Towards High Fidelity Mapping of Global Inland Water Quality Using Earth Observation Data

BY

JEREMY KRAVITZ

SUPERVISED BY

MARK MATTHEWS

STEWART BERNARD

SARAH FAWCETT

DISSERTATION SUBMITTED IN FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF OCEANOGRAPHY

UNIVERSITY OF CAPE TOWN

AUGUST 2020

DEDICATION

For my mom, Robin Kravitz.

Thanks for everything.

DECLARATION

The work in this thesis is based on research carried out at the Department of Oceanography, University of Cape Town. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced as otherwise in the text.

I confirm that I have been granted permission by the University of Cape Town's Doctoral Degrees Board to include the following publications in my PhD thesis, and where co-authorships are involved, my co-authors have agreed that I may include the publications:

Kravitz, J., Matthews, M., Bernard, S., & Griffith, D. (2020). Application of Sentinel 3 OLCI for chl-a retrieval over small inland water targets: Successes and challenges. *Remote Sensing of Environment*, *237*, 111562.

I hereby: (a) grant the University free license to reproduce the above thesis in whole or in part, for the purpose of research; (b) declare that: (i) the above thesis is my own unaided work, both in conception and execution, and that apart from the normal guidance of my supervisor, I have received no assistance apart from that stated below; (ii) except as stated below, neither the substance or any part of the thesis has been submitted in the past, or is being, or is to be submitted for a degree at this University or any other University. (iii) I am now presenting the thesis for examination for the Degree of PhD.

x: _____ Date:_____

Name: Jeremy Kravitz Student #: KRVJER001

"*But concerning vision alone is a separate science formed among philosophers, namely, optics, and not concerning any other sense… It is possible that some other science may be more useful, but no other science has so much sweetness and beauty of utility. Therefore it is the flower of the whole of philosophy and through it, and not without it, can the other sciences be known.*"

- Roger Bacon, *The Opus Maius of Roger Bacon*, 1928

ABSTRACT

This body of work aims to contribute advancements towards developing globally applicable water quality retrieval models using Earth Observation data for freshwater systems. Eutrophication and increasing prevalence of potentially toxic algal blooms among global inland water bodies have become a major ecological concersn and require direct attention. There is now a growing necessity to develop pragmatic approaches that allow timely and effective extrapolation of local processes, to spatially resolved global products. This study provides one of the first assessments of the state-of-the-art for trophic status (chlorophyll-a) retrievals for small water bodies using Sentinel-3 Ocean and Land Color Imager (OLCI). Multiple fieldwork campaigns were undertaken for the collection of common aquatic biogeophysical and bio-optical parameters that were used to validate current atmospheric correction and chlorophyll-a retrieval algorithms. The study highlighted the difficulties of obtaining robust retrieval estimates from a coarse spatial resolution sensor from highly variable eutrophic water bodies. Atmospheric correction remains a difficult challenge to operational freshwater monitoring, however, the study further validated previous work confirming applicability of simple, empirically derived retrieval algorithms using top-of-atmosphere data. The apparent scarcity of paired in-situ optical and biogeophysical data for productive inland waters also hinders our capability to develop and validate robust retrieval algorithms. Radiative transfer modeling was used to fill this gap through the development of a novel synthetic dataset of top-of-atmosphere and bottom-of-atmosphere reflectances, which attempts to encompass the immense natural optical variability present in inland waters. Novel aspects of the synthetic dataset include: 1) physics-based, two-layered, size and type specific phytoplankton IOPs for mixed eukaryotic/cyanobacteria

assemblages, 2) calculations of mixed assemblage chl-a fluorescence, 3) modeled phycocyanin concentration derived from assemblage based phycocyanin absorption, 4) and paired sensor-specific TOA reflectances which include optically extreme cases and contribution of green vegetation adjacency. The synthetic bottom-of-atmosphere reflectance spectra were compiled into 13 distinct optical water types similar to those discovered using in-situ data. Inspection showed similar relationships and ranges of concentrations and inherent optical properties of natural waters. This dataset was used to calculate typical surviving water-leaving signal at top-of-atmosphere, as well as first order calculations of the signal-to-noise-ratio (SNR) for the various optical water types, a first for productive inland waters, as well as conduct a sensitivity analysis of cyanobacteria detection from top-of-atmosphere. Finally, the synthetic dataset was used to train and test four state-of-the-art machine learning architectures for multi-parameter retrieval and cross-sensor capability. Initial results provide reliable estimates of water quality parameters and inherent optical properties over a highly dynamic range of water types, at various spectral and spatial sensor resolutions. It is hoped the results of this work incrementally improves inland water Earth observation on multiple aspects of the forward and inverse modelling process, and provides an improvement in our capabilities for routine, global monitoring of inland water quality.

ACKNOWLEDGMENTS

My deepest gratitude belongs to the most important women in my life, my mother and my wife. My mother for always believing in me, and for the unwavering support provided in both the ups and downs. My wife for constantly pushing me to become the best I can be, while constantly being my number one fan. Your irritating motivation and positivity provided the backbone for this work.

Thank you to my supervisor Dr. Mark Matthews, for guiding me through out these years, and introducing me to the world of inland water optics. None of this work would have been possible without your invaluable knowledge and feedback.

Thank you to the CSIR crew, Dr. Stewart Bernard, Dr. Lisl Lain, and Derek Griffith, for whose work much of the radiative transfer modeling in this thesis is based on. Your valuable insights into radiative transfer modeling allowed this work to become so much more.

Many, many thanks to my UCT supervisor, Dr. Sarah Fawcett, for your support throughout these years, always having my back in times of need, and early career guidance. I am deeply appreciative.

Thanks to Zimbini Faniso for your help with fieldwork activities.

Thanks to my co-workers and office-mates for the chats, the laughs, the drinks, and the pranks.

And thank you to my funding sources, the Water Research Commission (K5/2158 and K5/2458), the National Science Foundation, CyanoLakes (pty) Ltd., and CSIR.

TABLE OF CONTENTS

1

1 CHAPTER 1: INTRODUCTION

1.1 Problem Statement

Degradation of our planet's coastal and inland water resources due to anthropogenic perturbations at both local and global scales continues to place human health at substantial risk (Peters & Meybeck, 2000; Watson et al., 1998). Of the Earth's global water resources, only less than 1% are available as liquid surface freshwater. There is a staggeringly disproportionate relationship between the small extent of this invaluable resource, and the amount of pressure we place on this resource in terms of industry, agriculture, and human survival. The consequences and implications of reduced water quality and an increased predominance of eutrophication is now well understood (Smith et al., 2003; Sukenik et al., 2012; Hudnell, 2010; O'neil et al., 2012), and numerous reviews have been published emphasizing the necessity for improved management practices moving forward (Padedda et al., 2017; Schindler et al., 2012; Van Ginkel, 2012). The common consensus is global freshwater resources will become further limited and tainted as water quality degradation and eutrophication continue.

The majority of literature to-date generally focuses on point-based in-situ and laboratory studies to help us understand local responses to external pressures at the process level. There is now a growing necessity to develop pragmatic approaches which allow timely and effective extrapolation of local processes, to spatially resolved global products to promote operational and sustainable resource policy management. Advancements in space-based or airborne technology to monitor the Earth's surface have improved dramatically over the last few decades where routine measurements of oceanographic or terrestrial bio-geophysical characteristics can be acquired with high certainty. However, there is scarcity of research and development in regards to application to inland water environments. The recently released 2018 United States National Academies' Decadal Survey, "Thriving on Our Changing Planet: A decadal Strategy for Earth Observation from Space", establishes

the priorities addressed by advisory scientific panels to develop a framework for the coming decade on how to utilize space-based measurements in a manner that is relevant for scientific and socioeconomic goals outlined by varying scientific communities in Earth observation. These science and application priorities include water and energy cycles, ecosystem change, improved forecasting, reducing uncertainties in products, and Earth surface dynamics and hazards. It is the goal of the scientific community to achieve breakthroughs in these fields and improve knowledge transfer for cost-effective benefits to society.

Based on the priority science applications laid out in the Decadal Survey, the Surface Biology and Geology (SBG) targeted observable was established to improve measurements of the Earth's surface and identified as *Designated,* in which the observable is expected to be implemented as instruments or missions. Among the considerations for the SBG group, water quality and water use are of primary importance and are connected to one or more science application objectives classified as Most Important or Very Important from each of the priority panels and feeds into the Earth Science and Applications from Space (ESAS) 2017 major integrating themes such as water cycle, carbon cycle, and extreme events.

Although recent advancements in sensor technology and algorithm development have allowed for improved measurements of coastal and inland waters (Palmer et al., 2015b; Matthews et al., 2012; Smith et al., 2018; Hu et al., 2009), significant limitations still exist in the capability of modern day ocean color sensors to retrieve viable bio-geophysical data with high certainty, especially with regards to inland waters (Blix et al., 2018). The optical complexity surrounding these ecosystems, combined with fine-scale horizontal and vertical heterogeneity, induce large errors in retrieved products from current sensors and retrieval algorithms (Kudela et al., 2015). As outlined in the Decadal Survey, this

hinders our capability to develop and execute global baseline studies pertaining to the state of our inland and coastal water resources, as well as utilizing archival imagery to understand global trends and develop predictive capabilities in regards to poor water quality. It is therefore imperative to develop suitable algorithms for atmospheric correction and optical constituent retrieval for current and planned missions, with a full understanding of the uncertainties and limitations involved.

1.2 Scientific Background

A dedicated optical sensor to adequately retrieve water leaving radiances for transitional coastal environments as well as inland waters would require very demanding opto-mechanical characteristics (Mouw et al., 2015; Muller-Karger et al., 2018). Current terrestrial dedicated sensors such as Landsat-8 Operational Land Imager (L8-OLI) and Sentinel-2 Multispectral Imager (S2-MSI) may have sufficient spatial resolution in the 10 m to 60 m range (Turpie et al., 2015) to map fine-scale horizontal diversity of coastal and inland ecosystems, however these sensors are thought to lack adequate spectral resolution needed to estimate biodiversity of coastal organisms and habitats (Muller-Karger et al., 2018). Current global ocean dedicated satellites such as the MODerate resolution Imaging Spectroradiometer (MODIS), the Visible Infrared Imaging Radiometer Suite (VIIRS), and Sentinel-3 Ocean and Land Color Imager (S3-OLCI), have improved spectral resolution to further characterize organism functional types and physiology, however lack the spatial characteristics necessary to understand fine-scale spatial dynamics. To partially fill this gap, NASA had proposed the Hyperspectral Infrared Imager (HyspIRI) mission to address the needs of the previous Decadal Survey (2007) that allowed study of the utility of an imaging spectrometer from the visible to shortwave infrared and a multispectral imager in the thermal infrared. The HyspIRI study provided valuable research and data that is being leveraged for future mission studies. While planned missions have potential to improve

monitoring efforts of transitional waters, significant research and development is still required to understand the complex optical relationships in these ecosystems, and transfer this knowledge to application in the form of high performance retrieval algorithms.

The challenges facing optimum radiance retrieval in coastal and inland waters have recently been well demonstrated (Moses et al., 2009; Lavender et al., 2005; Matsushita et al., 2015). Atmospheric correction remains one of the most difficult challenges. Traditional image-based methods for obtaining relevant atmospheric variables over ocean pixels generally fail in the presence of highly productive or turbid water (Moses et al., 2009). Multiple atmospheric corrections have been recently developed to address these shortcomings (Keukelaere et al., 2018; Steinmetz et al., 2011), however, none have yet to be shown to produce enough consistent, high quality results which could be used for systematic ecological observations (Kutser et al., 2018; Keukelaere et al, 2018; Xue et al., 2019). Another observing challenge for such waters involves having a highly scattering atmosphere over a non-uniform reflecting surface. When this occurs, the spectral signal from highly reflecting surfaces such as dry vegetation, sand, and snow, becomes scattered into the sensor field-of-view (FOV) for a target water pixel and cause spectral perturbations in the retrieved signal (Bulgarelli et al., 2017). There has been some limited success in the development for correcting what is now termed the Adjacency Effect (AE) (Keukelaere et al., 2018), however, major shortcomings still exist to have confidence in operational procedures.

The optical complexity of the water leaving radiance signal for productive and turbid waters also poses significant challenges for deconvolving the signal due to non-covarying optical constituents such as inorganic substances and dissolved organic matter which highly influence the signal in the shorter wavelengths of the visible spectrum and contaminate the signal associated with

phytoplankton absorption properties (Dall'Olmo et al., 2005). Several pigment retrieval algorithms have been developed for coastal and inland waters which take advantage of the spectral features which manifest in the longer red and NIR portions of the spectrum, which are more dominated by algal spectral properties. Current models to derive chl-a concentrations from radiometric data can generally be divided into three major classes: 1. Analytical, 2. Semi-analytical, or 3. Derivative or spectral shape (Stumpf et al., 2016). Analytical and semi-analytical models generally require very high quality atmospherically corrected reflectance data since they essentially follow radiative transfer theory for their derivation, and have been shown to perform very poorly in coastal and inland waters where atmospheric correction is unreliable (Palmer et al., 2015a; Binding et al., 2011). For multispectral sensors such as S3-OLCI and the now defunct Medium Resolution Imaging Spectrometer (MERIS), which are arguably the sensors with spectral band characteristics most capable to sufficiently study high biomass, optically complex waters (Matthews et al., 2012; Palmer et al. 2015a), simple empirically based derivative-type models, which are quite insensitive to poor atmospheric correction, and can be utilized at top-of-atmosphere (TOA), have been most successful (Binding et al., 2011; Palmer et al., 2015c, Matthews et al., 2012). However, these models generally require local tuning and can involve high uncertainty in the presence of variable atmospheres or diverse algal populations involving cyanobacteria.

The development of robust water quality algorithms also depends on high quality in-situ data collected for calibration and validation purposes. Fine-scale horizontal and vertical variability of productive waters (Kutser et al., 2004; Kutser et al., 2008), and the substantial increase in cyanobacteria blooms of inland waters (O'neil et al., 2012) make the collection of high-quality coincident measurements with satellite overpass very difficult and error prone. Consequently, trustworthy in-situ data for productive coastal and inland waters are extremely limited compared to

combined global datasets for ocean calibration and validation, which critically hinders our capability for robust algorithm development.

1.3 Objectives and Thesis Structure

This body of this work aims to address many of the issues outlined above hindering the progression of inland water observation. The main objective is to extend our current limited capacity for inland water Earth Observation towards fully realized global monitoring efforts. The evolution of this work is explored through smaller, more focused studies with specific aims, which will be presented as individual chapters. Each chapter, other than the introduction and concluding chapters, are represented as self-contained published, or publishable research works.

The aims to be addressed include:

Aim 1: Provide the first comprehensive validation of Sentinel-3 OLCI capabilities for retrieving chl-a from small, productive inland water targets. This work is described in Chapter 2, and includes assessments of the current-state-of-the-art for atmospheric correction and chl-a retrieval models. Further contributions include quantification of the adjacency effect for a small water target, and a novel duplicate pixel correction, as well as discussion on the limitations of OLCI and its current capacity for adequate trophic status retrieval.

Aim 2: To develop a unique, state-of-the-art synthetic dataset of above-water reflectances with paired biogeophysical concentrations and inherent optical properties, to be used for future model calibration and sensitivity analysis. This work is provided in Chapter 3 and describes the radiative transfer modeling and parameterization for the aquatic synthetic dataset. The chapter discusses the novelty of the dataset as the first to more accurately define mixed assemblage waters involving

cyanobacteria. This includes novel calculations for the modeling of chl-a fluorescence, phycocyanin concentration, and application of an Equivalent Algal Populations model to define phytoplankton inherent optical properties.

Aim 3: To expand on the utility of a global, empirically derived chl-a retrieval processor, through the validation of the maximum peak height algorithm (MPH), which was identified as the most optimal performing model in chapter 2 for Sentinel-3 OLCI, using the synthetic dataset. This research is provided in Chapter 4 and includes information on the parameterization of an atmospheric radiative transfer model which was used to model above-water reflectances from Chapter 3 to top-of-atmosphere (TOA). Using this combined bottom-of-atmosphere (BOA) and TOA dataset, the relative extent of the surviving water-leaving signal reaching TOA is investigated along with first order estimates of typical signal-to-noise-ratio (SNR) for productive inland waters. The investigation includes an assessment of the MPH model to the synthetic global dataset by sensitivity analysis of the detection of cyanobacteria, and subsequent estimation of chl-a.

Aim 4: Explore the capability of the current state-of-the-art in machine learning (ML) to derive pertinent water quality parameters and optical properties using Earth Observation data using the synthetic dataset developed in Chapters 3 and 4. Discussed in Chapter 5, this work includes the parameterization and training of four ML models for multi-parameter retrieval, and cross-sensor applicability. The results of a large scale cross-validation of retrieval models is presented, as well as an investigation into the application to Earth Observation imagery.

An overall synthesis of the main findings from this body of work and conclusion statements will then conclude the thesis in Chapter 6.

2

2 CHAPTER 2: APPLICATION OF SENTINEL 3 OLCI FOR CHL-A RETRIEVAL OVER SMALL INLAND WATER TARGETS: SUCCESSES AND CHALLENGES

This chapter is published as:

Kravitz, J., Matthews, M., Bernard, S., & Griffith, D. (2020). Application of Sentinel 3 OLCI for chl-a retrieval over small inland water targets: Successes and challenges. *Remote Sensing of Environment*, *237*, 111562.

2.1 Introduction

Over recent decades, the eutrophication of global inland water bodies has been greatly accelerated by an increasing anthropogenic influence (Smith et al., 2014; Van Ginkel, 2008). The ramifications of eutrophication and reduced water quality have been highlighted in recent years (Harding, 2015; Smith et al., 2003; Sukenik et al., 2012; Yang et al., 2008; Oberholster and Ashton, 2008; Hudnell, 2010; O'neil et al., 2012; Graham et al., 2004;), emphasizing the necessity for improved management practices (Van Ginkel, 2012, Forsberg, 1998; Padedda et al., 2017; Schindler et al., 2012). Recent advancements in the remote sensing of coastal, estuarine and inland water bodies have provided a positive outlook for improved monitoring efforts (Palmer et al., 2015). With recent advancements in high spatio-temporal resolution space-based sensors, operational near-real-time monitoring of inland water resources is now achievable. When supported appropriately by in-situ monitoring programs, remote sensing can provide valuable aid for strategic planning and cost reduction measures by local managers. This is especially empowering for developing nations and regions where human and scientific resources are limited.

The Medium Resolution Imaging Spectrometer (MERIS) on board the European Space Agency (ESA) ENVISAT platform provided unprecedented capability to monitor coastal and inland water systems from 2002 until 2012 (Matthews et al., 2014; Binding et al., 2018; Palmer et al., 2014; Binding et al., 2011), and the Ocean and Land Color Instrument (OLCI) on board the Sentinel-3 satellite is designed to build on that success as part of the European Copernicus programme for earth observation solutions from satellite (Donlon et al., 2012). The first in a constellation of four satellites, Sentinel-3A was successfully launched in February 2016, followed by Sentinel-3B which was successfully launched in April 2018. OLCI, like its predecessor MERIS, has ideal spatial, spectral and radiometric

characteristics that enable quantitative, routine estimates of water quality variables for larger water bodies. OLCI has a number of improvements when compared to MERIS including an increase from 15 to 21 spectral bands, improved signal-to-noise ratio (SNR), tilted cameras for mitigation of sun-glint, global spatial resolution of 300m at full resolution, and improved coverage. Full OLCI instrument and band characteristics can be found at https://sentinel.esa.int/web/sentinel/missions/sentinel-3. With these improved capabilities there is substantial motivation to develop suitably calibrated and validated OLCI products to monitor the risks of eutrophication and toxic algal blooms which typically present challenging targets for remote sensing but have significant economic and societal impacts.

While OLCI represents many improvements over MERIS, some fundamental challenges remain regarding the remote sensing of small and/or productive water bodies. A fundamental challenge for MERIS was the handling of atmospheric effects over these targets, and these are well documented (Moses et al., 2009b; Lavender et al., 2005; Wang & Shi, 2008; Matsushita et al., 2015; Palmer et al., 2015; Binding et al., 2011). The best results were generally obtained by methods developed for land applications, spatially interpolated over nearby water bodies (Guanter et al., 2010), or by radiative transfer solver methods such as the vector version of the second simulation of a satellite signal in solar spectrum (6SV) (Matthews et al., 2010; Giardino et al., 2007). Similar findings are being reported for OLCI, with 6SV performing best over Chinese inland waters (Xue et al., 2019; Shen et al., 2017). Methods incorporating short-wave-infra-red (SWIR) bands have proven successful with other sensors such as the Moderate Resolution Imaging Spectroradiometer (MODIS) and Landsat 8 (Wang & Shi, 2007; Vanhellemont & Ruddick, 2015), and while OLCI lacks far-SWIR bands to appropriately apply these methods, the inclusion of a new 1020nm band has shown potential in atmospherically correcting OLCI red/near-infrared (NIR) bands in highly scattering coastal waters (Gossn et al., 2019; Delgado et al., 2018). A synergistic approach combining OLCI and Sentinel-3's far-SWIR bands on its

Sea and Land Surface Temperature Radiometer (SLSTR), has also proven to be feasible (Bi et al., 2018). The spatial interpolation method developed by Guanter et al., (2005) was adapted for use with Landsat 8 Operational Land Imager (OLI) and Sentinel 2 Multispectral Instrument (MSI) with good results for inland waters (Keukelaere et al., 2018) and recently adapted for use with OLCI, although validation results have yet to be published.

Another major challenge to the remote sensing of inland water bodies is spectral perturbation caused by light scattered into the sensor field of view (FOV) by highly reflecting surfaces near the target water body, known as adjacency effect (AE) (Bulgarelli et al., 2017). The associated error in derived remote sensing reflectances and pigment concentrations may be large (e.g. Bulgarelli and Zibordi, 2018; Odermatt et al., 2008), and while AE has been widely examined (Santer & Schmechtig, 2000; Righter et al., 2006; Sei, 2007; Belanger et al., 2007; Bulgarelli et al., 2014; Bulgarelli et al., 2017; Bulgarelli & Zibordi, 2018), it remains a complex issue that is difficult to resolve. Some proposed corrections for coastal areas, such as Improved Contrast between Ocean and Land (ICOL) have been shown to have little to no effect for productive inland water applications (Matthews et al., 2012; Odermatt et al., 2009; Binding et al., 2011). The SIMEC adjacency correction, having had some previous success with MERIS (Sterckx et al., 2015) will be further explored in this study in conjunction with the image correction for atmospheric effects (iCOR) processor for OLCI.

The difficulty inherent in obtaining accurate water-leaving radiances also affects Level-3 (L3) satellite data products, of which the concentration of the photosynthetic pigment chl-a is the most important: it is integral to any water quality monitoring initiative (Schalles, 2006) as the standard in quantifying phytoplankton biomass as an indicator of trophic status. Given the challenges of atmospheric correction over productive inland waters, simple, empirically parameterized models based on the

absorption and scattering properties of chl-a in the red and NIR wavelengths have been most successful in estimating chl-a concentration from MERIS (Binding et al., 2011; Matthews et al., 2012; Palmer et al., 2015; Moses et al., 2009b), and have recently been applied to OLCI data (e.g. Moses et al., 2019, Xue et al., 2019, Smith et al., 2018) as well. These models generally take the form of band ratio or band difference (derivative) algorithms which - while typically requiring local tuning - are robust, and can retrieve chl-a concentrations with confidence from MERIS without a full atmospheric correction (Matthews, 2011).

This study provides the first comprehensive validation of OLCI inland water Rrs and OLCI-derived chl-a, using a variety of bio-optical algorithms and AC procedures. Four South African water bodies of varying size and optical nature were used as validation targets. The primary aim of this study is to evaluate the utility of OLCI radiometry over small inland water targets. Specific objectives are to 1) quantify adjacency effects for small water targets, 2) evaluate four current and freely available atmospheric correction algorithms (or models or approaches) applicable to inland water monitoring systems, and 3) evaluate a suite of chl-a retrieval algorithms suited to turbid and productive waters. Further contributions are the application of a novel duplicate pixel correction, and a discussion on the limitations and advantages of using OLCI as a successor to MERIS for the remote sensing of small water targets.

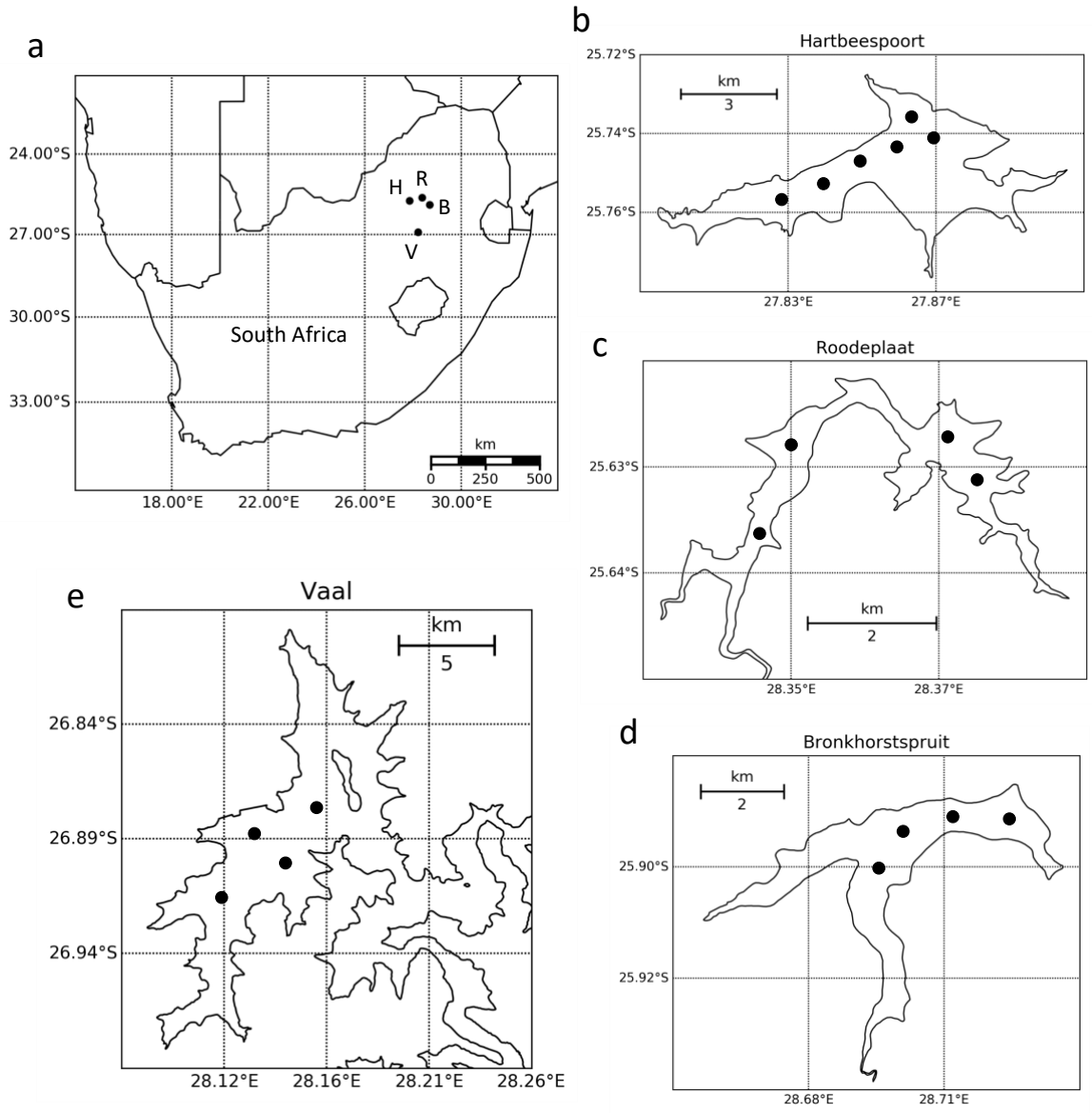

Figure 1: Geographical locations of dams and sample points (a), South Africa (b), Hartbeespoort dam (c), Roodeplaat dam (d), Bronkhorstspruit dam (e), Vaal dam.

2.2 Methods

2.2.1 In-Situ Measurements

2.2.1.1 Biogeophysical Data

Sample points for Hartbeespoort dam, Roodeplaat dam, Bronkhorstspruit dam, and Vaal dam were chosen to represent pixels minimizing spectral contamination from nearby land or floating aquatic vegetation induced by AE (Fig. 1). Exact locations may have changed during different fieldwork campaigns due to environmental or logistical reasons, but the total number of points were kept to no greater than four per reservoir per sampling day, to maximize the number of validation points while minimizing the time between sampling and satellite overpasses (usually less than 2 hours). Roodeplaat and Bronkhorstspruit, both with a maximum water surface area in the main basins of less than 2km^2, are likely spectrally contaminated to some degree by AE in their entirety, and approach the limit of observation with OLCI: an extreme test case for observing small eutrophic waters. Hartbeespoort and Vaal are larger reservoirs and a greater number of pixels with lower AE may be identified. Sampling in Hartbeespoort was hampered in 2017 by the growth and proliferation of water hyacinth, which covered greater than 30% of the area of the reservoir making it impossible to sample certain points.

In situ measurements consisted of biogeophysical data collected on the same day as Sentinel-3A overpasses with the interval between overpass and collection being no greater than two hours. Measurements for water quality and clarity, and visual information on weather and water surface conditions were collected by a small boat at each station. Water clarity was determined by Secchi Disk. Water samples are collected by first rinsing a 25 L black bucket three times with surface water followed by the actual collection of a surface water sample. The bucket is then sealed with a lid and shaken to allow adequate mixing of the sample. For chl-a and total suspended solids (TSS) analysis,

two well-rinsed 1 L bottles were used to collect water from the surface of the bucket, taking care to minimize bubbles. Samples for phytoplankton identification and enumeration were collected in the same manner in a well-rinsed 100 ml bottle. To-date, there has not been a standardized protocol for productive inland water collection for cal/val purposes of optical and Earth observation data. This methodology for sample collection was chosen, particularly the artificial mixing of the water sample, in attempt to roughly homogenize the top optical layer of the water column (typically between 0.1 and 2.0 m, for these water types (Matthews et al. 2013)). This homogenization was intended to standardize the 25 L water sample for the various quantitative measurements conducted in the lab as mentioned here, as well as other biogeophysical and optical parameters which are not described or presented in this research, which included laboratory measurements from a Sea-Bird Scientific AC-S spectral absorption and attenuation sensor, and an ECO BB9 optical backscattering sensor (data not shown). Chl-a and TSS analysis took place no later than one day from collection. Samples for phytoplankton enumeration were preserved with formaldehyde until analysis one to two months later by Rand Water Analytical Services, South Africa.. Chl-a analysis was performed using the spectrophotometric method and measurements of TSS were performed gravimetrically.

2.2.1.2 Shipborne Radiometric Data

During field campaigns, measurements of upwelling radiance from the water (L_u) and downwelling sky radiance (L_{sky}) were collected concurrently with water samples using an ASD-FR Field Spectroradiometer 3 (Analytical Spectral Devices, Boulder, CO USA). Measurements of a white Spectralon reflectance plaque with a known bidirectional reflectance distribution function (BRDF) were used to normalize the uncalibrated radiance measurements for downwelling irradiance, (E_s). Methods for the derivation of these quantities can be found in Mueller et al., (2003), with a brief overview in Appendix A Remote sensing reflectance (R_{rs}) was then derived using (Mobley, 1999)

$$R_{rs}\,(\lambda) = [L_u\,(\lambda) - \rho L_{sky}] / E_s\,(\lambda) \quad (2.1)$$

where ρ is the reflectance of skylight from the water surface and estimated using approximations based on wind speed using Mobley (1999).

2.2.1.3 Buoy Radiometric Data

An autonomous buoy developed by the Council for Scientific and Industrial Research (CSIR), South Africa, was moored at Roodeplaat Dam for three months to collect continuous radiometric measurements for validation purposes. The radiometric payload consists of one Trios RAMSES ACC downwelling irradiance sensor (E_s) and two Trios RAMSES ARC upwelling radiances sensors at 0.15 m (L_{uz1}) and 0.76 m (L_{uz2}) depths. The radiometric sensors measuring the upwelling radiance were directed towards nadir. Other sensors include a Trios chl-a fluorometer, an ultrasonic anemometer, pitch and roll sensors, and a temperature sensor (Fig. 2). The control operations for the buoy system were set to collect measurements from all instrumentation in 5-minute bursts every 15 minutes and were battery powered, charged by two solar panels attached above the water surface. The data were transmitted via Global System for Mobile Communications (GSM) telemetry for real-time data acquisition and equation (1) was used to calculate R_{rs}. For detailed information on the derivation of variables used in equation (1), including correcting for errors induced by instrument self-shading, please see processing guidelines provided in Mueller et al., (2003) and Antoine et al., (2008).

2.2.1.4 Aerosols

Aerosol loading – the highly variable presence of fine scale particles in the atmosphere - can have large implications on water leaving radiance retrieval (Bassani et al., 2015), and was therefore measured at time of OLCI overpasses at wavelengths 440, 500, 550, 675, 870 and 936 nm using a

Microtops sunphotometer. AOT at 550 nm (aot550) was derived using the Angstrom law to quantify the aerosol-driven attenuation of light in the vertical column of the atmosphere: higher AOT's represent higher attenuation and lower visibility (Wang and Christopher, 2003). An AERONET station located at CSIR Pretoria, South Africa (GPS: 25°45'25.2"S, 28°16'48"E) which houses a Cimel CE318 robotic sunphotometer, was used to acquire atmospheric data on overpass dates where Microtops II data was not available, or as a comparison.

