NASA Bio-optical Marine Algorithm Dataset (NOMAD) which was compiled from multiple cruises. I was aboard, and collected data on, the ACE cruise whereas the rest of the data is publicly available and downloadable on the SeaWiFS Bio-optical Archive and Storage System (SeaBASS) website [68]. Each of these cruises and datasets are covered in more detail in this chapter.

Table 3 gives the complete list of all of the coincident data obtained from all of the cruises. This dataset allows for the creation of models that try to infer pigment and particle size information from absorption.

Table 3: All coincident cruise data used in this thesis.

Dataset Name	Data Type A	Intersection Data B	Total Samples
ACE	Absorption	Pigments	122
NOMAD	Absorption	Pigments	241
Tara	Absorption	Pigments	156

3.2 Tara Cruise Data

One of the datasets used in this paper comes from the Tara Oceans expedition, a 3-year long expedition from 2009 to 2013 around the world, covering a 140 000km oceanic route spanning the Indian, Atlantic and Pacific oceans. Both pigment and absorption data from this cruise can be downloaded from the SeaBASS website [68].

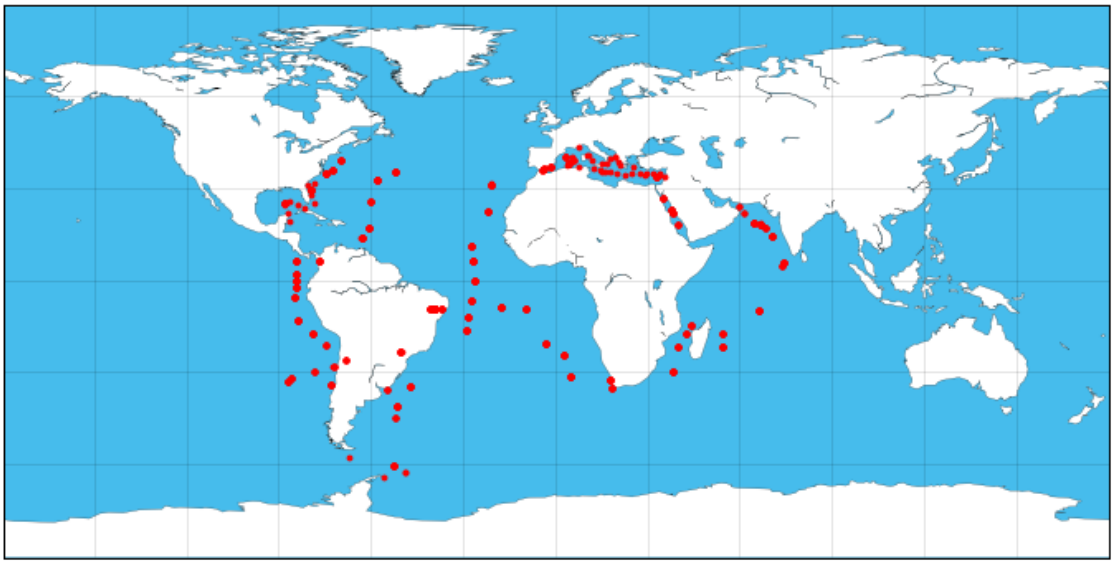


Figure 14: Sample collection points from the Tara Oceans cruise.

3.2.1. Data Preparation

The downloaded dataset contained 976 HPLC pigment samples and 636868 hyperspectral absorption samples gathered with an AC-S sensor, from 403.9nm – 733.2nm. Since the pigment samples are taken at specific stations, as shown in Figure 14, and the absorption samples are collected every minute, some pre-processing had to be performed in order to link the pigment samples with the closest absorption readings.

The downloaded data was in the form of separate Comma Separated Value (CSV) files, split by cruise leg and by reading type (absorption or HPLC), totalling 1194 separate files. In order to easily interrogate the data and perform spatial queries, the CSV files were first imported into a Postgres [69] database. The PostGIS extension [70] was installed in Postgres to facilitate working with spatial data such as latitude and longitude coordinates.

Figure 15 illustrates the high-level pre-processing steps that were taken to extract the spatio-temporally related data. Separate Python scripts were written to import the pigment and absorption data into the Postgres database, with the schema shown in Figure 16, before a Structured Query Language (SQL) query, Figure 17, was run to combine the datasets.

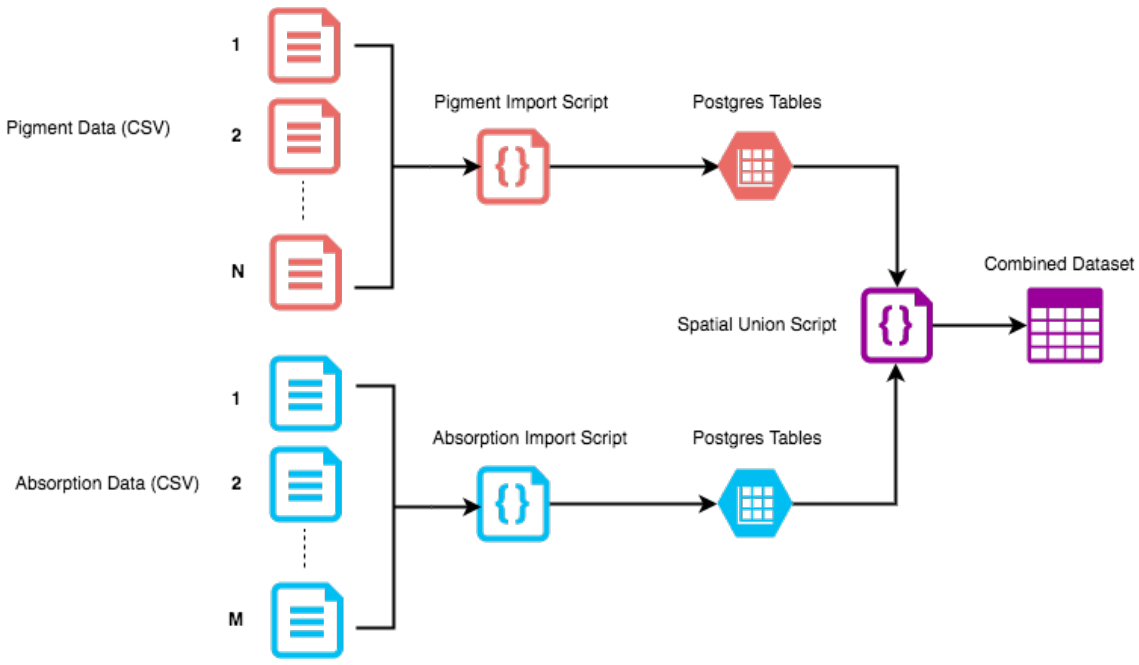


Figure 15: Tara dataset pre-processing flow.

Within the CSV data files there were some inconsistencies around the naming of fields, and not all of the fields were always present. It was for this reason that the values of each sample were stored as a set of key/values pairs. The Entity Relationship Diagram (ERD) in Figure 16 shows this relationship, where the tables "*optics_readings*" and "*pigment_readings*" are the parent records, and the tables "*optics_reading_data*" and

"*pigment_reading_data"* are the children records holding the key/value pairs for each of the samples. This structure ensured that even if there were inconsistencies in the data, all of the data for a sample would still be stored. Once the records had been saved into the database it was possible to get a distinct list of all of the fields so that a clean-up could be performed to ensure that the field names were consistent.

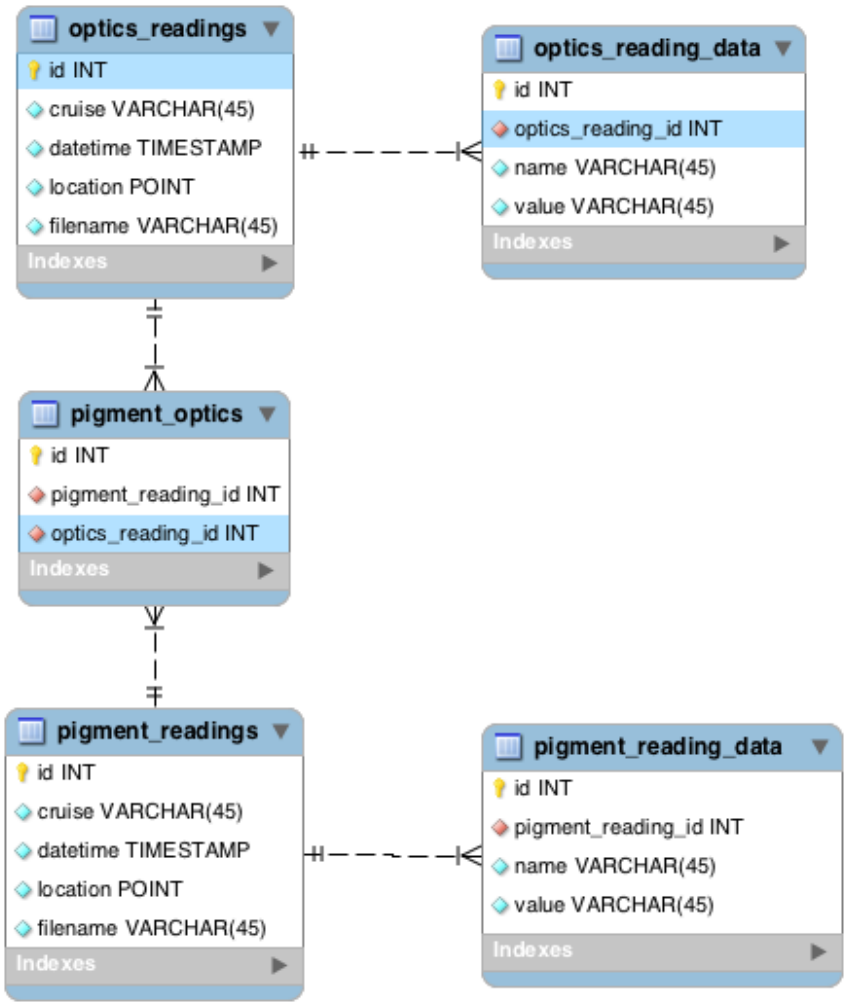


Figure 16: Tara database ERD for storing absorption and pigment data.

Storing the data in this format, instead of in a properly normalised table where all of the sample fields are table headings, had its drawbacks when it came to performance and query simplicity. The "*optics_reading_data"* table ultimately had over 100 million records in it and was therefore very slow to query.

Once all of the fields were cleaned up and made consistent, the SQL query shown in Figure 17 was developed to link the pigment data to the absorption data. This query would populate the bridge table, *"pigment_optics",* with all pigment and absorption records that were taken within 1km of each other and were no more than 30 minutes apart, and where the sample was only taken in the top 5m of the surface.

```
1   INSERT INTO pigment_optics
2       (optics_reading_id, pigment_reading_id)
3       SELECT
4           or1.id AS optics_reading_id,
5           pr.id  AS pigment_reading_id
6       FROM pigment_readings pr
7       INNER JOIN
8           (SELECT
9               pigment_reading_id,
10              datetime,
11              location
12          FROM pigment_reading_data d
13          LEFT JOIN pigment_readings r
14              ON d.pigment_reading_id = r.id
15              WHERE  name = 'depth'
16              AND value :: FLOAT <= 5
17          ) AS inn
18      ON pr.id = inn.pigment_reading_id
19      INNER JOIN optics_readings or1
20      ON Abs(Extract(epoch FROM inn.datetime - or1.datetime)) <= 1800
21      AND St_distance(inn.location, or1.location) < 1000;
```

Figure 17: SQL query for linking spatially and temporally related records.

Due to the high frequency sample rate of the AC-S, in comparison to the pigment samples taken, there were typically many AC-S readings for every pigment reading. The average value of these AC-S records was taken so that there could be one absorption reading for every pigment sample. This data was then exported to a CSV file where it was later analysed.

3.3 NOMAD Data

NOMAD is a high quality, in situ bio-optical dataset compiled by the "NASA Ocean Biology Processing Group" in order to facilitate research in ocean colour algorithm development. This dataset is publicly available on the SeaBASS website [68] and contains both absorption and pigment data for a number of its samples. Details on how this data was collected, cleaned and standardised can be found in the report by Werdell and Bailey [71].

As this data was already of a useable standard, the only pre-processing required was to select the records that had both absorption and pigment information available. A total of 241 usable samples were extracted, collected from various parts of the Atlantic and Pacific oceans, as illustrated in Figure 18.

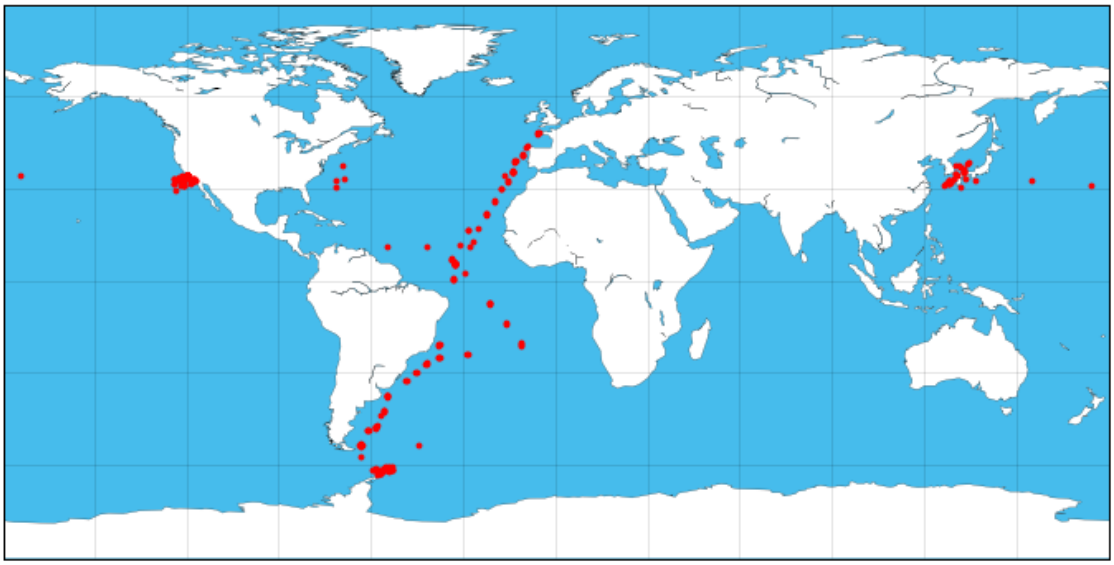


Figure 18: Used sample collection points from the NOMAD dataset.

The only drawback of this dataset is that, in order to be aligned with the spectral resolution of ocean colour algorithms, only 21 nominal wavelengths from 405nm – 683nm were chosen to represent their absorption data, as shown by Werdell and Bailey [71]. Even though this stands in stark contrast to the 401 and 161 wavelengths provided by the ACE and Tara cruises respectively, it is a more realistic representation of what might be available when working on satellite-based ocean colour algorithms. To this end it is important to ensure that the algorithms used in this research are capable of obtaining reasonable results even with this low-resolution data, if they are to be of any value in satellite-based ocean optics research.

3.4 ACE Cruise Data

On the 20th of December 2016 I was privileged enough to join the ACE expedition, a three-month long research expedition in the Southern Ocean funded by the Swiss Polar Institute (SPI). I was a part of the bio-optics team, which was lead and organised by David Antoine of Curtain University and the Southern Ocean Climate & Carbon Observatory (SOCCO) department of the Council for Scientific and Industrial Research (CSIR).

3.4.1. Sample Collection

We were a 4-person team, working 6-hour shifts every day, and collected various optical and in situ samples from the Southern Ocean, as can be seen in Figure 19. The samples that were of relevance to this project were particulate absorption, HPLC pigment and particle size information. Between 500ml and 2000ml of seawater was filtered onto 25mm GF/F filters with the use of a vacuum pump. The samples were then flash frozen in liquid nitrogen before being stored at -80°C so that they could be processed in a lab once ashore.