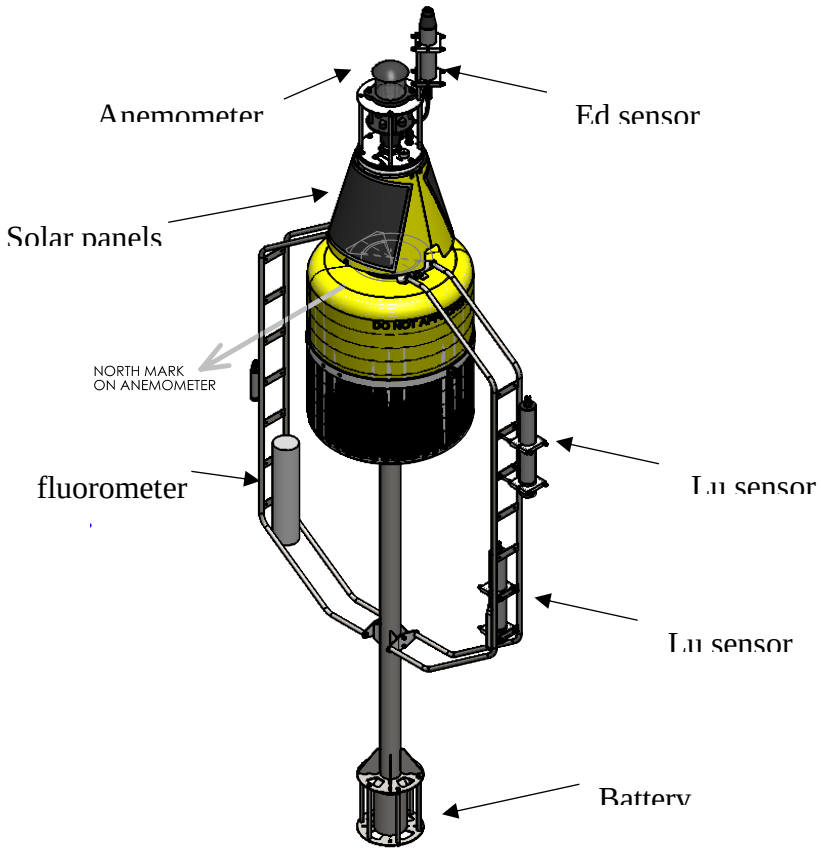

Figure 2: The autonomous radiometric buoy displaying instrument payload. (credit: CSIR)

2.2.2 Modeling to top-of-atmosphere

In simple terms, the radiance received by an optical sensor at TOA from a water target can be defined as (Bulgarelli et al., 2014)

$$L_{tot} = L_{path} + L_{BG} + tL_{water} \quad (2.2)$$

where L_{tot} is the total radiance received by the sensor, L_{path} is the path radiance which defines the photons scattered into the instantaneous FOV by the atmosphere alone, L_{BG} is the background radiance from neighboring pixels which are diffusely scattered into the sensor FOV, L_{water} is the water leaving radiance at the sensor, and t is the diffuse transmittance. L_{BG} is considered as the radiance introduced due to AE. A radiative transfer modeling exercise was performed to quantify the error introduced by the presence of background radiation for target water pixels in small water basins using OLCI. MODTRAN 5.0 radiative transfer software was used to propagate matchup R_{rs} measurements to OLCI at-sensor radiance using aerosol optical properties measured in situ and obtained from the AERONET station in Pretoria. Total spectral radiance reaching the satellite sensor was computed using the following steps:

1) The atmospheric model was compiled in MODTRAN using available Microtops and Aeronet data (AOT, single scattering albedo (SSA), water vapor, Angstrom extinction coefficient, and asymmetry parameter).
2) An area-averaged (~10km diameter) surface reflectance around the dam was retrieved from an S3 image and used to compute E_s at bottom-of-atmosphere (BOA) in MODTRAN.
3) Calculated E_s at BOA was used to compute L_w at BOA via R_{rs} measurements measured in-situ around the time of overpasses using equation (1).

4) L_{sky} is calculated using an upward-looking MODTRAN run in the correct viewing geometry and added to L_w to obtain total upwelling radiance (L_u) above water at BOA. Water-surface reflectance was interpolated from the Mobley (2015) look-up tables (assumed spectrally invariant for convenience).
5) L_u at BOA was multiplied by atmospheric path transmittance (t) provided by MODTRAN to obtain water-target radiance at TOA.
6) The atmospheric path radiance (L_{path}) is added to L_u from previous step to obtain the total spectral radiance seen by satellite at TOA (L_{tot})
7) All computations up to this point performed at full MODTRAN 5 spectral resolution. The OLCI spectral response functions (SRFs) were then applied to compute channel radiances which were then compared to OLCI measurements at selected pixels in the product.

In the modeled TOA radiance, all terms in equation (2.2) are accounted for except L_{BG}, the radiance introduced from AE. Thus, when comparing modeled versus measured TOA radiances, assuming atmospheric variables used to compile MODTRAN are the true quantities at scene acquisition, the majority of error should be due to either (1) the radiometric differences between the OLCI sensor and the in-situ radiometer whose data were used to generate the modeled TOA radiance, (2) errors due to comparison of a spatial averaged 300 m x 300 m OLCI pixel to a point measurement, and (3) the background radiance. While errors from situations (1) and (2) are hard to quantify and account for, we can assume the a significant portion of the total error is due to the missing background radiance term.

2.2.3 OLCI Image Processing

2.2.3.1 Pre-processing

- While the opto-mechanical characteristics of OLCI are based on the design of ENVISAT MERIS, the push-broom imaging spectrometer of OLCI is tilted off-nadir in a westerly direction by 12.6˚, in an attempt to mitigate the effects of sun glint. The footprint of an OLCI charge-coupled device (CCD) pixel on the ground is approximately 300 m / cos(θ), which is an underestimate due to the curvature of the Earth. Therefore, at the Western edge of the image swath, where the observation zenith angle (OZA) could reach upwards of 55-60˚, the footprint of a CCD pixel becomes closer to 600 m. The level-1 (L1) product is a 300 m resolution standard grid across the whole swath, and consequently when larger ground sampling distances (GSD's) are resampled back down to a 300 m standard resolution, duplicate pixels result. In order to correct for the increased number of duplicate pixels with increasing OZA, a novel, image-wide cubic interpolation scheme was implemented to each OLCI L1b band. In this process, the duplicate pixels are first flagged using the duplicate pixel mask in the OLCI L1b product. Flagged pixels are then interpolated over using neighboring non-duplicate pixels by cubic spline method. OLCI L1b duplicate corrected subsets were then processed using the four different full atmospheric corrections and two partial atmospheric corrections. For the iCOR full atmospheric correction method, no pre-processing was performed as the processor requires the unaltered OLCI L1b file as input. See Figure 3 for a visualization of the finalized image-processing chain.

2.2.3.2 Atmospheric corrections

Full atmospheric corrections over small eutrophic water bodies are extremely challenging due to the complex nature of the in-water optical signal and interference from AE. Most South African inland water bodies have a high particle load, resulting in reflectances that are much higher than those of

clear water. This allows spectral features to be identified at TOA, avoiding the need for a full AC and allowing simpler, faster models for pigment retrieval. However, uncertainties due to a variable atmosphere must be taken into consideration. A "partial atmospheric correction" was performed to test this method, using the bottom-of-Rayleigh reflectance (BRR) processor to account for the effects of molecular Rayleigh scattering and gaseous absorption in the red and NIR bands (Matthews et al. 2012) where important spectral information pertaining to pigments can be identified. The Radiance-To-Reflectance Processor was also used to convert OLCI L1b radiances to normalized apparent reflectances. Atmospheric corrections are assessed qualitatively by comparing in situ R_{rs} measurements with atmospherically corrected R_{rs} and quantitatively through band-wise performance using R^2, root mean square error (RMSE), bias, and relative error (RE).

For full atmospheric correction of OLCI L1b radiances to derive R_{rs} , four current and freely available methods were used:

(a) *Case 2 Regional CoastColour (C2RCC):* Developed for MERIS, this is a Neural Network approach where the atmospheric correction and derivation of inherent optical properties are coupled and solved for simultaneously (Doerffer & Schiller, 2007), and can be used in both forward and inverse directions to retrieve water-leaving radiances or atmospheric parameters, respectively (Doerffer & Schiller, 2008). C2RCC is a major revision of the original C2R processor developed through subsequent modifications and testing and is now incorporated for use with the Sentinel constellation of satellites, supported by ESA's CoastColour project aimed at expanding MERIS and Sentinel capabilities to coastal and inland waters. C2RCC incorporates the same foundational technology as in C2R but has expanded to include a 5-component bio-optical model, a coastal aerosol model, and expanded bio-optical training ranges (Brockmann

et al., 2016). In 2019, an alternative NN for more extreme waters was released for OLCI and this processor was run through the open source Sentinel application platform (SNAP) software version 7.0.

(b) *6SV1:* The Second simulation of a Satellite Signal in the Solar Spectrum (6S) is an advanced radiative transfer code based on successive orders of scattering (SOS) approximations designed to simulate reflection observed by a satellite sensor for a target at bottom of atmosphere using a coupled atmosphere-surface system (Vermote, 1997). The code has been used successfully for atmospheric correction over water bodies for a variety of multispectral sensors (Matthews et al., 2010; Giardino et al., 2014; Martins et al., 2017). For the validation of OLCI products, the 6SV version 1.1 was applied using the Py6S Python programming language interface (Wilson, 2013). OLCI L1b TOA radiances were collected for in situ match-up points along with sensor geometries for each point. Depending on the time of year of the fieldwork campaign, either a mid-latitude summer or mid-latitude winter atmospheric model was used with the measured AOTs. The code was run for each band of OLCI using defined spectral response functions with outputs in reflectances. Reflectances were further divided by pi to obtain R_{rs} to be intercomparable between instrument and algorithm outputs.

(c) *POLYMER:* The POLYnomial based algorithm applied to MERIS (POLYMER) is a spectral optimization approach which uses a polynomial to model separate spectral influence from atmosphere and sun glint (Steinmentz et al., 2011). After a Rayleigh correction, POLYMER decomposes the total signal into a water reflectance spectrum, a spectrally smooth function for the atmosphere, and everything else which is “non-water”. POLYMER V4.7 was implemented using Python, adapted for use with Sentinel 3 OLCI, and some minor adjustments were made for inland water application.

(d) *iCOR:* The iCOR atmospheric correction, previously known as OPERA (Sterckx et al., 2015), uses sun and sensor geometry, aerosol optical depth, ozone, water vapor, and elevation to derive atmospheric parameters from pre-computed MODTRAN-5 Look-Up-Tables (LUTs). The correction is a completely image-based processor which works over land and water. AOT at 550nm (aot550) is derived over land from spectral endmember inversions (Guanter et al., 2007). These values are then interpolated over water bodies or to coastal regions. Water vapor is retrieved using bands situated within the water vapor absorption features and elevation is derived using a digital elevation model (DEM). iCOR was initially developed for high spatial resolution multispectral sensors (De Keukelaere et al., 2018), and the developers have recently released an OLCI compatible processor. iCOR also includes an optional adjacency correction known as the SIMilarity Environment Correction (SIMEC) which estimates and removes contamination from surrounding pixels using the NIR similarity spectrum (Ruddick et al., 2006). As with C2RCC, iCOR is provided as a plugin for SNAP.

2.2.3.3 Models for chl-a estimation

The Gons et al. (2008) band ratio method as well as the two and three band ratio (Moses et al., 2009; Dall'Olmo et al., 2005; Gilerson et al., 2010) semi-analytical algorithms were used, denoted as G08, M09-2B, and D05-3B, respectively. G08 was only applied to atmospherically corrected data, while M09-2B and D05-3B were also applied to TOA data.

Derivative or spectral shape (SS) type models measure the height or depth of a spectral peak or trough over/under a baseline established by two surrounding bands. The models used in this study exploit either the inelastic feature associated with the solar-induced chlorophyll-a fluorescence peak located at 681nm (Gower et al., 1980; Gower et al., 1999; Gitelson et al., 1994) identified as the fluorescent

line height (FLH), or the elastic scattering feature that forms between 700-710nm (Schalles et al., 1998; Gower et al., 1994; Alikas et al., 2010) identified as the maximum chlorophyll index (MCI). Both models take the form

$$\text{Chl-a} \propto R(\lambda_2) - R(\lambda_1) + (R(\lambda_1) - R(\lambda_3)\frac{(\lambda_2 - \lambda_1)}{(\lambda_3 - \lambda_1)}) \quad (2.3)$$

Where $R(\lambda_2)$ is the radiance or reflectance peak/trough at λ_2 and $R(\lambda_1)$ and $R(\lambda_3)$ are the neighboring radiance or reflectance values at λ_1 and λ_3, respectively, used to create a baseline for λ_2. For OLCI, the FLH model uses band 10 (681nm), and MCI uses band 11 (709m) for λ_2 , respectively. The maximum peak height (MPH) algorithm is similar to MCI and FLH, however, first determines the highest peak in the red/NIR region and measures the height accordingly (Matthews et al., 2012; Matthews et al., 2015). The MPH algorithm flags each pixel as dominated by cyanobacteria or algae and calculates chl-a concentration individually based on empirically derived relationships (Matthews et al., 2012; Matthews & Odermatt, 2015). The MPH algorithm has already been locally tuned for South African waters and thus the calibrated algorithm was used here only with BRR and identified as MPH15.

The normalized difference chlorophyll index (NDCI) uses the spectral band difference at 709nm and 685nm but normalizes by the sum of the two reflectances to reduce differences in solar and atmospheric variability (Mishra & Mishra, 2012). The MCI, FLH, and NDCI models were applied to both atmospherically corrected and TOA data. Lastly, the chl-a concentration product from the C2RCC neural network is also retrieved and compared with in-situ data.

Other than MPH15, G08, and C2RCC, all models were initially locally tuned and tested using ordinary least squares regression. Linear and non-linear trends were fitted to the data and the one producing

the highest R^2 was used as the final model. Due to the low number of matchups, Leave One Out Cross Validation (LOOCV) was used to test predictive capability of the final models. Models were evaluated based on R^2, slope, relative RMSE (RMSE divided by mean of true values multiplied by 100) bias, and relative error (RE) from the comparison of resulting predicted vs measured chl-a concentrations.

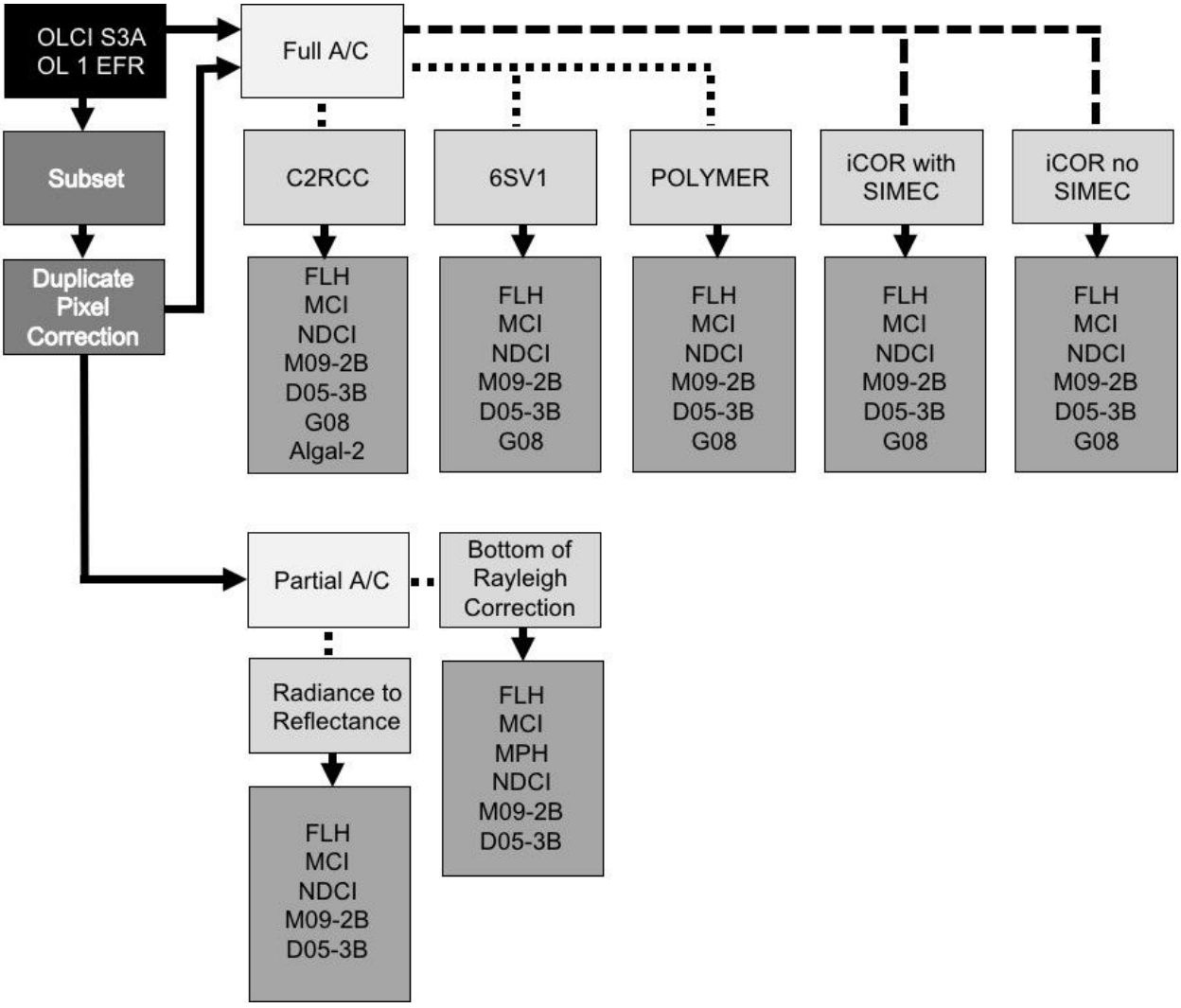

Figure 3: The processing chain applied for OLCI L1b imagery.

2.3 Study site characteristics

2.3.1 Water quality

The Hartebeespoort, Roodeplaat, Bronkhorstspruit and Vaal dams are situated in the Gauteng Province of South Africa. A summary of basic catchment characteristics of the water impoundments can be found in Appendix A.1. These small, semi-high-altitude dams supply the region with water for domestic, irrigation and agricultural purposes, and are used for public recreation and water sports. The Gauteng Province is one of the most densely populated regions of South Africa, leading to large implications for anthropogenic eutrophication within local inland water reservoirs. All four dams regularly see hypertrophic chl-a concentrations exceeding 30 mg/m^3 threshold (Matthews, 2014; DWAF, 2002), and rarely see optical depths of greater than a few meters - typically in the tens of cm's during mild bloom conditions to less than 1 cm in hyper-scum conditions (Matthews et al., 2013). The reservoirs were visited between June 2016 and April 2017 for the purpose of collecting radiometric and biogeophysical validation data.

The optical properties of South African inland waters are generally not well characterized, but data from the current study indicate that the four reservoirs are quite similar with respect to the composition of their optically significant constituents, with phytoplankton and suspended sediments providing the bulk of the optical signal, thus classifying these reservoirs as "bright" case-2 waters. Summary statistics for in situ chl-a, total suspended solids (TSS), and secchi disk can be found in Appendix A.2. Matthews et al., (2013) provides a comprehensive insight into typical bio-optical characteristics for inland water bodies of South Africa.

Calculated cell counts for algal abundances collected during the course of this study are presented in Appendix A.3, giving context to the optical targets in question. Algal blooms in all four reservoirs were predominantly caused by the Cyanophyta, *M. aeruginosa,* with lesser proportions of Chlorophyceae

and Bacillariophyceae appearing in the assemblages. Average cell counts for Cyanophyceae in Hartbeespoort and Bronkhorstspruit were around 70,000 cells/ml, while Roodepaat and Vaal had much higher densities of roughly 200,000 cells/ml and 1,200,000 cells/ml respectively. The Vaal dam experienced widespread cyanobacterial blooms (including surface scum conditions) on isolated occasions at the time of sampling, reaching a maximum *Microcystis sp* abundance of almost 6 million cells/ml.

High cell abundances were generally associated with high chl-a concentrations reaching up to roughly 500 mg/m^3 in both Roodeplaat and Vaal dams (likely underestimated as the contribution of floating algal mats is difficult to quantify). However, Roodeplaat is generally considered one of the most eutrophic inland water bodies in South Africa alongside Hartbeespoort (Torien, Hyman, and Brewer, 1975). Chl-a concentrations in Hartbeespoort have historically ranged between 30mg/m^3 and over 20,000mg/m^3 in scum conditions (Matthews et al., 2013). More recently, a floating invasive plant, *Eichhornia crassipes,* commonly known as water hyacinth, has been a troublesome presence in many of South Africa's inland water systems (Cilliers, 1991; Coetzee and Hill, 2012), sometimes exceeding 60% of the water surface (Scott et al. 1979).

2.3.2 Optics

2.3.2.1 Radiometry

Radiometric measurements using the Trios sensors or the ASD field spectroradiometer could only be collected on clear days, compromising our ability to obtain consistent match up data. All reliable ASD collections are presented in Figure 4a for the four fieldwork sampling campaigns and for the period the radiometric buoy was in the water. Variability in the magnitude and shape of the reflectance spectra are clearly visible throughout the year. Figure 4b clearly shows the various optical water types

encountered in our study along with relevant OLCI bands located in regions of relevant spectral features. For logistical reasons, the majority of measurements were obtained at Roodeplaat dam where *Mycrocistis* colonies generally persist. The trough at 620 nm exhibited in the reflectances is therefore attributed to the strong absorption of the cyanobacterial pigment Phycocyanin (PC) (Simis et al. 2005). As concentrations of chl-a reach about 10 mg/m^3, the specific absorption peak of chl-a at 675 nm also begins to dominate the adjacent chl-a fluorescence peak around 681 nm and forms another trough (Brigidare et al., 1990). The troughs at 620 nm and 680 nm produce a prominent peak in the 655 nm region, which is also potentially aided by phycobiliprotein fluorescence at roughly 650 nm (Simis & Huot, 2012). A very strong peak is evident around 709 nm due to the phytoplankton absorption induced trough at 675 nm and the strong absorption due to water (Gitelson, 1992; Gons, 1999). Strong algal backscattering at high concentrations also causes the peak at 709 nm to become more pronounced.

Unfortunately, only one reliable reflectance measurement was obtained at Hartbeespoort dam, but this measurement displayed elevated reflectance in the blue/green region and a small chl-a fluorescence reflectance peak around 681nm – resembling a more oligotrophic, eukaryote-dominated (Case 1) water type. The Vaal dam experienced super-scum conditions which can be readily identified by the highly elevated signal in the red and NIR (Fig. 4a and 4b). This water body is inundated with illite material which results in reflectances typical of highly sedimented waters: a broad increase in reflectance between about 550 and 700 nm due to augmented/elevated backscatter from the increased particle load. However, reflectance peaks around 660 and 710 nm are still prominent due to high PC and chl-a concentrations.

Overall, reflectances in the blue spectral region were highly constrained due to strong absorption by chl-a as well CDOM and tripton. The variable presence of these other constituents generally results in poor performance of blue/green based chl-a algorithms (Gurlin et al., 2011; Gitelson et al., 2009). There is some reflection of sky light identifiable by enlarged reflectance at wavelengths less than 450 nm due to the appearance of overhead clouds or haze and smog effects from biomass burning in the region (Doxoran et al., 2004). This effect sometimes led to unreliable radiance measurements which had to be discarded. Overall the reflectance features are typical of eutrophic waters with high concentrations of cyanobacteria and algae (see Qi et al., 2014; Gitelson et al., 2008; Moses et al., 2009b).

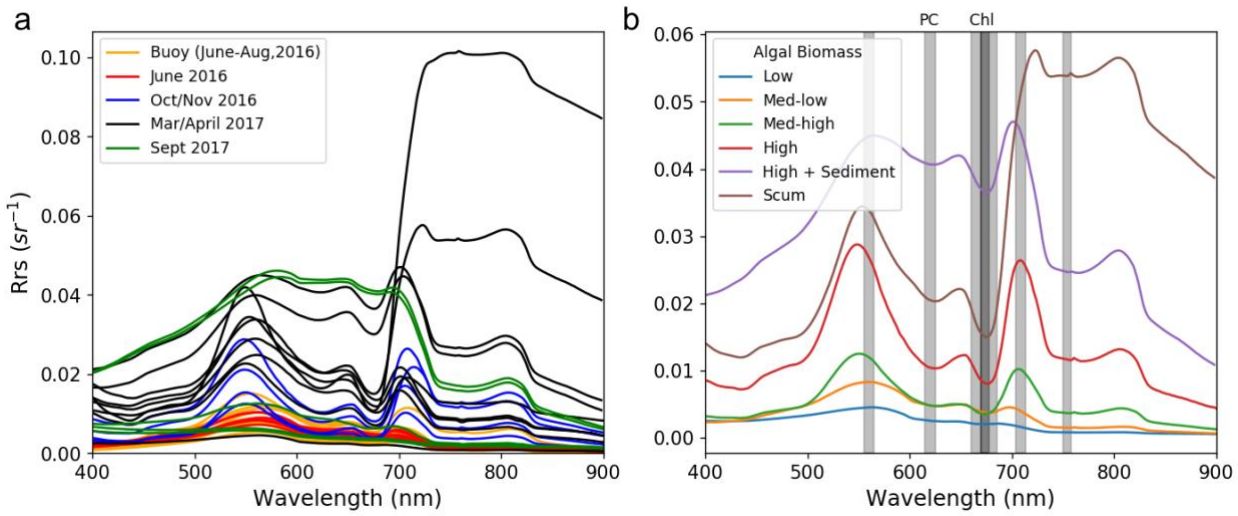

Figure 4: In situ collected R_{rs} data (a), all R_{rs} data from each field sampling campaign (b), selected R_{rs} data depicting shapes for various algal biomass with red/NIR OLCI bands in grey.

2.3.2.2 Aerosols

Figure 5a compares in situ Microtops AOT measurements with coincident Aeronet AOT measurements. AOT at 550nm showed high variability over the various sampling periods with the Microtops generally displaying a positive bias compared to Aeronet measurements. However, it must

be noted that Roodeplaat, Hartbeespoort, Bronkhorstspruit, and the Vaal are roughly 15, 42, 43, and 130 km away from the Aeronet station, respectively. In almost all cases, Microtops data were used for model processing as they were acquired at the site and at the specific time of overpass. Mismatch between the two data sources could be due to spatial heterogeneity or incorrectly calibrated instruments. It is acknowledged that this is a source of uncertainty when applying models requiring ancillary AOT.

Monthly means of aot550, columnar water vapor, and alpha (the Angstrom exponent, which gives an indication of aerosol particle size) are given in Fig. 5b for the duration of the fieldwork sampling campaigns. The monthly aot550 mean does not exhibit strong temporal or seasonal patterns throughout the sampling period but large standard deviations are evident. Columnar water vapor shows strong seasonal variability. The Angstrom exponent indicates a rather constant aerosol particle size distribution throughout the year.

Depending on time of year, the solar zenith angle (SZA) of this region generally lies between 30˚ and 65˚ with respect to the overpass time of OLCI. The solar azimuth angle (SAA) ranges from roughly 35˚ to 77˚. Due to the 12.5˚ tilt of the optical sensor, OZAs can vary substantially depending on the position of the target within the swath. OZA for our match up points ranged from 7˚ when the target is close to nadir, to 54˚ when closer to swath-edge.

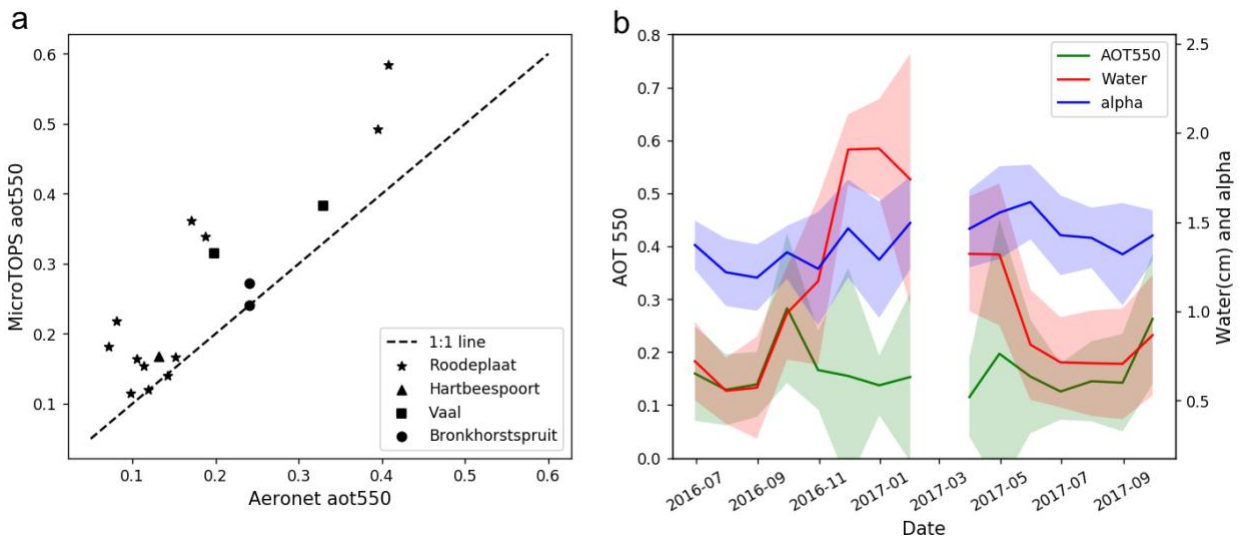

Figure 5: In situ atmospheric data (a), Comparison of aot550 values from Microtops sunphotometer and local Aeronet Station (b), Time series of aot550, water vapor, and alpha from Aeronet covering all field sampling campaigns.

2.4 Results

2.4.1 Atmospheric modeling

Quantitative comparisons between modeled and derived OLCI radiances are presented in Figs. 6 and 7. A negative bias persists in all channels, but this is small in the blue and green channels (around -5%) and good correlation is observed, with R^2 values above .9 and both RMSE and relative error averaging below 10%. Errors increase in the red channels, notably around the chl-a fluorescence region between 665nm and 681nm. The majority of mismatch occurred in the NIR channels, where negative bias increases to roughly 15%. RMSE and relative error increase dramatically above 700nm, reaching upwards of 90% and 40% respectively.

A qualitative matchup analysis for each dam is presented in Figure 8. Roodeplaat and Bronkhorstspruit dams present more extreme cases for testing OLCI's spatial capabilities, and show

a significant elevation of spectral signal in the red/NIR for the OLCI observed radiance compared to modeled radiance. This difference is attributed to primarily due to signal contamination from vegetation in nearby pixels. The single matchup point for Hartbeespoort exhibits good agreement, particularly in the green and red spectral regions. A slight over-propagation of modeled radiance occurs in the NIR. Matchups for the Vaal dam showed the poorest correlation. Modeled radiances exhibited great variability for Vaal dam, as shown by the standard deviation, particularly in the NIR region. The modeled radiances with elevated signal beyond 700 nm are typical of floating cyanobacteria mats of which the aquatic radiometric measurements were taken whereas the lower signal from OLCI is presumably due to averaging of the sub-pixel scale scum and non-scum features, as opposed to contamination from nearby vegetation pixels as in the case for the two smaller dams. Overall, the results confirm that OLCI measurements are indeed capturing significant background radiation in the red and NIR channels for small water targets such as Roodeplaat and Bronkhorstspruit. It is also noted that differences between simulated and measured radiances could very well be attributed to the simulation procedure as well as inaccurate atmospheric parameters. However, it is positive to see such good agreement in the blue and green channels, and that OLCI is correctly detecting all radiance contributions at TOA.

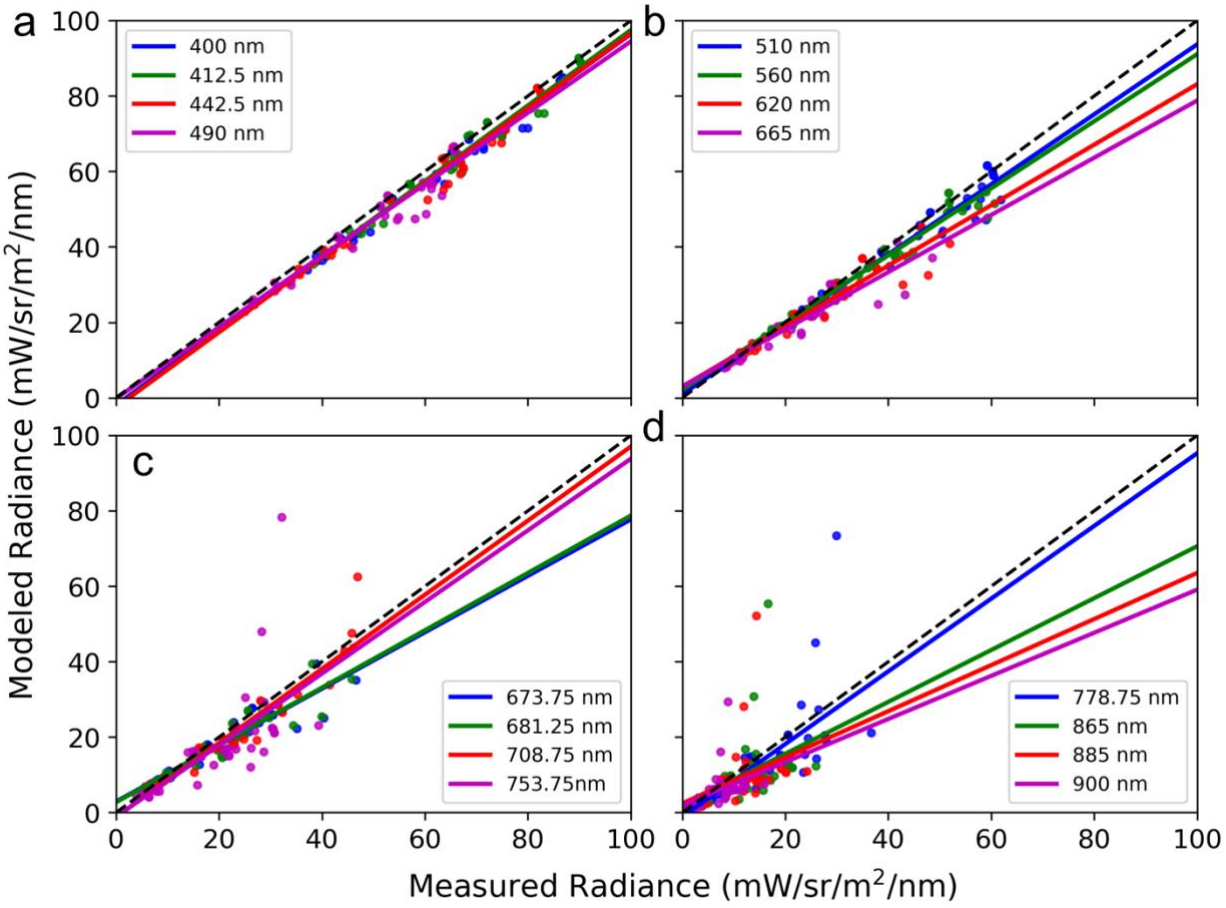

Figure 6: Scatterplot comparison between modeled and measured OLCI observed L1b radiances for specific OLCI channels (a), bands centered at 400 nm, 412.5 nm, 442.5 nm, 490 nm (b), 510 nm, 560 nm, 620 nm, 665 nm (c), 673.75 nm, 681.25 nm, 708.75 nm, 753.75 nm (d), 778.75 nm, 865 nm, 885 nm, 900 nm. Solid lines are linear best fits for each channel matchup.

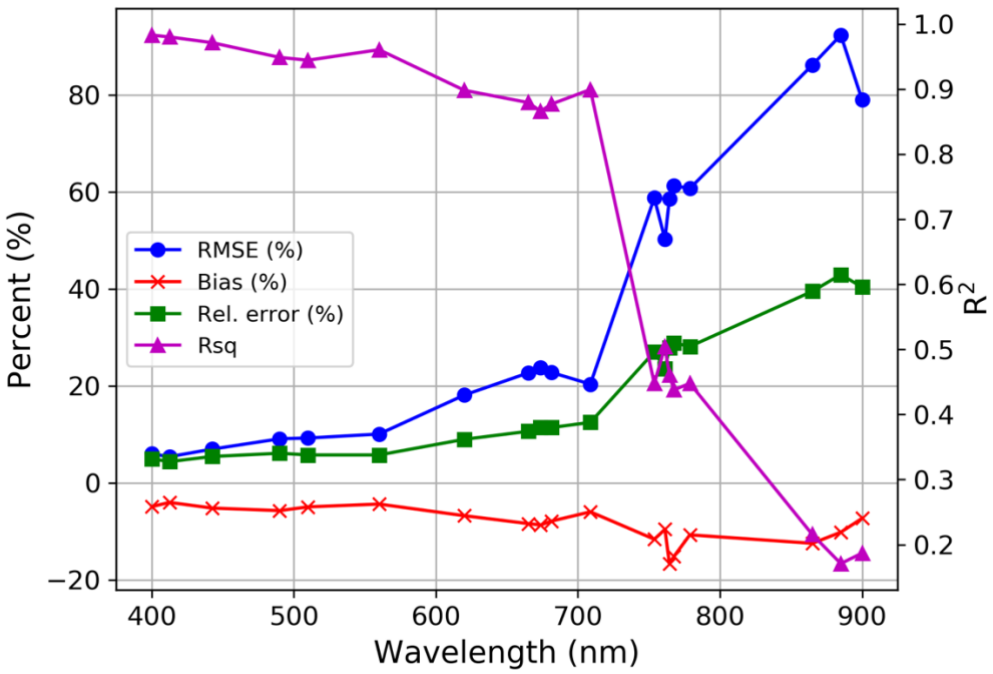

Figure 7: R^2, RMSE (%), bias (%), and RE (%) calculated from quantitative analysis between modeled and measured OLCI observed L1b radiances for specific OLCI channels.

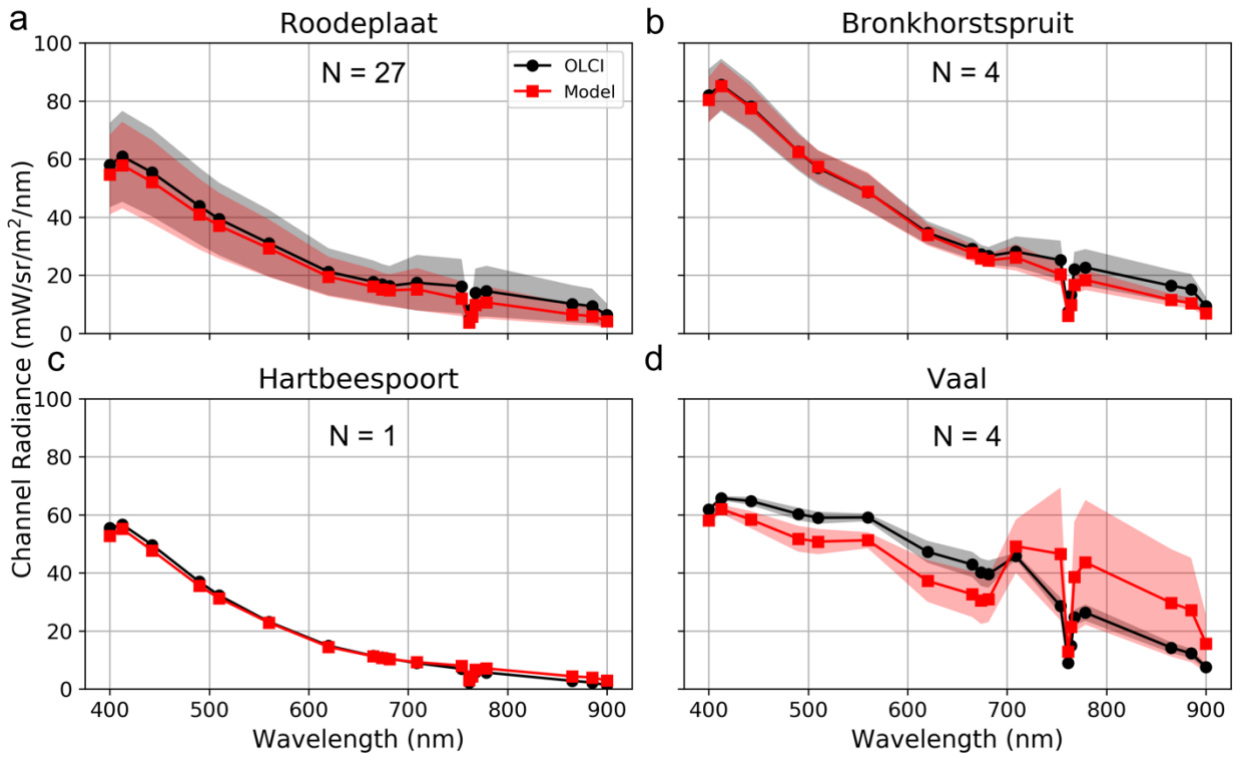

Figure 8: Qualitative comparison between modeled and measured OLCI observed L1b radiances (a), Roodeplaat (b), Bronkhorstspruit (c), Hartbeespoort (d), Vaal. Solid lines are means of total matchups per dam, shaded regions represent one standard deviation from the mean.

2.4.2 OLCI image processing

2.4.2.1 Pre-processing

Figure 9 displays the effect of GSD resampling for small water targets with increasing OZA and resulting duplicate pixels. The three different degrees of zenith angle shown correspond to OLCI L1b band 17 (865nm) on the dates indicated. The Vaal and Bronkhorstspruit dams lie on roughly the same longitude and thus result in similar OZAs. The increased number of duplicate pixels with increasing OZA reduces data quality for small targets and means that a cut-off OZA might need to be applied when viewing such small targets. In order to correct for the effect of duplicate pixels, an image-wide correction was applied which interpolates over them using a cubic spline. Although some

assumptions are made by interpolating data across pixels, the correction smooths out the image and provides a more realistic gradation of radiances rather than duplicates (Fig. 10).

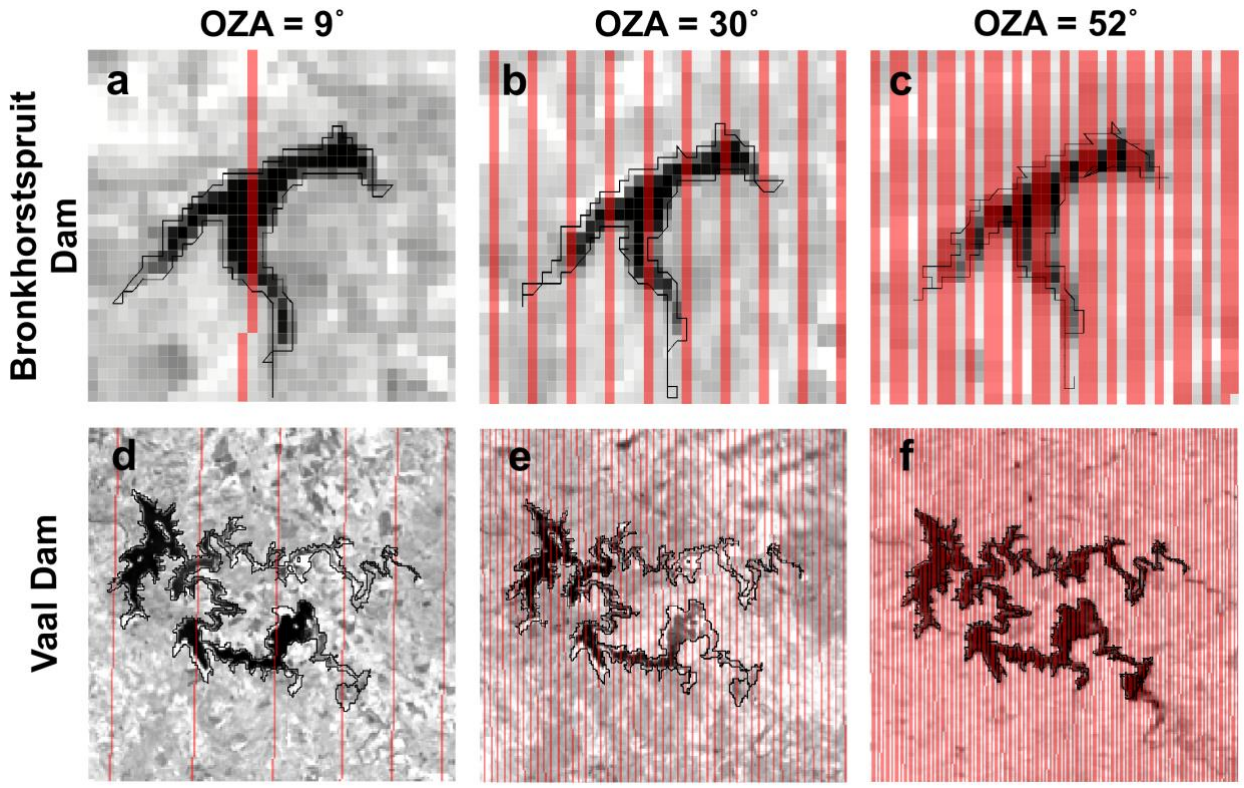

Figure 9: OLCI L1b band 17 with duplicate pixels highlighted in red for Bronkhorstspruit and Vaal dams at varying degrees of OZA on June 20, 2016, October 30, 2017, and April 3, 2017, respectively. For OLCI, an OZA of 9˚, 30˚, and 52˚ results in 3%, 25%, and 60% of the subset consisting of duplicate pixels.

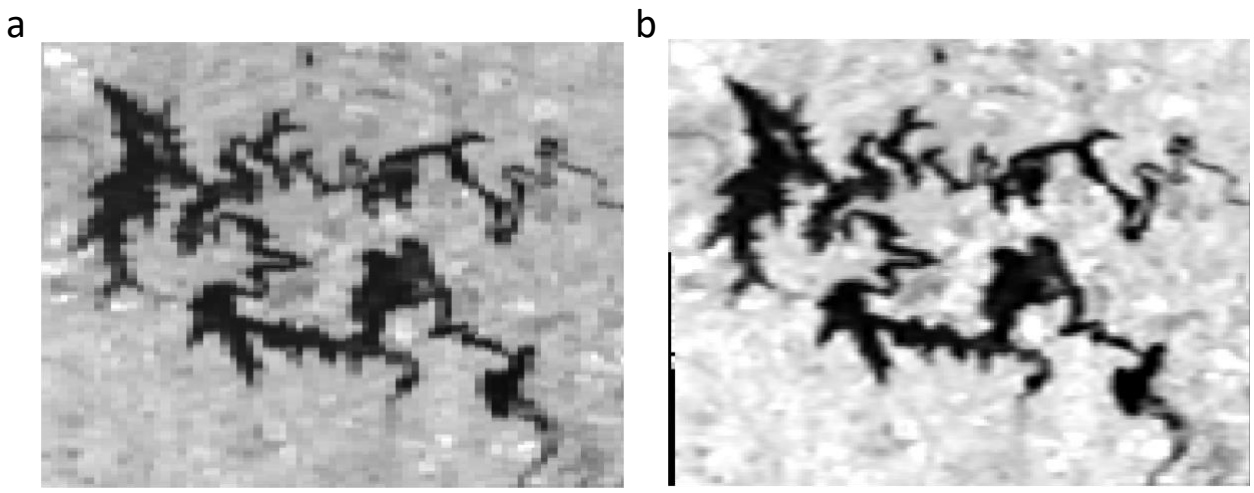

Figure 10: OLCI L1b band 17 of Vaal dam on April 3, 2017 with an OZA of 52˚ (a), uncorrected image (b), duplicate pixel corrected image.

2.4.2.2 Atmospheric corrections

For each study site, in situ R_{rs} measurements and matchup full or partially atmospherically corrected reflectance measurements for all sample points were averaged, with errors calculated as one standard deviation from the mean and interpreted as representing the temporal and spatial variability of each site (Figs. 11, 12). Matchup TOA reflectance, BRR, and in situ R_{rs} are normalized by their respective integrals and plotted against each other in Fig. 11 to qualitatively assess preservation of spectral shape. In the blue region, below 560nm, the BRR correction removes much of the Rayleigh scattering and forms a more defined green peak at 560 nm than TOA reflectance. Beyond 560 nm, BRR's are slightly elevated compared to TOA reflectance, however, many of the spectral features in the red and NIR appear consistent between the two.

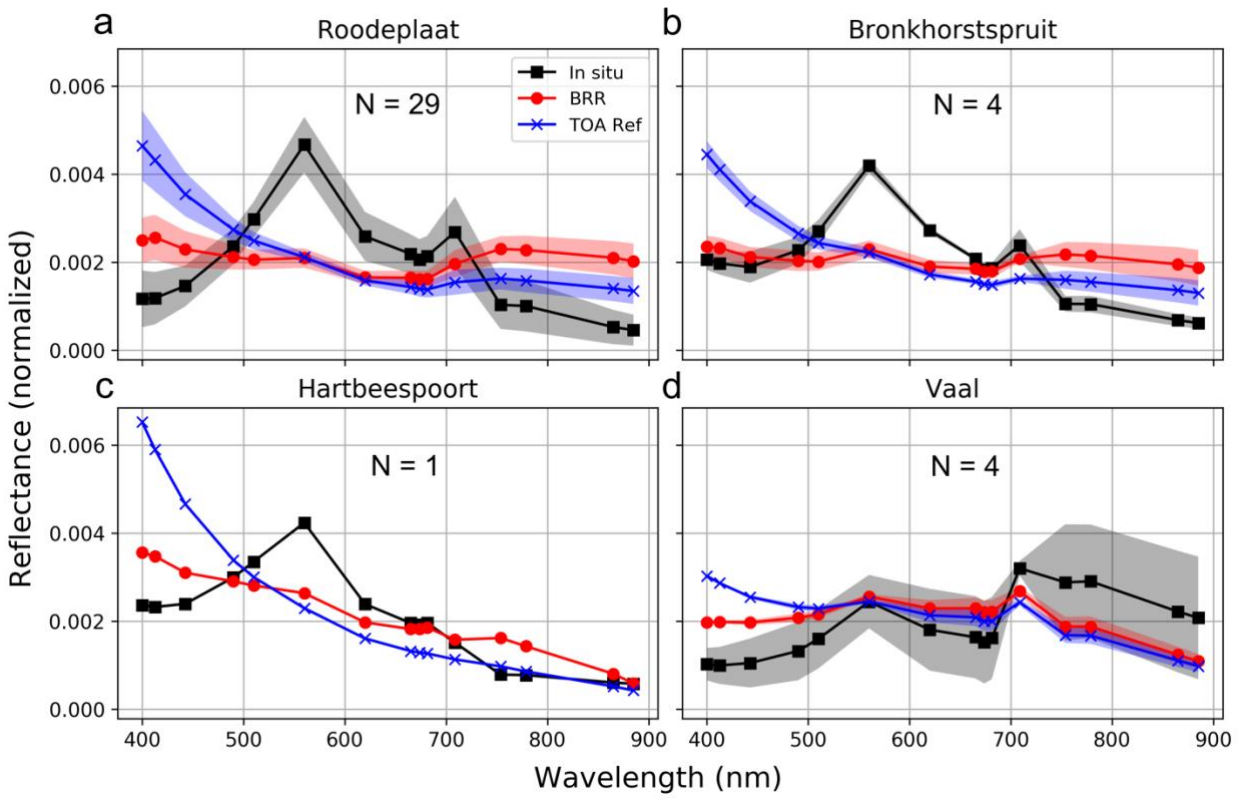

Figure 11: Qualitative comparison between matchup in situ R_{rs}, BRR, and TOA reflectance (a), Roodeplaat (b), Bronkhorstspruit (c), Hartbeespoort (d), Vaal. All reflectances are normalized by their respective integrals. Solid lines are means of all matchups for each dam, shaded regions are standard deviation from the mean.

There are notable differences in the capabilities of the four full atmospheric correction methods to characterize fully the high variability of in-situ radiometry. In similar style to Figure 8, Figure 12 displays a qualitative investigation for comparison between in situ and atmospherically corrected reflectances. C2RCC, POLYMER, and iCOR all produced some negative reflectances in the smaller dams of Roodeplaat and Bronkhorstspruit. 6SV1 mean derived spectra preserved spectral shapes quite well, except beyond 709nm which was elevated due to unaccounted background radiance. Standard deviations are quite high in smaller dams, indicating instances where the correction did not perform properly. POLYMER preserved the spectral shape of bloom waters much better than C2RCC,

but consistently overcorrected: negative reflectances are observed at the chl-a absorption peak around 675 to 680 nm and beyond 850 nm in smaller dams, with more encouraging results at Hartbeespoort and Vaal dam. iCOR preserved spectral shape similarly to 6SV1, however the application of SIMEC for these waters produced widely variable results throughout the spectrum, inducing large standard deviations per waveband and negative reflectances in smaller dams. Fig 12g and 12h indicate where models over-corrected in the NIR due to OLCI's inability to pick up the floating scum signal within the $300m^2$ pixel.

For a quantitative investigation into band-wise performance of atmospheric corrections to validate some of the qualitative findings, readers are directed to Figures 13 and 14. Some of the statistics may be misleading, such as a relatively high band-wise R^2 reached by C2RCC and POLYMER due to low measured in situ signal comparing favorably with the low signal from the failed corrections. More informative results could be deduced from the higher band-wise RMSE values, and negative biases. Overall, band-wise results proved most favorable for the NIR scattering peak at 709nm, which his promising for use with chl-a models utilizing this band. Results of iCOR retrieval of AOT at 550 nm were quite poor when compared to match-up AOT retrievals using Microtops (Fig. 15). An R^2 value of zero was retrieved with a RMSE of 74%.

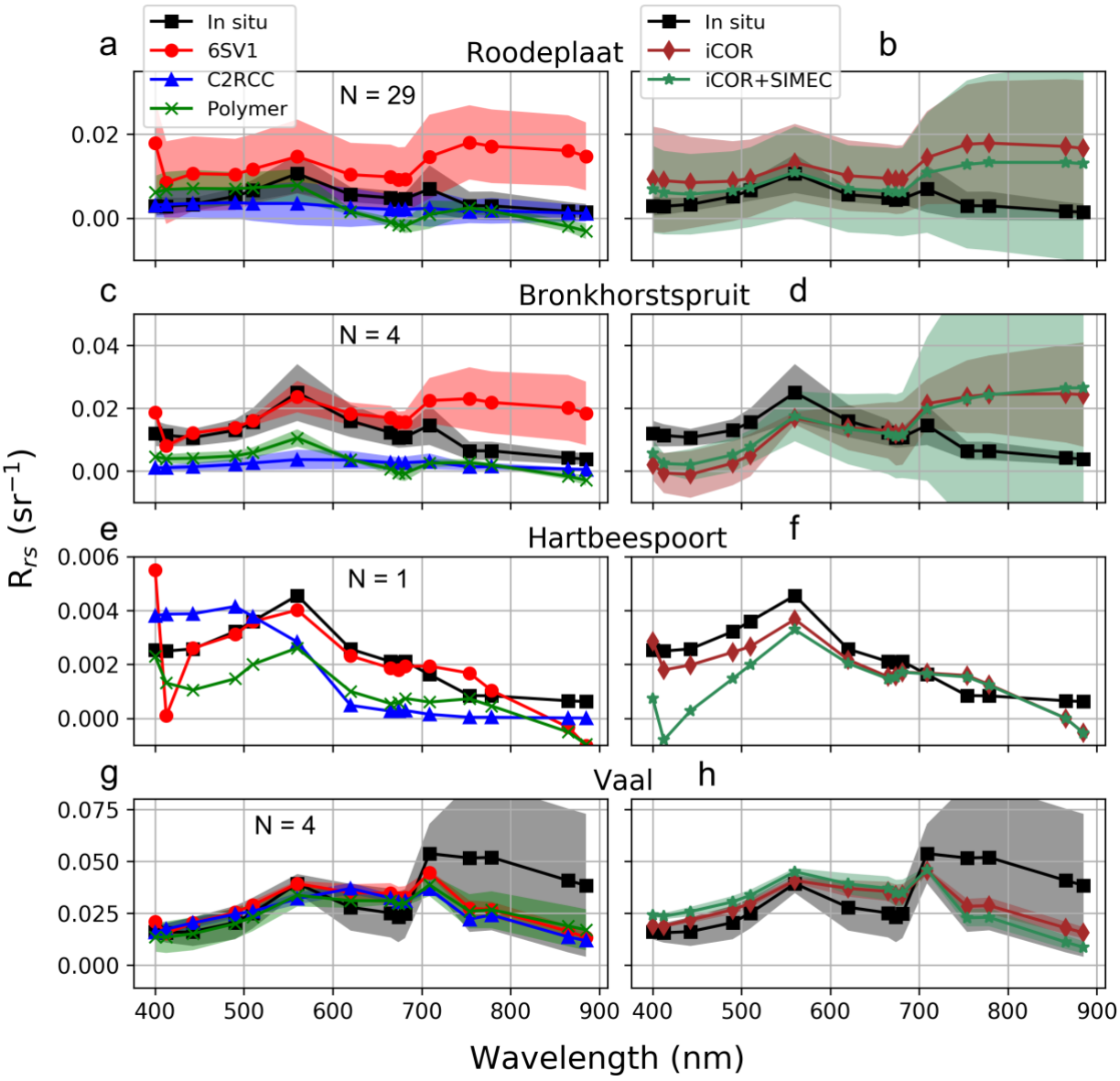

Figure 12: Qualitative comparison of in situ R_{rs} measurements against 6SV1, C2RCC, POLYMER, and iCOR corrected R_{rs} for Roodeplaat (a,b), Bronkhorstspruit (c,d), Hartbeespoort (e,f), and Vaal (g,h). Solid lines are mean reflectance for all matchups while shaded regions are one standard deviation from the mean.

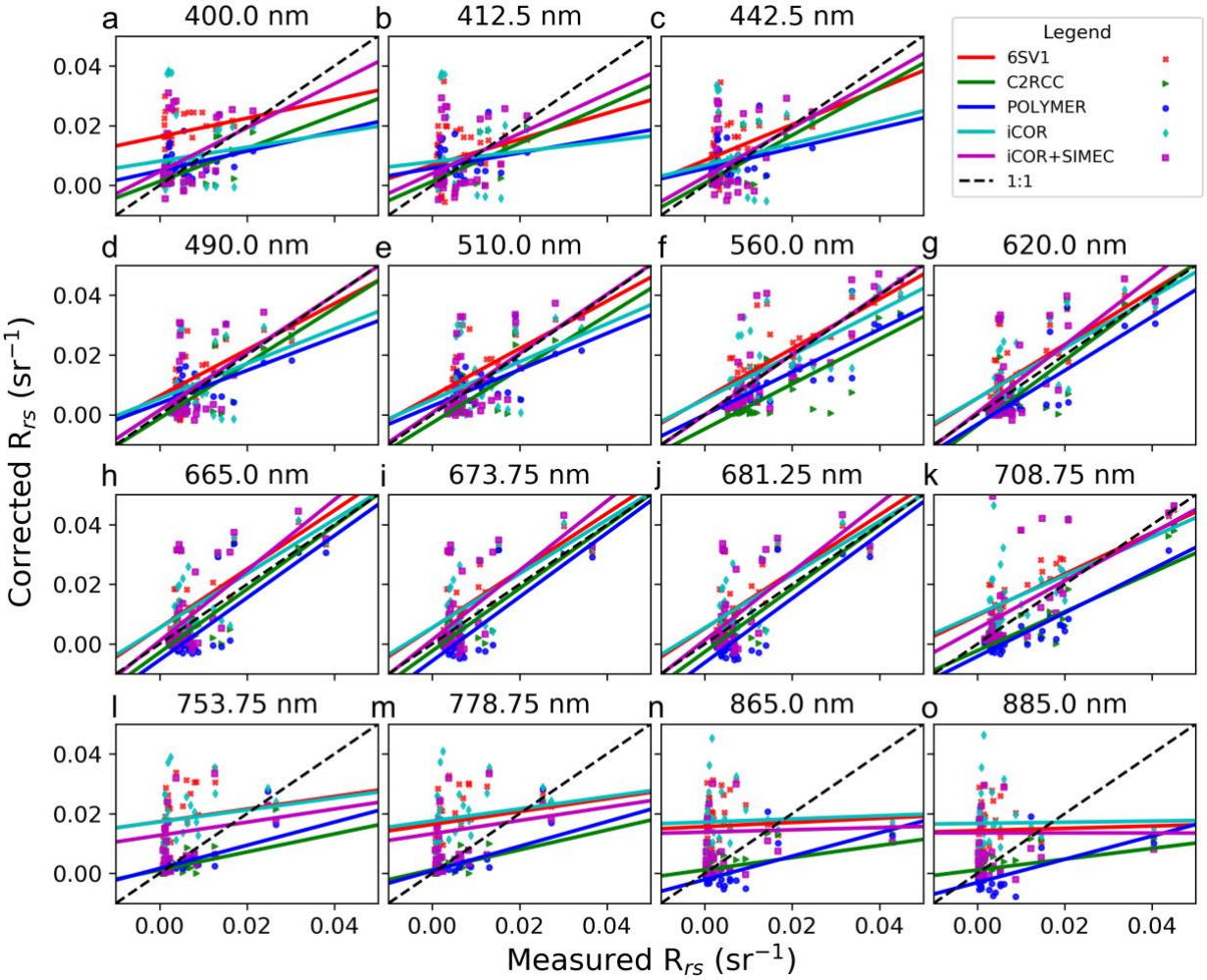

Figure 13: Scatterplot comparison between measured and atmospherically corrected R_{rs} at selected OLCI bands for 6SV1, C2RCC, POLYMER, and iCOR.

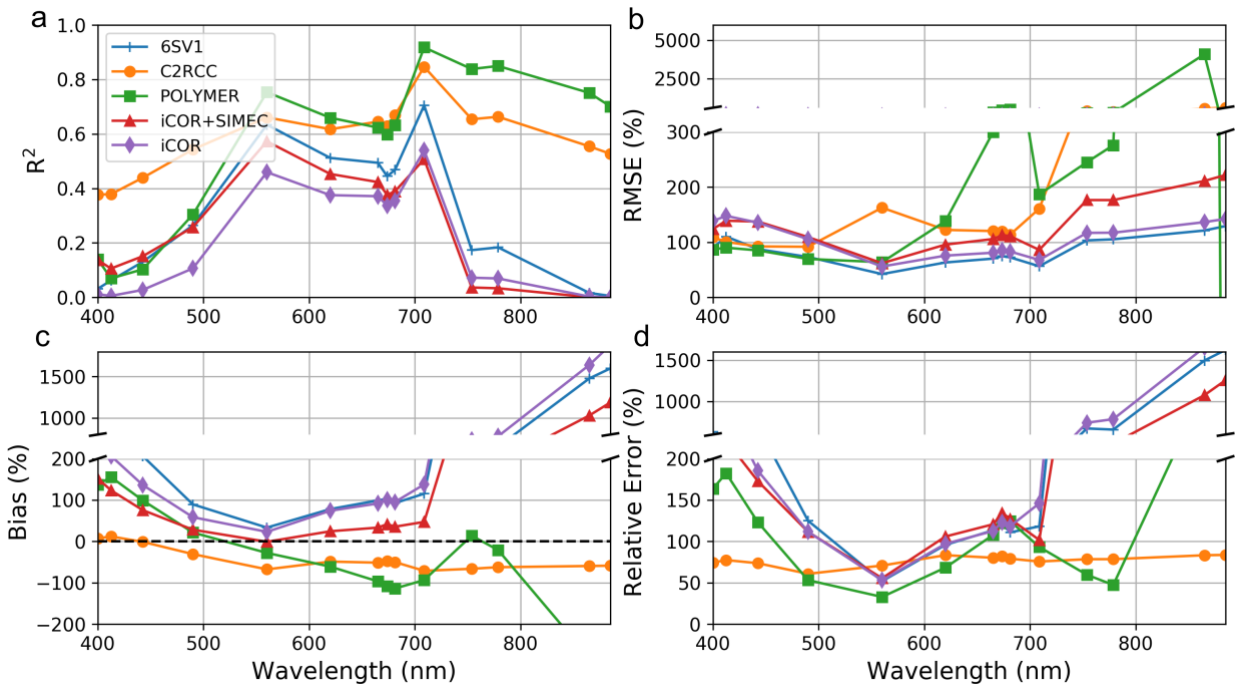

Figure 14: R^2 (a), RMSE (%) (b), bias (%) (c), and relative error (%) (d) of 6SV1, C2RCC, POLYMER, and iCOR derived R_{rs} when compared to in situ measurements for OLCI bands.

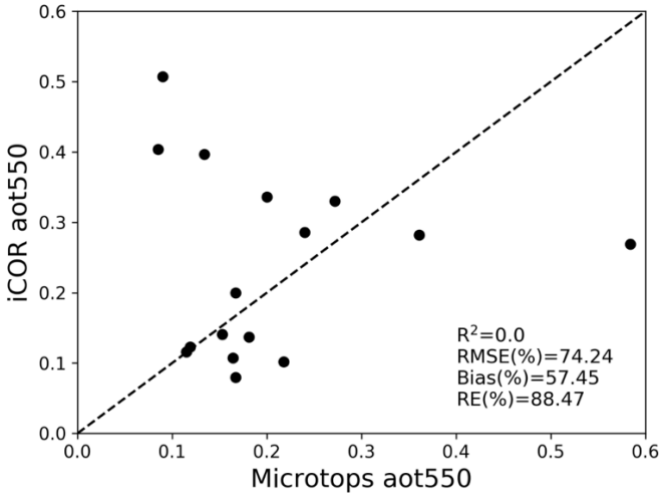

Figure 15: Comparison of matchup iCOR and Microtops derived AOT at 550 nm.

2.4.3 Chlorophyll-a models

The results for the local tuning of chl-a models are found in Appendix B.1. Linear and non-linear trends were tested for all models, but logarithmic and power trends were not performed in cases where negative indices were obtained. Aside from models applied to in situ R_{rs} data, quadratic polynomial functions resulted in highest R^2 values in all cases. LOOCV was then used to test the predictive ability of calibrated models for the estimation of chl-a concentrations. Scatter plots of most reasonable performing models can be found in Figure 16, otherwise readers are directed to Appendix B.2 for complete LOOCV matchup statistics.

At TOA, model performance was quite similar when applied to BRR and at-sensor reflectance.. Spectral shape type models performed best overall with MPH15 performing best with an R^2 0.55 and RMSE of 63.3 mg/m^3. A recalibrated MPH variable using only one calibration curve for both cyanobacteria and eukaryotes resulted in a smaller R^2 of 0.47, but a slightly lower RMSE of 61.3 mg/m^3. Results varied widely when chl-a models were applied to atmospherically corrected reflectances. Most models were unable to overcome the lack of preservation of spectral shape or handle negative reflectances.. Only derivative type models proved to show any predictive ability when applied to corrected reflectances, particularly MCI. Band ratio based models generally performed very poorly, unable to account for inconsistent corrections. However, the band ratios performed extremely well when applied to in situ R_{rs} data. The D05-3B and G08 algorithms matched with highest R^2 of 0.94, but D05-3B achieved an RMSE of 15.2 mg/m^3 while G08 received the worst RMSE of all models applied to in situ R_{rs} with 66 mg/m^3. When comparing chl-a retrievals with OZA (Fig. 16), there appeared to be no significant relationship between standard error of chl-a estimation and OZA for any of the models.

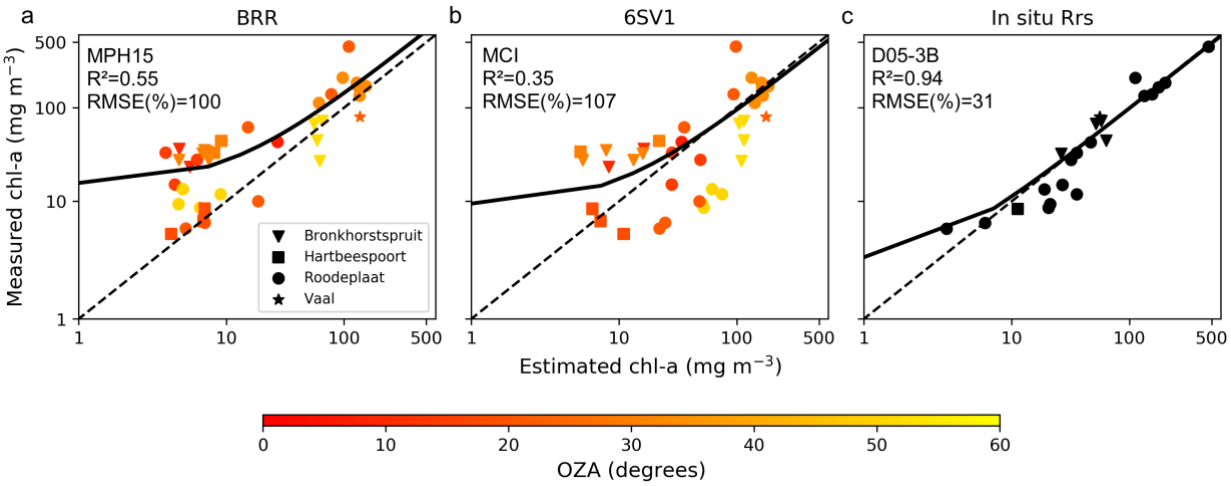

Figure 16: Best performing chl-a retrieval models for BRR (a), 6SV1 (b), and in situ R_{rs} (c) when compared to in situ chl-a. Limited performance statistics are provided by LOOCV, for full summary of predictive ability of all models, see Appendix B.

2.5 Discussion

2.5.1 Synopsis of chl-a and atmospheric correction models for OLCI application

The specific objective of this study, aside from quantifying adjacency effects, is to evaluate different atmospheric correction models and chl-a retrieval algorithms for OLCI radiometry. The above analysis of chl-a models using OLCI L1b versus in situ radiometry highlights the difficulties of retrieving accurate estimates of chl a concentrations in productive inland waters from satellite radiometry. This is largely due to the inaccuracies of atmospheric corrections. The discussion below details the shortcomings of the AC models used, the relative strengths and weaknesses of the chl-a algorithms, and an assessment of the suitability of OLCI radiometry for inland water monitoring (addressing some technical considerations as well as adjacency).

2.5.1.1 C2RCC

Even though the "Extreme Case" alternative NN of C2RCC was used, the poor results can still be mostly attributed to the limited training ranges of in-water and atmospheric optical properties, and inability of handling cyanobacteria bloom conditions. The NN was trained using data from the NOMAD database which is limited in its scope of bloom waters. The algorithm is also incapable of handling increased background radiance. Atmospheric optical properties over rural inland areas of South Africa are also more complex than those found over the coastal and ocean waters C2RCC was trained with, due to the increased presence of biomass burning aerosols. Other recent studies experienced similarly poor results from OLCI using C2RCC over bloom waters of the Baltic sea (Toming et al., 2017; Kutser et al., 2018) and Chinese lakes (Bi et al., 2018). The first iteration of the NN, the Case 2 Regional (C2R) and its associated eutrophic lakes processors (Doerffer & Schiller, 2008), have also been shown to perform poorly over bloom conditions in inland waters using MERIS (Binding et al., 2011; Palmer et al., 2015). Good agreement between C2R and in situ products have been found over less productive lakes in Italy (Giardino et al., 2010), perialpine lakes in the Alps (Odermatt et al., 2010), and other large lakes through Europe (Alikas et al., 2008). However, the inability for C2RCC to preserve any spectral shape in bloom waters in the red and NIR results in extremely poor performance of band ratio and band difference algorithms for chl-a estimation. Thus, using C2RCC should only be considered applicable to oligotrophic and mesotrophic (chl-a < 20mg/m^3) inland waters which are big enough to ensure minimal adjacency issues.