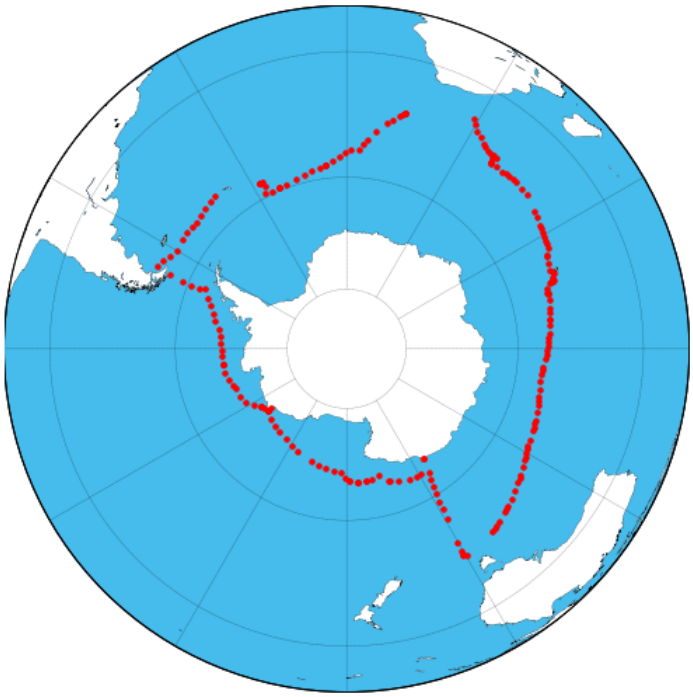


Figure 19: Used sample collection points from the ACE cruise.

3.4.2. Lab Processing

Once the samples had been collected on the cruise they needed to be processed in a lab in order to measure the particulate absorption and pigment concentrations. I was not involved in this process. The pigments were extracted via HPLC analysis by the Laboratoire d'Océanographie de Villefranche-sur-Mer (LOV) in France using the analytical procedure as described by Ras, Claustre and Uitz [42].

The particulate absorption data, in the range of 350nm – 750nm at a 1nm resolution, was extracted by one of the other team members, Thomas Ryan-Keogh, using the

methods outlined by Bricaud and Stramski [72] with further processing done by Charlotte Robinson from Curtain University, applying the method defined by Stramski, Reynolds and Kaczmarek [73].

3.5 Absorption Data Analysis

Figure 22 shows the particulate absorption spectra from all three cruises. This is where the low resolution of the NOMAD dataset can be seen in comparison to the ACE cruise data resolution. The combination of these datasets provided valuable spectral variability, covering a large range of absorption values.

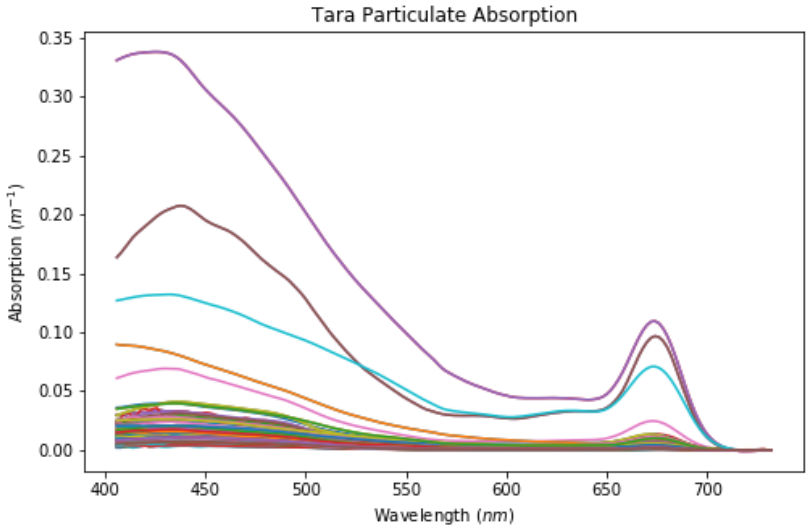


Figure 20: Particulate absorption of the Tara cruise dataset, where each sample is a different colour.

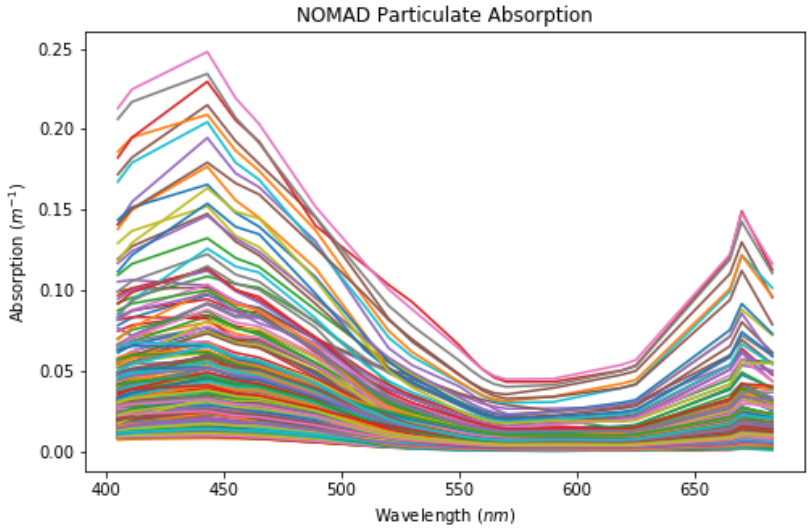


Figure 21: Particulate absorption of the NOMAD dataset, where each sample is a different colour.

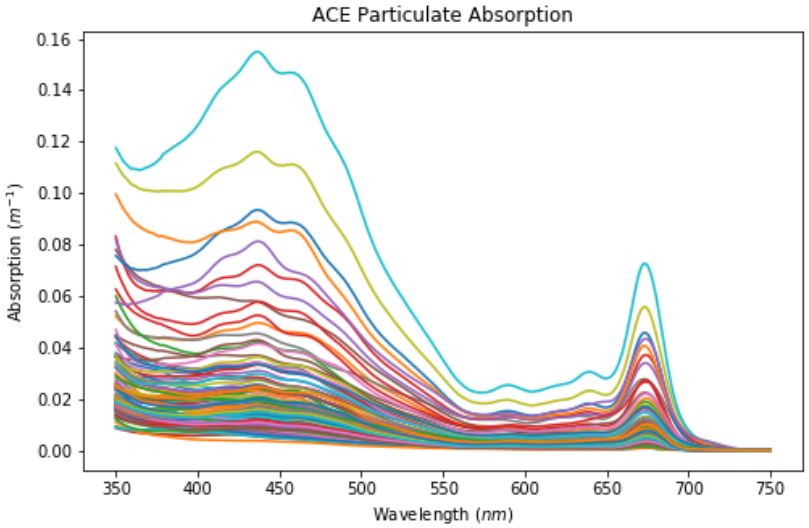


Figure 22: Particulate absorption of the ACE dataset, where each sample is a different colour.

Plotting a covariance matrix as a heat map is a convenient way of visually inspecting the dataset for joint variability. The lighter the heat map, the higher the joint variability, and the better linear dimensionality reduction techniques will work. The absorption data is very uniformly distributed with a very high covariance, as seen in Figure 23, where the lowest covariance between spectral bands is 94%.

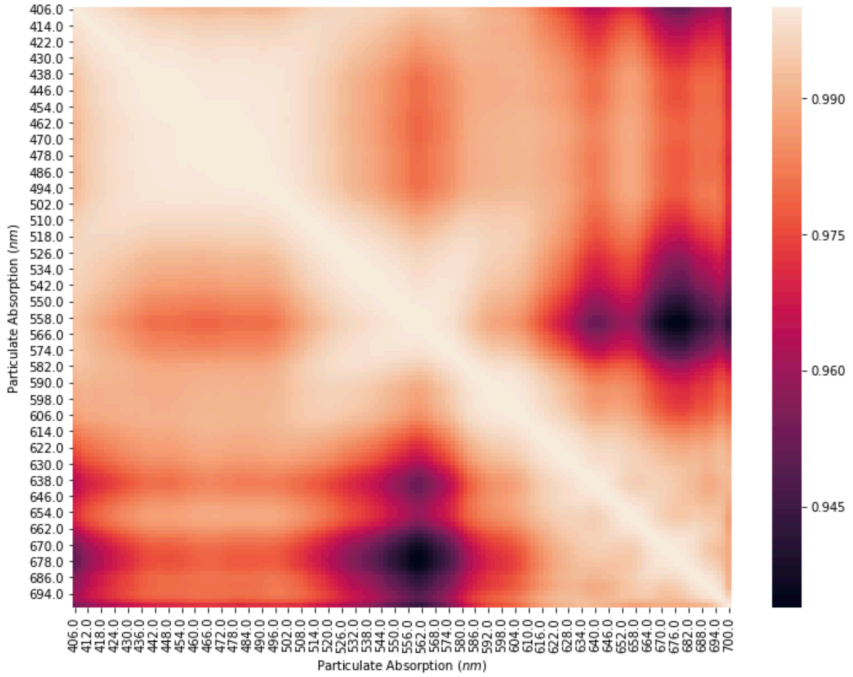


Figure 23: A heat map of the covariance matrix of absorption data.

3.5.1 PCA Applied to Absorption Data

When PCA was applied to the absorption dataset, over 99.5% of the variance is explained by the first principal component alone. This can be seen in Figure 24, where the contributions of the second and third principal components contain very little information in comparison to the first. As with the particle size data, the eigenvectors contain negative values, as seen in Figure 25. If the eigenvectors are merely used as an arbitrary set of vectors that best describe the components of the signal, the negative values, although nonsensical, make no difference to the effectiveness of the technique. In scenarios where non-negativity is a requirement, an alternative algorithm – Non-Negative Matrix Factorisation (NMF) – can be applied. This technique is used and discussed in Chapter 4, where the basis vectors can be interpreted as a meaningful descriptor of particle size contribution.

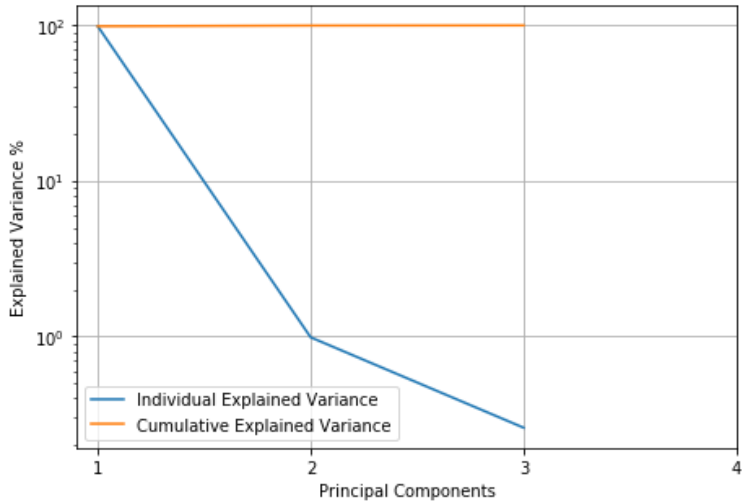


Figure 24: Variance explained by principal components in particulate absorption data.

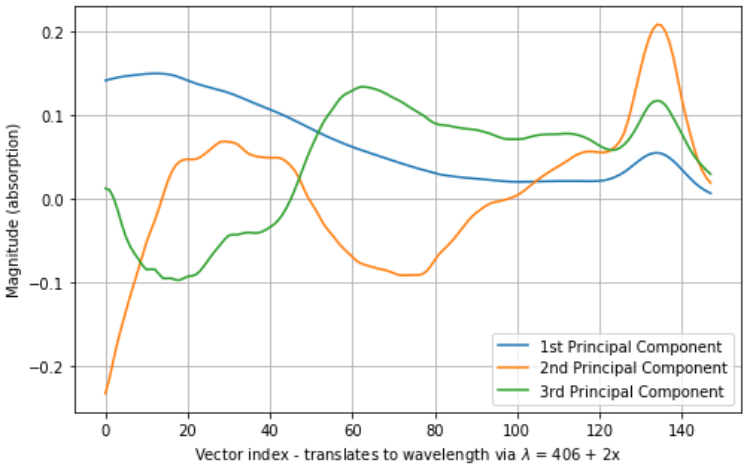


Figure 25: First three eigenvectors of the absorption data.

3.6 Pigment Data

The pigment data extracted on all of the cruises, via HPLC, provides concentrations of the pigments listed in Table 4. Section 2.2.2 provides further information on the HPLC process and what the detection wavelengths mean. As shown in later chapters, some of these pigments provide valuable information regarding the phytoplankton size, as certain pigments are only found in certain species.

Table 4: Pigments extracted via HPLC from the ACE cruise.

Pigment Name	Description	Detection wavelength (nm)
Chlorophyll-c3		450
Chlorophyll-c1+c2		450
Chlorophyllide-a	Chlda + Chlda-like	667
Peridinin		450
Phaeophorbid-a	Phda + Phda-like	667
19'-Butanoyloxyfucoxanthin		450
Fucoxanthin		450
Neoxanthin		450
Prasinoxanthin		450
Violaxanthin		450
19'-Hexanoyloxyfucoxanthin		450
Diadinoxanthin		450
Antheraxanthin		450
Alloxanthin		450
Diatoxanthin		450
Zeaxanthin		450
Lutein		450
Bacteriochlorophyll-a		770
Chlorophyll-b		450
Divinyl Chlorophyll-a		667
Chlorophyll-a	Chlorophyll-a + allomers + epimers	667
Total-Chlorophyll-a	Chla + DV Chla + Chlorophyllid-a	667
Phaeophytin-a	Phytna + Phytna-like	667
Total-Carotenes	Beta-carotene + Alpha-carotene	450

3.7 Adding Size Labels

As shown in Figure 2 in the Introduction chapter, phytoplankton range nine orders of magnitude in size and, as such, acquiring data that encompasses this full range is very challenging. Developing a model that estimates the particle size distribution that covers this full range is also very difficult, and impractical if any reasonable resolution is required. It is for these reasons that biologists tend to define phytoplankton in terms of size classes as opposed to a full distribution. These classes are, however, not very well-defined and slight variations of them exist [39]. One of the more common categorisations is the three-size class model as used by Uitz et al. [40]. The boundaries of the size classes, as shown in Figure 26, were grouped based on average sizes of species, as shown by Vidussi et al. [45].

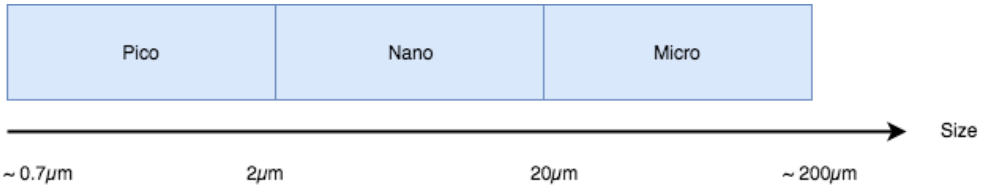


Figure 26: Size ranges of the three phytoplankton size classes (Pico, Nano and Micro).

Certain diagnostic pigments only occur in certain species, and these species fall into specific size ranges. The DPA technique was developed for estimating the percentage contributions of the three size classes (Pico, Nano and Micro) from their diagnostic pigment concentrations [25], [31]. It is unlikely that DPA has been applied to every phytoplankton species, but performance of this algorithm has been validated against both global in situ pigment datasets and a concurrent co-located satellite matchup dataset [25]. The following equations describe the relationships between diagnostic pigments and size fractions [40]:

$$f_{micro} = (1.41[fucoxanthin] + 1.41[peridinin])/wDP, \quad (3.1)$$

$$f_{nano} = (0.6[alloxanthin] + 0.35[BF] + 1.27[HF])/wDP, \quad (3.2)$$

$$f_{pico} = (0.86[zeaxanthin] + 1.01[Chl\ b + divinyl\ Chl\ b])/wDP, \quad (3.3)$$

$$wDP = 1.41[fucoxanthin] + 1.41[peridinin] + 0.6[alloxanthin] + 0.35[BF] + 1.27[HF] + 0.86[zeaxanthin] + 1.01[Chl\ b + divinyl\ Chl\ b]. \quad (3.4)$$

This was applied to all of the datasets individually in order to assess the distribution of size classes across the datasets, as shown in Figure 27. Critical findings were that the Tara cruise did not have a high number of Micro particles, and that the ACE cruise had very few Pico particles present. The Tara and ACE datasets were therefore combined in order to create a dataset that had a more even distribution of size classes. This merge was only possible because of the similar sample resolution of the absorption data. Having a more evenly distributed dataset ensured that when the absorption basis vectors were extracted, they were more representative of the absorption for the given size classes. The low sample resolution of the NOMAD dataset prohibited it from being merged with Tara and ACE datasets, so it was processed separately. This did not pose any problems though, as it was already an evenly-distributed dataset.