2.5.1.2 Polymer

Statistically, POLYMER performed equally as badly as C2RCC, but although negative values occurred, the preservation of spectral shape using POLYMER is promising for bloom waters.

The POLYMER model is an optimization approach which incorporates the use of all available bands in the vis/NIR. The spectrally smooth function that is removed from the total signal helps preserve the spectral shape, by removing effects of aerosols, sun glint, thin clouds, and even adjacent surfaces that contain a spectrally smooth albedo (such as snow/ice and sand). It seems, however, that there is a limitation when it comes to strong vegetation adjacency because the spectrally smooth function does not represent well the strong signal in the NIR from green vegetation. Roodeplaat and Bronkhorstspruit may be too spatially constrained for POLYMER, as much better performance was achieved at Hartbeespoort and Vaal dams. The good results from Vaal indicate that POLYMER can handle hypertrophic bloom conditions. However, strong green vegetation adjacency causes inaccurate polynomial fitting and overcorrection, resulting in negative reflectances and therefore poor chla retrievals. This overcorrection by POLYMER over inland waters has also been noted in China (Xue et al., 2019; Bi et al., 2018) and suggests that POLYMER may not be suitable for inland water retrievals. However, POLYMER is continuously being developed and future versions could address this.

2.5.1.3 6SV1

Chl-a models applied to 6SV1-corrected data produced the only reasonable estimates of chl-a concentrations from the four atmospheric corrections. The poor performance in the blue for 6SV1 can likely be attributed to poor characterization of aerosols using a generalized model from the radiative transfer code. Best results with 6SV1 were found using the biomass burning aerosol model, however, optical components of the local atmosphere in SA are not likely to be well characterized by a generalized model. The carbon-rich aerosols in South Africa tend to be highly absorbing with low local single scattering albedo (SSA), which is probably underestimated in the biomass burning model. Good results using 6SV1 have been seen before over South African (Matthews et al., 2010) and Italian

inland waters (Giardino et al. 2015) using MERIS and MODIS, and it has performed well with Sentinel 2 MSI over Amazonian waters (Martins et al., 2017). When compared with C2RCC and POLYMER, Xue et al., (2019) also report that 6SV1 produced more accurate reflectances over Chinese inland waters. A major drawback from using 6SV1 is its dependency on atmospheric parameters such as AOT as input, which hinders its ability to be used in near-real-time (NRT) operational processing chains unless combined with other methods of retrieving image-based AOT (Martins et al., 2017, Guanter et al., 2009), or in conjunction with Aeronet stations.

2.5.1.4 iCOR

iCOR is a fully self-contained method that derives AOT over land based on a method developed by Guanter et al., (2005) using a multiparameter endmember inversion technique and pre-defined default vegetation and soil spectra (Keukelaere et al., 2018). The derived land AOT's are then interpolated over inland water regions. Other atmospheric correction parameters are then derived from MODTRAN-generated LUTs. While 6SV1 performed best using a biomass burning aerosol model, iCOR uses a standard rural aerosol model, which again can be limiting in terms of actual in situ aerosol particle variability. The AOT retrieval method was also shown to perform quite poorly (Fig. 15) when compared to in situ Microtops measurements, likely contributing to the poor comparison of iCOR-derived and in situ measured reflectances.

More testing is needed to make definitive conclusions on applying SIMEC with iCOR. While chl-a retrieval results did not seem to be affected by SIMEC, validation of corrected reflectances in the smaller dams showed significantly higher RMSE in the red/NIR spectral regions, including more instances of negative reflectances. Due to these inconsistencies, it is not advisable to use SIMEC for OLCI when observing smaller water bodies where AE can be substantial. In larger lakes, the SIMEC

correction has been shown to improve the accuracy of atmospherically corrected reflectances using Sentinel 2 MSI and Landsat 8 OLI (Keukelaere et al., 2018). The fact that iCOR is also fully image based, with no additional user input, makes it ideal for operational processing, although the algorithm currently seems hindered by poor AOT retrieval in this region. Processing time may be a factor when batch processing.

2.5.1.5 Semi-analytical and band difference algorithms

The chl-a algorithm testing results suggest that in the absence of a high-quality atmospheric correction, band difference type algorithms are most robust. The locally tuned MCI algorithm produced reasonable chl-a predictions when applied to TOA or fully atmospherically corrected data, with the exception of the POLYMER and C2RCC cases. All empirical algorithms must be calibrated to the type of data they are applied to, but the adaptability of MCI to various data types is extremely useful in areas where high quality reflectance information is difficult to obtain. For this reason, MCI was commonly used as an effective tool for quantitative and qualitative mapping of bloom characteristics using MERIS (Alikas et al., 2010; Binding et al., 2011; Palmer et al., 2014; Gower et al., 2005). The ability to be easily applied to multiple sensors, such that the sensor includes a band at the NIR scattering peak, also makes it an ideal model for information continuity and has successfully been applied to OLCI here and in Xue et al., (2019). The MPH15 algorithm outperformed MCI when applied to BRR, most likely due to the peak switching nature of the model and having two different calibrations depending on if the pixel was flagged as dominated by cyanobacteria or algae. The majority of waters sampled contained a prominent NIR reflectance peak at 709nm. When this is the case, the MPH and MCI are measuring the same spectral feature, although with different calibration curves. Pitarch et al., (2017) found better success with MPH using one standard calibration curve with MPH instead of having separate cyanobacteria and eukaryotic curves. This was not the case with our

data (Table A.2) where a single calibration curve using the MPH index did not perform as well as the pre-calibrated MPH15 algorithm.

Band ratios can be simple yet very powerful tools as evidenced by the very good performance when applied to in situ R_{rs} here and in other studies using MERIS bands (Gurlin et al., 2011; Moses et al., 2009a; Moses et al., 2012). Band ratio models using OLCI bands tested quite well on synthetic datasets comprising reflectances typical of productive waters (Watanabe et al., 2018; Ligi et al., 2016). Semi-analytical algorithms, however, are still sensitive to imperfect atmospheric corrections or other uncertainties associated with sensor noise or sun glint. However, results in table A.2 show that band ratio models still perform better at TOA than on fully atmospherically corrected reflectances. This suggests that when band ratio models are applied to un-normalized TOA data with variable atmosphere, there is less error than when using an unreliable full atmospheric correction for these water types. Band difference algorithms are more tolerant of artifacts that may reduce the quality of reflectances, thus their ability to perform better on TOA reflectances (Hu et al., 2012; Mathews et al., 2011; Matthews et al., 2010; Stumpf et al., 2012; Palmer et al., 2015). Local tuning of band difference algorithms is recommended; however, their ease of implementation and low computational requirements make them ideally suited for operational monitoring of inland water bodies using OLCI. This analysis confirms the applicability of models such as the MPH, MCI and band ratio algorithms with OLCI as for MERIS.

2.5.2 Challenges and sources of error affecting OLCI

Inland water remote sensing has received more attention this decade due to a better global understanding of the threats and challenges that harmful algal blooms such as cyanobacteria pose. However, the methodology is still very error prone and it is important to understand the sources of

these uncertainties. A potentially large source of error with medium resolution sensors such as OLCI, MERIS or MODIS is the influence of horizontal heterogeneity due to bloom patchiness. Chl-a concentrations have been shown to vary by two orders of magnitude within a 300 x 300 m pixel and it is suggested that even a 30 x 30 m pixel could be too coarse depending on the bloom (Kutser et al., 2004). This type of uncertainty can be observed in the figures for qualitative atmospheric correction assessment at the Vaal dam (Figures 8d, 11d, 12g,h). In situ R_{rs} measurements over bloom waters where surface scums were evident exhibited highly elevated reflectance in the NIR, as is the case with dry vegetation. However, a spatial averaging of 300 m^2 reduces the signal from the scum as seen in the above-mentioned figures. For small water bodies, the validation error due to sub-pixel features (such as surface scum) is also compounded by the error due to adjacency effects from nearby vegetation. Accurately distinguishing between these two sources of error requires further inspection, potentially through parameterized modeling exercises.

Hu et al., (2009) suggested that for a surface slick to be detected by a satellite sensor, the width needs to be at least one fourth of the sensor pixel size. Thus, for a 300 x 300 m OLCI pixel, a surface scum must be at least 75 m in width to be detectable. In this investigation only one sample point per pixel was used for validation, whereas multiple sample points per pixel area would be much more representative and reduce the error due to bloom patchiness, although this can drastically reduce the number of matchup points. Using the 709 nm band of a Hyperspectral Tethered Surface Radiometer Buoy (Satlantic Inc.), Matthews et al., (2012) calculated a point sampling error in a hypertrophic reservoir of between 15% and 45%; however in scum conditions this can be considerably larger. The error can also be visualized in Figure 16 where a very high biomass sampling point at Roodeplaat (450 mg/m^3) was severely underestimated using OLCI, but was predicted quite accurately when the models were applied to in situ reflectance.

Vertical heterogeneity also poses a difficult issue in terms of in situ water sampling. Kutser et al., (2008) showed how the vertical distribution of cyanobacteria can have a significant impact on R_{rs} and chl-a retrievals. In highly stratified conditions, the top optical layer may only be a few centimeters thick, hence typical water sampling protocols cannot be applied. In this study, bucket samples were taken from the top 0.0 - 1 m of the water column, depending on the secchi depth, then mixed well. This procedure homogenizes the top surface layer and could introduce considerable error depending on the average top optical layer of the satellite pixel. The artificial mixing most certainly disrupted particle assemblages and aggregates from their natural state, however, the extent of the error this potentially introduces for validation purposes is unknown and requires further inspection. Although, considering the spatial averaging of OLCI's 300 m pixel size, the error due to water sample mixing is probably minimal compared to errors due to horizontal heterogeneity.

In this study, two very small dams were examined to test the capabilities of OLCI for spatially extreme cases. Roodeplaat and Bronkhorstspruit dams are roughly 4 and 8.5 km^2 in total surface area, respectively. Hestir et al., (2015) suggest that for a water pixel to be considered "pure" and not contain significant contamination from nearby vegetation, the water body should be four times the size of the pixel. At 300 m^2 resolution for OLCI, a potential water target should be at least 1200 m^2 with no mixed pixel interference. The largest pure water polygon at Bronkhorstspruit is roughly 1000 m^2, while Roodeplaat is 750 m^2. These are significantly smaller than what Hester et al., (2015) recommend and is why these dams exhibit strong adjacency signal in the NIR. At edge of swath for OLCI, where OZAs range between 40˚ and 60˚, the area required for a water target to be considered pure increases to between 1560 m^2 and 2400 m^2, respectively. Even though Roodeplaat and Bronkhorstspruit would be considered too small to measure using these guidelines with OLCI, reasonable chl-a retrievals were still possible with derivative models. The results of the modeling

study in section 4.1 showed that OLCI is indeed picking up significant background radiance in the red/NIR when observing small water targets. The surrounding bushveld in this region is primarily made up of green vegetation and thus contributes considerable background signal in the NIR (Bulgarelli et al., 2018). However, the blue and green spectral region was only minimally affected. Given the amount of contamination at longer wavelengths for these small targets, it should be noted that OLCI is still observing enough of the water signal to distinguish variations in spectral shapes. At Roodeplaat dam, in the presence of strong adjacency, the MPH algorithm was still able to detect concentrations < 10 mg/m^3 at BRR (Fig 16). The detection limit for the MPH algorithm with OLCI has been calculated as between 1 and 5 mg/m^3 depending on the variable fluorescence quantum yield of natural phytoplankton populations (Matthews, unpublished), but inland waters here in South Africa rarely get that low. It is recommended that for band difference algorithms, λ_3 be located farther into NIR closer to 900 nm, where the strong absorption due to water reduces the magnitude of reflectance induced by adjacency effect from nearby vegetation.

The 12.6˚ camera tilt aboard OLCI leads to OZA's near the edge of swath reaching upwards of 60˚, contrary to the 68.5˚ field of view over nadir aboard MERIS (roughly 35˚ at edge of swath). Increasing OZA from nadir to the edge of swath has the potential to increase the amount of adjacency influence (Bulgarelli and Zibordi, 2018). Royal et al., 1985 examined the effects of satellite viewing angles on reflectance measurements and found significant increase in intrinsic atmospheric reflectance with increasing OZA, with larger effects seen in shorter wavelengths due to increased Rayleigh scattering. They also found that the spectral contrast of neighboring targets will increase with increasing off-nadir viewing. Figure 17 shows the same matchup point at the radiometric buoy location at Roodeplaat dam for three dates with similar atmospheric and water optical properties, but with varying OZA. The three points were dominated by cyanobacteria and had similar chl-a concentrations.

While no quantitative analysis was undertaken, Figure 17 appears to validate some of the above-mentioned findings. There was a notable increase in stray-light with increasing OZA, as well as increased radiance in the blue region due to Rayleigh scattering. Although there did not appear to be a significant relationship between OZA and chl-a retrieval accuracy in our data (see Fig. 16), caution must be taken when examining targets near edge of swath and an OZA cutoff should potentially be taken into consideration for smaller water targets.

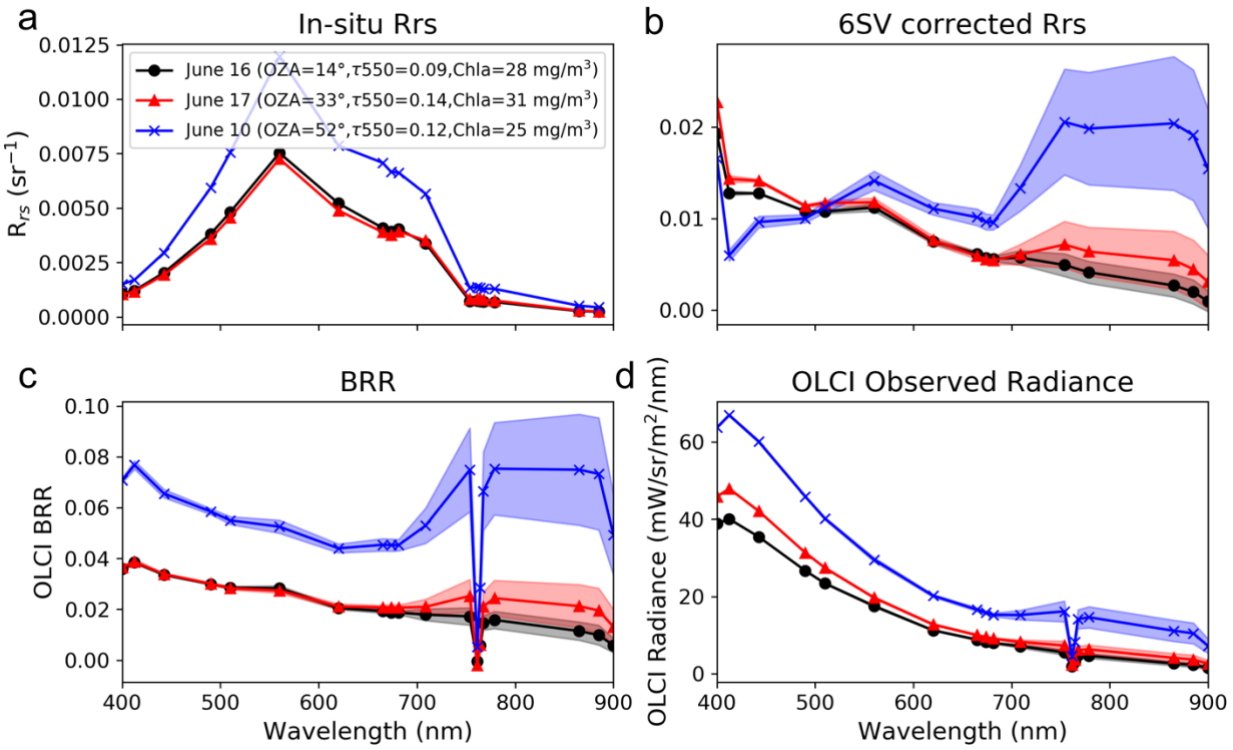

Figure 17: Effects of increasing OZA on OLCI derived reflectances and radiances for three dates of similar water and atmospheric optical properties (a), in situ R_{rs} (b), 6SV1 corrected R_{rs} (c), BRR (d), OLCI observed L1b radiance.

A robust method for retrieving atmospheric aerosol load for atmospheric correction is still one of the biggest challenges to remote sensing of productive inland waters. Algorithms which retrieve AOT information from the image itself, such as iCOR, would ideally be the most useful for an operational monitoring program. iCOR is quite computationally expensive, and other radiative transfer solver methods such as 6SV1 are able to run much faster. Aeronet stations are a useful source of relevant atmospheric variables for algorithms which require *a priori* atmospheric information, although one must be careful of the spatial heterogeneity of atmospheric aerosols over land. At very small spatial scales (<30 km) it can be assumed that aerosol properties are constant (Guanter et al., 2007), but opinions differ about larger scales and appear to be location dependent e.g. Holben et al., (1991) , Kumar et al., (2013).

2.6 Conclusion

A suite of semi-analytic band ratio and spectral shape band difference algorithms were applied to OLCI L1b TOA reflectance, BRR, and atmospherically corrected R_{rs} measurements to test the challenges and limitations of OLCI for monitoring small productive inland water systems. The above results indicate that OLCI will prove to be a valuable information source for monitoring inland water systems as long as one understands the uncertainties and limitations involved in individual applications. The 6SV1 model produced the best quantitative results as a full atmospheric correction for OLCI when compared to iCOR, C2RCC, and POLYMER. C2RCC should only be applied for pure water pixels in more oligotrophic to mildly mesotrophic waters. Empirically derived derivative models such as MPH and MCI provide a robust, easy and fast method of reliably retrieving information on trophic status, as was the case for MERIS. In the absence of a reliable atmospheric correction, these models provide adequate estimates of chl-a concentrations. Band ratios proved not to be as robust as band

difference models, and if the quality of the atmospheric correction is questionable, or the user intends on using TOA data such as TOA reflectance or BRR, derivative type models should be used (although band ratios work exceptionally well on in situ data). AE contributes considerable error to the red/NIR spectral region for these waters, and SIMEC was not able to properly account for this, thereby not improving retrieval results. Horizontal water heterogeneity pose an issue with OLCI and multiple sample points should be considered within each pixel if a validation exercise is to be undertaken in hypertrophic waters. Further research is required to understand the negative effects of retrieving information from OLCI for off-nadir targets. A simple exercise in this study verified the influence of amplified Rayleigh scattering and adjacency effects with larger OZA. A standard OZA cutoff around 40˚ would seem appropriate. If the goal is to fully map a water body for biogeophysical distributions, one should consider a duplicate pixel correction to provide a more realistic view of the distribution. This study has also proven the potential for OLCI to monitor very small water bodies such as Roodeplaat dam, with a surface area of roughly 4 km^2 and a largest detectable water polygon of 750 m^2. Although the entire surface area of the reservoir cannot be mapped with OLCI, a few valid pixels were proven to be attainable, as long as one takes precaution with mixed pixels and understands the uncertainties involved.

2.7 Appendix A

During field campaigns, radiometric data was collected concurrently with water samples using an ASD-FR Field Spectroradiometer 3 (Analytical Spectral Devices, Boulder, CO USA). Methods to derive remote sensing reflectance were derived from Mueller et al., (2003) and briefly described here. Measurements of a white Spectralon reflectance plaque with a known BRDF were used to normalize the uncalibrated radiance measurements for downwelling irradiance, $E_s(\lambda;\theta_o)$. The plaque was held horizontally and exposed to the sun, free from any shading or reflections by equipment, the boat, or

crew. Radiances of the plaque (S_g) were recorded using a viewing angle of $\theta = 40°$ away from the nadir and azimuthally away from the sun at $\boldsymbol{\phi}$ =135° and ten spectra were collected. Immediately following measurements of S_g, the radiometer was used to collect both water and sky radiance measurements, L_{sfc} ($\lambda,\theta,\varphi \in \Omega_{FOV};\theta_0$) and L_{sky} ($\lambda,\theta_{sky},\varphi_{sky} \in \Omega'_{FOV}; \theta_o$) respectively. Measurements were performed using the same zenith and azimuthal viewing angles as the plaque and again, 10 spectra were collected for each. Measurements were collected on the sunny side of the boat and precautions were taken not to collect water measurements in recently disturbed water from the boat or where there was excessive surface roughness. Once completed, a dark offset reading was taken and the whole process was repeated three times with the mean of the spectra being used. Water leaving radiance (L_w) was computed as:

$$L_w(\lambda,\theta,\varphi \in \Omega_{FOV};\theta_0) = F_L(\lambda)\,[S_{sfc}(\lambda,\theta,\varphi \in \Omega;\theta) - \rho S_{sky}(\lambda,\theta_{sky},\varphi_{sky} \in \Omega';\theta_0)] \quad \text{(A1)}$$

where $F_L(\lambda)$ is the instrument's unknown radiance response calibration factor, and $S_{sfc}(\lambda,\theta,\varphi \in \Omega;\theta)$ and $S_{sky}(\lambda,\theta_{sky},\varphi_{sky} \in \Omega';\theta_0)$ are the radiometer's measured responses. The reflectance of skylight from the water surface ρ is estimated using approximations based on wind speed using Mobley (1999) and a value of 0.028 was used. The radiance reflected from the plaque is scaled to estimate downwelling irradiance (Es) as:

$$E_S(\lambda;\theta_0) = \frac{\pi F_L(\lambda) S_g(\lambda,\theta_g,\phi_g \in \Omega_{FOV};\theta_0,\phi_0)}{R_g(\lambda,\theta_g,\phi_g \in \Omega_{FOV};\theta_0,\phi_0)} \quad \text{(A2)}$$

where $S_g(\lambda,\theta_g,\phi_g \in \Omega_{FOV};\theta_0,\phi_0)$ is the sensor response signal when the plaque is viewed at angles (θ_g ,φ_g) with the sun at (θ_0, φ_0) , and $R_g(\lambda,\theta_g,\phi_g \in \Omega_{FOV};\theta_0,\phi_0)$ is the plaque's bi- directional reflectance function (BRDF) for that sun and viewing geometry (including whatever is assumed regarding the contribution of sky irradiance to $Es(\lambda;\theta o)) = 1$ in this instance. When equations (1) and (2) are substituted into equation (3) to calculate remote sensing reflectance R_{RS}, the unknown

radiance response calibration factor $F_L(\lambda)$ cancels. Remote sensing reflectance is then calculated using water-leaving radiance as:

$$R_{RS}(\lambda, \theta, \varphi \in \Omega_{FOV}; \theta_0) = \frac{L_w(\lambda, \theta, \varphi \in \Omega_{FOV}; \theta_0)}{E_S(\lambda; \theta_0)} \quad \text{(A3)}$$

2.8 Appendix B

Table B.1

Summary of major characteristics of Hartbeespoort, Roodeplaat, Bronkhorstspruit, and Vaal Dams

Reservoir	***Hartbeespoort***	***Roodeplaat***	***Bronkhorstspruit***	***Vaal***
Location	25°43'30" S; 27°51' E	23°58' S; 27°43' E	25°53'25" S; 28°24'43 E	26°54'11" S; 28°08'31 E
Altitude	1,162m	1,314m	1,429m	1,486m
Catchment area	4,144km^2	668km^2	1,263km^2	37,100km^2
Volume	195 x 10^6 m^3	41.9 x 10^4 m^3	58.5 x 10^6 m^3	2,330 x 10^6
Surface area	20km^2	3.96km^2	8.5km^2	321.07km^2

Max Depth	32.5m	43m	19.5m	47m
Mean Depth	9.6m	10.6m	6.8m	22.5m

Table B.2

Summary statistics for all in-situ water sample points.

Dam	*N*	*Min*	*Max*	*Mean*	*Median*	*St. Dev*
Roodeplaat						
Secchi (cm)	23.0	50.0	420.0	132.9	92.5	106.7
Chl (ug/L)	33.0	5.1	449.0	118.5	109.9	113.8
TSS (mg/L)	30.0	0.8	58.5	18.7	17.0	14.6
Hartbeespoort						
Secchi (cm)	8.0	50.0	315.0	205.5	209.5	75.5
Chl (ug/L)	8.0	4.5	215.3	70.0	33.5	90.7
TSS (mg/L)	8.0	0.8	28.0	7.2	4.8	9.0

Bronkhorstspruit						
Secchi (cm)	12.0	90.0	240.0	150.0	142.5	54.4
Chl (ug/L)	12.0	23.3	72.0	37.5	31.9	16.1
TSS (mg/L)	12.0	4.0	13.0	8.2	8.3	3.0
Vaal						
Secchi (cm)	6.0	10.0	40.0	25.0	20.0	12.2
Chl (ug/L)	6.0	15.1	520.0	160.0	63.8	202.4
TSS (mg/L)	6.0	4.8	84.0	27.3	14.3	30.5

Table B.3

Cell count values for all in situ sample points in cells/ml

Dam	*N*	*Min*	*Max*	*Mean*	*Median*	*St. Dev*
Roodeplaat						

Microcystis sp	23	1555	557480	195274	181724	158519
BACILLARIOPHYCEAE	24	0	3258	679	20	970
CHLOROPHYCEAE	24	0	32942	8075	1810	10753
CRYPTOPHYCEAE	24	0	2534	472	362	629
CYANOPHYCEAE	24	0	557480	187839	179914	160420
EUGLENOPHYCEAE	24	0	724	136	0	234
TOTAL	24	1555	576666	197202	180819	164962
Hartbeespoort						
Microcystis sp	8	483	380617	62691	13300	130146
BACILLARIOPHYCEAE	8	186	4525	1674	1414	1313
CHLOROPHYCEAE	8	0	15204	5699	3439	6544
CRYPTOPHYCEAE	8	0	1812	729	647	651
CYANOPHYCEAE	8	483	424057	68214	13300	145281
EUGLENOPHYCEAE	8	0	362	88	0	164

TOTAL	8	2655	435109	76404	27394	146659
Bronkhorstspruit						
Microcystis sp	12	166	400372	72928	19471	120074
BACILLARIOPHYCEAE	12	80	2172	1107	1086	723
CHLOROPHYCEAE	12	0	8620	2769	1810	2735
CRYPTOPHYCEAE	12	0	690	260	264	231
CYANOPHYCEAE	12	166	406164	76582	30083	120232
EUGLENOPHYCEAE	12	0	362	30	0	105
TOTAL	12	1563	409784	80777	37904	120286
Vaal						
Microcystis sp	6	5068	5792003	1188829	145162	2289205
BACILLARIOPHYCEAE	6	0	12670	3600	1086	5058
CHLOROPHYCEAE	6	0	724	121	0	296

CRYPTOPHYCEAE	6	0	1810	463	121	718
CYANOPHYCEAE	6	11946	5792003	1192831	148782	2286814
EUGLENOPHYCEAE	6	0	0	0	0	0
TOTAL	6	11946	5806483	1197135	153488	2291758

2.9 Appendix C

Table C.1

All model calibration parameters for all datasets. Best performing regression model including calibration coefficients and R^2 are provided.

Source	***Model***	***a0***	***a1***	***a2***	***R^2***
TOA Ref					
MCI	Quadratic	-93927.30	7700.85	16.62	0.53
FLH	Quadratic	-254128.72	-12490.41	17.98	0.41
M09-2B	Quadratic	-394.14	1230.09	-785.11	0.47
D05-3B	Quadratic	-376.24	455.54	49.74	0.46

NDCI	Quadratic	-1062.04	814.88	50.95	0.46
BRR					
MPH	Quadratic	-169671.35	10815.89	7.54	0.59
MCI	Quadratic	-81652.52	7224.72	17.05	0.52
FLH	Quadratic	-248832.29	-12154.13	24.26	0.42
M09-2B	Quadratic	-230.96	816.97	-553.93	0.41
D05-3B	Quadratic	-157.18	301.59	32.65	0.37
NDCI	Quadratic	-701.31	677.11	33.85	0.39
6SV1					
MCI	Quadratic	-380798.73	16371.65	-0.55	0.45
FLH	Quadratic	-984461.35	-26157.31	9.85	0.36
M09-2B	Quadratic	-26.26	149.69	-86.85	0.16
D05-3B	Quadratic	-2.64	23.10	55.45	0.06

NDCI	Quadratic	-810.40	545.84	22.27	0.17
C2RCC					
MCI	Quadratic	-2985446.11	43220.99	46.82	0.26
FLH	Quadratic	8793486.33	-35616.18	54.59	0.20
M09-2B	Quadratic	-129.26	342.82	-124.58	0.19
D05-3B	Quadratic	-1273.43	268.11	94.12	0.17
NDCI	Quadratic	-35.48	189.45	84.07	0.18
Polymer					
MCI	Quadratic	-5.87	163.49	66.55	0.02
FLH	Quadratic	-1399.20	-14816.16	32.70	0.21
M09-2B	Quadratic	1.32	4.40	42.79	0.26
D05-3B	Quadratic	1.91	6.87	54.59	0.05
NDCI	Quadratic	-4.56	61.07	38.53	0.26

iCOR Simec					
MCI	Quadratic	-318399.03	13964.86	16.93	0.43
FLH	Quadratic	-1091613.83	-24578.92	23.75	0.35
M09-2B	Quadratic	-0.02	3.04	53.73	0.11
D05-3B	Quadratic	-0.03	3.57	57.83	0.12
NDCI	Quadratic	-14.86	77.03	49.73	0.11
iCOR no Simec					
MCI	Quadratic	-219327.72	12306.93	8.85	0.37
FLH	Quadratic	-552512.83	-19192.47	19.55	0.26
M09-2B	Quadratic	-8.34	89.95	-55.87	0.24
D05-3B	Quadratic	-10.52	66.82	28.29	0.16
NDCI	Quadratic	-364.27	399.62	9.82	0.17
Rrs					

MCI	Exponent	20.33	187.66	-	0.85
FLH	Exponent	17.46	-504.95	-	0.87
M09-2B	Linear	114.04	-80.59	-	0.91
D05-3B	Linear	211.22	36.21	-	0.95
NDCI	Exponent	33.44	4.11	-	0.91

Table C.2

Model validation and predictive capability results for all datasets including slope (M), Mean absolute error (MAE) in mg/m^3, root mean square error (RMSE) in mg/m^3, relative RMSE in percent, relative bias in percent, and relative error (RE) in percent.

Source	***R^2***	***M***	***MAE (mg/m^3)***	***RMSE (mg/m^3)***	***RMSE (%)***	***Bias (%)***	***RE (%)***
TOA Ref							
MCI	0.42	0.47	36.00	64.46	100.96	64.59	110.57
FLH	0.31	0.36	40.24	70.04	109.71	98.55	143.47

M09-2B	0.37	0.42	38.53	67.21	105.28	88.74	162.13
D05-3B	0.32	0.39	40.11	69.76	109.27	88.10	151.10
NDCI	0.36	0.42	39.03	67.79	106.17	93.45	161.61
BRR							
MPH	0.47	0.53	31.25	61.27	95.97	44.25	85.50
MPH15	0.55	1.29	31.77	63.63	99.67	-37.43	55.03
MCI	0.41	0.47	36.24	64.69	101.32	68.99	114.89
FLH	0.32	0.37	40.07	69.66	109.11	96.95	142.11
M09-2B	0.30	0.35	40.48	70.46	110.37	122.37	185.35
D05-3B	0.26	0.31	41.58	72.85	114.11	133.23	185.41
NDCI	0.29	0.35	40.01	71.00	111.20	126.51	186.22
6SV1							
MCI	0.35	0.41	38.70	68.08	106.63	83.53	129.91

FLH	0.26	0.32	42.03	72.64	113.77	112.73	152.95
M09-2B	0.07	0.13	47.98	82.68	129.51	208.45	269.17
D05-3B	0.00	0.03	59.44	89.79	140.63	229.43	314.74
NDCI	0.09	0.12	51.85	80.89	126.70	227.00	268.98
G08	0.05	0.24	58.14	102.34	160.30	250.32	289.97
C2RCC							
MCI	0.26	0.13	69.10	153.44	240.33	213.22	238.87
FLH	0.20	0.03	73.84	150.63	235.94	224.77	255.27
M09-2B	0.19	0.11	51.17	82.99	129.99	185.29	230.43
D05-3B	0.17	0.10	50.61	83.60	130.95	167.86	196.54
NDCI	0.18	0.12	50.27	82.60	129.37	192.33	228.65
G08	0.19	2.55	54.19	93.04	145.72	-46.08	114.45
C2RCC_Chl-a	0.15	1.08	51.82	89.22	139.74	-10.91	138.13

Polymer							
MCI	0.00	-241	61711	380025	595238	170055	170138
FLH	0.00	38616	9877303	60887518	95368895	-27209923	27210263
M09-2B	0.01	0.16	77.39	154.10	241.37	202.62	234.79
D05-3B	0.01	-0.04	63.09	93.61	146.62	219.65	249.28
NDCI	0.01	0.80	190.30	910.50	1426.13	-252.72	579.60
G08	0.02	0.07	88.73	177.51	278.03	34.27	179.08
iCOR Simec							
MCI	0.26	0.33	41.02	74.13	114.63	89.48	127.86
FLH	0.16	0.24	44.83	79.94	123.60	116.31	155.85
M09-2B	0.04	0.12	54.70	89.83	138.90	220.00	243.90
D05-3B	0.05	0.09	51.39	83.39	128.94	206.21	240.41
NDCI	0.02	0.31	79.48	197.40	305.22	-439.65	878.56

G08	0.10	0.02	502.14	1532.66	2369.78	331.39	547.90
iCOR no Simec							
MCI	0.26	0.32	39.99	72.64	113.77	140.65	174.40
FLH	0.17	0.22	42.71	76.87	120.40	182.78	212.51
M09-2B	0.00	0.07	64.01	176.73	276.82	261.50	281.63
D05-3B	0.01	0.07	52.35	93.74	146.83	228.18	251.90
NDCI	0.03	0.10	53.12	90.74	142.12	214.64	248.68
G08	0.08	0.23	65.00	121.15	189.75	275.78	299.63
Rrs							
MCI	0.73	0.66	33.33	53.91	68.53	48.60	73.78
FLH	0.75	0.69	31.79	52.21	66.37	34.53	61.69
M09-2B	0.87	0.83	23.64	37.54	47.73	14.64	38.66
D05-3B	0.94	0.97	15.20	24.48	31.13	29.16	45.41

NDCI	0.85	0.77	27.77	40.58	51.58	56.74	76.72
G08	0.94	1.99	37.36	66.08	84.00	-20.31	39.98

3

3 CHAPTER 3: A SYNTHETIC HYPERSPECTRAL LABELED DATASET FOR CALIBRATION OF REMOTE SENSING ALGORITHMS FOR PRODUCTIVE INLAND WATERS: PARAMETERIZATION AND ASSESSMENT

3.1 Introduction

Recent advancements in sensor technology and algorithm development have allowed for improved measurements of coastal and inland waters (Pahlevan et al., 2020; Hu et al., 2019; Palmer et al., 2015b; Matthews et al., 2012; Smith et al., 2018). However, significant limitations still exist in the capability of modern-day ocean color sensors to retrieve viable bio-geophysical data with high certainty, especially with regards to inland waters (Kravitz et al., 2020). The development of robust water quality retrieval algorithms depends on the collection of high quality in-situ data collected for calibration and validation purposes. Fine-scale horizontal and vertical heterogeneity of productive waters (Kutser et al., 2004; Kutser et al., 2008; Kravitz et al., 2020), and the substantial increase in cyanobacteria blooms of inland waters (O'neil et al., 2012) make the collection of high-quality coincident measurements with satellite overpass laborious and error prone. Consequently, trustworthy in-situ data for productive coastal and inland waters is limited compared to combined global datasets for ocean calibration and validation, which critically hinders our capability to execute global baseline studies, as well as identify global trends using archival imagery pertaining to the state of our inland and coastal water resources. It is therefore imperative to develop suitable algorithms for optical constituent retrieval for current and planned missions, with a full understanding of the uncertainties and limitations involved.