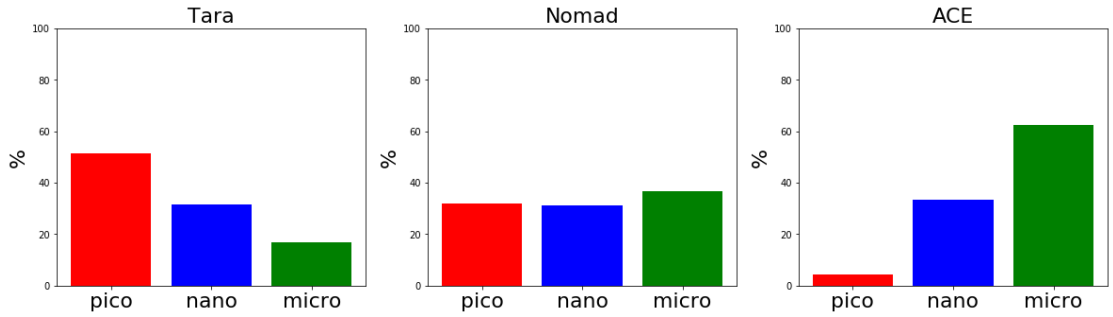


Figure 27: The percentage contribution of each size class across each of the datasets.

The absorption signals need to be normalised to account for biomass, since it is a first-order source of variability in the absorption signal. This can be done in one of two ways, either by dividing by total Chlorophyll-a or by dividing by the mean absorption [74], [47] as shown below:

$$a_p^*(\lambda) = \frac{a(\lambda)}{Tchla}, \quad (3.5)$$

$$a_{mean}^*(\lambda) = \frac{a(\lambda)}{< a(\lambda) >}, \quad (3.6)$$

where $Tchla$ represents the total Chlorophyll-a and $< a(\lambda) >$ the mean absorption.

After the size class contributions were calculated per sample, they were overlaid on top of their corresponding absorption signals so that the relationship between size and absorption can be seen. In Figure 28 and Figure 29 each line represents the absorption signal, of a given sample, which has been normalised by the total amount of Chlorophyll-a present in the sample, whereas in Figure 30 and Figure 31 the absorption signals were normalised by the mean absorption of the signal across all wavelengths. Each of the figures has been split into three charts, one for each of the size classes Pico, Nano and Micro, where the colour of the line represents the quantity of the given size class present in the sample, scaled between 0 and 1. Therefore, lines that are redder in colour represent absorption signals of samples containing a high relative percentage of the given size class, whereas the bluer lines represent a lower relative percentage.

When normalising by total Chlorophyll-a it can be seen that Pico particles have the largest absorption coefficients and that Micro particles have the smallest. When normalising by the mean absorption, the signals' shapes change rather drastically, becoming tightly grouped. Initially the pattern is the same: Pico particles have the largest absorption coefficients and Micro particles the smallest, but at approximately 520nm an inflection point exists after which the absorption contributions are inverted. This observation that normalised absorption varies with size is well documented [35], [30] and forms the basis for Model 1 described in Chapter 4.4.

An interesting observation is that the absorption boundaries by size are not nearly as well-defined in the Tara and ACE datasets as they are in the NOMAD dataset. This is probably because the NOMAD dataset is a specially curated dataset that has been selected to represent an even distribution of size classes so that it can be used in ocean colour algorithms.

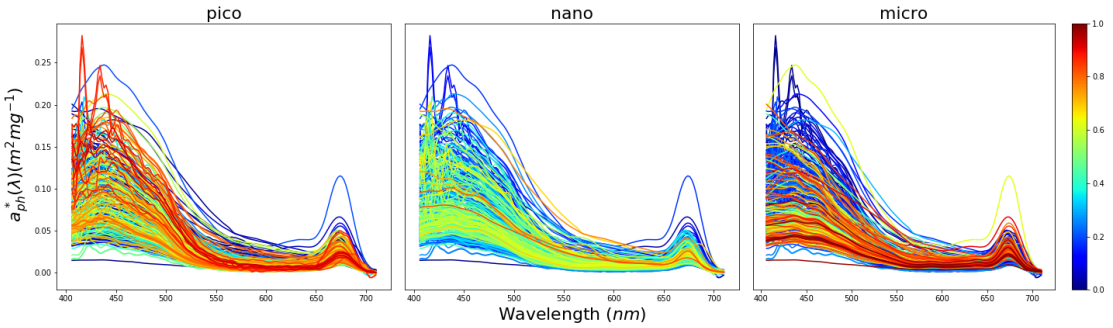


Figure 28: Absorption signals normalised by Chlorophyll-a, showing particle size contributions for the Tara and ACE datasets.

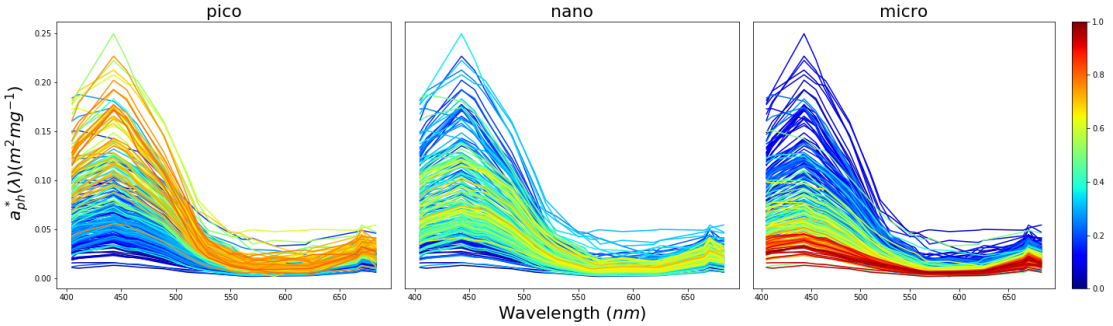


Figure 29: Absorption signals normalised by Chlorophyll-a, showing particle size contributions for the NOMAD dataset.

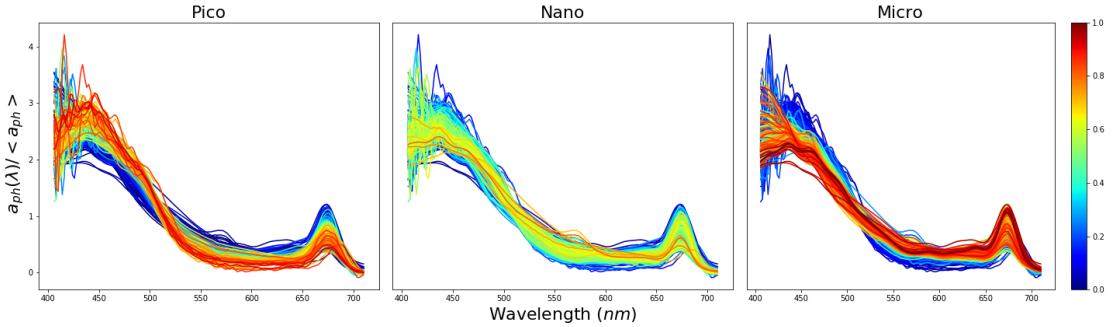


Figure 30: Absorption signals normalised by the mean, showing size contributions for the Tara and ACE datasets.

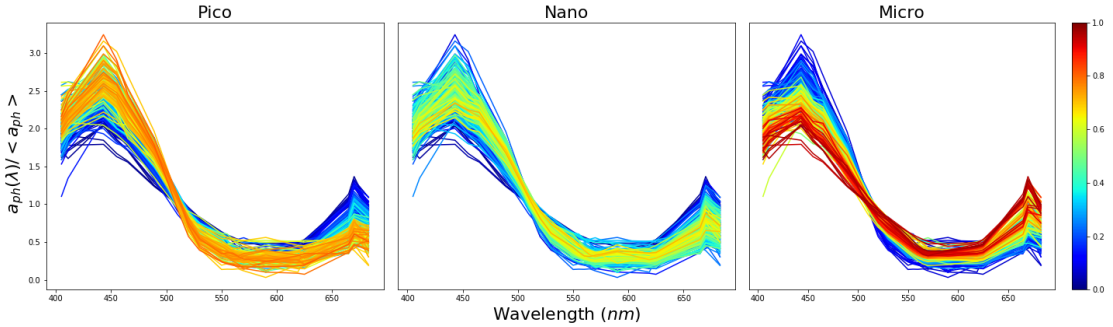


Figure 31: Absorption signals normalised by the mean, showing size contributions for the NOMAD dataset.

3.8 Summary of Data Used

Table 5 and Table 6 below show the total number of records available to per cruise as well as the number of intersecting records between the absorption and pigment datasets.

Table 5: Available absorption data per cruise.

Dataset	Absorption Range	Resolution	Samples
ACE	350nm – 750nm	1nm	254
Tara	403.9nm – 733.2nm	4nm	636868
NOMAD	405nm – 683nm	~13nm (21 bands)	766

Table 6: Available pigment data per cruise along with the total number of records that have coincident absorption records.

Dataset	Total Intersecting Samples	Total Samples
ACE	122	193
Tara	156	221
NOMAD	241	620

Chapter 4: Phytoplankton Size Class Models

The previous chapter discussed how the data was obtained, cleaned, and how the size class labels were estimated via DPA for use in model development. This chapter uses the processed data and size labels to generate a series of models that are ultimately combined into an ensemble model, the purpose of which is to estimate the size classes of phytoplankton from their optical properties. Four different models will be discussed in this chapter – three that are capable of independently estimating phytoplankton size classes, and the fourth being a final ensemble model that combines the first three to produce better results than any one of the models alone. Before the models are described in detail, a high-level overview provides context for each one.

4.1 Model Training

All of the models presented need to go through some sort of training stage, an optimisation process where the best possible values of the models' hyperparameters are found. Some of the models presented are multi-stage models, and as such have more than one training stage – typically where the output of one model is fed into the input of the next model.

All datasets were randomised and split into training and test sets of 60% and 40% respectively. This particular split ratio was chosen so that there would be enough samples in the test set to prevent high model variance even though the total number of data samples is low. The models were trained on the training set and then evaluated against the test set. This process prevented overfitting of the models, which would have otherwise diminished their ability to generalise well outside of the given dataset. This split also provides a common dataset against which all models can be evaluated, allowing for a direct comparison of results.

4.2 Model Evaluation

Before the details of the models are discussed, it is important to define some metrics by which they can be evaluated. The output of all of the models, disregarding intermediate outputs for the time being, are the particle size classes Pico, Nano and Micro. The size classes are always expressed as a percentage and sum up to 100%. In order for the performance of the models to be comparable, the same metrics will be used.

4.2.1 Baseline Mean Model

Once a method of evaluating error has been defined it is still not clear what error value equates to a useful model, as a Root Mean Square Error (RMSE) alone is not enough to determine whether a model is performing well. In order to put the accuracy into perspective, a simple baseline model was created. The baseline model simply returns the arithmetic mean of the Pico, Nano and Micro values of the given dataset:

$$\hat{y}_{pico} = \frac{1}{N}\sum_i^N S_i^{pico}, \tag{4.1}$$

$$\hat{y}_{nano} = \frac{1}{N}\sum_i^N S_i^{nano}, \tag{4.2}$$

$$\hat{y}_{micro} = \frac{1}{N}\sum_i^N S_i^{micro}, \tag{4.3}$$

where $\hat{y}_{pico}$, $\hat{y}_{nano}$ and $\hat{y}_{micro}$ represent the mean Pico, Nano and Micro estimates of the model and S_i^{pico}, S_i^{nano} and S_i^{micro} the i^{th} Pico, Nano and Micro records in the given training dataset, respectively.

4.2.2 Root Mean Square Error

The RMSE is an aggregation of the individual estimation errors into a single statistic and can be interpreted as the average prediction error. The output of RMSE is in the same units as the value being evaluated, making the results interpretable. The drawback of this method is that it cannot be used to compare models that are estimating different variables. RMSE is described as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_i^N (Y_i - \hat{Y}_i)^2}, \tag{4.4}$$

where $\hat{Y}_i$ is the i^{th} estimation and Y_i is the i^{th} value. The square error is averaged over all estimations N before being square-rooted to bring it back into the original units. The values of the size classes that are being estimated represent the normalised relative percentage of each of the size classes present in the given sample. Each size class contribution is always in the unit interval and the three size class contributions always sum to 1. The RMSE is therefore also in the unit interval and $RMSE \times 100\%$ represents the average % that the model was incorrect by. All of the particle size models will provide the RMSE for each of the size classes, as well as a combined RMSE, which is defined as follows:

$$RMSE_{total} = \sqrt{\frac{1}{3}(RMSE_{pico}{}^2 + RMSE_{nano}{}^2 + RMSE_{micro}{}^2)}. \quad (4.5)$$

This is only reasonable to assume if the relative errors of each of the size classes are normalised into the same interval. In this case the RMSE of each size class is in the range of $[0:1]$ and as such the $RMSE_{total}$ is also in the range $[0:1]$.

4.2.3 Coefficient of Determination

R^2 is a statistical measure of how well the regression line fits the data and is achieved by taking a ratio of the explained variation to the total variation as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y}_i)^2}, \quad (4.6)$$

where $\hat{Y}_i$ is the i^{th} estimation and Y_i is the i^{th} value, averaged over all data-points N. The output is a value between 0 and 100%, where 0% means that the model explains none of the variance around the mean and 100% indicates that all of the variance from the mean is explained. It is for this reason that the R^2 values for the mean models provided in the results are 0.

4.3 Overall Model Design

The final ensemble model is made up of three sub-models, each of which is independently capable of estimating the percentage abundance of each size class. Even though the final ensemble model achieves accuracies greater than any of the individual models, the intention is that these sub-models can be used in isolation.

Each model was designed to exploit a different relationship between the particulate absorption spectrum and the underlying size distribution of phytoplankton absorbing the light. This required a considerable amount of domain research and resulted in models that are meaningful and understandable by oceanographers and biologists. The alternative would be to develop some “black box” model like a Deep Neural Network, which might achieve a greater accuracy, but the underlying causality and interpretability of the model would be lost. This modular approach also allows for improvements in any one of the models as research and understanding of these relationships progresses, which would improve the overall accuracy of the model.

The first model extracts the absorption basis vectors per size class, which can be thought of as the “typical” absorption spectra for the given size class. These are later used to estimate the ratios of these basis vectors present in a given phytoplankton absorption signal, and in so doing, estimating the percentage of each size class present.

The second model is a multistage model that first estimates various pigments, through Gaussian decomposition followed by non-linear regression, so that the relative pigment concentrations (mainly Chlorophyll-a) can be used to estimate the percentage of each size class present.

The third model exploits the fact that the first principal component of the phytoplankton absorption signals, across all the datasets, explains most of the variability. An empirical equation was developed, which can parameterise a given absorption signal into a single parameter. This parameter is then used, through the use of SVR using the RBF kernel, to estimate the percentage of each size class present.

These three models are then linked together to create a final ensemble model. Figure 32 shows how these models are arranged and how the data flows through them.

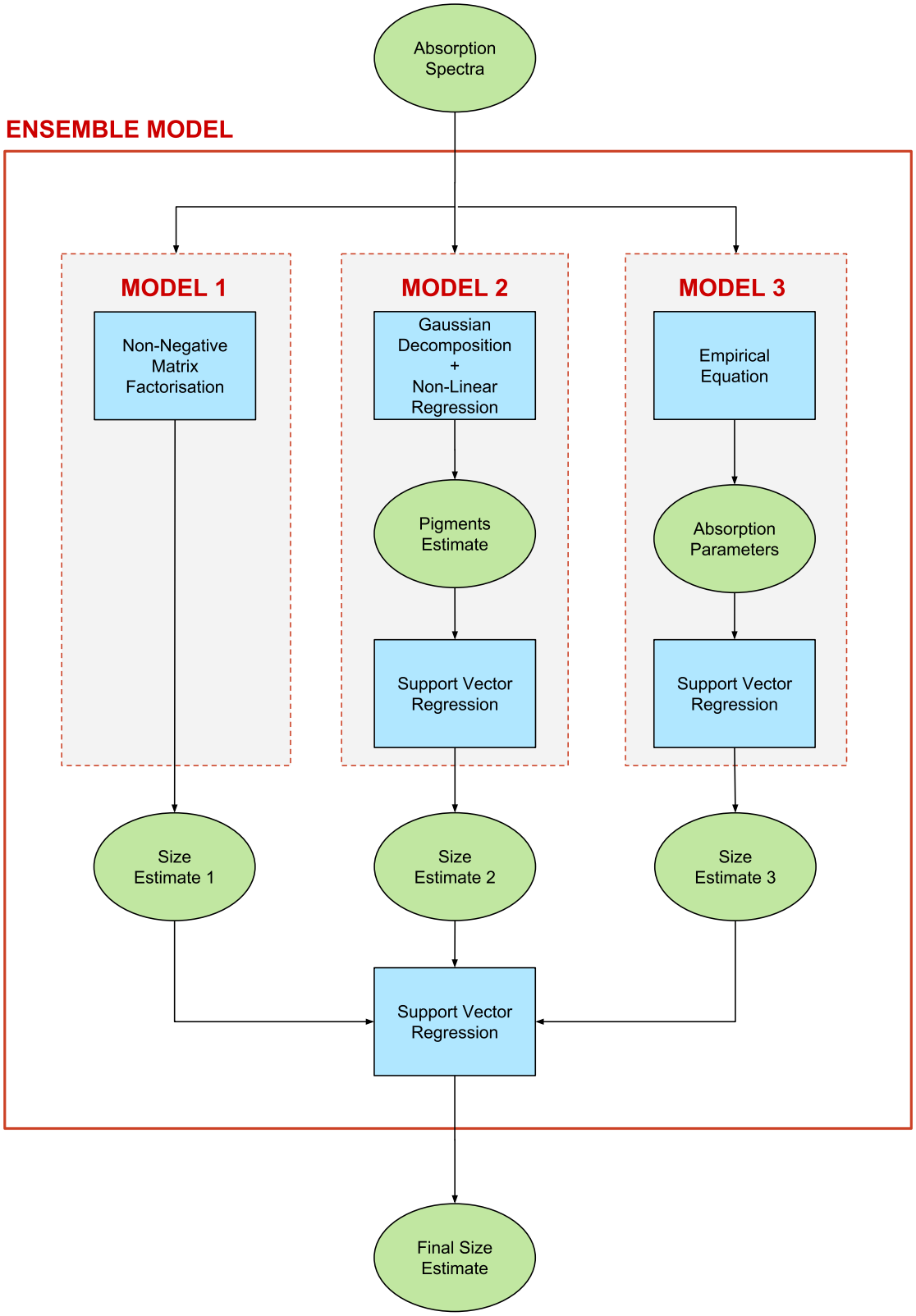


Figure 32: Ensemble model, composed of three sub-models, for estimating phytoplankton size classes from absorption.

4.4 Model 1: Size-classes via Matrix Factorisation

The previous chapter showed (Figure 28 to Figure 31) what the typical absorption spectra of the different size classes look like. This section extracts these "typical" absorption spectra, known as basis vectors, so they may be used for estimating how much a given absorption signal is comprised of these basis vectors. The ratios of the basis vectors contained within the absorption sample define the ratios of size classes present in the sample. It will then be shown, through a technique known as Semi-Supervised Learning (SSL) that the model's accuracy can be further improved upon by the use of unlabelled data.

4.4.1 Model 1: Aim

To develop a model that is capable of estimating phytoplankton size classes from absorption data, through the use of derived absorption basis vectors.

4.4.2 Model 1: Method

1. Randomly split the data into training and test data sets in a 60% to 40% ratio, respectively. This particular ratio was chosen due to the relatively low number of samples.
2. Calculate the absorption basis vectors through NMF.
3. Improve the basis vectors through semi-supervised learning.
4. Measure the baseline accuracy through the mean-size estimator.
5. Measure model accuracy and compare results.

4.4.3 Model 1: Data Used

The model was trained and evaluated with data from the combined ACE + Tara datasets and for the NOMAD datasets separately. Table 7 gives a breakdown of all the data used for training, training via SSL and evaluation.

Table 7: All the data used for the training, semi-supervised learning and evaluation of Model 1.

Dataset	Data Type A	Used For	Intersection Data B	Samples
ACE + Tara	Absorption	Training	Pigments	141
ACE + Tara	Absorption	Semi-Sup	-	300
ACE + Tara	Absorption	Evaluation	Pigments	95
NOMAD	Absorption	Training	Pigments	145
NOMAD	Absorption	Semi-Sup	-	300
NOMAD	Absorption	Evaluation	Pigments	98

4.4.4 Model 1: Calculating Basis Vectors

Given the absorption spectra of a water sample along with the coincident size class concentrations, basis vectors can be calculated such that each basis vector represents the characteristic absorption signatures of the Pico, Nano and Micro size classes. Estimating these basis vectors falls within the training phase of the model, and as such only the training data is used to calculate these basis vectors. Figure 33 shows the high-level input/output of the model, along with the data dimensions. The details of this model and factorisation process are explained in detail below.

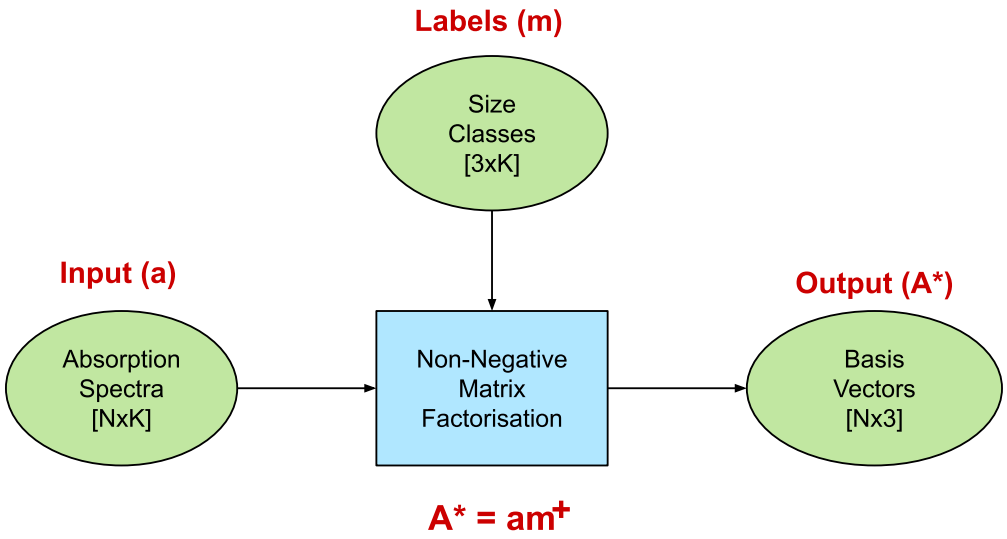


Figure 33: Factorising the absorption spectra into basis vectors representing the typical absorption signals of the given size classes.

The core of this model revolves around the fact that absorption is an additive model. In other words, in a given sample all of the constituents within the water that absorb the light, across the frequency spectrum, can be added together to reveal the total absorption by all constituents. This is useful as it can be used to decompose an absorption signal into the absorption of either the individual constituents or of some grouping thereof. The total absorption of seawater can be broken up as follows:

$$a(\lambda)_{total} = a(\lambda)_w + a(\lambda)_{par}, \tag{4.7}$$

where $a(\lambda)_{total}$ is the total absorption, $a(\lambda)_w$ is the absorption of pure water and $a(\lambda)_{par}$ represents the total particulate absorption and is made up of the following constituents:

$$a(\lambda)_{par} = a(\lambda)_{ph} + a(\lambda)_{CDOM} + a(\lambda)_{nap}, \tag{4.8}$$

where $a(\lambda)_{ph}$ represents the absorption by phytoplankton, $a(\lambda)_{CDOM}$ the absorption by colour dissolved organic matter and $a(\lambda)_{nap}$ the absorption by non-algal particles. The absorption by CDOM and NAP is often combined into a single quantity as they both absorb light in a very similar manner. This combined value is represented as $a(\lambda)_{CDM}$ and approximated as the following exponential, normalised by the absorption at 400nm [31]:

$$a(\lambda)_{CDM} = e^{-\alpha(\lambda-400)}. \tag{4.9}$$

Since the absorption spectra of water is known, it is removed from the absorption data upfront. The absorption by phytoplankton is then partitioned into contributions by size classes (Pico, Nano and Micro) to yield the following:

$$a(\lambda)_{par} = a(\lambda)_{pico} + a(\lambda)_{nano} + a(\lambda)_{micro} + a(\lambda)_{CDM}. \tag{4.10}$$

This can be represented in a more succinct way, where the total amount that each constituent contributes is accounted for as follows:

$$a(\lambda)_{par} = \sum_{i}^{M} m_i a_i^*(\lambda)\,. \tag{4.11}$$

Here m_i represents the amount of the constituent present and $a_i^*(\lambda)$ represents the absorption by the given constituent at the given wavelength. This can be further simplified and represented as a linear system:

$$\boldsymbol{a} = \boldsymbol{A}^*\boldsymbol{m}\,, \tag{4.12}$$

where total absorption can be represented as a column vector with *N* wavelengths:

$$\boldsymbol{a} = \ [a(\lambda_1),\ a(\lambda_2), \dots,\ a(\lambda_N)]^T, \tag{4.13}$$

and m is represented as a row vector of weights with *M* elements:

$$\boldsymbol{m} = [m_1, m_2, \dots, m_M]. \tag{4.14}$$

The value $\boldsymbol{A}^*$ represents a set of basis vectors, with shape $N \times M$, that holds the unique absorption signature for each of the M constituents (size classes) across the N wavelengths:

$$\boldsymbol{A}^* = [\boldsymbol{a_1^*}, \boldsymbol{a_2^*}, \dots, \boldsymbol{a_M^*}]. \tag{4.15}$$

Here each basis vector, represented as a column vector, contains the absorption coefficients for a single component at all wavelengths:

$$\boldsymbol{a_i^*} = \ [a_i^*(\lambda_1),\ a_i^*(\lambda_2), \dots,\ a_i^*(\lambda_N)]^T. \tag{4.16}$$

The final objective is to calculate the row vector of weights (m), as these weights indicate how much of the given constituent is present in the sample, but in order to calculate m we first need to solve for $\boldsymbol{A}^*$. By solving for $\boldsymbol{A}^*$, a set of basis vectors is acquired which represent the absorption signatures of the three size classes (Pico, Nano and Micro). The output of DPA is the fractional composition of the three size classes that can be used as the initial values for $\boldsymbol{m}$. These values along with the particulate absorption data are all that is required to calculate $\boldsymbol{A}^*$, if equation 4.12 is rearranged to make $\boldsymbol{A}^*$ the subject of the equation:

$$\boldsymbol{A}^* = \boldsymbol{a}\boldsymbol{m}^+, \tag{4.17}$$

where $\boldsymbol{m}^{+}$ is the Moore-Penrose inverse of $\boldsymbol{m}$, since $\boldsymbol{m}$ is not square.

The basis vectors are then calculated for the combined Tara + ACE dataset as well as for the NOMAD dataset, using both normalisation techniques described in equations 3.5 and 3.6. These basis vectors are shown in the plots of Figure 34 and Figure 35.

The NOMAD dataset shows a very clear delineation of size classes when normalising by total Chlorophyll-a. The combined dataset does not show this separation as clearly and has a rather large basis vector representing Micro particles. When normalising by the mean absorption, the inflection points at roughly 520nm can clearly be seen in both the combined and the NOMAD datasets.

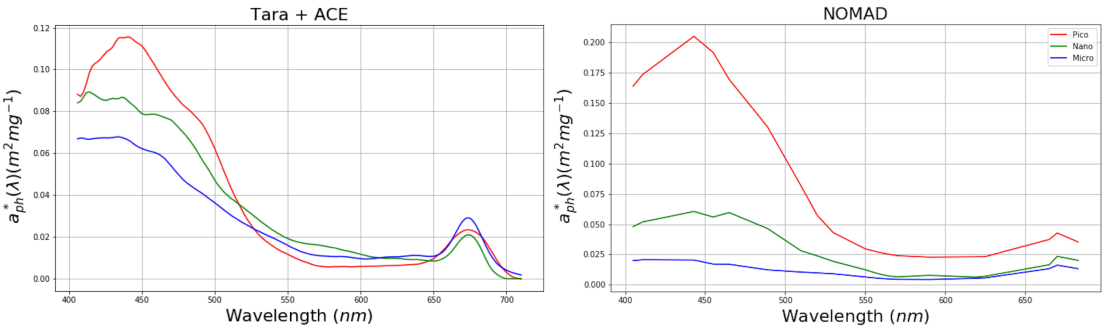


Figure 34: Basis vectors representing the specific absorption spectra for the size classes Pico, Nano and Micro when normalising by total Chlorophyll-a. The combined dataset is on the left and the NOMAD dataset on the right.

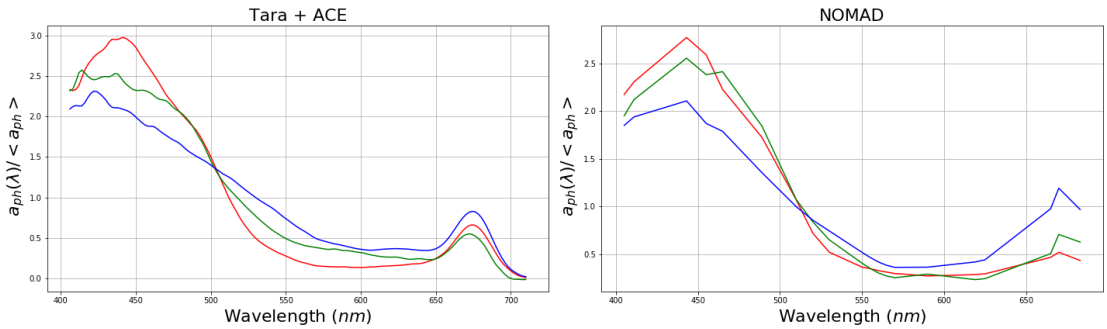


Figure 35: Basis vectors representing the specific absorption spectra for the size classes Pico, Nano and Micro when normalising by the mean absorption. The combined dataset is on the left and the NOMAD dataset on the right.

4.4.5 Model 1: Using the Basis Vectors to Estimate Size

The next step, after the basis vectors have been identified, is to try and estimate the ratio of particle size classes from their absorption signals. This factorisation is applied to the evaluation dataset so that the model's accuracy can be measured. This process,

shown in Figure 36, is very similar to the training step, shown in Figure 33, where size estimations are now the subject of the formula.