With recent increased attention placed on retrieving eutrophication metrics for inland water bodies, numerous studies have appeared attempting radiometric retrieval of chl-a or phycocyanin (PC), the diagnostic pigment within cyanobacteria, with varying degrees of success (see reviews by Ogashawara, 2020; Odermatt et al., 2012; Blondeau-Patissier et al., 2014; Matthews et al., 2011; Gholizadeh et al., 2016). Retrieval of chl-a concentration has been more developed, and generally

more robust for trophic delineation, however, PC is highly specific to cyanobacteria and acts as a better indicator of potential water toxicity (Stumpf et al., 2016). Although, highly resistant cyanobacteria cell walls make laboratory quantification of PC laborious and time-consuming. Numerous methodologies have been proposed for quantifying cellular PC concentrations (Sarada et al., 1999; Stewart et al., 1984; Wyman et al., 1986; Zhu et al., 2007; Viskari et al., 2003), however, due to lack of standardization of field methods, laboratory procedures, and analysis for mixed freshwater phytoplankton assemblages, it is difficult to conduct high impact optical sensitivity studies.

The Lake Bio-optical Measurements and Matchup Data for Remote Sensing (LIMNADES, http://www.limnades.org) is an online collection of global in-situ inland water optical and biogeographical data. The database currently consists of roughly 40,000 various data measurements from thousands of stations around the world. While numbers of coincident radiometric, optical, and biophysical samples are considerably lower, optical analysis and algorithm development for inland waters have already greatly benefited (Spyrakos et al., 2018; Pahlaven et al., 2020). Using the LIMNADES dataset, roughly 2000 co-located and coincident in-situ remote sensing reflectance (R_{rs}) – chl-a pairs were used to train a highly robust global chl-a retrieval model using machine learning (ML) architecture (Pahlaven et al., 2020). Artificial intelligence (AI) and associated Deep Learning architectures substantially benefit from greater volumes of high quality training data. Vastly more coincident reflectance – biophysical parameter pairs, PC in particular, are required to train new and improved multi-parameter inversions for synoptic image analysis at global scales.

Radiative transfer modeling (RTM) has proven instrumental to furthering our understanding of coastal aquatic optical relationships in the form of numerous parameterized case studies (Dall'Olmo et al., 2005;2006; Gilerson et al., 2007;2008, lain et al., 2014;2016, Evers-King et al., 2014). Few,

however, have expanded these analyses to cyanobacteria dominated inland waters (Mathews and Bernard, 2013; Kutser et al., 2006; Kutser, 2004; Metsamma et al., 2006). Others have utilized RTM to develop large synthetic datasets for which to train neural network (NN) retrieval models (Doerffer & Schiller, 2007; Brockmann et al., 2016; Hieronymi et al., 2017; Arabi et al., 2016; Fan et al., 2017). While few of these models such as the OLCI Neural Network Swarm (ONNS, Hieronymi et al., 2017) and Case 2 Regional Coast Color (C2RCC, Brockmann et al., 2016) include samples for extremely absorbing and scattering cases due to global instances of elevated colored dissolved organic matter (CDOM) and non-algal particles (NAP), the phytoplankton component of these models is not optimized for adequate pigment retrieval in optically complex eutrophic inland water (Palmer et al., 2015; Kravitz et al., 2020; Kutser et al., 2018). Fan et al., (2017) and C2RCC utilize chlorophyll-specific phytoplankton absorption (a*phy) measurements straight from the NASA bio-Optical Marine Algorithm Dataset (NOMAD), while ONNS uses five a*phy shapes derived from cluster and derivative analysis of various phytoplankton cultures (Xi et al., 2015). These studies rely heavily on phytoplankton absorption characteristics as the main driver for resulting functional type and biomass related differences in modeled reflectances. This generally can suffice for oligotrophic to mesotrophic water conditions, whereas recent research suggests this decoupling of the phytoplankton absorption and backscattering terms, or using backscattering relating only to gross particulate, is too simplistic for eutrophic conditions and generally underperforms in more productive waters (Lain et al., 2014; Lain et al., 2016).

The fundamental building blocks of aquatic RTM rely on accurate parameterization of the inherent optical properties (IOPs, i.e. absorption and scattering properties) of all light altering constituents in a volume of water. The entire angular structure of the light field can be determined from the absorption coefficient a(λ) and the Volume Scattering Function (VSF) $\beta(\lambda)$ (Mobley et al., 2002).

Normalization of the VSF by the scattering coefficient b(λ) provides the scattering phase function $\beta\hat{}(\lambda)$, critical in fully realizing the underwater light field. Several approximations of $\beta\hat{}(\lambda)$ exist for use in RTM, generally as simple functional forms for mathematical simplicity, or derived from Mie theory, which over-generalizes phytoplankton particles as spherical homogenous structures (Mobley et al., 2002). However, some studies characterizing the backscattering properties of various monospecific cultures have found a prominent deviation from the homogenous sphere model, and a poor simplification of the complex cellular structures found in bloom forming phytoplankton (Vaillancourt et al., 2004; Quirantes and Bernard, 2004; Whitmire et al., 2007; Zhou et al., 2012; Matthews and Bernard, 2013). This is specifically important for productive inland waters where blooms of the potentially toxic cyanobacterium *M. aeruginosa* are becoming more prevalent. Cyanobacteria, *M. aeruginosa* especially, have shown to be extremely efficient backscatterers (Zhou et al., 2012), which has been attributed to internal gas vacuoles associated with cyanobacteria (Matthews and Bernard, 2013). Due to strong effects on attenuation, rather than absorption, vacuolate induced spectral scattering (Walsby, 1994; Ganf et al., 1989) cannot be overlooked when parameterizing RTMs for inland water application. To address these over-simplifications, the Equivalent Algal Populations (EAP) model provides an alternative assemblage-based particle modeling approach, simulating phytoplankton IOPs derived from differences in cell and assemblage size distributions, dominant pigmentation, cell composition, and ultrastructure (Lain et al., 2014; Bernard et al., 2009).

Chl-a fluorescence has the potential to be an important information source of phytoplankton physiology, size, or identification (Behrenfeld et al., 2009; Greene et al., 1992), although to what extent still remains uncertain. While an integral part of phytoplankton physiology, fluorescence is often omitted in RTMs (as in the case of Hieronymi et al., (2017) and Fan et al., (2017)) due to

uncertainty in quantum yield efficiencies, or fluorescence is modeled as a simplistic gaussian term centered at 685 nm (Huot et al., 2007; Gilerson et al., 2007). Complexity of the fluorescence signal increases in mixed cyanobacteria assemblages due to the fact sun induced chlorophyll fluorescence (SICF) is vastly reduced in cyanobacteria, with most chl-a pigments located in the non-fluorescing photosystem I (PSI) (Simis and Huot, 2012). This has yet to be addressed in RTM for eutrophic inland waters, where the red edge and near infrared (NIR) spectral relationships are considered most valuable in aquatic pigment retrieval models.

The objective of this study is to develop a unique, state-of-the-art synthetic dataset of paired R_{rs} and pigment/IOP combinations using a radiative transfer model. This work aims to begin to simulate the immense natural optical variability of inland waters and to address the issues described above. The dataset includes novel physics-based, two-layered, size and type specific phytoplankton IOPs for mixed eukaryotic/cyanobacteria assemblages along with novel calculations of mixed assemblage chl-a fluorescence. Accurate modeling of paired PC concentration also adds immense value to the dataset. The parameterization of the dataset and subsequent evaluation in terms of natural biophysical variability is discussed in conjunction with most current literature values. The dataset will be further used for analysis of optical relationships, and to be used in the training of advanced machine learning architectures, and is available to download as open access.

3.2 Methods: Parameterization of radiative transfer model

In order to be consistent with natural optical relationships, four datasets were compiled based on the domination of a particular optical constituent. The first dataset is modeled as typical Case 1 waters where water and phytoplankton provide the bulk of the optical signal and depicts more oligotrophic conditions. The bio-optical model in this dataset closely follows that of Lee (ed.) in which other optical

constituents co-vary with phytoplankton biomass. The other three datasets resemble that of cyanobacteria dominated inland waters, CDOM dominated waters, and inorganic sediment dominated waters where more complex optical relationships persist and optical constituents do not tend to co-vary (Brewin et al., 2017). A four-component bio-optical model was used to generate the IOPs of these hypothetical inland water cases to be used in the RTM (Gilerson et al., 2007; IOCCG, 2006):

$$a(\lambda) = a_w(\lambda) + a_g(\lambda) + a_{phy}(\lambda) + a_{nap}(\lambda) \quad (3.1)$$

Where $a_w(\lambda)$, $a_g(\lambda)$, $a_{phy}(\lambda)$, and $a_{nap}(\lambda)$ represent the spectral absorptions of water, a combined CDOM/detritus term, phytoplankton, and non-algal particles (NAP), respectively. Refer to Appendix B for a full list of definitions of symbols and units used throughout this manuscript. Except for the Case 1 dataset, which is defined solely on C_{chl} and relationships defining how other constituents co-vary, the other three datasets are defined by independent values of C_{chl}, the concentration of nonalgal particles (C_{nap}), and the absorption of CDOM at 440 nm ($a_g(440)$). Great care was taken to be sure that constituent ranges were appropriate and based on natural populations from the LIMNADES in-situ inland water dataset. A table of mode and standard deviations used for the lognormal distributions within each dataset can be found in Appendix A. To generate synthetic datasets representative of natural waters, values of all constituents were randomly chosen out of the described lognormal distributions. Derivation and equations used in modeling components other than phytoplankton are common to other studies which have parameterized models for Case 2 waters (Gilerson et al., 2007; Bukata et al., 1995; Twardowski et al., 2001) and can be found in Appendix B.

3.2.1 Phytoplankton component

The total spectral phytoplankton component in Eq. 1 is modeled as a product of C_{chl} and the chlorophyll-specific absorption spectrum.

$$a_{phy}(\lambda) = C_{chl} * a^*_{phy}(\lambda) \quad (3.2)$$

Where $a^*_{phy}(\lambda)$ is the spectral specific chlorophyll absorption spectrum in m^2/mg. Phytoplankton chlorophyll specific IOPs (SIOPs) for this work are based on the physics-based two-layered spherical Equivalent Algal Population (EAP) model, where population-specific refractive indices are used to derive IOPs and thus not independent of each other (Lain et al., 2018; Bernard et al., 2009). The two-layered spherical geometry consists of a core sphere, acting as the cytoplasm, and a shell sphere acting as the chloroplast. The EAP model calculates, from first principles, biophysically linked phytoplankton absorption and scattering characteristics from particle refractive indices reflecting the primary light-harvesting pigments of various phytoplankton groups (Lain et al., 2018; Lain et al., 2014). IOPs are calculated at 5nm spectral resolution between 200 and 900 nm and integrated over an entire equivalent size distribution represented by effective diameters (D_{eff}) between 1 μm and 50 μm (Bernard et al., 2007; Lain et al., 2017). For a hypothetical Eukaryotic population, refractive indices are derived from blooms in the Benguela off of Southern Africa which is typically dominated by chlorophyll-a (chl-a) and carotenoid pigments fucoxanthin and peridinin, which are the main light harvesting pigments in diatoms and dinoflagellates, respectively. Considering minimal differences within carotenoid pigment refractive indices and absorption, these two groups were combined into a generalized set of chl-a – carotenoid IOPs (Organelli et al., 2017; Bernard et al., 2009). The model has been consistently validated and can be considered an accurate phytoplankton model for coastal

and inland waters (Lain et al., 2017; Evers-King et al., 2014; Mathews and Bernard, 2013; Smith et al., 2018).

The EAP two-layered sphere model has also been used to derive IOPs for the particularly troublesome cyanobacteria *M. aeruginosa* (Matthews and Bernard, 2013). In this instance, the core layer is assigned to a highly scattering vacuole, while the shell layer acts as the chromatoplasm. *M. aeruginosa* is modeled with and D_{eff} of 5 μm. Derivation of the complex refractive indices, influence of gas vacuolation, and tuning of the two-layered model for cyanobacteria can all be found in Matthews and Bernard (2013). IOPs for the cyanobacteria *Aphanizomenon, Anabaena cirinalis* and non-vacuolate *Nodularia spumigena* which were measured in laboratory were also included in the dataset (Kutser et al., 2006). The final phytoplankton SIOPs used in the RTM can be found in Figure 1.

In order to account for optical variation due to mixed populations, the $a^*_{phy}(\lambda)$ term in Eq. 3.2 is modeled as an admixture of eukaryotic and cyanobacteria SIOPs based on a series of weighting factors. Total $a^*_{phy}(\lambda)$ is therefore calculated as the sum of the cyanobacteria and eukaryotic populations:

$$a^*_{phy}(\lambda) = S_f\,(a^*_{cy}(\lambda)) + (1 - S_f)\,(a^*_{euk}(\lambda)) \quad (3.3)$$

where S_f = [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1.0], a^*_{cy} is the chlorophyll specific absorption of the cyanobacteria population and a^*_{euk} is the chlorophyll specific absorption for the carotenoid containing eukaryotic population. Total scattering and backscattering coefficients of the phytoplankton component ($b_{phy}(\lambda)$ and $b_{bphy}(\lambda)$, respectively) are calculated in a similar manner using EAP derived spectral chlorophyll-specific scattering and backscattering terms (Appendix B).

The admixture weighting factor and input D_{eff} for the eukaryotic population were also randomly varied for the RTM, however, with some constraints. Several studies have shown that for natural populations of oligotrophic to mesotrophic waters, a^*_{euk} tends to decrease with increasing C_{chl} (Bricaud et al., 1995; Babin et al., 1996). This rule is not as strict in more complex inland and coastal waters, but rough relationships have been found (Matthews et al., 2013). Due to the nature of the EAP model, the magnitude of resulting SIOPs are highly dependent on the particle size. Thus, to generalize this natural relationship in our RTM, input phytoplankton SIOPs of the carotenoid containing population were constrained in the model by D_{eff} as: $5 < D_{eff} < 20$ μm for $0 < C_{chl} < 20$ mg/m^3 , $15 < D_{eff} < 35$ μm for $20 < C_{chl} < 50$ mg/m^3 , and $30 < D_{eff} < 45$ μm for $C_{chl} > 50$ mg/m^3. Ranges for appropriate cyanobacteria admixture weighting must also be comparable to natural variations as a function of phytoplankton biomass. Randomization of weighting factors was constrained based on in-situ phytoplankton abundance and biomass collected from South African inland waters between 2016-2018 (Fig. 2). Information regarding collection and analysis of field data can be found in Chapter 2. From the figure, it was roughly assumed that if cyanobacteria are a part of the phytoplankton population, they will tend to dominate at higher biomass (i.e. it is rare to find low fractions of cyanobacteria as C_{chl} rises to extremely hypertrophic levels, if cyanobacteria are present). *M. aeruginosa* has been known to form extremely high biomass blooms with potential to form floating scum mats which can reach C_{chl} upwards of 20,000 mg/m^3 (Matthews et al., 2013). This is reflected in the RTM, however, only *M. aeruginosa* is included above a C_{chl} greater than 500 mg/m^3, as no other data exists confirming blooms of that extent for other species.

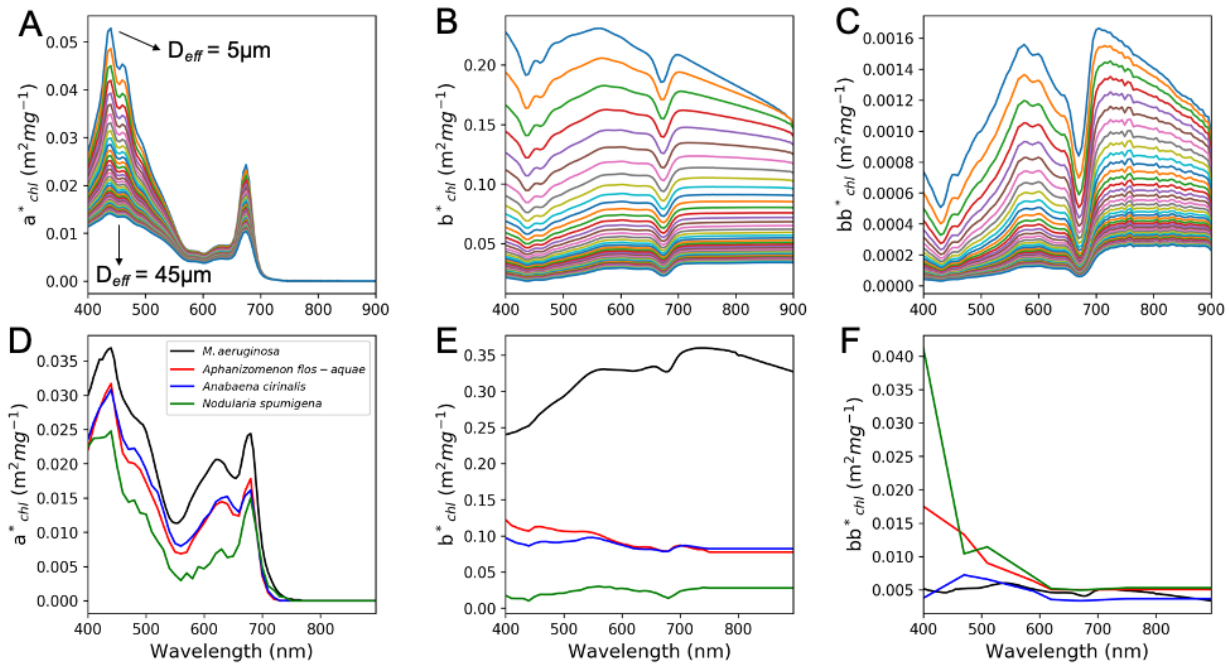

Figure 1: Phytoplankton SIOPs used in the RTM, A), B), C), are eukaryotic specific spectral absorption, scattering, and backscattering, respectively. Different colors indicate increasing effective diameter (D_{eff}) from top spectra to bottom spectra. D), E), F), are cyanobacteria specific spectral absorption, scattering, and backscattering, respectively. Eukaryotic and *M. aeruginosa* spectra modeled from EAP 2-layer code, while Aphanizomenon, Anabaena, and Nodularia taken from Kutser et al., (2006).

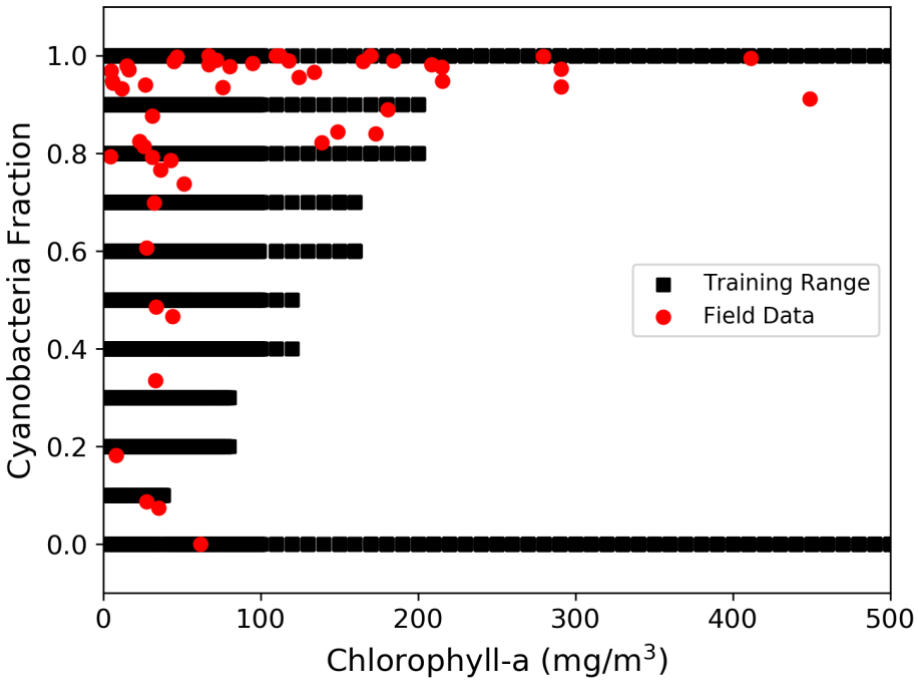

Figure 2: Cyanobacteria abundance fraction as a function of chl-a concentration. Red dots are field measurements from South African inland waters, black lines are constrained ranges used in the RTM.

3.2.2 Chl-a fluorescence

Previous modeling of chl-a fluorescence has been defined as a simple gaussian shape centered around 685 nm with a full width half max (FWHM) of 25 nm. The magnitude of the depth-integrated radiance contribution by chl-a fluorescence at 685 nm has been calculated as (Huot et al., 2005; Huot et al., 2007):

$$L_f(685) = 0.54 L_f^-(685) = 0.54 \frac{1\emptyset_f}{4\pi C_f} Q_a^*[Chl] \int_{400}^{700} \frac{a_{phy}^*(\lambda) E_o^-(\lambda)}{K(\lambda)+K_{Lu}(685)} d\lambda \quad (3.4)$$

Refer to Appendix B for definitions of symbols and units. For natural coastal and cyanobacteria dominated waters this is an oversimplification in two ways. First, this equation is assuming a purely eukaryotic, photosynthetic carotenoid containing phytoplankton assemblage. In other words, it is

assuming that the modeled population contains all of the intracellular chl-a in the fluorescing photosystem II (PSII). Emission spectra of chl-a are a response to photosynthetic pigments that harvest light in PSII. However, cyanobacteria generally only contain roughly 10-20% of total cellular chl-a in PSII, with no accessory chlorophylls or carotenoids, with the remaining cellular chl-a located in non-fluorescing photosystem I (PSI) (Simis and Huot, 2012, Bryant, 1986, Johnsen and Sakshaug, 1996).

Cyanobacteria are comprised of phycobilin pigments (i.e. Phycocyanin (PC)) located in supramolecular phycobilisomes (PBS) which harvest light for PSII (Campbell et al., 1998). While the major chl-a fluorescence emission sits around 685 nm, emission for phycobilin pigments generally resides between 620 – 660 nm (Seppala et al., 2007). The ability to quantify and model phycobilin pigment fluorescence to R_{rs} is an active area of research, but currently not with high enough confidence to include in the current modeling.

The second oversimplification pertains to how the shape of gaussian fluorescence emission is usually modeled:

$$L_f(\lambda) = L_f(685)\exp\left(-4\log(2)\left[\frac{(\lambda-685)}{25}\right]\right) \quad (3.5)$$

Where the model assumes a single gaussian peak centered at 685nm with FWHM of 25 nm. In reality, chl-a fluorescence does indeed have a major fluorescence emission around 685 nm but also an adjacent vibrational satellite emission centered around 730 – 740 nm (Govindjee, 2004 and references therein). Although generally smaller amplitude due to increased absorption from water farther into the near-infrared (NIR), it can potentially contribute to the water leaving radiance.

To overcome these simplifications, firstly, fluorescence amplitude is calculated at both 685 nm and 730 nm. Secondly, the integration of absorbed radiation over the visible spectrum is separated into a carotenoid containing eukaryotic component, in which 100% of the chl-a is assumed to be contained in PSII, and a cyanobacteria component, in which only 15% of chl-a is assumed to be contained in PSII (Eqs. 3.6 and 3.7). Considering the increased attenuation of upwelling radiance in the presence of cyanobacteria due to elevated scattering relative to eukaryotic phytoplankton (Matthews and Bernard, 2013), total attenuation at both 685 nm and 730 nm (C_{tot} (λ)) is considered to more appropriately define loss of the upwelling fluorescence signal.

$$L_f(685) = 0.54L_f^-(685) = 0.54\frac{1\emptyset_f}{4\pi C_f}Q_a^*[Chl]\int_{400}^{700}\frac{(a_{Euk}^*(\lambda)(1-S_f)E_o^-(\lambda)) + (a_{Cy}^*(\lambda)(S_f)(0.15)E_o^-(\lambda))}{K(\lambda)+C_{tot}(685)}d\lambda$$

(3.6)

$$L_f(730) = 0.54L_f^-(730) = 0.54\frac{1\emptyset_f}{4\pi C_f}Q_a^*[Chl]\int_{400}^{700}\frac{(a_{Euk}^*(\lambda)(1-S_f)E_o^-(\lambda)) + (a_{Cy}^*(\lambda)(S_f)(0.15)E_o^-(\lambda))}{K(\lambda)+C_{tot}(730)}d\lambda$$

(3.7)

The actual spectral emission signal is then modeled as a double gaussian function with the 730 nm peak modeled with a FWHM of 50 nm:

$$L_f(\lambda) = L_f(685)\exp\left(-4\log(2)\left[\frac{(\lambda-685)}{25}\right]\right) + L_f(730)\exp\left(-4\log(2)\left[\frac{(\lambda-730)}{50}\right]\right) \quad (3.8)$$

For each individual modeled sample, fluorescence quantum yield (FQY) is randomly chosen between 0.005% - 1% simulating ranges typically found in these waters (Gilerson et al., 2007; Gilerson et al., 2008; Babin et al., 1996; Behrenfeld et al., 2009). The reabsorption coefficient of algal cells (Q_a^*) was randomly chosen between 0.3 – 0.6 (Babin et al., 1996). Figure 3 acts as a visualization of changes in

red/NIR spectral peaks using the described modeling approaches for a C_{chl} of 10 mg/m^3 with an FQY of 1% and similar contributions of non-algal constituents.

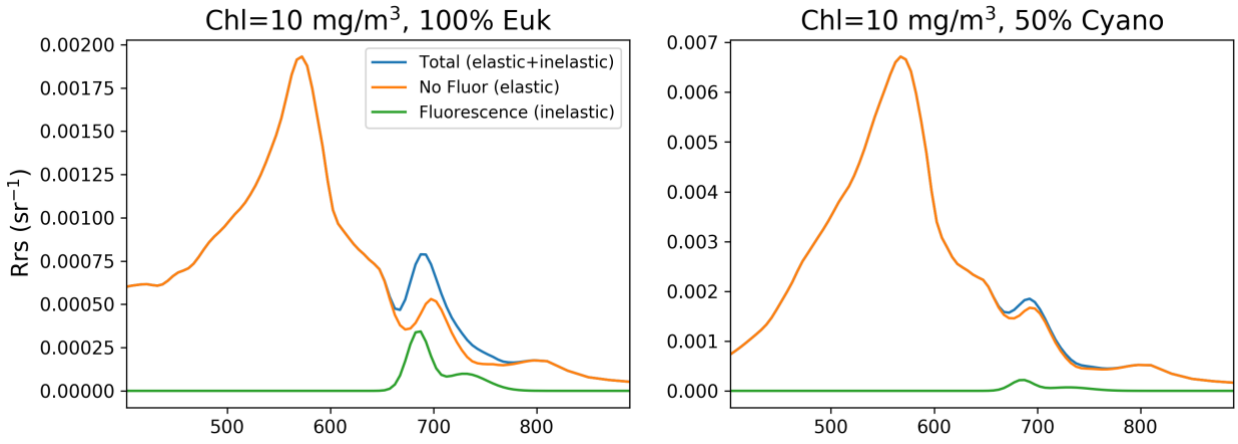

Figure 3: Red/NIR spectral changes with and without modeled fluorescence included for a 100% eukaryotic population and a 50/50 eukaryotic/cyanobacteria population.

3.2.3 Phycocyanin concentration

While the ECOLIGHT radiative transfer code allows you to define C_{chl} as an input to the model, C_{pc} must be modeled independently. The calculation of C_{pc} can be accomplished as follows (Simis et al., 2005):

$$C_{pc} = a_{pc}(620) \,/\, a^*_{pc}(620) \qquad (3.9)$$

Where $a_{pc}(620)$ is the total absorption due to PC at 620 nm, and $a^*_{pc}(620)$ is the specific absorption coefficient of PC at 620 nm. Quantifying the amount of absorption solely due to PC at 620 nm requires removing the effect of all other optical constituents and pigments at 620 nm. This was completed following similar logic as Yacobi et al., (2015) to remove the absorption due to Chl-a and its accessory pigments, Chl-b and Chl-c using calculated pigment ratios from Yacobi et al., (2015) and calculated specific pigment absorption ratios in Bidigare et al., (1990).

$$a_{Chla}(620) = a_{cy}(675) * [a^*_{Chla}(620)/ a^*_{Chla}(675)] \quad (3.10)$$

$$a_{Chlb}(620) = [a_{cy}(620) - a_{Chla}(620)] * [\text{Chl-b/Chl-a}] * [a^*_{Chlb}(620)/ a^*_{Chla}(620)] \quad (3.11)$$

$$a_{Chlc}(620) = [a_{cy}(620) - a_{Chla}(620)] * [\text{Chl-c/Chl-a}] * [a^*_{Chlc}(620)/ a^*_{Chla}(620)] \quad (3.12)$$

Where $a_{cy}(\lambda)$ is the phytoplankton absorption from PC containing cyanobacteria. The [Chl-b/Chl-a] and [Chl-c/Chl-a] ratios used in equations 3.9 and 3.10 are 0.44 and 0.059, respectively, from median values calculated from cyanobacteria dominated waters in Yacobi et al., (2015). The $[a^*_{Chla}(620)/ a^*_{Chla}(675)]$, $[a^*_{Chlb}(620)/ a^*_{Chla}(620)]$, and $[a^*_{Chlc}(620)/ a^*_{Chla}(620)]$ terms were calculated from measured unpackaged specific absorption values in Bidigare et al., (1990) resulting in values of 0.179, 0.64, and 1.14, respectively. The $a_{pc}(620)$ term is then calculated as follows:

$$a_{pc}(620) = a_{cy}(620) - [a_{Chla}(620) + a_{Chlb}(620) + a_{Chlc}(620)] \quad (3.13)$$

While adopting an appropriate standard $a^*_{pc}(620)$ value is still debated (Simis et al., 2005; 2007; Mishra et al., 2013; Matthews et al., 2013; Jupp et al., 1994), all studies generally rely on a fixed $a^*_{pc}(620)$ value for PC estimation models. Considering that $a^*_{pc}(620)$ has the potential to vary by a factor of 60 (see Table 4 in Yacobi et al., 2015) in nature, this is a dramatic oversimplification, especially for lower C_{pc} or when cyanobacteria is not the dominant species. Figure 4a displays PC:Chl-a plotted against $a^*_{pc}(620)$ for data from both Simis et al., (2005; denoted S05) and Matthews et al., (2013, denoted M13). A strong non-linear relationship is apparent (R^2=0.73) and is used in conjunction with each sample's admixture to define a sample specific $a^*_{pc}(620)$.

Considering the common consensus that a PC:Chl-a ratio >=0.5 (mg/m^3) implies a cyanobacteria dominant water target (Simis et al., 2005; Yacobi et al, 2015, Hunter et al., 2010), the admixture and

PC:chl-a does not scale linearly (i.e. an admixture of 0.5 does not equal PC:chl-a of 2, for a range of 0 to 4). Thus, our admixture of 0 - 1 was non-linearly scaled to a PC:Chl-a between 0 – 4, where an admixture of 0.6 (60% dominance by Cyanobacteria in population) is equal to PC:Chl-a of 0.5 (Fig. 4b). The per sample $a^*_{pc}(620)$ was then calculated as:

$$a^*_{pc}(620) = 0.0093(S_{ad})^{-0.717} \quad (3.14)$$

Where S_{ad} is the scaled admixture parameter. Equation 3.7 can then be used to calculate a final PC concentration.

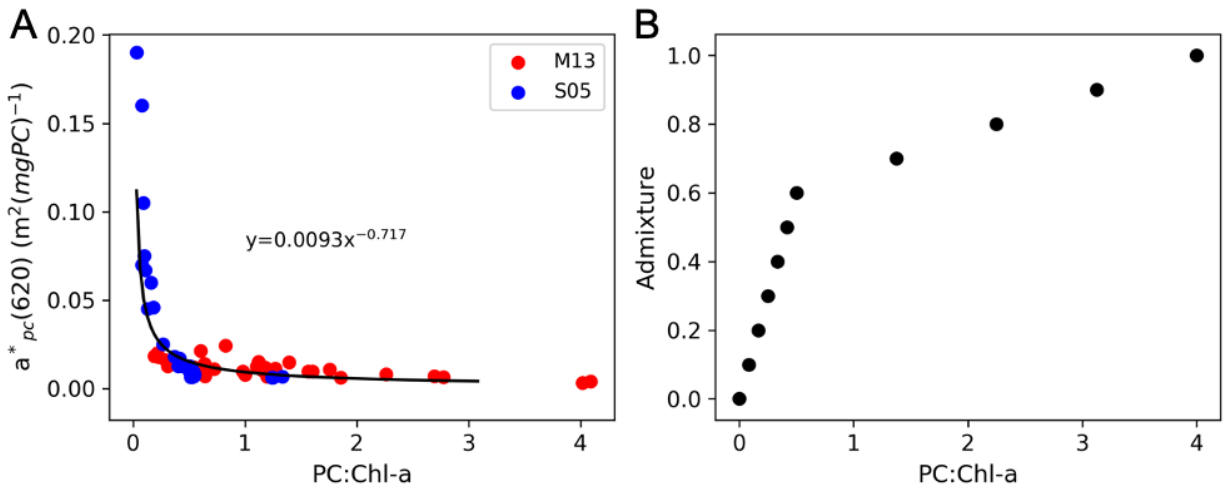

Figure 4: A) $a^*_{pc}(620)$ plotted as a function of PC:chl-a from two different sources along with best fit line in black, B) Visual of admixture scaling where an admixture of 0.6 (60% dominance by Cyanobacteria in population) is equal to PC:Chl-a of 0.5.

3.3 Evaluation of synthetic dataset

3.3.1 Remote sensing reflectance

3.3.1.1 Pre-processing

Roughly 70,000 R_{rs} spectra were modeled with coincident C_{chl}, C_{pc}, and suite of associated IOPs. As a way of dissecting the modeled data and validating it with variables from natural conditions, a clustering procedure was undertaken to identify distinct optical water types (OWTs) within the dataset. Clustering of water types on the basis of optical properties has been a common technique since the 1970's as method to direct the application of Earth observation (EO) for aquatic purposes (Prieur and Sathyendranath 1981; Lubac and Loisel, 2007; Vantrepotte et al., 2012; Moore et al., 2001; 2009; 2014; Spyrakos et al., 2018). Clustering of optical data has historically been beneficial for demonstrating underlying bio-optical relationships and variability and guiding the development and application of retrieval models. In order to be consistent with previous clustering applications in coastal and inland waters, the functional data analysis (FDA) approach of Spyrakos et al., (2018) is closely followed, and briefly discussed here. A full analysis of historical clustering techniques is beyond the scope of this paper, and readers are directed to Spyrakos et al., (2018), and references therein, for a more comprehensive overview of clustering approaches. A comprehensive guide to FDA can also be found in Ramsay (2006).