The total concentration of a given particle size class is always positive and therefore a non-negativity constraint is added to the factorisation process. This technique is known as Non-negative Matrix Factorisation (NMF) and is found by solving the following minimisation problem:

$$\min \| \boldsymbol{A}^{*}\boldsymbol{m} - \boldsymbol{a} \|^{2}, \quad \boldsymbol{m} \in \mathbb{R}\,[0,1]. \tag{4.18}$$

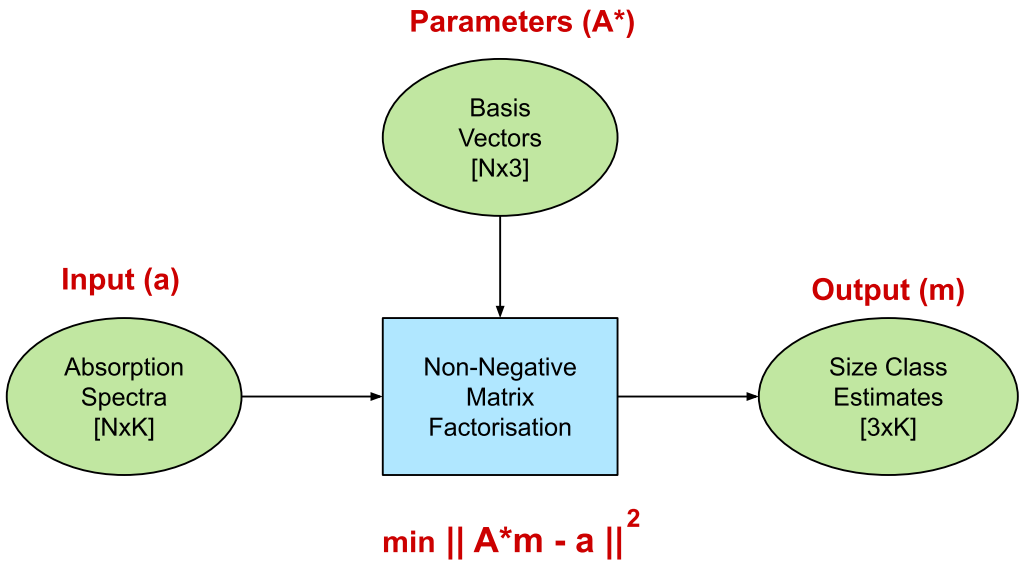


Figure 36: Using the derived basis vectors to estimate the contributions per size class of a given absorption signal.

The value m represents the percentage composition by each size class in the given absorption signal, and therefore also requires the following sum-to-one constraint so that it can be compared to the size classes labels generated by DPA:

$$\sum_{i}^{M} m_i = 1. \tag{4.19}$$

4.4.6 Model 1: Improving Performance via Semi-Supervised Learning

Due to the limited amount of absorption data with size labels (as a result of lacking coincident pigment measurements), and the abundant availability of extra unlabelled absorption data, Semi-Supervised Learning (SSL) was employed to make use of this unlabelled data and further increase the accuracy of the model.

SSL is an iterative process of labelling batches of unlabelled data with the trained model and then incorporating this labelled batch into the original training data for the model to be trained again. The training data grows with the addition of the newly labelled data, but the accuracy is constantly checked against the original 40% that was allocated as the test set. This can be repeated until the model converges, assuming it does, and the accuracy stabilises [75].

A number of considerations need to be made when training a model through SSL. The batch size chosen is probably the single most important variable to be set. If the batch size is too large it can cause the model to end up in a negative feedback loop, causing the accuracy to get progressively worse. If the batch size is too small one could run into performance issues and might have to retrain the model many times over. The optimal batch value for this model was set to 3, which was discovered through a process of trial and error. The next consideration is the ratio of labelled to unlabelled data. If the amount of data is insufficient the initial accuracy of the model might not be high enough to create the positive feedback needed to increase the accuracy of the model.

SSL is not guaranteed to work with all models. An example of where it would not work is a simple linear regression model. In this case the estimated labels will fall exactly on the fitted line, and as a result the addition of new data points will not move the fitted line and improve the model's accuracy. The reason that SSL can be used to improve the accuracy of Model 1 is because NMF can be shown to be the same as "k-means" clustering, where each basis vector represents the centroid of the cluster. With the addition of new data points the centroid position will constantly shift, and if the unlabelled data is from the same distribution as the labelled data, this might help the convergence of the centroid position.

Figure 37 and Figure 38 show how the RMSE of the model changes with the introduction of data batches. With each additional data batch, the available data for the model to use grows by 3 samples. Initially, the batches consist only of labelled data,

until the labelled data runs out, at which point unlabelled data is introduced (the red dashed lines in Figure 37 and Figure 38). It is at this point that you would typically stop training if you were not using SSL. The batch of unlabelled data is labelled by the currently trained model, and included into the training set, before the model is retrained. It is at this point that the accuracy of the model is measured against the original test data set and the next batch of unlabelled data is introduced. In both the combined ACE + Tara and NOMAD datasets, the same trend can be seen: the addition of unlabelled absorption data improves the accuracy of the model. In Figure 37 it appears as though the model is performing very well after the first iteration, but at this point the model as has only been trained with 3 samples, and still has very high variance, as seen by the large jumps in the next couple data batch iterations.

The training error of the model was measured using 10-fold cross validation and is represented as “CV error” in the plots. K-fold Cross Validation is the process of partitioning the training data into *k* partitions, such that each partition will be the test set on a given iteration, with the remaining data used as the training data. The results of the *k* model-evaluations are then averaged, resulting in less biased results.

The green dashed line represents the “mean model” test error which provides a benchmark, i.e. if the test error is not lower than the baseline error then it is not performing well.

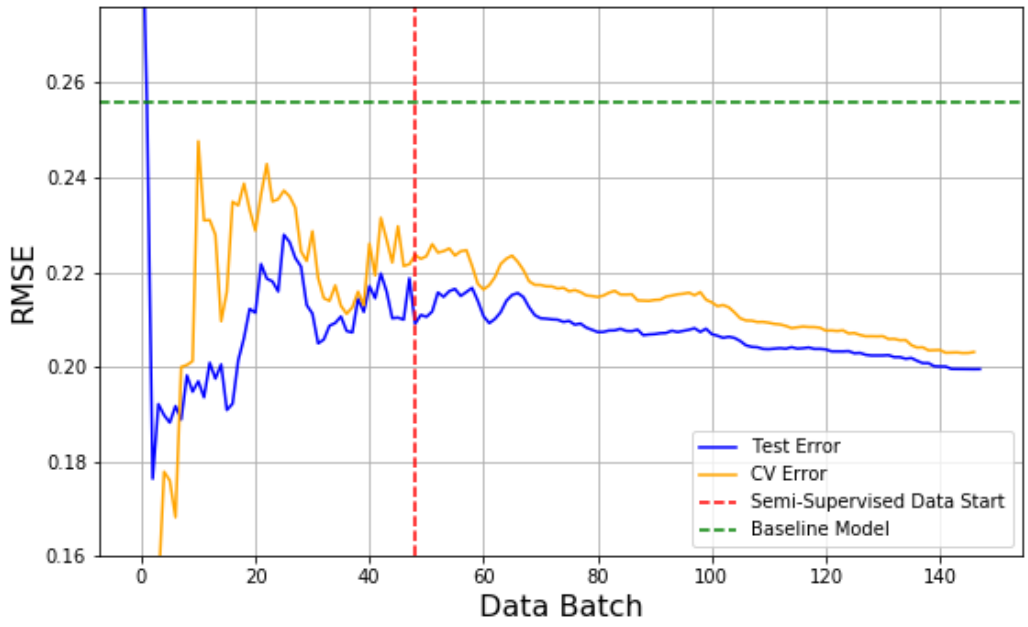


Figure 37: Training and test error of Model 1 when using the combined dataset.

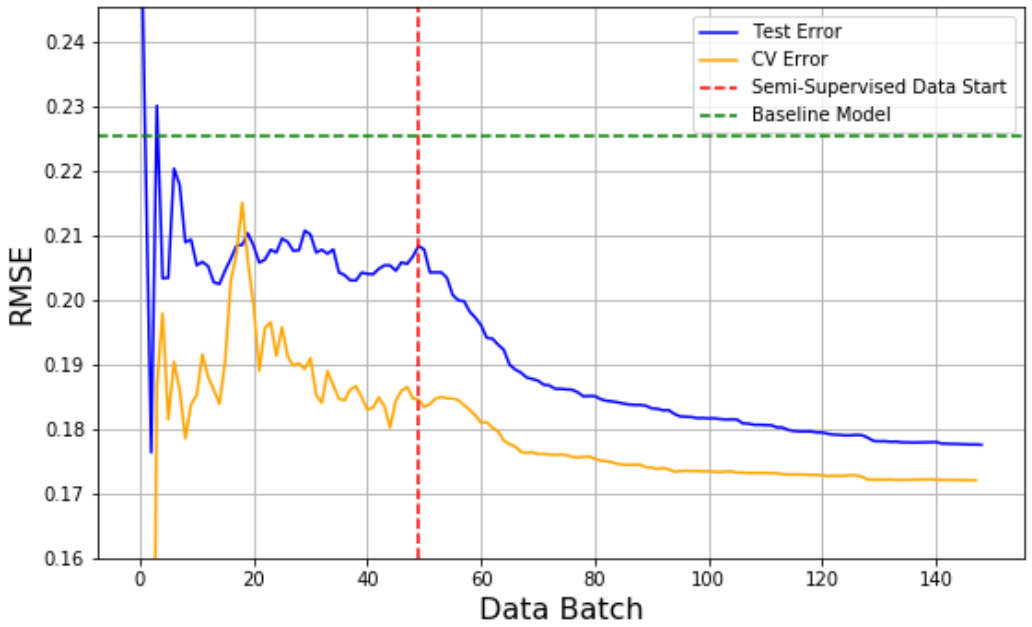


Figure 38: Training and test error of Model 1 when using the NOMAD dataset.

4.4.7 Model 1: Results

The matrix factorisation model has demonstrated that it is possible to factorise the absorption signal into a set of basis vectors, which can then be used to estimate the concentration of size classes present. It was further demonstrated that SSL can be used as an effective method for improving performance in scenarios where unlabelled data is available.

Table 8 shows that, for both datasets, the model is better than the baseline model (which only estimates the mean value) and how the error decreases after performing SSL. Since the size classes are in the unit interval, $RMSE \times 100\%$ represents the average % that the model was incorrect by for the given size class.

The R^2 values of the estimated Nano size class was consistently lower than the other size classes. These poor results could be attributed to ambiguities discovered with the marker pigment Fucoxanthin, which was originally used as a marker pigment for Diatoms [45], a species found within the Micro group, but was later found to be a precursor to the pigment 19'-Hexanoyloxyfucoxanthin [34], used as a marker pigment for Nano phytoplankton.

Table 8: Model 1 results including the baseline mean model for comparison.

		RMSE				R2		
Dataset	**Technique**	**Total**	**Pico**	**Nano**	**Micro**	**Pico**	**Nano**	**Micro**
ACE + Tara	Mean Model	0.256	0.281	0.158	0.307	0	0	0
ACE + Tara	NMF	0.21	0.197	0.240	0.19	0.687	0.142	0.694
ACE + Tara	Semi-Supervised	0.199	0.188	0.228	0.179	0.65	0.036	0.669
NOMAD	Mean Model	0.225	0.235	0.154	0.273	0	0	0
NOMAD	NMF	0.206	0.244	0.207	0.157	0.374	0.23	0.731
NOMAD	Semi-Supervised	0.178	0.2	0.173	0.157	0.37	0.075	0.673

4.5 Model 2: Gaussian Decomposition and Regression

The next model consists of two fundamental stages. The first stage is concerned with the estimation of pigment concentrations and the second stage estimates size class concentrations from the derived pigment concentrations.

In Model 1 the absorption signal was decomposed into basis vectors per size class, and in Model 2 the signal is decomposed into absorption by pigments and NAP, where each pigment is represented as a Gaussian band centred at a specific wavelength.

This two-stage model estimates the size class concentrations, first from Chlorophyll-a alone, and then from all of the estimated pigments. This is shown in Figure 39 as the "Primary Model Flow" and "Extended Model Flow".

4.5.1 Model 2: Aim

To develop a model that is capable of estimating phytoplankton size classes from their absorption spectra by estimating specific pigment concentrations that can then be used as a proxy to infer size class concentrations.

4.5.2 Model 2: Method

1. Split data into training and test data sets in a 60% to 40% ratio, respectively.
2. Decompose the absorption signal into Gaussian bands at specific wavelengths.
3. Train a series of non-linear regression models to estimate the Chlorophyll-a and accessory pigment concentrations from the Gaussian peaks.
4. Measure pigment estimation accuracy and show results.
5. Train an SVR model, using an RBF kernel, to estimate the size class concentrations from measured Chlorophyll-a.
6. Measure size class accuracy and show results.
7. Combine the models and measure the absorption to size class performance (Primary Model Flow).
8. Update the SVR model to estimate the size class concentrations from all the estimated pigment concentrations (Extended Model Flow).
9. Measure size class accuracy and show results.

4.5.3 Model 2: Data Used

The first stage model, which estimates pigment concentrations, and the final end-to-end evaluation, which estimates size classes from the absorption signal, is trained and

evaluated with data from the combined ACE + Tara datasets and for the NOMAD datasets separately. The second stage model, that estimates size classes from Chlorophyll-a, is trained and evaluated with pigment data gathered on all of the cruises. This combined pigment dataset does not require any intersecting absorption readings and is therefore much larger. Table 9 gives a breakdown of all the data used in these models.

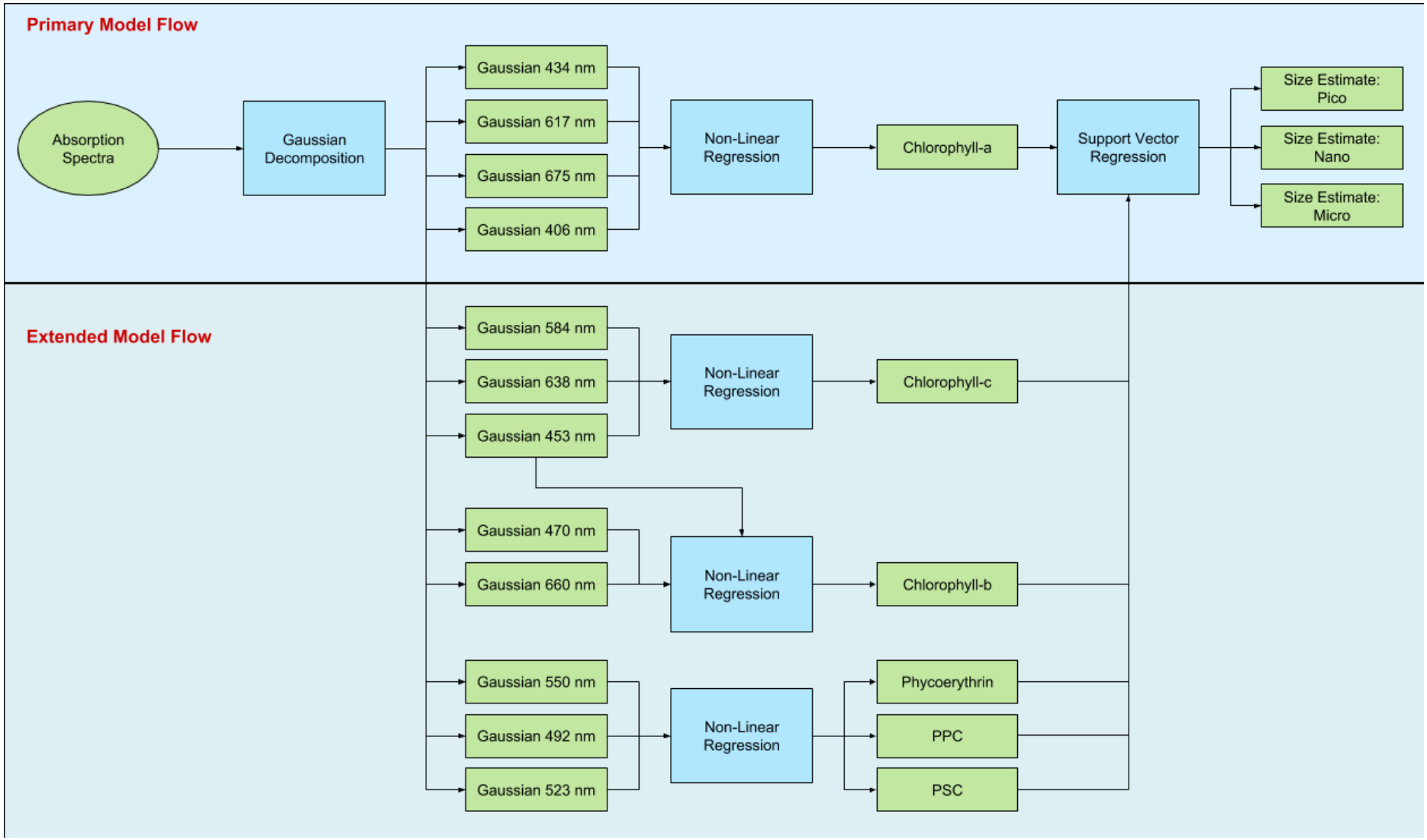


Figure 39: Model 2 process flow, estimating size classes by first estimating pigment concentrations.

Table 9: All of the data used for training the pigment estimation model and the size class concentration model.

Dataset	Data Type A	Used For	Intersection Data B	Samples
All	Pigments	Training	-	632
All	Pigments	Evaluation	-	271
ACE + Tara	Absorption	Training	Pigments	141
ACE + Tara	Absorption	Evaluation	Pigments	95
NOMAD	Absorption	Training	Pigments	145
NOMAD	Absorption	Evaluation	Pigments	98

4.5.4 Model 2: Estimating Pigments from Absorption

This technique is based on the fact that the constituents of an absorption signal are additive and that pigments typically have very specific and narrow absorption spectra. The absorption signal is decomposed into a set of Gaussian bands, at predefined wavelengths, which are then regressed against known pigment concentrations. This technique has been successfully applied in numerous studies [76] and [77], but the specific methods described in Chase et al. 2013 are used in this work [27].

The NOMAD and the combined Tara + ACE absorption datasets were decomposed into twelve Gaussian bands. The peak locations and widths of these Gaussian bands are based on laboratory measurements of the absorption spectra of the various pigments and were further finetuned against in situ data by Chase et al. 2013 [27]. The Gaussian peaks and widths are shown in Table 10.

The peak locations are far more specific than the detection wavelengths used in HPLC, as shown in Table 4, as this technique relies only on the absorption spectra in order to separate out the constituent pigments. HPLC, on the other hand, does not have the risk of spectral ambiguity as the pigment present at the sensor is a function of time, as each pigment takes a different amount of time to pass through the adsorption material.

Given that the size labels used in this research were estimated through DPA, it would be logical to estimate the specific pigments so that the sizes can be calculated through DPA. This is, however, unfortunately not entirely possible because of overlap in the

absorption spectra for some of these diagnostic pigments, thereby making it impossible to estimate all of the pigments from the absorption signal. This is especially apparent with the Photoprotective Carotenoids (PPC) and the Photosynthetic Carotenoids (PSC).

Table 10: The peak locations and widths of the 12 Gaussian bands used to reconstruct the absorption signal, along with the pigments responsible for absorption at these wavelengths [27].

Pigments	Peak Location (nm)	Standard Deviation (nm)
Chlorophyll-a&c	406	16
Chlorophyll-a	434	12
Chlorophyll-b&c	453	12
Chlorophyll-b	470	13
PPC	492	16
PSC	523	14
Phycoerythrin	550	14
Chlorophyll-c	584	16
Chlorophyll-a	617	13
Chlorophyll-c	638	11
Chlorophyll-b	660	11
Chlorophyll-a	675	10
PPC = α-carotine + β-carotine + zeaxanthin + alloxanthin + diadinoxanthin. PSC = 19′-hexanoyloxyfucoxanthin+fucoxanthin+19′-butanoyloxyfucoxanthin+peridinin		

The total particulate absorption can be represented by the sum of the absorption of all of the pigment Gaussian bands:

$$a_p^*(\lambda) = \sum_i^{12} a^i{}_{gauss}(\lambda) + a_{NAP}(\lambda)\,, \tag{4.20}$$

where $a_p^*(\lambda)$ represents the particulate absorption, $a^i{}_{gauss}(\lambda)$ the i^{th} Gaussian band and where $a_{NAP}(\lambda)$ represents the absorption by Non-Algal Particles (NAP), as defined by the following exponential:

$$a_{NAP}(\lambda) = a_{NAP}(400nm)\, e^{-0.01*(\lambda-400nm)}, \tag{4.21}$$

where $a_{NAP}(400nm)$ represents the peak absorption at 400nm and is in effect a scaling parameter defining the overall contribution of the NAP signal in the absorption signal.

A linear non-negative least-squares algorithm was used to find the best fitting magnitudes of the Gaussian bands, using the widths specified in Table 10, and the NAP signal for the datasets. An example of a decomposed signal, along with the reconstructed signal can be seen in Figure 40.

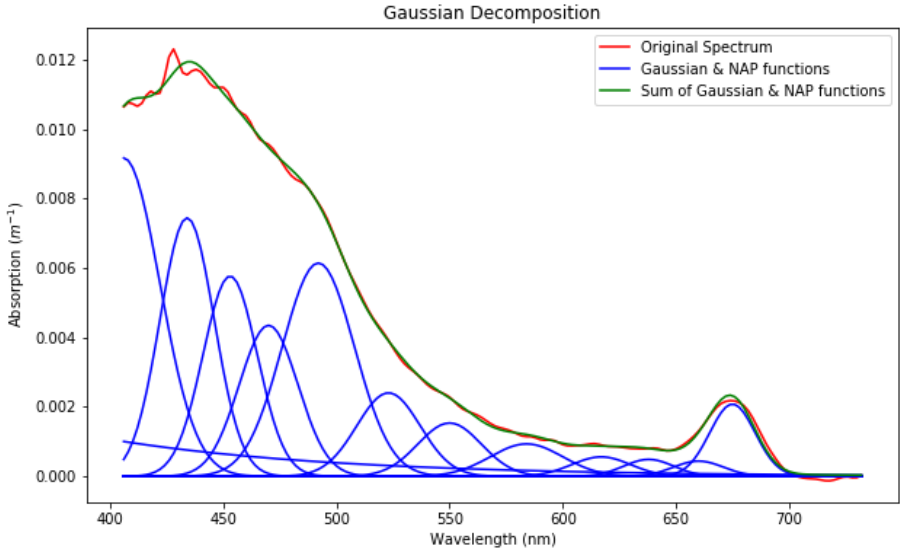


Figure 40: Absorption signal decomposed into a series of Gaussian bands and NAP function.

In order to estimate the pigment concentrations from their Gaussian amplitudes, the following equation describes the exponential relationship between a given gaussian band peak and each of the pigments:

$$\log[a^{i}{}_{gauss}(\lambda)] = A_{ij} + B_{ij}\log\,[pigment_j]. \tag{4.22}$$

This can be rearranged such that the concentration of a given pigment can be solved:

$$pigment_j = \left(\frac{a^{i}{}_{gauss}(\lambda)}{e^{A_{ij}}}\right)^{1/B_{ij}}. \tag{4.23}$$

The values of A_{ij} and B_{ij} define the relationship between the i^{th} Gaussian peak and the j^{th} pigment concentration. These values are calculated by non-linear regression before the estimated pigment concentrations can be found.

4.5.5 Model 2: Pigment Estimation Results

Strong correlations were found between the Gaussian peaks and the key Chlorophyll pigments. Chlorophyll-a, the primary pigment used in photosynthesis, was best estimated with the Gaussian peaks centred at 617nm and 675nm, with R^2 values of 0.962 and 0.951 respectively, for the NOMAD dataset. This is almost certainly because this is the only pigment that absorbs light at these wavelengths. Chlorophyll-b is the pigment with the lowest correlation to any one Gaussian peak, with the highest R^2 value of 0.5 for the absorption peak at 470nm. This can be attributed to the high number of pigments which absorb light at this wavelength, as seen in Figure 7 of Chapter 2.

When using a combination of peaks to estimate these pigments, the RMSE reduced for both the NOMAD and the combined Tara + Ace datasets. As a result, these estimated pigments were used in the estimation of size classes in the next stage of the model. The complete set of estimated pigment results can be seen in Table 11 and Table 12. Since the size classes are in the unit interval, $RMSE \times 100\%$ represents the average $\%$ that the model was incorrect by for the given size class.

Table 11: Single wavelength pigment estimation results with NOMAD on the left and the combined dataset on the right.

		NOMAD		Tara + ACE	
Estimated Pigment	**λ**	**RMSE**	**R^2**	**RMSE**	**R^2**
Chlorophyll-a	434	0.725	0.906	0.165	0.924
Chlorophyll-a	617	0.532	0.962	0.261	0.885
Chlorophyll-a	675	0.498	0.951	0.172	0.920
Chlorophyll-b	470	0.075	0.5	0.083	0.000
Chlorophyll-b	660	0.081	0.268	0.083	0.098
Chlorophyll-c	638	0.257	0.481	0.121	0.778
Chlorophyll-c	584	0.201	0.696	0.126	0.735
PPC	492	0.1	0.797	0.066	0.596
PSC	523	0.259	0.935	0.195	0.834

Table 12: Multi-wavelength pigment estimation results with NOMAD on the left and the combined dataset on the right.

		NOMAD		Tara + ACE	
Estimated Pigment	**λ**	**RMSE**	**R^2**	**RMSE**	**R^2**
Chlorophyll-a	406, 434, 617, 675	0.490	0.956	0.150	0.937
Chlorophyll-b	453, 470, 660	0.061	0.657	0.083	0.119
Chlorophyll-c	453, 584, 638	0.196	0.712	0.104	0.784

4.5.6 Model 2: Estimating Size from the Derived Pigments

Once the pigment estimation model was complete these pigments were used to estimate the size classes through SVR using an RBF kernel. The value of the hyperparameter C, see equation 2.2, was set to 1000 and was chosen based on it being the point just before the model begins to overfit, as shown in Figure 41.

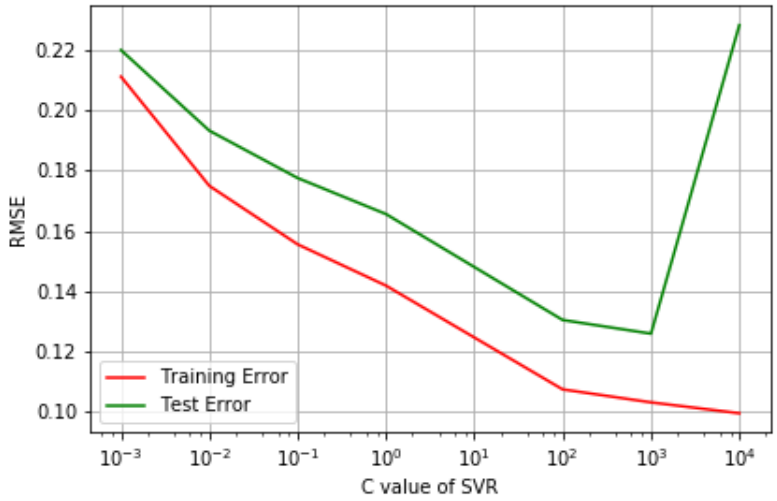


Figure 41: Training and test errors of the SVR model against the hyperparameter C, where derived pigments are regressed against particle size.

As shown in Figure 39, two approaches were attempted. The primary flow only made use of estimated Chlorophyll-a to infer size classes, since Chlorophyll-a is the pigment for which the most amount of data is available. This model is simplistic yet practical. The extended model flow makes use of the rest of the derived pigments. As expected, the results when using all of the estimated pigments are better than when only Chlorophyll-a. Table 13 shows the complete set of results for this model, where it can be seen that it performs considerably better than the baseline mean model. It is also worth noting that this model performs better than Model 1 for both of the datasets. Since

the size classes are in the unit interval, $RMSE \times 100\%$ represents the average % that the model was incorrect by for the given size class.

Table 13: Size class estimation results for Model 2.

		RMSE				R^2		
Dataset	**Technique**	**Total**	**Pico**	**Nano**	**Micro**	**Pico**	**Nano**	**Micro**
ACE + Tara	Mean Model	0.256	0.281	0.158	0.307	0	0	0
ACE + Tara	Chlorophyll-a	0.233	0.26	0.168	0.259	0.178	0.101	0.292
ACE + Tara	All Pigments	0.174	0.175	0.158	0.187	0.614	0.09	0.63
NOMAD	Mean Model	0.225	0.235	0.154	0.273	0	0	0
NOMAD	Chlorophyll-a	0.177	0.177	0.146	0.204	0.456	0.18	0.465
NOMAD	All Pigments	0.126	0.128	0.108	0.139	0.723	0.537	0.751

4.6 Model 3: Empirical Equation for Absorption

Due to the regular structure in the absorption signal it was decided that a simple empirical equation could be used to approximate the signal, as a means of dimensionality reduction for training a particle size model. This section describes such an equation, which reduces the particulate absorption signal into a single parameter, before using the parameter to estimate particle size through the use of SVR, using the RBF kernel.

4.6.1 Model 3: Aim

To develop a model that is capable of estimating phytoplankton size classes from absorption data through the use of an empirical equation and SVR.

4.6.2 Model 3: Method

1. Derive an empirical equation capable of parameterising the absorption signal.
2. Estimate the optimal values for the parameters within the equation via a linear least squares approximation, based on the non-intersecting absorption data.
3. Measure the total explained variance of the approximated signals.
4. Split absorption/pigment intersection data into training and test data sets in a 60% to 40% ratio, respectively.

5. Train an SVR model, using an RBF kernel, with the calculated signal parameters and the known particle sizes, for the training data of both the NOMAD and combined datasets.
6. Measure size class accuracy and show results.

4.6.3 Model 3: Data Used

For each of the cruises there are two sets of data used: the intersecting and the non-intersecting absorption/pigment data. As per the other models, the size classes Pico, Nano and Micro are estimated through DPA using the pigment data. The non-intersecting absorption data is data that has no pigment contribution for the same sample. This absorption data is what is used in defining the empirical equation and finding the optimal values for the parameters within the equation.

Table 14: All the data used for training the pigment estimation model and the size class concentration from Chlorophyll-a model.

Dataset	Data Type A	Used For	Intersection Data B	Samples
ACE + Tara	Absorption	Empirical Equation	-	344214
ACE + Tara	Absorption	Training	Pigments	141
ACE + Tara	Absorption	Evaluation	Pigments	95
NOMAD	Absorption	Empirical Equation	-	766
NOMAD	Absorption	Training	Pigments	145
NOMAD	Absorption	Evaluation	Pigments	98

4.6.4 Model 3: Defining the Empirical Equation

The PCA results showed that over 99.5% of the variance is captured in the first principal component, and as such, it would be a good candidate to base the empirical equation on. The eigenvector representing the first principal component was first adjusted into a range between [0-1] by applying the following normalisation:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \tag{4.24}$$

where $x = (x_1 \dots x_n)$, and z_i is the i^{th} normalised value. The signal has two distinct peaks, at roughly 440nm and 675nm respectively. This is mainly as a result of

absorption by Chlorophyll-a, although many other pigments also affect the spectral shape of the absorption signal. Based on the aforementioned characteristics of the signal, an additive empirical equation was created with the following form:

$$f(\lambda) = A \cdot (a_1 X_1 + a_2 X_2 + m\lambda), \tag{4.25}$$

$$X_1 = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{-(\lambda-\mu_1)^2}{2\sigma_1^2}}, \tag{4.26}$$

$$X_2 = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{\frac{-(\lambda-\mu_2)^2}{2\sigma_2^2}}, \tag{4.27}$$

where A represents a scaling value of the entire function, equivalent to the eigenvalue in PCA and, X_1 and X_2 represent two Gaussian functions, one for each of the peaks, each with their own scaling values a_1 and a_2. A linear slope parameter m was introduced to apply a tilt to the Gaussians to introduce a bit of flexibility and reduce the bias of the function. Within the Gaussian distributions μ_1 and μ_2 represent the positions of the peaks and σ_1^2 and σ_2^2 control the width of the bell curve. This is the first step in parameterising the absorption data signal, but further analysis showed that the number of parameters could be further reduced.

This equation was fitted to the signal using a linear least squares technique, so that the parameters could be assessed. It was found that the second peak in the signal is only 27% of the size of the first peak and that the optimal Gaussian width of the first peak is $\sim 70.71nm$ while the optimal width of the second peak is $\sim 10nm$. The best fitting peaks of the two Gaussians were at 430nm and 672nm and not the typical 440nm and 675nm peaks of Chlorophyll-a. Substituting these values back into equation 4.25 yields the following:

$$a(\lambda) = A \cdot \left(e^{\frac{-(\lambda-430)^2}{10^4}} + 0.27 e^{\frac{-(\lambda-672)^2}{200}} + 10^{-4}\lambda \right), \quad \{\lambda \in \mathbb{R}: 400 < \lambda < 700\}. \tag{4.28}$$

This equation was fitted to all of the non-intersecting absorption data, finding the optimal value of A in each case, after which R^2 was used to evaluate the performance. The final R^2 score was 0.991. This simple equation, with a single variable, is capable of

explaining 99% of the variance in the absorption data. Figure 42 shows an example of how well the approximated signal matches the original.