Prior to clustering, all R_{rs} spectra were normalized by their respective integrals, as a way to standardize amplitude variation attributed to concentrations of optically significant constituents. Each spectrum was deconvolved into 26 cubic basis functions (Fig. 6), of which a linear combination results in a smoothed R_{rs} Spectra. The same *B*-spline representation was used here as in Spyrakos et al., (2018) except with the inclusion of one extra knot in the 800-900 nm region. The actual clustering by *k*-means is then performed on the 26 basis coefficients from the cubic functions. This acts as a

method of dimensional reduction which removes excessive local variability, reduces noise in the data, keeps independence among variables, and allows for a customizable smoothing approach through number and placement of knots.

Various methods exist for determining the optimal number of clusters to be used in a clustering analysis. It can be heavily debated which technique is most appropriate for different data types and results can be quite subjective. Thus, there is no definitive solution to obtain the most accurate number of clusters for a dataset. Certain statistical testing methods such as the gap statistic (Tibshirani et al., 2001) are quite common and used in Spyrakos et al., (2018), however, are computationally expensive when dealing with large datasets. Other more direct methods of determination by optimizing a certain criterion such as sum of squares are generally more subjective but require vastly less processing time. These include the elbow and silhouette methods. Multiple approaches were attempted with our modeled dataset, however, no consistent number of clusters was obtained and results generally varied between 9 – 20 clusters. Spyrakos et al., (2018) defined 13 inland OWTs using the gap statistic, while 21 clusters defined the combined inland and coastal dataset. Subsequently, we set initial optimal number of clusters to be 17, to be followed by a manual inspection.

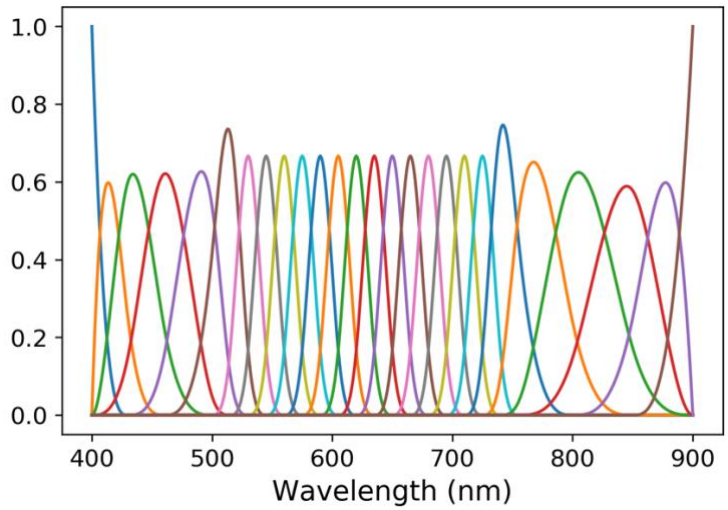

Fig. 5: The 26 *B*-spline basis functions used for curve fitting.

3.3.1.2 Clustering analysis

k-means was used to cluster the dataset of basis coefficients into 17 distinct OWTs. Further inspection of clusters was performed by interpretation of functional box plots of each cluster (Sun and Genton, 2011). Functional box plots are similar to the classical box plot, however, each observation is interpreted as a real function. Ordering of each function is defined by band depth, a metric determining the centrality of each curve. A smaller band depth rank defines curves which are more central to each cluster, with the smallest rank identifying the median curve. Similar to classical box plots, band depths are used to define the central 50% of the data (i.e. interquartile range (IQR)). The maximum outlying envelope is generally defined as 1.5*IQR, however, a factor of 3 was used here to include more extreme cases. Curves are defined as outliers based on proportion of curve which resides outside the 3*IQR range.

After inspection of the 17 functional box plots, two clusters which identified unrealistic reflectances were removed and two different pairs of clusters were combined due to no statistical difference being

found (data not shown), resulting in 13 distinct clusters (Fig. 6). Median curves as defined by the most central function are located in Figure 7. It is important to note that the aim of this paper was not necessarily to define and produce the most optimal OWTs for inland waters. The clustering analysis in this research was used as method to inspect the results of the radiative transfer modeling as a means to provide substance for discussion on the modeling approach. The following sections will include further inspection of each water type and discussion of the model and steps moving forward.

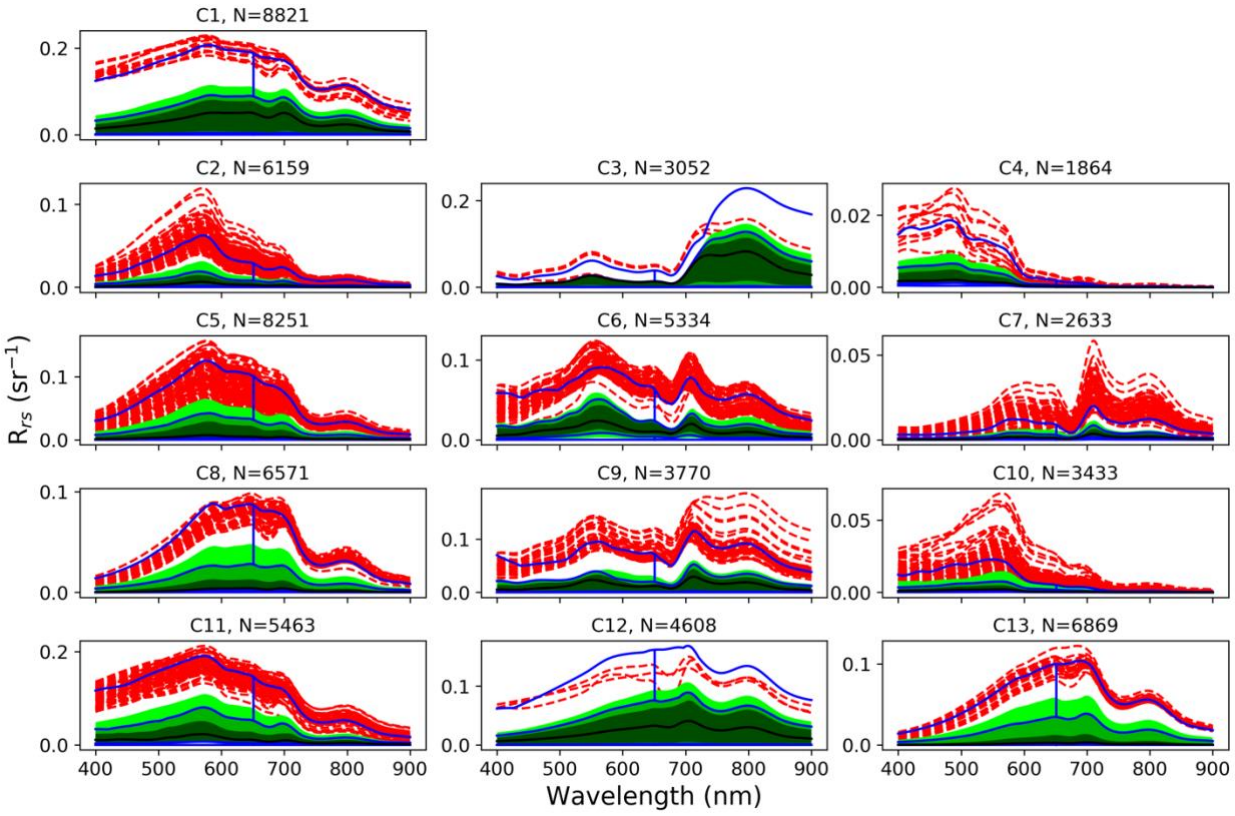

Fig. 6: Functional box plots of the final 13 defined OWTs using synthetic R_{rs} spectra. Dark green shaded regions represent the central 25% of spectra based on calculated band depth, medium green represents the middle 50%, light green the middle 75%. Red spectra are classified as outliers.

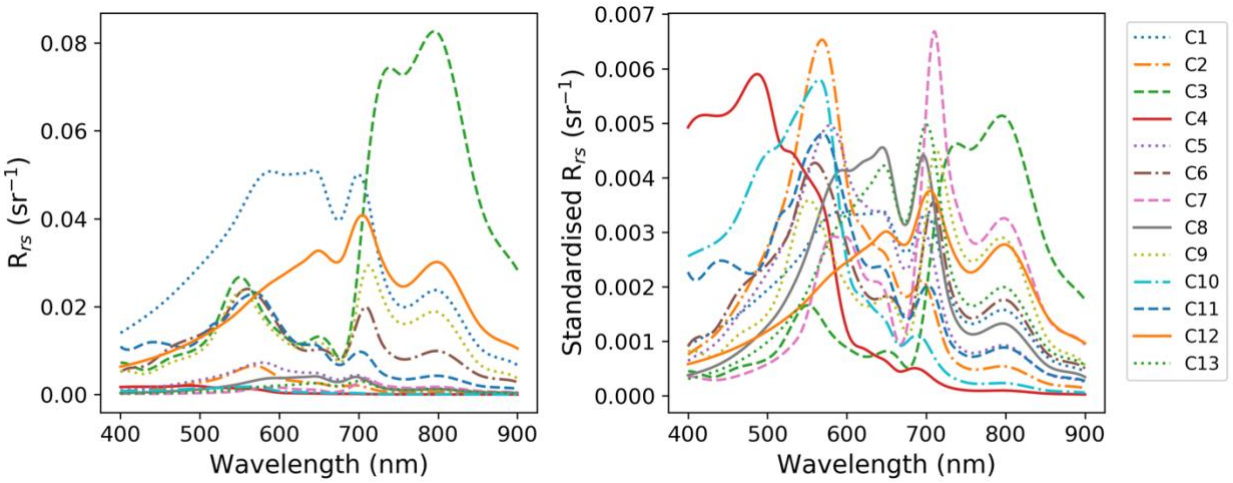

Fig. 7: The median synthetic R_{rs} spectra for each cluster defined by band depth using functional data analysis (left), as well as standardized (right).

3.3.1.3 Cluster descriptions

The clustering analysis produced 13 median reflectances with visibly distinct shapes and widely varying IOPs. Figs. 8, 9, and 10 display box plots for various concentration and IOP parameters for each cluster, while Fig. 11 displays a ternary plot of percent of total absorption at 440 nm for phytoplankton, CDOM, and NAP. The 13 distinct water types were derived from four datasets of modeled R_{rs} based on IOPs from clear waters, high biomass cyanobacteria dominated waters, NAP dominated waters with associated extreme cases, and CDOM dominated waters with associated extreme cases. Although the dataset was modeled to reflect natural conditions of inland water ecosystems, there are certainly cases which could be attributed to coastal marine ecosystems.

Cluster 1 contains roughly equal contribution of phytoplankton and CDOM absorption at 440 nm, however, with a more dominant contribution from NAP. Cluster 1 is not as extreme in terms of NAP concentration as Cluster 12 (98±118 versus 220±198 mg/m^3), but would still be considered an extreme case. Value ranges identified here are derived from log-normal distributions. Pigment

concentrations are moderate but can vary quite dramatically (Chl-a = 90±311, PC = 110±960) and cyanobacteria dominance is usually quite low (PC:Chl-a = 0.19±0.57).

Cluster 2 represents mesotrophic to eutrophic waters (Chl-a = 19.9±45) with very low NAP contribution. Absorption at 440 nm is also slightly more dominated by CDOM. PC:Chl-a ratios tend to be a bit higher with the potential to be greater than 0.5 (0.45±0.72). A smaller and variable reflectance peak is apparent in the red/NIR which can shift depending on biomass and cyanobacteria contribution. This class is also more dominated by the cyanobacteria Anabaena than other cyanobacteria species.

Cluster 3 represents extremely high biomass waters dominated by cyanobacteria and includes hyper-scum situations. This water type is almost 100% dominated by phytoplankton and reflectances almost resemble that of dry vegetation with a dramatic increase in reflectance in the NIR. Some variable NAP and CDOM can persist, however, reflectance shape is almost purely defined by the high cyanobacteria biomass. Extremely high C_{pc} (4008±4730 mg/m^3) induces a rough peak around 650 nm, however, the blue-green spectral region is overshadowed by the extremely high scattering in the NIR. PC:Chl-a of 2.92±0.64 indicates a complete dominance of specifically Microcystis.

Cluster 4 represents more clear, oligotrophic type waters with low biomass, and low presence of cyanobacteria, if any. The small and variable presence of NAP tends to contribute more to the backscattering signal than phytoplankton. There is also a small contribution from CDOM. These reflectances depict your general Case 1 water type signals with high reflectance in the blue, and very low reflectance in the red/NIR due to minimal scattering from such low particle loads, and domination of water absorption. Although low in biomass, the variable presence of CDOM and Cnap of inland

waters make this cluster still slightly more turbid than your average oligotrophic ocean case. Average D_{eff} of eukaryotes were a bit smaller with 16.3±5.3 μm.

Cluster 5 depicts a common water type for inland waters comprising of moderate contributions of phytoplankton and CDOM with a smaller contribution from NAP. Three pigment induced relfectances peaks are generally visible in less extreme cases. The green and NIR reflectance peaks are caused by strong absorption of CDOM and chl-a in the blue, and a smaller peak around 650 nm caused by the combined absorption due to PC at 620 nm and chl-a at 675 nm. PC:Chl-a ratios tend to mostly be below 0.5, thus these waters are not generally dominated by cyanobacteria, but are usually present in some quantity. Average C_{chl} and C_{pc} were similar with 46.4±154 and 58±464 mg/m^3, respectively, while C_{pc} had a much higher magnitude of variation.

Cluster 6 denotes inland waters which are generally dominated by cyanobacteria (PC:Chl-a= 0.96±0.8) with moderately high biomass but has not reached extreme conditions as in Cluster 3, but more productive than Cluster 5. A strong scattering induced reflectance peak is usually apparent around 710 nm. Lower proportions of CDOM and NAP make this cluster dominated by phytoplankton absorption and scattering properties. Large eukaryotic populations generally persist with mean D_{eff} 35±7.2 μm.

Cluster 7 represents a high biomass eukaryotic bloom with little to no contribution from cyanobacteria (mean PC:Chl-a = 0.12±.5). A strong NIR reflectance peak is usually apparent, but the absence of PC vastly reduces the peak seen around 650 nm in cyanobacteria dominated waters. D_{eff} for these waters are generally quite high (35.8±6 μm). This cluster can typically be found in more productive coastal regions such as the Benguela off the coast of Southern Africa (Lain et al., 2014).

Cluster 8 is similar to Cluster 13, however, with slightly less CDOM absorption at 440 nm (6.1±4.2 m^{-1}) and slightly less contribution of NAP. Cluster 8 also includes slightly smaller eukaryotes on average, thus leading to slightly higher phytoplankton absorption at 440 nm. Due to these differences, a stronger green peak can be apparent.

Cluster 9 depicts a common high biomass cyanobacteria water type with pigment concentrations generally in the hundreds of mg/m^3. Average a_{pc}^{*}(620) below 0.01 also result in generally higher PC:Chl-a ratios of 1.5±1.1 signifying consistent dominance by cyanobacteria. While median reflectance shows similar shape to that of Cluster 6, the red/NIR reflectance peak is much more pronounced and higher in magnitude than the green reflectance peak due to increased algal backscattering.

Cluster 10 depicts oligo – mesotrophic waters with relatively higher contributions from CDOM. Due to the competing nature of the chl-a fluorescence peak at 685 nm from the eukaryotic population and the strong scattering reflectance peak in the NIR from cyanobacteria, a small and variable red/NIR peak is observed. The nature of this peak is very dynamic depending on biomass and contribution of non-fluorescing cyanobacteria (Gilerson et al., 2007). A strong peak in the green is the product the high CDOM absorption in the blue and strong absorption from water in the red/NIR. Cluster 8 contained a eukaryotic population with a mean D_{eff} of 14.5±5.7 μm.

Cluster 11 is another common case similar to Cluster 5, however, with less contribution from CDOM, creating roughly equal contribution with phytoplankton absorption. The main difference with this cluster is that the cyanobacteria population is more dominated by Nodularia and Aphanizomenon,

while other clusters are generally quite equally distributed over the four cyanobacteria species with a few (i.e. Cluster 3) more dominated by Microcystis. This creates slight differences in reflectance shape near the 620 nm PC absorption peak and visible differences in the blue in conjunction with lower CDOM.

Cluster 12 is almost completely dominated by NAP, contributing roughly 75% of light absorption at 440 nm and variable smaller fractions of absorption due to phytoplankton and CDOM. Although the absorption in the blue is dominated by NAP, C_{chl} were still moderately high with a mean of 74$\pm$188 mg/m^3. Average C_{pc} were much lower, although again were much more variable (19$\pm$344 mg/m^3). In less extreme cases where scattering from the NAP signal is not as strong, the phytoplankton reflectance peak can be visible around 710 nm, although with increasing NAP dominance (mean Cnap: 212$\pm$198 mg/m^3), this peak becomes less defined.

Cluster 13 is strongly dominated by CDOM absorption at 440 nm with values averaging 10.1$\pm$6.4 m^{-1}. Moderately high Cnap (27$\pm$68.3 mg/m^3) also elevates the signal due to high scattering from NAP, while phytoplankton optical properties are least dominant with low C_{pc}. Moderate C_{chl} of 37.6$\pm$50 mg/m^3 still triggers the red/NIR reflectance peak to appear, which ends up being slightly more pronounced due to high CDOM absorption at shorter wavelengths.

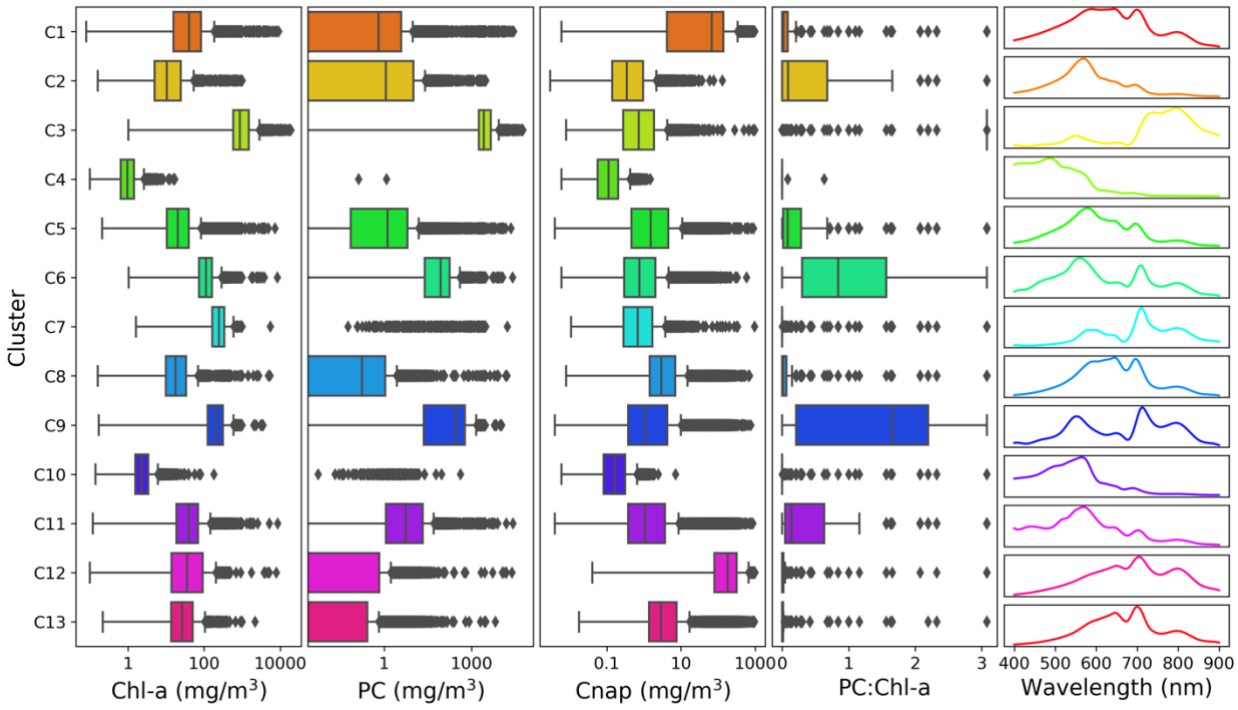

Fig. 8: Ranges of C_{chl}, C_{pc}, C_{nap}, and PC:chl-a for each defined synthetic cluster along with median R_{rs} spectra.

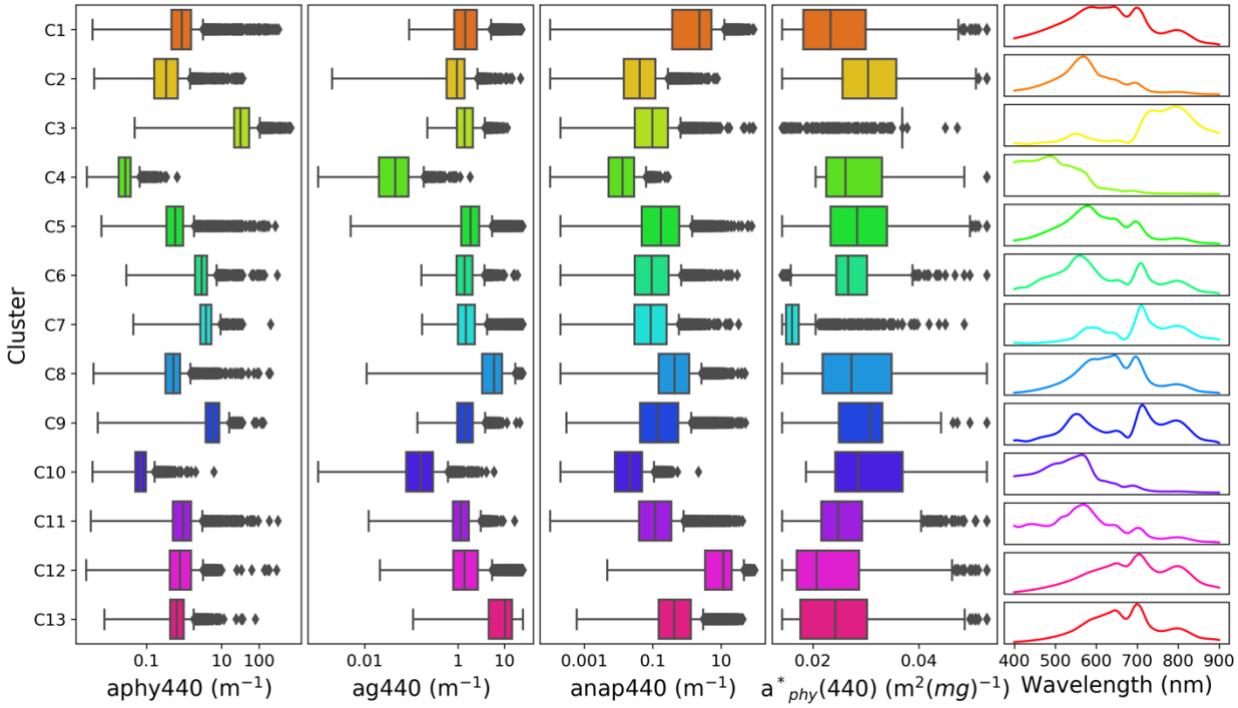

Fig 9: As Fig. 8, but with $a_{phy}(440)$, $a_g(440)$, $a_{nap}(440)$, and $a^*_{phy}(440)$.

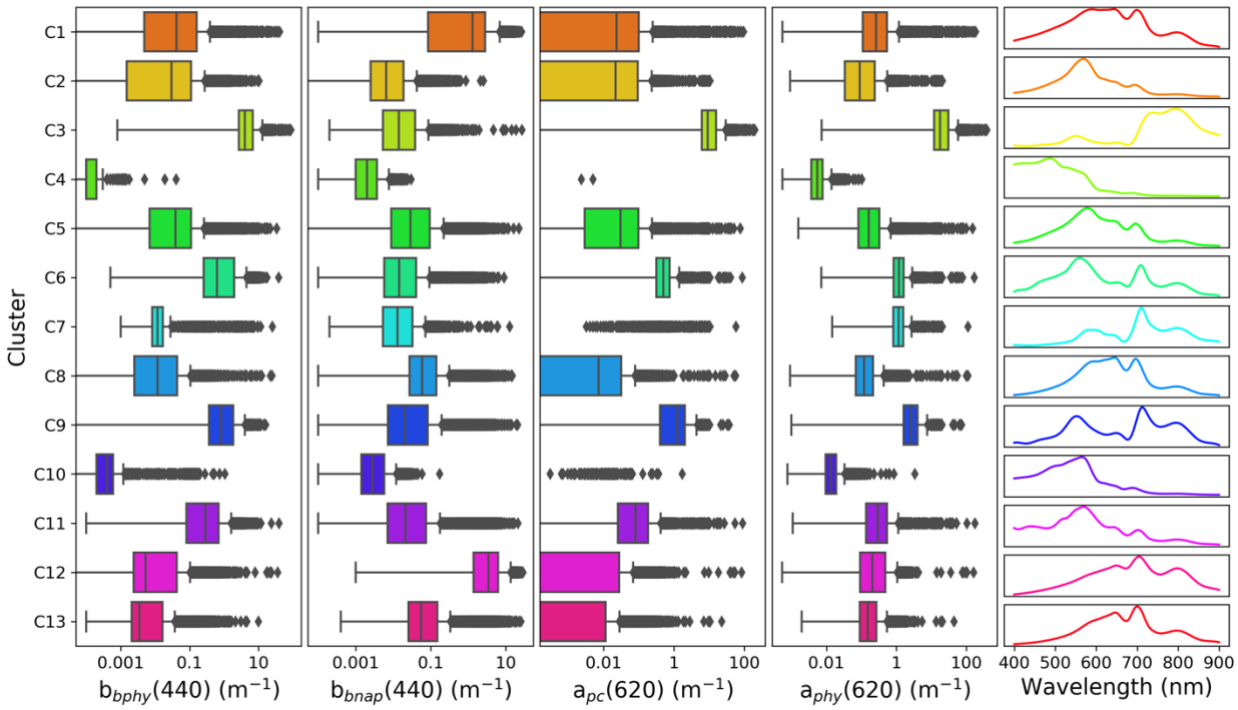

Fig. 10: As Fig. 8, but with $b_{bphy}(440)$, $b_{nap}(440)$, $a_{pc}(620)$, and $a_{phy}(620)$.

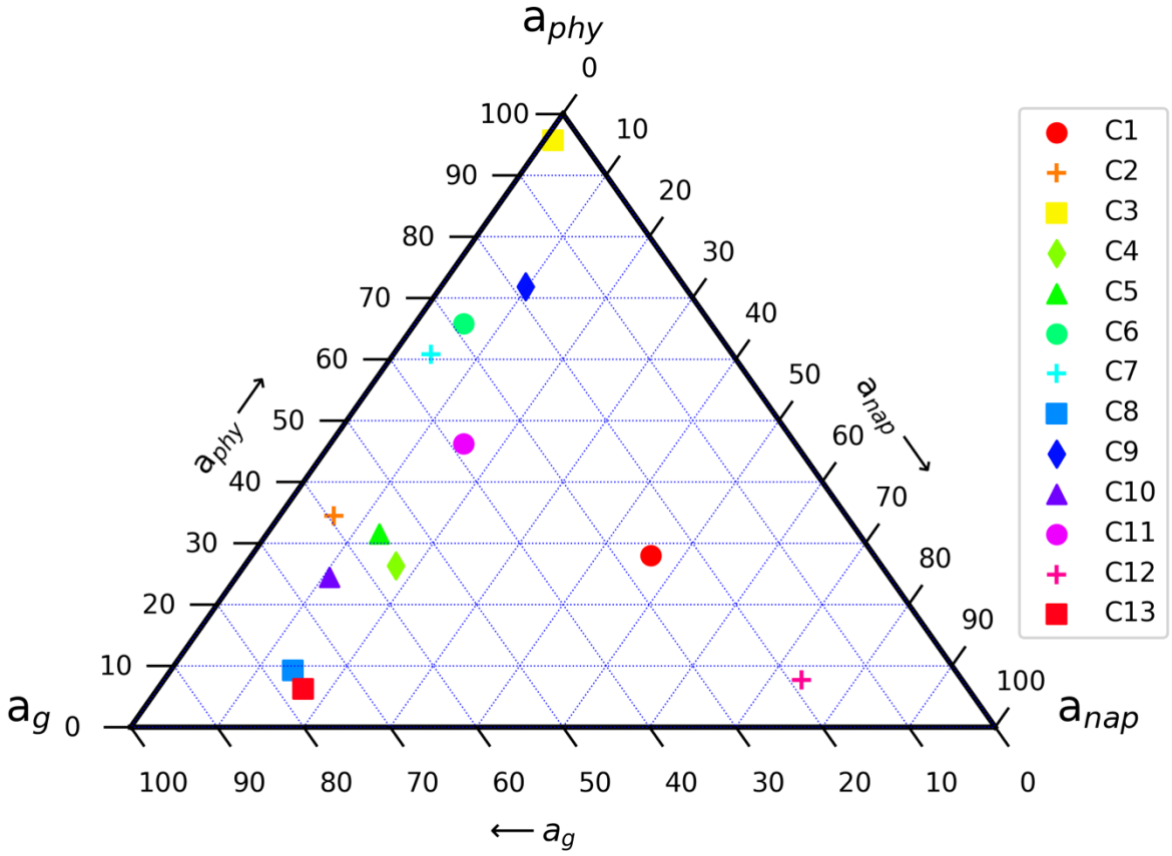

Fig. 11: Percent contribution of absorption at 440nm for the 13 defined clusters.

3.3.2 Assessment of constrained phytoplankton biomass

Little information exists in the literature regarding isolated SIOPs for eukaryotes in mixed cyanobacteria assemblages. Combined $a^*_{phy}(440)$ values for three eutrophic reservoirs in South Africa (M13, Matthews et al., 2013) are shown plotted against C_{chla} along with results from the RTM in Fig. 12. The data from M13 are from an extremely hypertrophic cyanobacteria dominated reservoir (squares), a reservoir of mixed assemblage (circles), and a eukaryotic dominated reservoir (triangles). Previous models, more suited to oligotrophic and mesotrophic waters, generally have a heavy dependence on cell size for $a^*_{phy}(\lambda)$ which is reasonable for low biomass waters with chl-a <10 mg/m^3, but for high biomass waters, this relationship is not as robust (Chrichton et al., 2013).

Although not strong, there is a significant power law relationship between the mixed and eukaryotic assemblage reservoir's $a^*_{phy}(440)$ and C_{chla}. The majority of variability is roughly below chl-a of 50 mg/m^3, while above this, $a^*_{phy}(440)$ stays relatively stable. This follows common relationships of oligotrophic waters for which cell size generally increases with increasing biomass, subsequently decreasing a^*_{phy}. The overwhelming presence of *M. aeruginosa* in cyanobacteria dominated waters poses an exception to this relationship. Microcystis is a relatively small cell (~5 μm), and thus with increasing dominance by *M. aeruginosa*, it follows that a^*_{phy} would be consequently increased. This can theoretically be visualized in Fig. 12 where changes in color indicate change in cyanobacteria admixture. This is also somewhat validated by the plotted in-situ data, in which elevated cyanobacteria contributions (as indicated by black markers), generally constitute a higher a^*_{phy} value at a given C_{chl}. The a^*_{phy} of the mixed and eukaryotic assemblages in M13 are a bit lower at higher C_{chl}

than what was used the RTM, most likely signaling flattening of blue absorption peak due to extensive pigment packaging effects at higher biomass (Morel and Bricaud, 1981; Bricaud et al., 1995).

The synthetic a^*_{phy} are derived from the two-layered EAP model which includes spectrally variable phase functions which can contribute appreciable effects on R_{rs} with smaller cell sizes or increased biomass, when scattering contributes more to the water leaving signal (Lain et al., 2018). The synthetic a^*_{euk} are also derived assuming purely a carotenoid containing (primarily fucoxanthin and peridinin) diatom/dinoflagellate mixture. This is most certainly an oversimplification of the variability in $a^*_{euk}\ (\lambda)$ shape of global natural waters, especially in non-bloom waters, and the EAP model should be used to define IOPs of other commonly occurring groups to be incorporated into the model. Although, the synthetic dataset should still be considered a major step forward for modeling of inland waters, and should encompass most 1st order high biomass IOP variability (Lain et al., 2018).

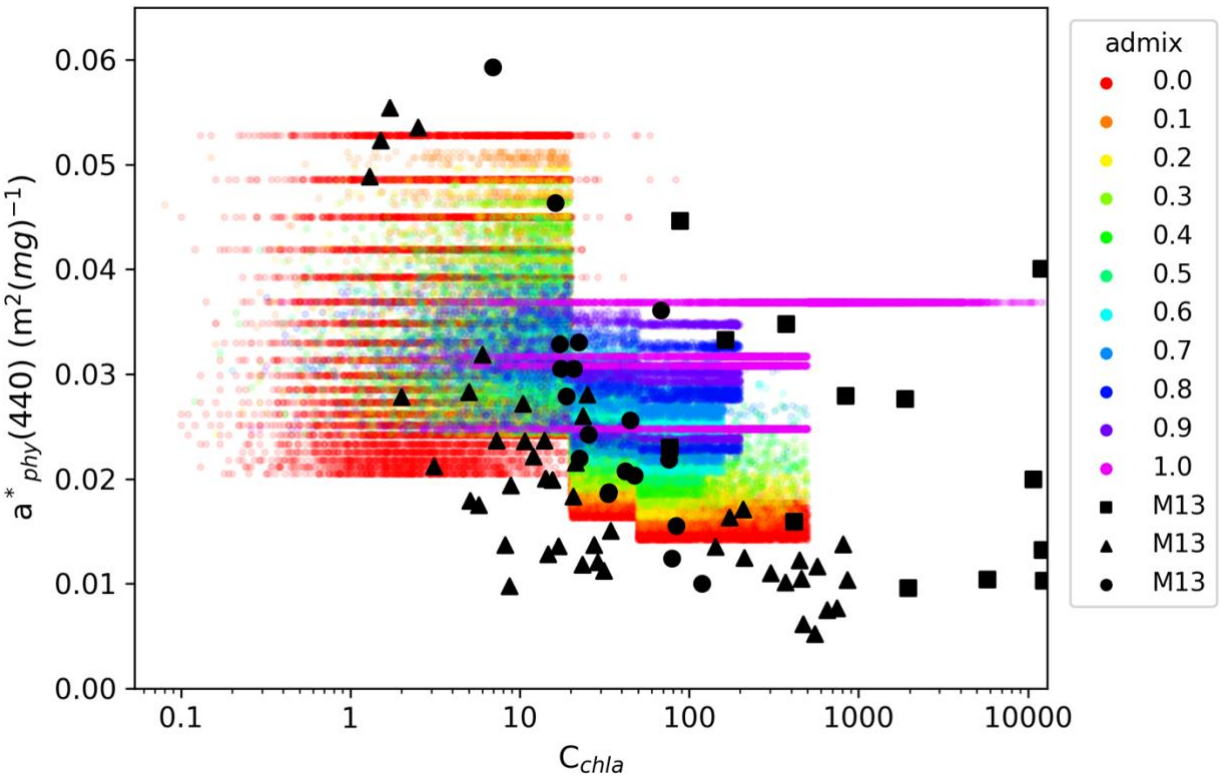

Fig. 12: $a^*_{phy}(440)$ plotted as a function of chl-a concentration for the synthetic data, with colors signifying cyanobacteria admixture. Black points are in-situ values from Matthews et al., (2013) where squares = hypertrophic cyanobacteria dominated waters, circles = mixed assemblage waters, and triangles = eukaryotic assemblage waters.