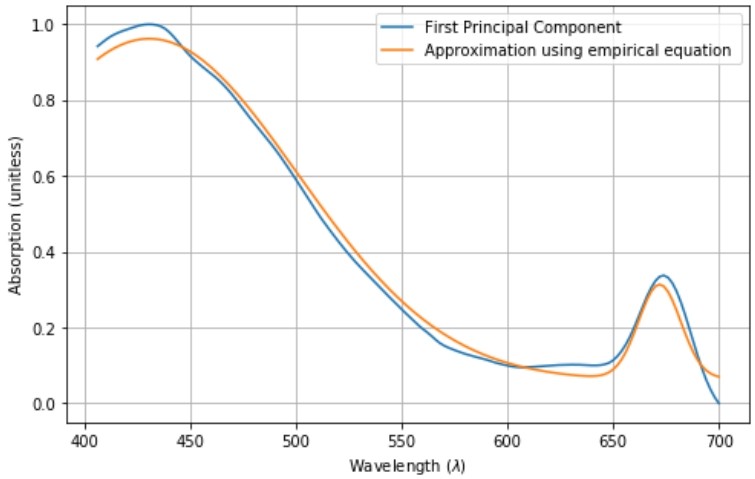


Figure 42: The empirical equation approximating the first principal component.

4.6.5 Model 3: Estimating Size Classes

Once this parameterisation technique had been defined it could be used in the estimation of size classes. For each of the absorption signals in the intersection datasets the parameter A from equation 4.25 was estimated and then regressed against the particle size values using SVR.

When configuring SVR, the hyperparameter C, refer to equation 2.2, was chosen such that the model was not overfit. This can be seen in Figure 43 where the test error begins to deteriorate after C=100, even though the training error continues to decrease. It was for this reason that a value of 100 was chosen.

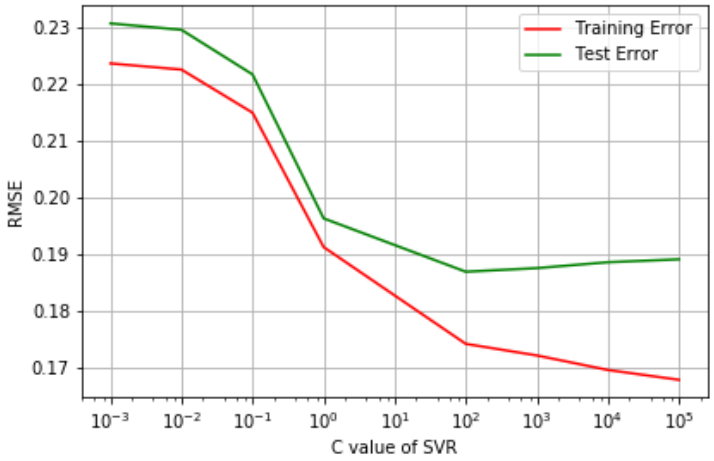


Figure 43: Training and test errors of the SVR model against the hyperparameter C, for the empirical model, regressed against particle size.

4.6.6 Model 3: Results

When estimating size classes, the model outperformed the mean model for both the NOMAD and the Tara + ACE combined datasets. Far better performance was achieved for the NOMAD dataset, with an improvement of 17% over the mean model, whereas for the noisier combined dataset a gain of only 9% was obtained. It is shown that the empirical model does not perform as well as the previous two models but considering the relative simplicity by comparison it is still reasonably effective.

Since the size classes are in the unit interval, $RMSE \times 100\%$ represents the average $\%$ that the model was incorrect by for the given size class.

Table 15: Model 3 results.

		RMSE				R^2		
Dataset	**Technique**	**Total**	**Pico**	**Nano**	**Micro**	**Pico**	**Nano**	**Micro**
ACE + Tara	Mean Model	0.256	0.281	0.158	0.307	0	0	0
ACE + Tara	Empirical + SVR	0.234	0.275	0.146	0.258	0.176	0.126	0.304
NOMAD	Mean Model	0.225	0.235	0.154	0.273	0	0	0
NOMAD	Empirical + SVR	0.187	0.191	0.144	0.218	0.325	0.12	0.399

4.7 Model 4: Ensemble

The final ensemble model combines the previous three models into a single model. This is achieved by combining all of the size estimate outputs from each of the models and using them as the input into another SVR model using an RBF kernel. This is done in order to increase the overall predictive performance, obtaining better results than any one of the models individually.

The optimal value for the hyperparameter C of the SVR algorithm was set to 1 based on the overfitting point, as identified in Figure 44.

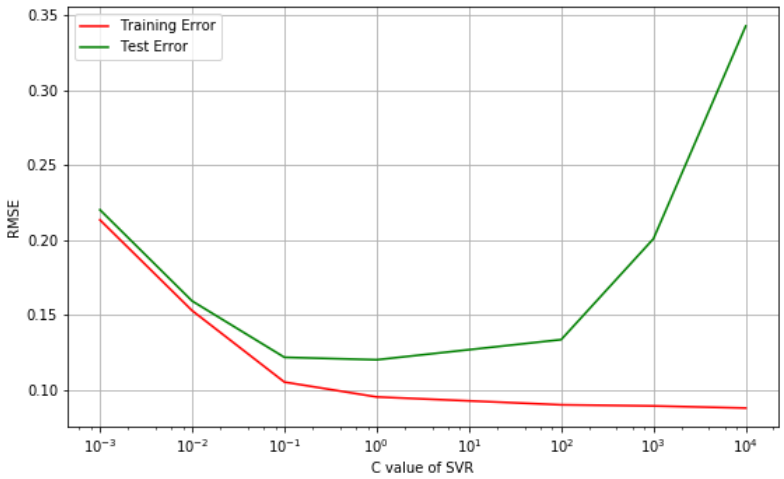


Figure 44: Training and test errors of the SVR model against the hyperparameter C, for the ensemble model, regressed against particle size

It can be seen in Table 16 that the ensemble model achieved far better results than any of the previous models, for both datasets. The RMSE went down and the R^2 value went up, showing a lower error and a better correlation with the true size classes.

Since the size classes are in the unit interval, $RMSE \times 100\%$ represents the average $\%$ that the model was incorrect by for the given size class.

Table 16: Ensemble Model 4 results.

		RMSE				R^2		
Dataset	**Technique**	**Total**	**Pico**	**Nano**	**Micro**	**Pico**	**Nano**	**Micro**
ACE + Tara	Mean Model	0.256	0.281	0.158	0.307	0	0	0
ACE + Tara	Ensemble	0.136	0.133	0.129	0.146	0.792	0.234	0.771
NOMAD	Mean Model	0.225	0.235	0.154	0.273	0	0	0
NOMAD	Ensemble	0.12	0.126	0.105	0.127	0.729	0.535	0.784

4.8 Final Results

The results of all four models developed in this research can be found in Table 17. It can be seen that for both datasets the final ensemble model performs better than any of the other models, with a 47% improvement over the Mean Model for both datasets.

Table 17: All of the results combined for comparison.

		RMSE				R^2		
Dataset	**Technique**	**Total**	**Pico**	**Nano**	**Micro**	**Pico**	**Nano**	**Micro**
ACE + Tara	Mean Model	0.256	0.281	0.158	0.307	0	0	0
ACE + Tara	NMF	0.21	0.197	0.240	0.19	0.687	0.142	0.694
ACE + Tara	Semi-Supervised	0.199	0.188	0.228	0.179	0.65	0.036	0.669
ACE + Tara	Chlorophyll-a	0.233	0.26	0.168	0.259	0.178	0.101	0.292
ACE + Tara	All Pigments	0.174	0.175	0.158	0.187	0.614	0.09	0.63
ACE + Tara	Empirical + SVR	0.234	0.275	0.146	0.258	0.176	0.126	0.304
ACE + Tara	Ensemble	0.136	0.133	0.129	0.146	0.792	0.234	0.771
NOMAD	Mean Model	0.225	0.235	0.154	0.273	0	0	0
NOMAD	NMF	0.206	0.244	0.207	0.157	0.374	0.23	0.731
NOMAD	Semi-Supervised	0.178	0.2	0.173	0.157	0.37	0.075	0.673
NOMAD	Chlorophyll-a	0.177	0.177	0.146	0.204	0.456	0.18	0.465
NOMAD	All Pigments	0.126	0.128	0.108	0.139	0.723	0.537	0.751
NOMAD	Empirical + SVR	0.187	0.191	0.144	0.218	0.325	0.12	0.399
NOMAD	Ensemble	0.120	0.126	0.105	0.127	0.729	0.535	0.784

In order to put these results into perspective, Table 18 provides the results obtained by other researchers using different techniques. All of these models were run against the NOMAD dataset, making the results directly comparable. The only RMSE results that perform better than the ensemble model are those of Hu et al. [55], using a Random Forest model, which obtains results that are 0.6% better. The R^2 results of the ensemble model when compared against the other published results show that the Micro size class is particularly well represented and attains a better fit than any of the other published results. When looking at the R^2 results of the Nano size class, only the Random Forest results of Hu et al. [55] perform better. The R^2 results of the Pico size class perform better than Wang et al. [15] and Zhang et al. [31] but, again, not as well as any of the model results of Hu et al. [55].

Table 18: The results of the Ensemble models compared to the results published by other researchers, using various techniques for the NOMAD dataset.

		RMSE				R^2		
Reference	**Technique**	**Total**	**Pico**	**Nano**	**Micro**	**Pico**	**Nano**	**Micro**
	Ensemble	0.120	0.126	0.105	0.127	0.729	0.535	0.784
Wang et al. [15]	PCA + Regression	0.124	0.109	0.128	0.134	0.548	0.280	0.608
Zhang et al. [31]	SVD	0.184	0.18	0.17	0.2	0.504	0.270	0.593
Hu et al. [55]	Neural Network	0.134	0.12	0.13	0.15	0.77	0.45	0.72
Hu et al. [55]	Random Forest	0.114	0.10	0.11	0.13	0.82	0.56	0.78
Hu et al. [55]	SVM	0.124	0.11	0.12	0.14	0.80	0.48	0.74

Chapter 5: Conclusion and Recommendations

5.1 Conclusion

This research evaluated a number of techniques for the estimation of phytoplankton size classes (Pico, Nano and Micro) from their Inherent Optical Properties. Some of the data used was collected on the ACE cruise, while the other datasets – NOMAD and Tara – are publicly available and were obtained online. These datasets contain both absorption data and coincident pigment data. The true size classes of the samples were not measured directly but were estimated through a technique known as Diagnostic Pigment Analysis.

As a precursor to the model development, the structure of the absorption and particle size data was analysed. Through the use of a covariance matrix it was shown that the absorption data is highly structured, and that dimensionality reduction techniques would work well on it. It was then confirmed, through the use of PCA, that the dimensionality could be drastically reduced, with 99% of the variance explained by the first principal component alone. The highly structured nature of the absorption data led to the development of an empirical equation, modelled against the first principal component, that obtained an R^2 score of 0.991 with a single scaling parameter.

Four models were then developed in order to estimate the phytoplankton size classes, as provided by DPA, from the absorption data. It was shown that all of the developed models performed better than the baseline model, which only estimates the mean values per size class. Additionally, the results of the final ensemble model are comparable to, and perform better than, most other published models on the NOMAD dataset.

The first model showed that a set of basis vectors, representing the typical absorption signal per size class, could be used to decompose an absorption signal into a set of relative abundances for each of the size classes. Although this technique worked relatively well, and showed an RMSE improvement of 18% over the baseline model for the combined dataset and 8% for the NOMAD dataset, it did not perform as well as some of the other models. It was then shown that the accuracy of the model could be improved upon by using SSL, with the addition of unlabelled data. The improved model

showed an RMSE improvement of 22% over the baseline model for the combined dataset and 21% for the NOMAD dataset.

The second model made use of Gaussian decomposition to first estimate the relative concentrations of pigments present in the sample, before utilising SVR to estimate size class concentrations. When only making use of Chlorophyll-a for the estimation of size classes, the model's performance was poorer than when all of the derived pigments were used. Other than the final ensemble model, this model performed the best with an RMSE improvement of 32% over the baseline model for the combined dataset and 44% for the NOMAD dataset.

The third model made use of an empirical equation, where the absorption was parameterised into a single value. This value was then modelled, through the use of SVR, to the particle size as provided by DPA. Even though this parameterisation captures 99.1% of the variance in the absorption signal, it was only capable of an RMSE improvement of 9% over the baseline model for the combined dataset and 17% for the NOMAD dataset. These results are not as good as the others but considering the simplicity of the model it may still have practical applications.

The fourth and final model combined all of the previous models into an ensemble model, where the outputs of the other models are used as the input to SVR. This model performed very well and has results comparable with the best published techniques. An RMSE improvement of 47% over the baseline model was achieved for both the combined dataset and the NOMAD dataset.

This thesis therefore shows that particle size and pigment information can be inferred from the particulate absorption signal, and that the results obtained are comparable to other published results.

5.2 Recommendations for Future Work

I believe that there is a lot of potential for spatio-temporal models, developed using Bayesian networks or something similar. These models could show how phytoplankton numbers or sizes change with respect to space and time. These probabilistic-type models could be very valuable for inferring information in areas of the ocean where readings do not exist, but where adjacent records do.

Some of the variables that were not considered in this thesis but that do affect the absorption signal are temperature and light exposure. All of the models in this paper would be improved if these variables could be included.

Bibliography

[1] C. S. Reynolds, *The ecology of phytoplankton*. Cambridge, 2006.

[2] F. B. Metting, "Biodiversity and application of microalgae", *Journal of Industrial Microbiology & Biotechnology*, vol. 17, no. 5–6, pp. 477–489, 1996.

[3] D. G. Mann and P. Vanormelingen, "An inordinate fondness? The number, distributions, and origins of diatom species", *Journal of Eukaryotic Microbiology*, vol. 60, no. 4, pp. 414–420, 2013.

[4] M. Guiry, "How many species of algae are there?", *Journal of Phycology*, vol. 48, 2012.

[5] R. Lindsey and M. Scott, "Phytoplankton", 2010. [Online]. Available: https://earthobservatory.nasa.gov/Features/Phytoplankton/. [Accessed: 21-Jan-2018].

[6] Z. V. Finkel, J. Beardall, K. J. Flynn, A. Quigg, T. A. V Rees, and J. A. Raven, "Phytoplankton in a changing world: Cell size and elemental stoichiometry", *Journal of Plankton Research*, vol. 32, no. 1, pp. 119–137, 2010.

[7] J. Aiken, S. Alvain, R. Barlow, H. Bouman, A. Bracher, and R. J. W. Brewin, "IOCCG Report: Phytoplankton Functional Types from Space", 2014.

[8] A. Longhurst, S. Sathyendranath, T. Platt, and C. Caverhill, "An estimate of global primary production in the ocean from satellite radiometer data", *Journal of Plankton Research*, vol. 17, no. 6, 1995.

[9] C. L. Sabine, R. A. Feely, N. Gruber, R. M. Key, K. Lee, and J. L. Bullister, "The Oceanic Sink for Anthropogenic CO2", *Science*, vol. 305, no. 5682, pp. 367–371, 2004.

[10] U. Riebesell, K. G. Schulz, R. G. J. Bellerby, M. Botros, P. Fritsche, and M. Meyerhöfer, "Enhanced biological carbon consumption in a high CO_2 ocean", *Nature*, vol. 450, p. 545, Nov. 2007.

[11] M. Reimer, E. Mikolajewicz, and U. Winguth, "Future ocean uptake of CO_2: interaction between ocean circulation and biology", *Climate Dynamics*, vol. 12, no. 10, pp. 711–722, 1996.

[12] S. L. Deppeler and A. T. Davidson, "Southern Ocean Phytoplankton in a Changing Climate", *Frontiers in Marine Science*, vol. 4, no. February, 2017.