3.3.3 Assessment of phytoplankton fluorescence in the RTM

As a method to qualitatively validate our modeled fluorescence amplitudes, they were compared with previously validated simplified fluorescence equations for calculating fluorescence amplitudes in low non algal particle concentration (0 < Cnap < 1 mg/m^3), medium concentration (1 < Cnap < 10 mg/m^3), and high concentration (Cnap > 10 mg/m^3) (Fig. 13) (Gilerson et al. 2007, Gilerson et al., 2008; Mishra et al., Eds. 2017). The figure also depicts a 50/50 eukaryotic/cyanobacteria population and a 100% cyanobacteria population. The validated equations (black lines in Fig. 13) are built assuming a 1% FQY and compared with the portion of our data equaling 1% FQY (colored dots), with lower FQYs plotted in the background (faded black dots). The close relationships for the 100% eukaryotic population gives us confidence in our fluorescence calculations. Following the theoretical considerations described in section 2.2 for mixed assemblage waters, the decrease in fluorescence amplitude with increasing cyanobacteria fraction can also be visualized in Fig. 13.

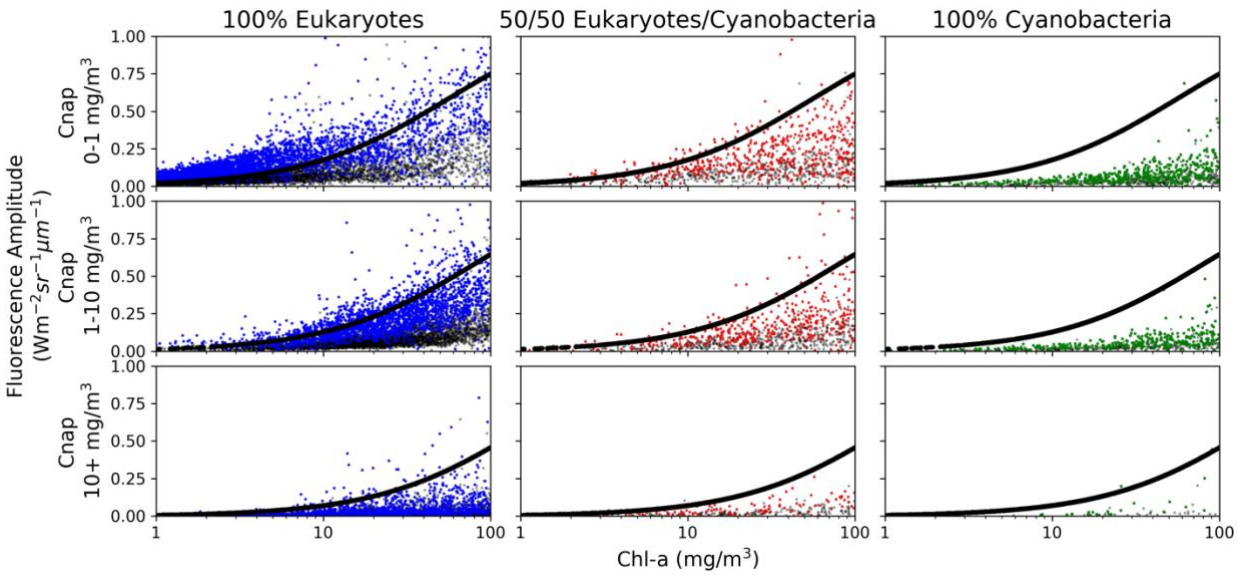

Figure 13: Fluorescence amplitude calculated in radiance units for synthetically derived spectra. Columns represent different phytoplankton assemblage, while rows signify increasing NAP concentrations. Black lines are fluorescence amplitude calculated from simplified models defined in Gilerson et al., (2007) for varying C_{nap} ranges.

While variability in fluorescence amplitude for oligotrophic marine waters can mostly be attributed to variations in high FQY variability (Roesler and Perry, 1995; Fisher and Kronfeld 1990), variability in coastal and inland waters can be more attributed to stronger absorption from CDOM and NAP (Gilerson et al., 2007; 2008, Huot et al., 2013). Gilerson et al., (2008) proposed that variability of FQY for more turbid, productive waters is potentially much smaller than previous literature estimates, and suggest that an FQY of 1% is very good estimate for these waters, and an FQY of 2% would begin to produce unrealistic reflectances. For our dataset, FQY was varied between 0.005-1% and should encompass natural variability for various light and pigment packaging conditions. It is evident in Fig. 13 how the contribution of cyanobacteria to the population can drastically alter resulting fluorescence amplitudes at the chl-a emission peak near 685 nm. Seppälä et al., (2007) showed how

C_{chla} was more related to PC fluorescence than it was to chl-a fluorescence in Baltic Sea waters where cyanobacteria persist, and chl-a fluorescence was lower in areas of higher PC. This agrees well with Fig. 13 where chl-a fluorescence decreases as cyanobacteria contribution increases. Employing a standard factor of 15% of total cellular chl-a contained in cyanobacteria PSII may be an oversimplification, as this value most certainly will vary between species and environmental conditions, however, the assumption seems reasonable enough until further laboratory analysis can be conducted.

No previous modeling work has incorporated the smaller, satellite chl-a fluorescence emission peak near 730 nm. The amplitude is quite low, and would most likely only be visible by a satellite with very high signal-to-noise-ratio (SNR) and in more oligotrophic conditions. For inland waters, where high turbidity and potential for extremely high biomass persist, even the primary fluorescence emission peak will rarely be visible, or reliable to obtain concentration estimates, however, could be useful when developing advanced AI approaches which utilize the entire spectrum to estimate relevant biogeochemical variables.

3.3.4 Assessment of modeled PC concentrations

Figure 14 depicts relationships between $a^*_{PC}(620)$, PC:Chl-a, C_{pc}, and C_{chl} for the four cyanobacteria species included in our modeled dataset. The figure depicts similar relationships found in other literature of natural values (Simis et al., 2005; 2007, Yacobi et al, 2015, Simis and Kauko, 2012, Matthews et al., 2013; Li et al, 2015; Mishra et al., 2013), and we can conclude that our modeled C_{pc} should be reasonable starting point until further concrete relationships can be found.

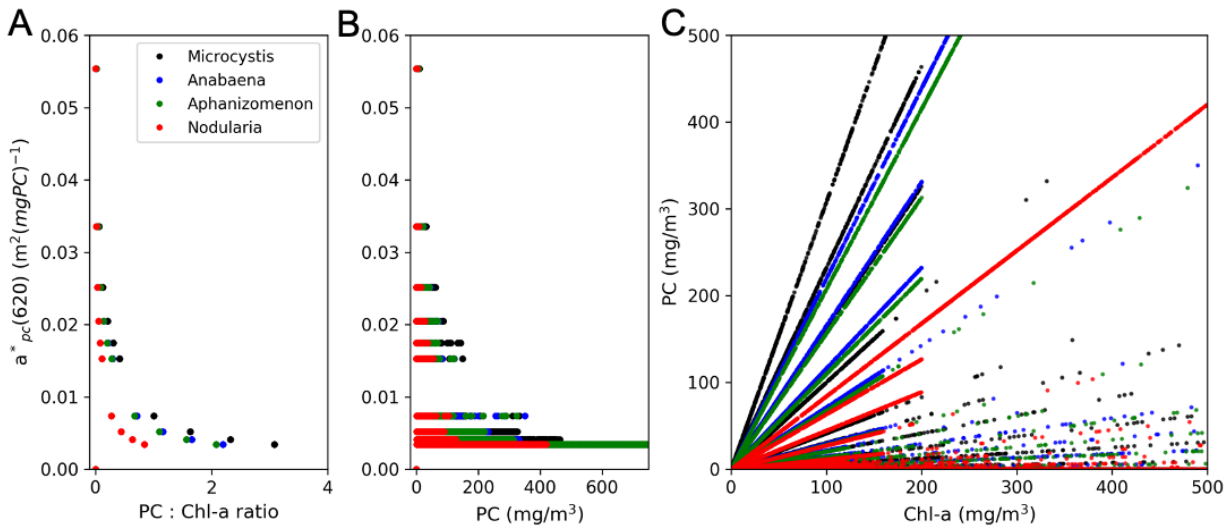

Figure 14: A) $a^*_{PC}(620)$ plotted as a function of PC:chl-a for modeled synthetic data, B) $a^*_{PC}(620)$ plotted as a function of PC concentration, C) PC plotted as function of chl-a concentration.

Traditionally, models which estimate PC concentration from radiometric data employ a fixed $a^*_{PC}(620)$, a parameter critical to the calculation of the C_{pc}. Simis et al., (2005) used an average value of 0.0095 $m^2(mg\ PC)^{-1}$ calculated from their in situ data while Matthews et al., (2013) found mean $a^*_{PC}(620)$ between various inland water bodies to range between 0.0072 and 0.0122. These methods only correct for absorption at 620 nm due to chl-a, and not other accessory pigments, thus more recent literature suggests that at low PC concentrations (<50 mg/m^3), estimated $a_{pc}(620)$ is not fully corrected for other pigment or constituent absorptions, resulting in overestimated C_{pc} (Simis et al., 2007; Yacobi et al., 2015). Yacobi et al., (2015) found that with C_{pc} > 10 mg/m^3, $a^*_{PC}(620)$ tended to hover around .007 $m^2(mg\ PC)^{-1}$, however noted that this value it still potentially too high. Other more recent literature suggests $a^*_{PC}(620)$ values between 0.004-0.005 $m^2(mg\ PC)^{-1}$ (Li et al., 2015; Mishra et al., 2013; Simis and Kauko, 2012; Li et al., 2012; Jupp et al., 1995). These values fall more in line with our modeled values which resulted in a mean and median $a^*_{PC}(620)$ of 0.013±0.017 and 0.0041, respectively, indicating that our calculated $a^*_{PC}(620)$ are within reasonable values. Above a modeled

PC concentration of 50 mg/m^3, mean and median $a_{PC}^*(620)$ stabilized to 0.0042±0.002 and 0.0034, respectively. The majority of variability in modeled $a_{PC}^*(620)$ was found below a PC:Chl-a of 0.5, or a C_{pc} < 50 mg/m^3, aligning with findings from the literature (Yacobi et al., 2015, Mishra et al., 2013). Resulting PC:Chl-a of the modeled synthetic data ranged between 0-3 mg/m^3. $a_{pc}(620)$ was calculated from the absorption due to the PC containing phytoplankton population only, thus there should be no interference from NAP or CDOM absorption. Absorption at 620 nm due to Chl-a and its accessory pigments are only roughly accounted for here using proxies based on measured literature values, however, it should estimate their contribution reasonably enough until further laboratory measurements can be made.

While information regarding the source of variability of $a_{PC}^*(620)$ in nature has not been clearly defined, we can assume that 1st order variation can be a result of variable algal/cyanobacteria composition and biomass effects. Thus, varying $a_{PC}^*(620)$ based on cyanobacteria dominance according to the admixture for each sample should be a reasonable assumption. In the literature, a combination of employing a fixed $a_{PC}^*(620)$ with generally overestimated $a_{pc}(620)$ results in a dramatic increase in error of PC retrieval when PC:Chl-a < 0.5 (Li et al., 2015; Simis et al., 2005; Randolph et al., 2008; Yacobi et al., 2015; Hunter et al., 2010), or when C_{pc} < 50 mg/m^3 (Ruiz-Verdu et al., 2008; Simis et al., 2005; Yacobi et al., 2015). By employing a model which estimates a variable $a_{PC}^*(620)$ based on cyanobacteria dominance, more appropriate values can be attributed to situations at lower PC concentrations. We assumed that a PC:Chl-a of 0.5 is the tipping point to when cyanobacteria become dominant in a population, thus scaling the admixture of our samples in a way which maximizes $a_{PC}^*(620)$ variability below PC:Chl-a = 0.5, or 60% dominance of cyanobacteria based on admixture. This is obviously assuming that PC:Chl-a is fully dependent on cyanobacteria

dominance which is not the case in natural conditions (Bryant, 1981; Grossman et al., 1993). However, as a 1st order level of variability, the resulting C_{pc} should be reasonable to natural values.

In Figure 14, it is evident how changes in cyanobacteria species can have dramatic effects on PC absorption and resulting PC concentrations. When comparing Nodularia and Microcystis, Nodularia has a much smaller specific absorption at 620 nm (Fig.1). Thus when modeling $C_{pc,}$ absorption due to PC at 620 nm will be much smaller, leading to much lower PC:Chl-a ratios. No current literature exists examining changes in PC:Chl-a for various cyanobacteria species, but the results here seem logical following radiative transfer theory.

3.4 Summary and Conclusions

A state-of-the-art synthetic dataset of R_{rs} measurements with associated IOPs and optical constituent concentrations was developed using novel techniques suited to high biomass complex optical systems and cyanobacteria dominated waters. The parameterization of the RTM describing the synthetic dataset utilizes our most current understanding of optical properties and relationships relating to eutrophic and cyanobacteria dominated waters. The RTM defining the synthetic dataset is built upon the physics-based two-layered spherical Equivalent Algal Population model which formulates the SIOPs of the phytoplankton component on population-specific refractive indices, and thus are not independent of each other. Support for the inclusion of spectrally variably phytoplankton scattering for productive waters has been previously validated (Lain et al., 2017), and highlights the significance of phytoplankton backscattering in PFT driven problems (Lain et al., 2018). Novel calculations of chl-a fluorescence within mixed cyanobacteria populations now more accurately define red/NIR relationships in lower biomass conditions. While variable FQY and a^*_{phy} may be the

driving factors in fluorescence amplitude variability for Case 1 waters, C_{nap} (Gilerson et al., 2007) and cyanobacteria admixture drive fluorescence variability in Case 2 and inland waters.

The quantification of in-situ PC concentration using laboratory-based methods is ill-defined, arduous, and not currently standardized as part of an operational protocol. There is consequently a scarcity of reliable PC concentrations paired with radiometric data. This synthetic dataset aims to partially close this information gap and includes paired modeled PC concentrations based on composition based $a^*_{pc}(620)$. R_{rs} spectra modeled through the RTM were compiled into 13 distinct OWTs using a functional data analysis and k-means clustering approach. Although OWTs are not exactly the same as those discovered using only in-situ data in Spyrakos et al., (2018), this was to be expected considering roughly 60,000 more spectra were used in the clustering analysis here. Even still, inspection of OWTs show similar relationships and ranges as described in Spyrakos et al., (2018). While the data collected and compiled as part of the LIMNADES in-situ dataset took over 10 years to produce along with millions of accumulated dollars, similar results were able to be achieved using a dataset 30x greater, developed using a personal laptop and commercial software. Laboratory studies of in-situ fieldwork data will always be crucial to better inform the RTM, and a deeper examination of the LIMNADES in-situ dataset is required to improve future versions of the synthetic dataset. Future work would aim to expand the EAP model to more PFTs and cyanobacteria species as well as include an atmospheric RTM to model up to at-sensor radiances. With the advent of high powered computing technology and cloud computing, we can translate our current limited capacity for algorithm development, into a Big Data problem, in which advanced machine learning architectures can be used to train robust retrieval models which can be used between sensors, and on archival data. With a combination of current and past sensor spatial resolutions ranging from 10 m to 4 km scales, a synergistic evaluation of water constituents, with known uncertainties, will allow for an

unprecedented global snapshot of fine scale ecological dynamics of coastal and inland waters. This synthetic dataset acts as the first step towards this goal. It is by no means, a fully comprehensive inclusion of all possible natural values and relationships found in inland waters, however, works as a proof-of-concept to show the capability of these techniques to create accurate simulations of real-world hypertrophic aquatic environments.

3.5 Appendix A

Table A1: Modes and standard deviations for lognormal distributions of the four synthetic datasets

Dataset	***Chl-a***		***Cnap***		***ag 440***	
	mode	st. dev	mode	st. dev	mode	st. dev
Case 1	1	5	-	-	-	-
ISM	3	300	50	200	0.7	1.2
CDOM	10	40	1	4	5	10
Cyano	5	1000	0.1	5	1	1

3.6 Appendix B

Table B1: List of equations and relationships used for bio-optical modeling

Absorption	**Parameters**
$a(\lambda) = a_w(\lambda) + a_g(\lambda) + a_{phy}(\lambda) + a_{nap}(\lambda)$	
$a_g(\lambda) = a_g(440)^{(-S_g(\lambda-440))}$	S_g = 0.012 – 0.021
$a_{nap}(\lambda) = a_{nap}(440)^{(-S_{nap}(\lambda-440))}$	S_{nap} = 0.007 – 0.015

$a_{nap}(440) = C_{nap} * \ a^*_{nap}(440)$	a^*_{nap} = 0.02 – 0.3 m^2/g
$a_{phy}(\lambda) = \left([Chl] * \ a^*_{cy}(\lambda) * \ S_f\right) + ([Chl] * a^*_{euk}(\lambda) * (1 - S_f))$	S_f = [0.1,0.2,0.3…1.0]
Scattering	
$b(\lambda) = \ b_w(\lambda) + b_{phy}(\lambda) + b_{nap}(\lambda)$	
$b_{nap}(\lambda) = b_{nap}(550)(^{550}/_{\lambda})^{\gamma_2}$	γ_2 = 0.5 – 2.0
$b_{nap}(550) = C_{nap} * \ b^*_{nap}(440)$	b^*_{nap} = 0.5 – 1.0 m^2/g
$b_{phy}(\lambda) = \left([Chl] * \ b^*_{cy}(\lambda) * \ S_f\right) + ([Chl] * b^*_{euk}(\lambda) * (1 - S_f))$	S_f = [0.1,0.2,0.3…1.0]
Backscattering	
$b_b(\lambda) = \ b_{bw}(\lambda) + b_{bphy}(\lambda) + (\tilde{b_{bnap}} * \ b_{nap}(\lambda))$	$\tilde{b_{bnap}}$ = 0.02
$b_{bphy}(\lambda) = \left([Chl] * \ b^*_{bcy}(\lambda) * \ S_f\right) + ([Chl] * b^*_{beuk}(\lambda) * (1 - S_f))$	S_f = [0.1,0.2,0.3…1.0]

Table B2: Definition of symbols

Chl-a	Chlorophyll-a
PC	Phycocyanin
PC:Chl-a	PC to chl-a ratio
C_{chl}	Chlorophyll-a concentration (mg/m^3)

C_{pc}	Phycocyanin concentration (mg/m^3)
NAP	Nonalgal particles
R_{rs}	Remote sensing reflectance
C_{nap}	Nonalgal particles concentration (mg/m^3)
SIOPs	Specific inherent optical properties
S_f	Admixture weighting factor
S_{ad}	Scaled admixture parameter
$a^*_{phy}(\lambda)$	Spectral specific absorption of phytoplankton (m^2/g)
$L_f(685)$	Amplitude of fluorescence peak at 685 nm (Wm^{-2}sr^{-1}μm^{-1})
$\emptyset_f$	Fluorescence quantum yield
C_f	Coefficient for assumed fluorescence shape
Q_a^*	Algal reabsorption coefficient
$E_o^-(\lambda)$	Downwelling irradiance just below the surface (μmol phot/m^{2}s)
$K(\lambda)$	Attenuation coefficient in fluorescence excitation zone
$K_{Lu}(\lambda)$	Upwelling attenuation coefficient
$c_{tot}(\lambda)$	Total attenuation at 685 nm (m^{-1})
$a_{pc}(\lambda)$	Absorption due to PC (m^{-1})
$a^*_{pc}(\lambda)$	Specific absorption of PC (m^2/g)
$a_{Chla}(\lambda)$	Absorption due to chl-a (m^{-1})
$a^*_{Chla}(\lambda)$	Specific absorption of chl-a (m^2/g)

$a_{Chlb}(\lambda)$	Absorption due to chl-b
$a^*_{Chlb}(\lambda)$	Specific absorption of chl-b (m^2/g)
$a_{Chlc}(\lambda)$	Absorption due to chl-c
$a^*_{Chlc}(\lambda)$	Specific absorption of chl-c (m^2/g)
$a_{phy-cy}(\lambda)$	Absorption due to phytoplankton from cyanobacteria component (m^{-1})
$a(\lambda)$	Total spectral absorption coefficient (m^{-1})
$a_w(\lambda)$	Water absorption spectrum (m^{-1})
$a_g(\lambda)$	CDOM spectral absorption (m^{-1})
$a_{phy}(\lambda)$	Phytoplankton spectral absorption (m^{-1})
$a_{nap}(\lambda)$	Nonalgal particle spectral absorption (m^{-1})
$a_g(440)$	CDOM absorption at 440 nm (m^{-1})
S_g	CDOM spectral slope
S_{nap}	Nonalgal particle spectral slope
$a_{nap}(440)$	Nonalgal particle absorption at 440 nm (m^{-1})
$a^*_{nap}(440)$	Specific absorption of nonalgal particles at 440 nm (m^2/g)
$a^*_{cy}(\lambda)$	Spectral specific absorption of cyanobacteria (m^2/g)
$a^*_{euk}(\lambda)$	Spectral specific absorption of eukaryotes (m^2/g)
$b(\lambda)$	Total spectral scattering coefficient (m^{-1})
$b_w(\lambda)$	Water scattering spectrum (m^{-1})
$b_{phy}(\lambda)$	Phytoplankton scattering spectrum (m^{-1})

$b_{nap}(\lambda)$	Nonalgal particle scattering spectrum (m^{-1})
$b_{nap}(550)$	Nonalgal scattering at 550 nm (m^{-1})
γ_2	Nonalgal exponent parameter
$b^*_{nap}(440)$	Specific scattering of nonalgal particles at 440 nm (m^2/g)
$b^*_{cy}(\lambda)$	Spectral scattering of cyanobacteria (m^2/g)
$b^*_{euk}(\lambda)$	Spectral scattering of eukaryotes (m^2/g)
$b_b(\lambda)$	Total spectral backscattering (m^{-1})
$b_{bw}(\lambda)$	Water backscattering spectrum (m^{-1})
$b_{bphy}(\lambda)$	Phytoplankton backscattering spectrum (m^{-1})
$\tilde{b_{bnap}}$	Backscattering ratio for nonalgal particles
$b^*_{bcy}(\lambda)$	Spectral backscattering of cyanobacteria (m^2/g)
$b^*_{beuk}(\lambda)$	Spectral backscattering of eukaryotes (m^2/g)

4

4 CHAPTER 4: SENSITIVITY ANALYSIS AND REFORMULATION OF THE MAXIMUM PEAK HEIGHT (MPH) ALGORITHM AGAINST A GLOBAL SYNTHETIC DATASET

4.1 Introduction

The intensified eutrophication of inland water compoundments (Ho et al., 2019) continues to place animal and human lives at risk (Falconer and Humpage, 2006; Bláha et al., 2009; Oberholster and Ashton, 2008). Subsequently, there has been increasing focus in recent years on the development of pigment retrieval models using ocean color satellites (Matthews et al., 2012,2015; Moses et al., 2009a, 2009b; Dall'Olmo and Gitelson, 2005, Gilerson et al., 2010; Simis, 2005; Qi et al., 2014). The majority of these studies utilized imagery from the ENVISAT Medium Resolution Imaging Spectrometer (MERIS), which can be argued as the sensor with the most ideal spatial, spectral, and radiometric characteristics for routine monitoring of larger inland water bodies (Matthews, 2014; Binding et al., 2018; Palmer et al., 2015c). As part of the European Copernicus programme for earth observation, the Sentinel-3 Ocean and Land Color Instrument (OLCI) was designed to build on the success of the now heritage MERIS (Donlon et al., 2012). Envisioned as a constellation of four satellites, the first two satellites (S3-A, S3-B) have launched successfully and preliminary validation over water targets suggest OLCI performance is similar to that of MERIS (Kravitz et al., 2020, Smith et al., 2018)

Calibration and validation of retrieval models generally involve local parameterization of common empirical or semi-analytical algorithms (Stumpf et al., 2016; Hestir et al., 2015). Considering the growing necessity for routine, operational processing of productive waters, it is now timely and necessary to develop pragmatic approaches for resolved global products. There has been significant steps made in this regard in evaluating the use of red-NIR semi-analytic chl-a retrieval models for a variety of locales and conditions including Chesapeake Bay, Lake Kinneret, Nebraska Lakes, the Azov Sea, and Taganrog Bay (Gitelson et al., 2009, 2011a; Gurlin et al., 2011; Moses et al., 2009; Gilerson

et al., 2010; Yacobi et al., 2011). Results show the capability of chl-a retrieval using similarly calibrated, empirically based band-ratio models on atmospherically corrected or in-situ remote sensing reflectance (R_{rs}) measurements over a variety of conditions.

Currently, the state-of-the-art does not allow for operational atmospheric correction over productive inland water bodies with high enough confidence for medium resolution sensors (Xue et al., 2019; Shen et al., 2017). Baseline type algorithms, which have proven to be robust and relatively insensitive to poor atmospheric correction, have been utilized on partially corrected bottom-of-Rayleigh reflectance (BRR) in attempt to bypass the requirement for a full atmospheric correction (Matthews et al., 2012; Binding et al., 2011; Palmer et al., 2015c). A BRR correction accounts for the effects of molecular Rayleigh scattering and gaseous absorption in the red and NIR bands, while leaving the more optically complicated aerosols. The Maximum Peak Height (MPH) algorithm (Matthews et al., 2012), is a shifting baseline model which has been validated using a global dataset of BRR measurements at 40 different lakes around the world (Matthews et al., 2014) and has been identified as the most optimal chl-a retrieval model for hypertrophic inland waters using OLCI (Kravitz et al., 2020). A unique feature of the MPH model flags each pixel as dominated by either cyanobacteria or algae and calculates the chl-a concentration individually based on empirically derived relationships. Other model features include flags for adjacency and floating aquatic algae. The cyanobacteria pixel flagging is fundamental to the operation of the MPH model and retrieval of chl-a, however, there has been almost no literature on the validation of the accuracy of these flags, compared to the chl-a product, and how an uncorrected atmospheric signal can impact these accuracies.

Considering the use of top-of-atmosphere (TOA) and BRR data is becoming more prevalent, relatively few studies exist which investigate the actual fraction of the isolated water-leaving signal and

associated Signal-to-Noise (SNR) values that reach the satellite sensor over inland water bodies. Utilizing TOA data is theoretically more feasible with turbid waters due to an elevated water signal from increased particulate backscattering, compared to “darker” oligotrophic waters which are dominated by water absorption. It is quite often cited that of the total radiance signal reaching a satellite signal over water, roughly only 10% of the signal is due to the upwelling water-leaving radiance (L_w), whereas the contribution from atmospheric aerosols make up the majority of the signal. Martins et al., (2017) presented a localized modeling study which found L_w had potential to reach ~43% of the total signal for red-edge bands of Sentinel-2 MSI over turbid lakes in the Amazon. It is important to understand the extent of the water signal at TOA and its sensitivity to certain water and atmospheric parameters, in order to more thoroughly evaluate models which utilize TOA data.

This study intends to expand on the utility of a global MPH processor, validating MPH flags and chl-a retrieval against a large state-of-the-art synthetically created dataset of OLCI BRR measurements which includes the extremely wide variability of water and atmospheric optics for productive inland waters. The main objectives of this study are three-fold: 1) to examine the relative extent of the L_w signal for at sensor radiances of OLCI specific bands along with first order estimates of typical SNR ranges, 2) to provide an accuracy analysis for MPH adjacency and cyanobacteria flags, and 3) provide a predictive capability analysis of the MPH for global inland waters using the synthetic dataset. The aquatic radiative transfer modeling is described in the previous chapter, while the atmospheric radiative transfer modeling will be established here The chapter finishes with a discussion on the capability using TOA data as a proxy for R_{rs} measurements and whether band difference models utilized on BRR data can account for the vast inherent variability of inland aquatic optics, as well as the capability of developing a global model.

4.2 Methods

4.2.1 Synthetic dataset

4.2.1.1 Aquatic radiative transfer modeling

The methodology describing the aquatic radiative transfer modeling is fully described in Chapter 3, and briefly discussed here. The R_{rs} dataset is defined by a four-component bio-optical model to derive inherent optical properties (IOPs) for a hypothetical global dataset of inland water conditions. The dataset was built by combining four separately created datasets defined by the relative contributions of three dominant optical constituents, namely, chlorophyll-a concentrations (C_{chl}), the concentration of nonalgal particles (C_{nap}), and the absorption of CDOM at 440 nm (a_g (440)). The phytoplankton component of the dataset was modeled using the physics-based, two-layerd spherical Equivalent Algal Populations (EAP) model, where population-specific refractive indices are used to derive IOPs and thus not independent of each other (Lain et al., 2018; Bernard et al., 2009). IOPs are calculated at 5nm spectral resolution between 200 and 900 nm and integrated over an entire equivalent size distribution for the eukaryotic population (Bernard et al., 2007; Lain et al., 2017), and modeled with an effective diameter (D_{eff}) of 5 μm for the vacuolate *M. aeruginosa* (Matthews and Bernard, 2013). IOPs for the cyanobacteria Aphanizomenon and Anabaena cirinalis and non-vacuolate *Nodularia spumigena* which were measured in laboratory were also included in the dataset (Kutser et al., 2006). The ECOLIGHT radiative transfer code version 5.0 (Sequoia Scientific, 2008) was then used to generate nadir-viewing remote sensing reflectances (R_{rs}) from built IOPs using wavelength-specific Fournier Forand phase functions. Novel calculations of sun-induced chlorophyll fluorescence (SICF), which take into account the uneven proportions of intracellular chl-a in the fluorescing photosystem II (PSII) for eukaryotic and cyanobacteria populations are also included in the model, as well as modeled

calculations of Phycocyanin (PC) concentration. A cluster analysis of the data found 13 distinct optical water types which are described in Chapter 3.

4.2.1.2 Atmospheric radiative transfer modeling

The MODTRAN 5.0 radiative transfer software was used to propagate both L_w and R_{rs} from the aquatic modeling to OLCI at-sensor radiances. The radiance received by an optical sensor can be defined in simple terms as (Bulgarelli et al., 2014):

$$L_{tot} = L_{path} + L_{BG} + tL_u \quad (4.1)$$

where L_{tot} is the total radiance received by the sensor, L_{path} is the path radiance which defines the photons scattered into the instantaneous FOV by the atmosphere alone, L_{BG} is the background radiance from neighboring pixels which are diffusely scattered into the sensor FOV, L_u is the combined sky reflected and water leaving radiance at the sensor, and t is the diffuse transmittance. L_{BG} is considered as the radiance introduced due to the adjacency effect (AE), which can lead to large errors in derived products if inter-pixel non-uniformity is very large as in the case for neighboring vegetation, sand, or snow (Bulgarelli et al., 2017). Optical properties for a hypothetical atmospheric column for defining the radiative transfer model (RTM) were compiled from level-2 (L2) derived products from the global Aerosol Robotic Network (AERONET) database (https://aeronet.gsfc.nasa.gov/). The parameters which were directly varied for the RTM were aerosol optical thickness at 550 nm (AOT550), the angstrom extinction coefficient (Ext), single scattering albedo (SSA), the altitude of the hypothetical water target (Alt), water vapor (H2O), and percent adjacency of green grass vegetation (Adj). A tropospheric canned model was used to define the initial Mie-generated phase functions and asymmetry parameter, while Ext, SSA, and AOT550 were used to tweak the model based on randomly selected values from the L2 AERONET database.

The ranges for these parameters can be seen in Figure 1. For each aquatic R_{rs} measurement, two random atmospheres were modeled, and for each atmosphere, a second run was performed with a random contribution of green grass adjacency between 0.5% - 50%, totaling four atmospheric radiative transfer runs per R_{rs} measurement. Spectral radiance reaching the satellite sensor was calculated as follows:

1. The weighted mean of mixed spectral albedo curves was computed based on the Adj parameter.
2. The atmospheric model was compiled in MODTRAN by tweaking the standard tropospheric canned model using randomly selected parameters (AOT550, SSA, H20, Ext, Alt, Adj).
3. L_u and L_w from Ecolight output was multiplied by atmospheric path transmittance (t) from MODTRAN output to obtain L_u and L_w at TOA (L_uTOA and L_wTOA, respectively).
4. Total radiance at TOA (L_{tot}TOA) was calculated by multiplying L_uTOA by the MODTRAN derived atmospheric L_{path}, which is the radiance contribution from a scattering atmosphere.
5. All computations up to this point performed at full MODTRAN 5 spectral resolution. The OLCI spectral response functions (SRFs) were then applied to compute channel radiances.
6. Fraction of surviving L_w reaching the satellite sensor was calculated as L_wTOA / L_{tot}TOA.

To further correct for molecular (Rayleigh) scattering effects and convert to a Bottom of Rayleigh Corrected Reflectance (further denoted as R_{rc}), an analytical derivation was used as in Hu et al., (2004):

$$R_{rc} = \pi L_t^* / (F_0 cos\theta_0) - R_r \quad (4.2)$$

Where L_t^* is the calibrated at-sensor radiance after adjustment for ozone and gaseous absorption, F_0 is the extraterrestrial solar irradiance, θ_0 is the solar zenith angle, and R_r is the Rayleigh reflectance.

R_r can separately be derived through an RTM such as 6SV (Vermote et al., 1997; Hu et al., 2009), however, is derived analytically here for computational efficiency:

$$R_r = (\tau_r * P_r) / 4 * cos\theta_0 * cos\theta_v \quad (4.3)$$

Where θ_v is the sensor zenith angle and τ_r, P_r are the spectral Rayleigh optical thickness and Rayleigh phase function, respectively. Further derivation of τ_r and P_r can be found in Appendix A.

4.2.2 SNR calculation

The band-wise SNR was calculated as (Qi et al., 2017):

$$SNR = L / \sigma \quad (4.4)$$

Where L is the signal radiance in mW $cm^{-2}\mu m^{-1}sr^{-1}$, and σ is the sensor noise. The total sensor noise is the sum of the photon (shot) noise, dark noise, readout noise, and digitization noise. The photon noise arises due to quantum fluctuations sensed at a given exposure level, or variations in photons detected at the sensor over unit time. It is considered physically as the minimum noise level possible for a specific imaging scenario. Shot noise is generally the dominant source of noise, especially when the signal exceeds 10^4 electrons and the data essentially becomes shot noise limited (Moses et al., 2012). Thus, only shot noise was considered in this study and calculated following steps in Moses et al., (2012) for OLCI configuration. SNR was calculated for the total at-sensor radiance signal as well as for isolated L_w, where L_w was considered as the "real" incoming signal and noise was calculated on $L_{tot.}$ This provides more relevant information for observing aquatic systems, where we are only concerned about the small fraction of L_w, and arguably more insightful than L_{tot} (Kudela et al., 2019). A full analysis of signal noise and subsequent error analysis is not the main focus of this chapter, but

rather to introduce the utility of a synthetic dataset for SNR studies and inspect typical SNR values for L_{tot} and L_w at TOA for complex inland waters.