[13] T. S. Kostadinov, S. Milutinovi, I. Marinov, and A. Cabré, "Carbon-based phytoplankton size classes retrieved via ocean color estimates of the particle size distribution", *Ocean Science*, vol. 12, no. 2, pp. 561–575, 2016.

[14] E. Boss, M. S. Twardowski, and S. Herring, "Shape of the particulate beam attenuation spectrum and its inversion to obtain the shape of the particulate size

distribution", *Applied Optics*, vol. 40, no. 27, pp. 4885–4893, 2001.

[15] S. Wang, J. Ishizaka, T. Hirawake, Y. Watanabe, Y. Zhu, and M. Hayashi, "Remote estimation of phytoplankton size fractions using the spectral shape of light absorption", *Optics Express*, vol. 23, no. 8, p. 10301, 2015.

[16] T. Kameda and J. Ishizaka, "Size-Fractionated Primary Production Estimated by a Two-Phytoplankton Community Model Applicable to Ocean Color Remote Sensing", *Journal of Oceanography*, vol. 61, no. 4, pp. 663–672, Aug. 2005.

[17] K. A. Miklasz and M. W. Denny, "Diatom sinking speeds: Improved predictions and insight from a modified Stoke's law", *Limnology and Oceanography*, vol. 55, no. 6, pp. 2513–2525, 2010.

[18] J. Happel and H. Brenner, *Low Reynolds number hydrodynamics: with special applications to particulate media*. Springer Netherlands, 1983.

[19] G. Woodward, B. Ebenman, M. Emmerson, J. M. Montoya, J. M. Olesen, and A. Valido, "Body size in ecological networks", *Trends in Ecology & Evolution*, vol. 20, no. 7, pp. 402–409, Jan. 2005.

[20] F. S. Scharf, F. Juanes, and R. A. Rountree, "Predator size - Prey size relationships of marine fish predators: Interspecific variation and effects of ontogeny and body size on trophic-niche breadth", *Marine Ecology Progress Series*, vol. 208, pp. 229–248, 2000.

[21] S. Jennings, K. J. Warr, and S. Mackinson, "Use of size-based production and stable isotope analyses to predict trophic transfer efficiencies and predator-prey body mass ratios in food webs", *Marine Ecology Progress Series*, vol. 240, pp. 11–20, 2002.

[22] A. Morel and L. Prieur, "Analysis of variations in ocean color", *Limnology and Oceanography*, vol. 22, no. 4, pp. 709–722.

[23] H. R. Gordon, D. K. Clark, J. W. Brown, O. B. Brown, R. H. Evans, and W. W. Broenkow, "Phytoplankton pigment concentrations in the Middle Atlantic Bight: comparison of ship determinations and CZCS estimates", *Applied optics*, vol. 22, no. 1, pp. 20–36, Jan. 1983.

[24] C. Mobley, *Light and Water: Radiative Transfer in Natural Waters*. Academic Press, 1994.

[25] R. J. W. Brewin, S. Sathyendranath, T. Hirata, S. J. Lavender, R. M. Barciela, and N. J. Hardman-Mountford, "A three-component model of phytoplankton size class for the Atlantic Ocean", *Ecological Modelling*, vol. 221, no. 11, pp. 1472–1483, 2010.

[26] Sea-Bird Scientific, "ac-s In-Situ Spectrophotometer Datasheet." 2017.

[27] A. Chase, E. Boss, R. Zaneveld, A. Bricaud, H. Claustre, and J. Ras,

"Decomposition of in situ particulate absorption spectra", *Methods in Oceanography*, vol. 7, pp. 110–124, 2013.

[28] A. Bracher, M. H. Taylor, B. Taylor, T. Dinter, R. Röttgers, and F. Steinmetz, "Using empirical orthogonal functions derived from remote-sensing reflectance for the prediction of phytoplankton pigment concentrations", *Ocean Science*, vol. 11, no. 1, pp. 139–158, 2015.

[29] T. Varunan and P. Shanmugam, "A model for estimating size-fractioned phytoplankton absorption coefficients in coastal and oceanic waters from satellite data", *Remote Sensing of Environment*, vol. 158, pp. 235–254, 2015.

[30] A. C. Brito, C. Sá, V. Brotas, R. J. W. Brewin, T. Silva, and J. Vitorino, "Effect of phytoplankton size classes on bio-optical properties of phytoplankton in the Western Iberian coast: Application of models", *Remote Sensing of Environment*, vol. 156, pp. 537–550, 2015.

[31] X. Zhang, Y. Huot, A. Bricaud, and H. M. Sosik, "Inversion of spectral absorption coefficients to infer phytoplankton size classes, chlorophyll concentration, and detrital matter", *Applied Optics*, vol. 54, no. 18, p. 5805, 2015.

[32] J. Uitz, H. Claustre, A. Morel, and S. B. Hooker, "Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll", *Journal of Geophysical Research: Oceans*, vol. 111, no. 8, 2006.

[33] R. J. W. Brewin and T. Hirata, "Case Study 9: Detecting Phytoplankton Community Structure from Ocean Colour", *Handbook of Satellite Remote Sensing Image Interpretation: Applications for Marine Living Resources Conservation and Management*, pp. 125–140, 2011.

[34] T. Hirata, N. J. Hardman-Mountford, R. J. W. Brewin, J. Aiken, R. Barlow, and K. Suzuki, "Synoptic relationships between surface Chlorophyll-a and diagnostic pigments specific to phytoplankton functional types", *Biogeosciences*, vol. 8, no. 2, pp. 311–327, 2011.

[35] G. Wang, W. Cao, G. Wang, and W. Zhou, "Phytoplankton size class derived from phytoplankton absorption and chlorophyll-a concentrations in the northern South China Sea", *Chinese Journal of Oceanology and Limnology*, vol. 31, no. 4, pp. 750–761, 2013.

[36] W. H. Slade and E. Boss, "Spectral attenuation and backscattering as indicators of average particle size", *Applied Optics*, vol. 54, no. 24, p. 7264, 2015.

[37] T. S. Kostadinov, D. A. Siegel, and S. Maritorena, "Retrieval of the particle size distribution from satellite ocean color observations", *Journal of Geophysical Research*, vol. 114, no. C9, p. C09015, Sep. 2009.

[38] T. S. Kostadinov, D. A. Siegel, and S. Maritorena, "Global variability of

phytoplankton functional types from space: Assessment via the particle size distribution", *Biogeosciences*, vol. 7, no. 10, pp. 3239–3257, 2010.

[39] C. B. Mouw, N. J. Hardman-Mountford, S. Alvain, A. Bracher, R. J. W. Brewin, and A. Bricaud, "A Consumer's Guide to Satellite Remote Sensing of Multiple Phytoplankton Groups in the Global Ocean", *Frontiers in Marine Science*, vol. 4, no. February, 2017.

[40] J. Uitz, Y. Huot, F. Bruyant, M. Babin, and H. Claustre, "Relating phytoplankton photophysiological properties to community structure on large scales", *Limnology and Oceanography*, vol. 53, no. 2, pp. 614–630, 2008.

[41] M. V Moreno-Arribas and M. C. Polo, "CHROMATOGRAPHY | High-performance Liquid Chromatography", in *Encyclopedia of Food Sciences and Nutrition (Second Edition)*, Second Edition., B. Caballero, Ed. Oxford: Academic Press, 2003, pp. 1274–1280.

[42] J. Ras, H. Claustre, and J. Uitz, "Spatial variability of phytoplankton pigment distributions in the Subtropical South Pacific Ocean: Comparison between in situ and predicted data", *Biogeosciences*, vol. 5, no. 2, pp. 353–369, 2008.

[43] J. T.O. Kirrk, *Light and Photosynthesis in Aquatic Ecosystems*. Cambridge, 2007.

[44] A. Bricaud, H. Claustre, J. Ras, and K. Oubelkheir, "Natural variability of phytoplanktonic absorption in oceanic waters: Influence of the size structure of algal populations", *Journal of Geophysical Research-Oceans*, vol. 109, 2004.

[45] F. Vidussi, H. Claustre, B. B. Manca, A. Luchetta, and M. Jean-Claude, "Phytoplankton pigment distribution in relation to upper thermocline circulation in the eastern Mediterranean Sea during winter", *Journal of Geophysical Research*, vol. 106, pp. 939–956, 2001.

[46] A. Chazottes, A. Bricaud, M. Crépon, and S. Thiria, "Statistical analysis of a database of absorption spectra of phytoplankton and pigment concentrations using self-organizing maps.", *Applied Optics*, vol. 45, no. 31, pp. 8102–8115, 2006.

[47] A. M. Ciotti, M. R. Lewis, and J. J. Cullen, "Assessment of the relationships between dominant cell size in natural phytoplankton communities and the spectral shape of the absorption coefficient", *Limnology and Oceanography*, vol. 47, no. 2, pp. 404–417, 2002.

[48] H. R. Gordon and D. K. Clark, "Remote sensing optical properties of a stratified ocean: an improved interpretation", *Appl. Opt.*, vol. 19, no. 20, pp. 3428–3430, Oct. 1980.

[49] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers",

IBM Journal of Research and Development, vol. 3, no. 3, pp. 210–229, Jul. 1959.

[50] S. Angra and S. Ahuja, “Machine learning and its applications”, *International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, no. IEEE, 2017.

[51] T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed. Springer, 2009.

[52] R. J. W. Brewin, N. J. Hardman-Mountford, S. J. Lavender, D. E. Raitsos, T. Hirata, and J. Uitz, “An intercomparison of bio-optical techniques for detecting dominant phytoplankton size class from satellite remote sensing”, *Remote Sensing of Environment*, vol. 115, no. 2, pp. 325–339, 2011.

[53] Z. Li, L. Li, K. Song, and N. Cassar, “Estimation of phytoplankton size fractions based on spectral features of remote sensing ocean color data”, *Journal of Geophysical Research: Oceans*, vol. 118, no. 3, pp. 1445–1458, 2013.

[54] M. Belgiu and L. Drăguţ, “Random forest in remote sensing: A review of applications and future directions”, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[55] S. Hu, H. Liu, W. Zhao, T. Shi, Z. Hu, and Q. Li, “Comparison of Machine Learning Techniques in Inferring Phytoplankton Size Classes”, *Remote Sensing*, vol. 10, no. 3, 2018.

[56] S. Haykin, *Neural Networks and Learning Machines*, vol. 3. Pearson Prentice Hall, 2008.

[57] F. Rosenblatt, “The Perceptron : A Probabilistic Model for Infomation Storage and Organization in the Brain”, *Psychological Review*, vol. 65, no. 6, pp. 386–408.

[58] Z. Alom *et al.*, “A State-of-the-Art Survey on Deep Learning Theory and Architectures”, pp. 1–67, 2019.

[59] I. Goodfellow *et al.*, “Generative Adversarial Nets”, in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[60] D. Meyer, “Support Vector Machines”, vol. 1. FH Technikum Wien, Austria, pp. 1–8, 2018.

[61] B. Hammer, M. Strickert, and T. Villmann, “Relevance LVQ versus SVM”, in *Artificial Intelligence and Soft Computing - ICAISC 2004*, 2004, pp. 592–597.

[62] C. O. Neil and R. Schutt, *Doing data science*, vol. 51, no. 12. O’Reilly, 2014.

[63] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.

[64] S. Bernard, F. Shillington, and T. Probyn, “The use of equivalent size

distributions of natural phytoplankton assemblages for optical modeling.", *Optics express*, vol. 15, no. 5, pp. 1995–2007, 2007.

[65] J. E. Hansen and L. D. Travis, "Light scattering in planetary atmospheres", *Space Science Reviews*, vol. 16, no. 4, pp. 527–610, 1974.

[66] L. Moberg, B. Karlberg, K. Sørensen, and T. Källqvist, "Assessment of phytoplankton class abundance using absorption spectra and chemometrics", *Talanta*, vol. 56, no. 1, pp. 153–160, 2002.

[67] A. Jammalamadaka, S. Joshi, S. Karthikeyan, and B. S. Manjunath, "Discriminative Basis Selection Using Non-negative Matrix Factorization", *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010.

[68] "SeaBASS." [Online]. Available: https://seabass.gsfc.nasa.gov/.

[69] University of California at Berkeley, "PostgreSQL", 2019. [Online]. Available: https://www.postgresql.org/. [Accessed: 25-Aug-2019].

[70] "PostGIS: Spatial and Geographic objects for PostgreSQL", 2019. [Online]. Available: https://postgis.net/. [Accessed: 25-Aug-2019].

[71] P. J. Werdell and S. W. Bailey, "An improved in-situ bio-optical data set for ocean color algorithm development and satellite data product validation", *Remote Sensing of Environment*, vol. 98, no. 1, pp. 122–140, 2005.

[72] A. Bricaud and D. Stramski, "Spectral absorption coefficients of living phytoplankton and nonalgal biogenous matter: A comparison between the Peru upwelling area and the Sargasso Sea", *Limnology and Oceanography*, vol. 35, no. 3, pp. 562–582, 1990.

[73] D. Stramski, R. A. Reynolds, S. Kaczmarek, J. Uitz, and G. Zheng, "Correction of pathlength amplification in the filter-pad technique for measurements of particulate absorption coefficient in the visible spectral region", *Applied Optics*, vol. 54, no. 22, pp. 6763–6782, 2015.

[74] H. Sosik, "Characterizing seawater constituents from optical properties", *UNESCO*, pp. 281–329, 2008.

[75] L. Metz, N. Maheswaranathan, B. Cheung, J. Sohl-Dickstein, and G. Brain, "Learning To Learn Without Labels", *Google Brain*, vol. 1, no. 2, pp. 2–6, 2018.

[76] N. Hoepffner and S. Sathyendranath, "Effect of pigment composition on absorption properties of phytoplankton ", *Marine Ecology Progress Series*, vol. 73, pp. 11–23, 1991.

[77] S. E. Lohrenz, A. D. Weidemann, and M. Tuel, "Phytoplankton spectral absorption as influenced by community size structure and pigment composition", *Journal of Plankton Research*, vol. 25, no. 1, pp. 35–61, 2003.

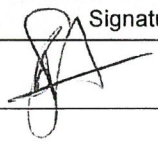
APPLICATION FORM

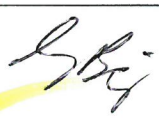
Please Note:
Any person planning to undertake research in the Faculty of Engineering and the Built Environment (EBE) at the University of Cape Town is required to complete this form **before** collecting or analysing data.The objective of submitting this application *prior* to embarking on research is to ensure that the highest ethical standards in research, conducted under the auspices of the EBE Faculty, are met. Please ensure that you have read, and understood the **EBE Ethics in Research Handbook**(available from the UCT EBE, Research Ethics website) prior to completing this application form: http://www.ebe.uct.ac.za/usr/ebe/research/ethics.pdf

APPLICANT'S DETAILS		
Name of principal researcher, student or external applicant		David Berliner
Department		Electrical Engineering
Preferred email address of applicant:		dsberliner@gmail.com
If a Student	Your Degree: e.g., MSc, PhD, etc.,	MSc
	Name of Supervisor (if supervised):	Dr Fred Nicolls
If this is a researchcontract, indicate the source of funding/sponsorship		Self Funded
Project Title		The classification of phytoplankton functional types from IOP data

I hereby undertake to carry out my research in such a way that:

- there is no apparent legal objection to the nature or the method of research; and
- the research will not compromise staff or students or the other responsibilities of the University;
- the stated objective will be achieved, and the findings will have a high degree of validity;
- limitations and alternative interpretations will be considered;
- the findings could be subject to peer review and publicly available; and
- I will comply with the conventions of copyright and avoid any practice that would constitute plagiarism.

SIGNED BY	Full name	Signature	Date
Principal Researcher/ Student/External applicant	David Berliner		**03 Apr 2017**

APPLICATION APPROVED BY	Full name	Signature	Date
Supervisor (where applicable)	Dr Fred Nicolls		Click here to enter a date.
HOD (or delegated nominee) Final authority for all applicants who have answered NO to all questions in Section1; and for all Undergraduateresearch (Including Honours).	E BOSE Click here to enter text.		3/4/17 Click here to enter a date.
Chair : Faculty EIR Committee For applicants other than undergraduate students who have answered YES to any of the above questions.	Click here to enter text.		Click here to enter a date.