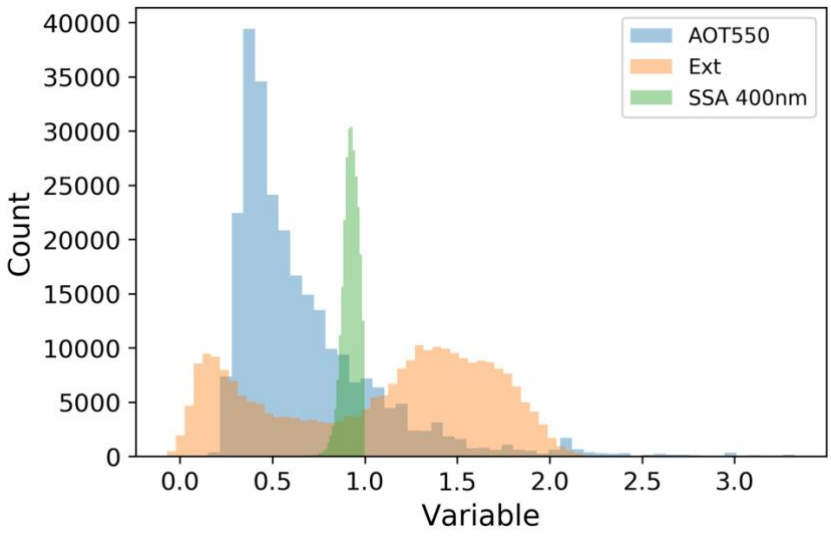

Figure 1: Histograms for angstrom extinction coefficient (Ext), single scattering albedo at 400 nm (SSA), and aerosol optical thickness at 550 nm (AOT550) used as input into MODTRAN.

4.2.3 MPH Algorithm

The Maximum Peak Height (MPH) algorithm is a modified-baseline subtraction algorithm which calculates the spectral derivative of the maximum peak height in the red and NIR region. The MPH variable is calculated by:

$$MPH = \rho_{BRmax} - \rho_{BR664} - ((\rho_{BR885} - \rho_{BR664}) \times (\lambda_{max} - 664) / (885 - 664)) \quad (4.5)$$

where ρ_{BRmax} and λ_{max} are respectively the BRR magnitude and position of the largest peak value from OLCI bands 10 (681nm), 11 (709nm), and 12 (753nm). The MPH algorithm acts a conglomerate of other indices relevant to productive waters by concentrating on three specific regions of the signal. The first case relates to low to medium biomass conditions with chl-a concentrations generally less

than 20 mg m^{-3}. These would be considered oligotrophic to mesotrophic conditions where eukaryotic phytoplankton sun-induced chlorophyll fluorescence (SICF) induce a peak near 681 nm. The algorithm then takes the shape of the Fluorescence Line Height (FLH) method (Gower et al., 1999), measuring the height of the peak at 681 nm. The second case relates to high biomass eutrophic/hypertrophic conditions with chl-a concentrations larger than 20 mg m^{-3} and when cyanobacteria tend to dominate. At these concentrations, the strong absorption by chl-a at 675 nm begins to mask the contribution produced by SICF. This combined with the high absorption due to water in the NIR creates a scattering-induced reflectance peak near 709 nm. In this instance, the MPH algorithm emulates the Maximum Chlorophyll Index (MCI) (Gower et al., 2005) and other reflectance/scatter line height algorithms calculating the spectral derivative at the 709 nm band. The third case for the MPH algorithm concerns extremely high biomass conditions and is used to identify surface scum conditions and floating vegetation by utilizing the 754 nm band. Due to the vacuolate-induced buoyancy of cyanobacteria, dense floating algal mats can occur with low wind activity. In these cases, the reflectance peak will shift from 709 nm to farther in the NIR, simulating a spectral signature of dry vegetation. This condition emulates the Floating Algae Index (FAI, Hu et al., 2010), calculating the height of the peak at 754nm.

The MPH model also calculates a series of flags denoting particular occurrences. The cyanoFlag identifies pixels which are dominated by cyanobacteria by a set of conditions:

cyanoFlag = (SICF < 0) & (SIPF > 0) & (BAIR > 0.002)

where SIPF is the sun induced phycobiliprotein fluorescence, which measures the height of the 665 nm peak over a baseline between 620 nm and 681 nm, and BAIR is the backscatter and absorption induced reflectance, otherwise known as the MCI. A negative value for SICF emulates the

cyanobacteria index (CI) model (Wynne et al., 2010), providing an index for cyanobacteria abundance. This metric has been commonly used in conjunction with SIPF to reduce the number of false positives (Lunetta et al., 2015), such that increased PC, which strongly absorbs at 620 nm (Simis et al., 2005), will enhance the peak at 665 nm.

The MPH model also includes a flag separating submerged from floating aquatic vegetation. This includes both floating cyanobacteria scums and aquatic macrophytes, however, specifically cyanobacteria scum is flagged if estimated C_{chla} > 350 mg/m^3.

floatingFlag = (λmax = 754 nm) & (MPH > 0.05) & (NDVI > 0.2)

Laslty, the MPH model includes a flagging procedure for adjacency as follows:

adjacencyFlag = (λmax = 754 nm) & (MPH <= 0.02) & (NDVI <= 0.2)

where NDVI is the normalized difference vegetation index, which is used to separate adjacency from floating vegetation. Depending on the accuracy of an applied land/water mask, there could potentially be mixed land/water pixels near the edges of the water body, or in the presence of floating aquatic material. These cases would not necessarily be considered true adjacency, however, the adjacencyFlag would treat mixed pixels the same. This caveat is ignored for the practical applications of this research.

The output of the MPH flags will be compared to the values in the synthetic dataset using a confusion matrix, which is typically used for classification algorithms and takes the following form:

Table 1: Typical form of a confusion matrix

	Real: True	**Real: False**
Predicted: True	True Positive (TP)	False Positive (FP)
Predicted: False	False Negative (FN)	True Negative (TN)

Where true and false represent a Boolean type classification for when the specific flag is raised. Further metrics of performance can then derived from the confusion matrix and are further explained in Appendix A.

The predictive capability of MPH will be assessed by comparing chl-a concentrations from the synthetic dataset and derived from MPH. Three versions of the MPH will be analyzed. The first will be the operational form of MPH from Matthews et al., (2015), the second will be MPH using re-calibrated models for cyanobacteria or eukaryotic dominated spectra, and the third will be an MPH using a single calibration. The predictive capability of the models will be assessed by comparing the calculated linear and log-transformed root mean squared error (RMSE and RMSELE, respectively), relative RMSE (rRMSE), bias, median absolute percent error (MAPE), and r-squared.

4.3 Results and Discussion

4.3.1 Surviving L_w at TOA

The average percent contributions of the surviving water signal of L_{tot} for the 13 OWTs derived in Chapter 3 for specific visible and NIR bands are displayed in Fig. 2. Please refer to the cluster analysis in Chapter 3 for basic cluster characteristics. The high inter and intra-variability of the percent contribution of the L_w signal is extremely evident. A common characteristic among the clusters is the

relatively low contribution from the 443 nm band. This region encompasses high amounts of absorption amongst the different optical constituents as well as significant interference from atmospheric aerosols. Consequently, this band only reaches above 20% contribution amongst the whole dataset only in extremely scattering conditions with relatively low amounts of blue absorption due to phytoplankton and CDOM. Other commonalities that exist is the general increase in surviving aquatic signal with increased inorganic sediment as well as increased phytoplankton component. This can also be visualized in Fig. 3, which displays the TOA L_w fraction as a function of chl-a concentration and of varying cyanobacteria admixtures for a "common" aquatic case and three extreme cases. In-situ samples from Chapter 2 are also plotted as validation of the synthetic data. For the common case, a general increase in TOA fraction exists with increasing chl-a fraction for each wavelength, although the 443 nm band is greatly subdued. A greater increase in L_w TOA fraction when cyanobacteria make up more the total algal biomass can also be visualized if Fig. 3 for the common case. When cyanobacteria dominate, L_w TOA fractions have the potential to reach 40% for red/NIR bands with chl-a concentrations as low as 10 mg/m^3, while maxing out at an average of roughly 60% for the 709 nm band just above 100 mg/m^3. When eukaryotic algae dominate, average surviving L_w TOA fraction only exceeds 20% for the NIR bands and at highly elevated chl-a concentrations. In Fig. 2, this relationship can also be visualized when comparing subdued L_w TOA fractions of cluster 7, which represents high biomass eukaryotic algae blooms, versus clusters 3,6, and 9 which contain much higher PC:Chl-a ratios. This illustrates how the increased backscattering from cyanobacteria IOPs can lead to drastic increases in strength of the signal. This especially relates to the NIR scattering peak, which are less affected by high phytoplankton absorption in the blue, as well as increased absorption due to water further into the NIR.

For the high C_{nap} case in Fig. 3, a flattening of the TOA fraction across all chl-a concentrations is observed, reaching TOA fraction averages above 60% for the red/NIR bands. Contrary to common cases, where we see a greater increase in L_w TOA fraction when cyanobacteria make up more of the total algal biomass, this relationship breaks down in extreme C_{nap} cases. The broad increase in backscattering signal from the inorganic component contaminates much of the signal induced from cyanobacteria. The high absorption in the blue/green bands for elevated CDOM results in similar reduction of surviving L_w signal as when elevated AOT at 550 exists. Below 100 mg/m^3, no bands exhibit average L_w TOA fractions above 20%, with blue bands almost non-existent at TOA. It is interesting to note the influence of CDOM absorption even into the red/NIR bands, considering the exponentially decreasing nature of the signal towards longer wavelengths. Elevated AOT causes severe attenuation in the atmosphere, drastically reducing surviving water-leaving signal throughout the entire visible spectrum.

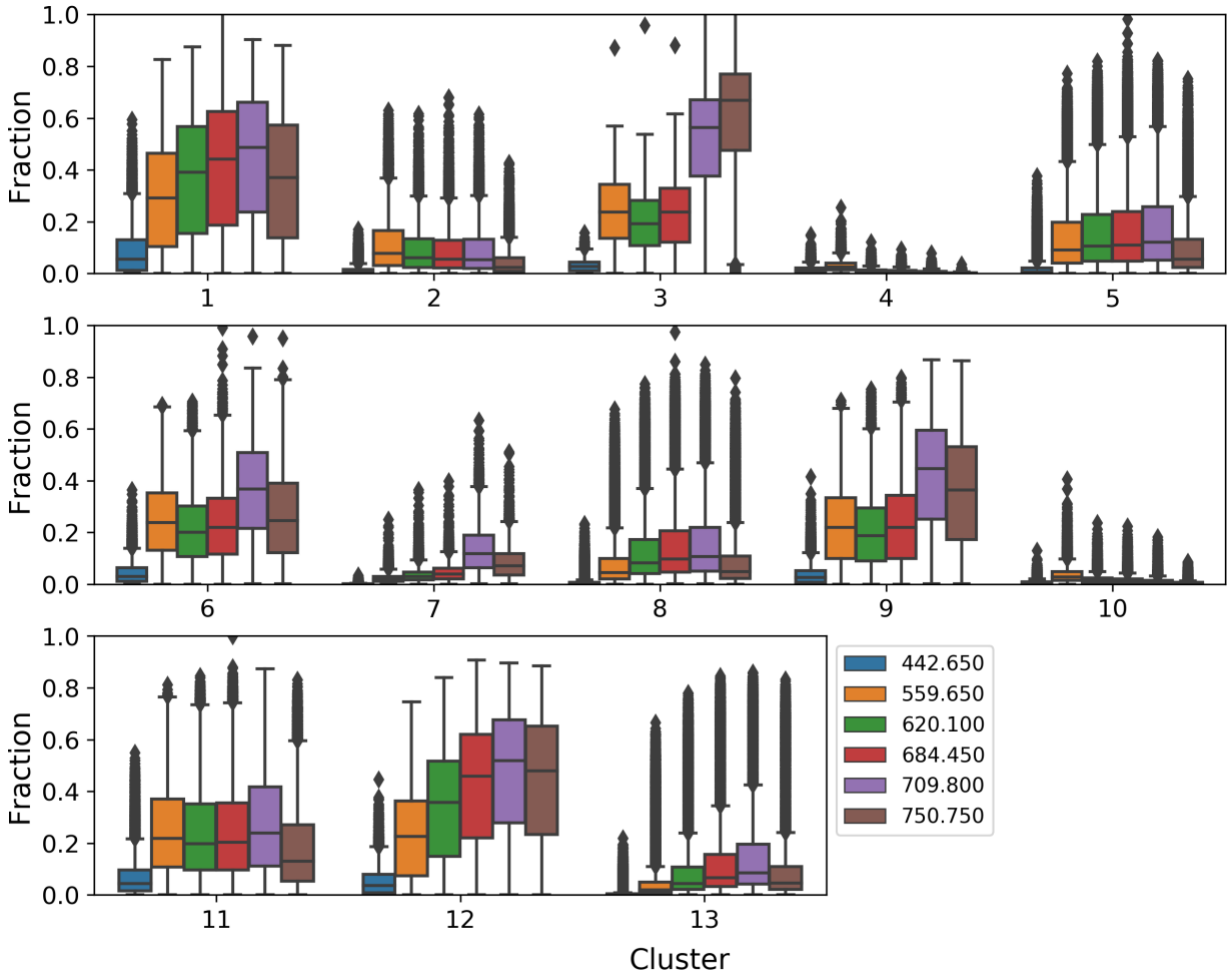

Fig. 2: Fraction of surviving L_w of L_{tot} for specific wavelengths for each derived OWT from chapter 3. Wavelengths are derived from MODTRAN spectral resolution.

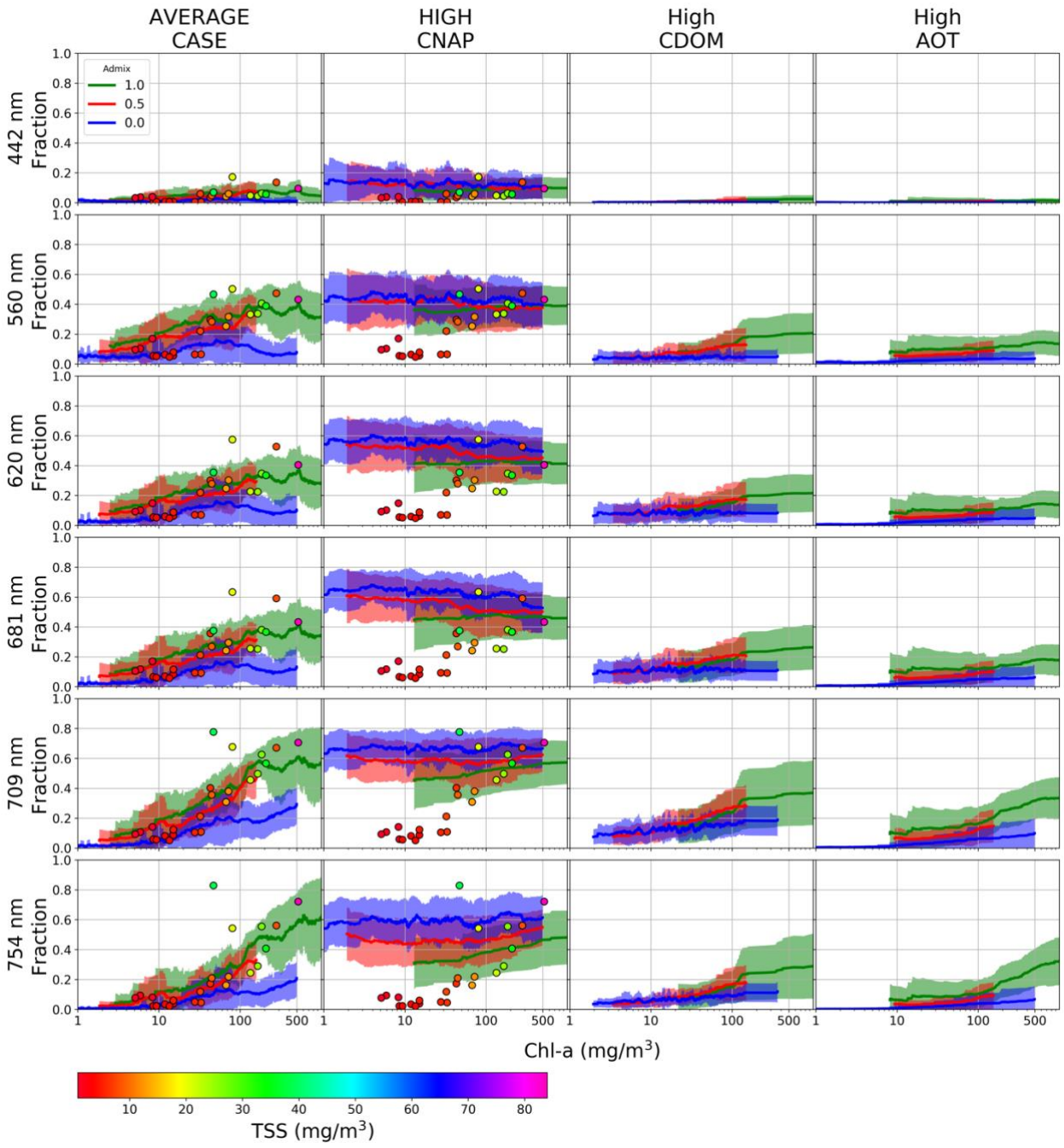

Fig. 3: Fraction of surviving L_w of L_{tot} for three cyanobacteria admixture cases as a function of chl-a for an average case (C_{nap} < 10 mg/m^3, CDOM < 1 m^{-1}, AOT550 < 1), a high C_{nap} case (C_{nap} > 10 mg/m^3), a high CDOM case (CDOM > 5 m^{-1}), and a high AOT550 case (AOT550 > 1). Admixture represents the cyanobacteria contribution from 0 to 100% of the phytoplankton IOPs. In-situ data points are from fieldwork described in Chapter 2.

4.3.2 SNR

The system SNR was calculated for both L_{tot} and L_w radiances for OLCI bands (Figs. 4 and 5). Numerous investigations have concluded that errors in both atmospheric correction and geophysical retrievals only become acceptable (<100% relative error) at an SNR of 300- 500 at visible wavelengths and > 100 at NIR wavelengths for water quality applications (Moses et al., 2015; Wang and Gordon; 2018; Qi et al., 2017; Jorge et al., 2017). For at-sensor SNR calculated on L_{tot}, all bands in the vis/NIR are, on average, above this threshold for all synthetic OWTs (Fig. 4). Across all water types, SNR is generally maximum in the blue spectral region, capable of reaching SNR values between 2000 – 3000. SNR values for Red/NIR bands generally reside between 500 – 1000 across water types, however, becoming > 1000 when PC:Chl-a ratios are higher or in high inorganic sediment cases.

SNR calculated on at-sensor L_w for the synthetic dataset can be visualized in Fig. 5. Considering that, in this instance, the effect of the atmosphere is removed, we see similar commonalities among OWTs as in Fig. 2, where higher at-sensor SNR exists in cases of elevated inorganic sediment as well as higher PC:Chl-a ratios. In these scenarios, other than the blue band, the SNR in the visible and NIR regions ($SNR_{vis/NIR}$) are on average above 100. While, for the most clear waters, ($SNR_{vis/NIR}$) values fluctuate between 1 and 10. Intermediate OWTs see ($SNR_{vis/NIR}$) values between 10 and 100. Although, amongst all water types, there is potential for ($SNR_{vis/NIR}$) to reach values below two, which is approaching sensor noise (Moses et al., 2012). It is important to note that this is not a comprehensive sensitivity analysis of sensor SNR, but rather an inspection of first order variability amongst varying inland water OWTs, and to test the capabilities of using a purely synthetic dataset to perform larger scale SNR studies in the future. While most studies only consider SNR on the entire TOA signal, it has been recommended that we switch concentration to SNR on solely the L_w signal (Kudela et al., 2019), considering this is the portion of signal we are most interested in. This initial investigation suggests

that sensor sensitivity for these water types is not of major concern for the OLCI sensor considering the strength of the water-leaving signals, and future sensor design considerations could potential invest more resources into spatial or radiometric categories.

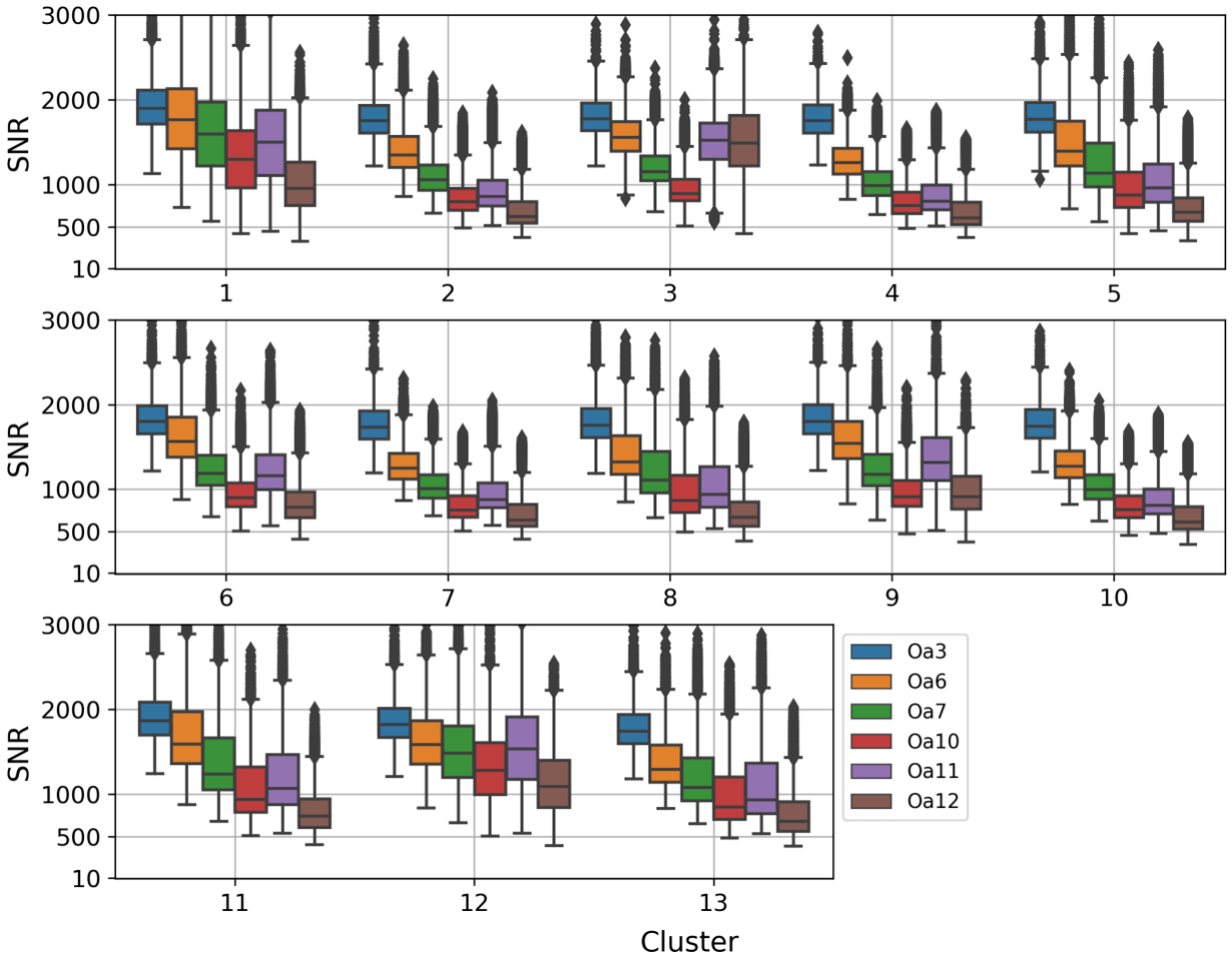

Fig. 4: First order variability of SNR of L_{tot} for six different OLCI bands for the 13 OWTs derived in chapter 3.

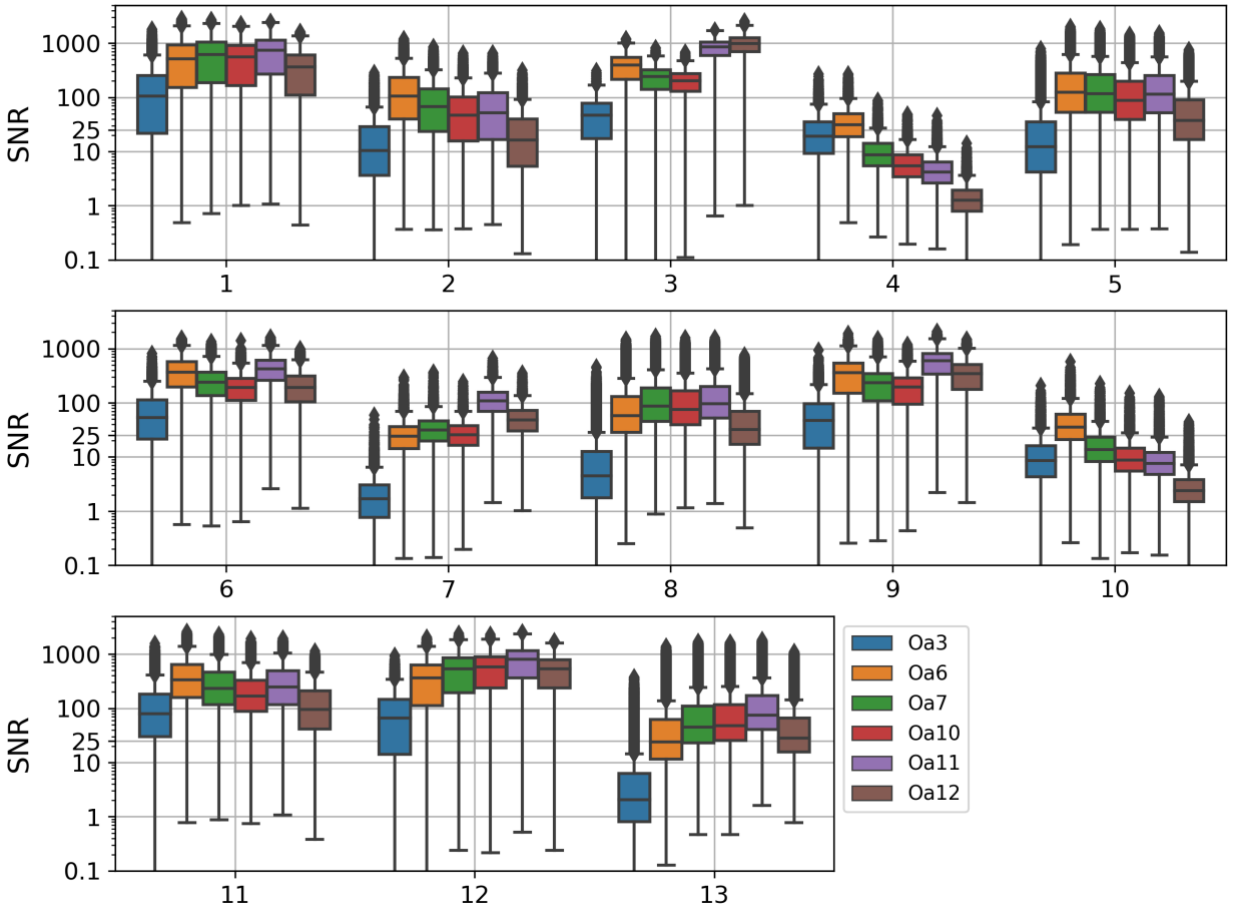

Fig. 5: First order variability of SNR of L_w for six different OLCI bands for the 13 OWTs derived in chapter 3.

4.3.3 Max lambda

To derive chl-a concentration from an aquatic reflectance, the MPH model first calculates the maximum lambda in the red/NIR spectral region. This step decides which band will be utilized to calculate the peak height which can drastically influence the resulting chl-a concentration. The ranges of max lambda for each OWT based on R_{rs} data are show in Fig. 6. For what can be considered more common Case 2 OWTs (OWTs 2,5,11), we see the max lambda, on average, switching between OLCI bands 10 (681 nm) and 11 (709 nm) between 15- 30 mg/m^3 when applied to R_{rs} data. This relationship still holds quite well even in OWTs of elevated C_{nap} or CDOM (OWTs 1,8,12,13). The significant

deviations to this rule are apparent in OWTs 3,4,6,7, and 9. OWT 6 and 7 represent high biomass cyanobacteria dominated and eukaryotic dominated assemblages, respectively. Thus, these OWTs never exhibit a max lambda of 681 nm, nor do these clusters ever reach scum conditions, and the 709 nm band remains the max lambda for the entire cluster. OWT 3 and 9 represent extremely high biomass waters, including the possibility for scum conditions, thus the max lambda can potentially shift to the 754 nm band. OWT 4 represents our oligotrophic water type where the 681 nm band dominates as the max lambda, although a few cases of elevated C_{nap} can shift the max lambda to 709 nm at very low biomass.

When inspecting max lambda at BRR (Fig. 7), we can see how the inclusion of a variable atmosphere can influence the shifting of max lambda. Excluding the 754 nm band, we see roughly similar patterns for the common Case 2 OWTs as on the R_{rs} data, although the general range for the average switch between 681 nm and 709 nm occurs at a bit higher chl-a concentration between 20 – 40 mg/m^3. Ranges for the 681 nm band also vary much more dramatically at BRR and can have significant presence even in OWTs 6,7, and 9 which are considered high biomass. At BRR, there is also much more potential for max lambda to shift to 754 nm, even without the presence of highly concentrated surface blooms. It must also be noted that these observations are from inspecting the inter-quartile ranges (middle 50%) of each cluster, of what we are considering is the “average” chl-a range. Within each cluster, ranges can vary wildly depending on the combination of highly scattering, or highly absorbing optical constituents.

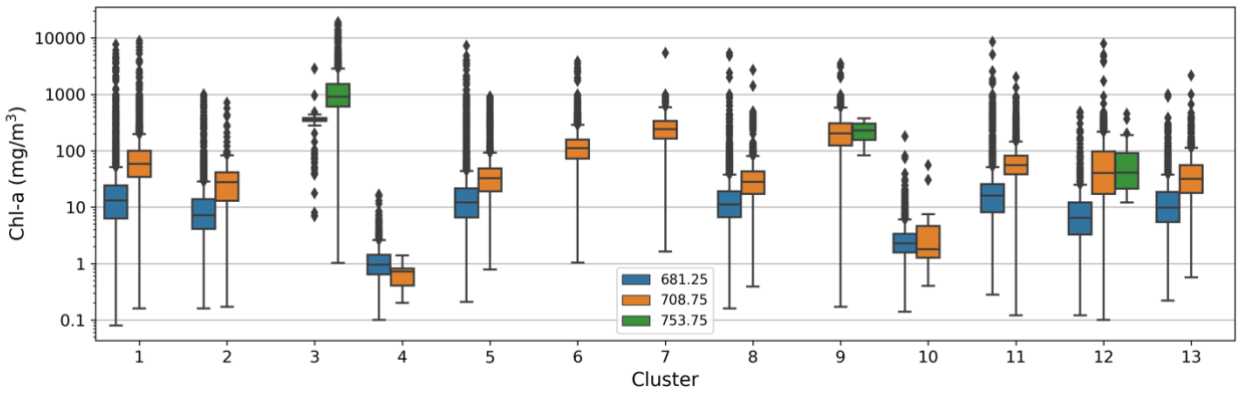

Fig. 6: Ranges of chl-a concentrations for the maximum red or NIR R_{rs} band utilized in the MPH separated by OWT.

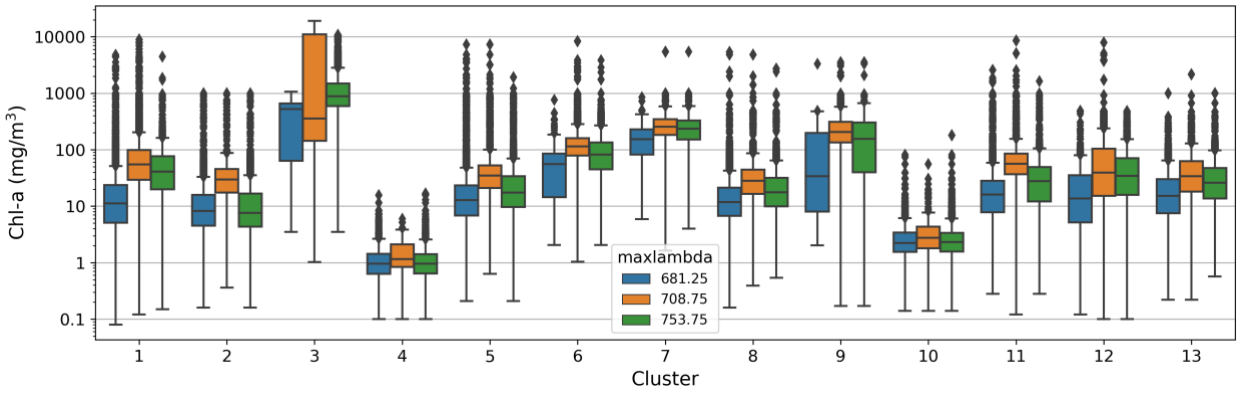

Fig. 7: Ranges of chl-a concentrations for the maximum red or NIR BRR band utilized in the MPH separated by OWT.

4.3.4 Adjacency

When using TOA reflectance data to derive biogeophysical variables from near-coastal or inland aquatic sites, the adjacency effect (AE), caused by spectral perturbation by light scattered into the sensor field of view (FOV) of an aquatic target by neighboring "bright" pixels such as vegetation, sand, or snow, can lead to significant errors in biogeophysical retrievals (Bulgarelli et al., 2017, 2018; Chapter 2). While attempts have been made at correcting for AE in the context of applying a full

atmospheric correction (Sterckx et al., 2015; Keukelaerie et al., 2018), no correction exists for applying retrieval models to BRR. The MPH does, however, produce an adjacency flag which signals where significant adjacency exists and thus, where results could be questionable. Average BRR spectra with and without green vegetation adjacency can be seen in Fig. 8. It is quite apparent how the magnitude of the difference between average BRR spectra with and without adjacency included varies by OWT. OWTs which include high cyanobacteria biomass or high inorganic particle loads (clusters 3,9,12), and thus, elevated signal in the red/NIR due to increased scattering, do not seem to be as affected by varying magnitudes of adjacency than when particle load is low. The visually elevated NIR signal contributed from adjacency suggests this is the reasoning behind dramatically more instances of max lambda at 754 nm in the BRR dataset in Fig. 7.

The performance of the MPH classification flag for adjacency is shown in Fig. 9. See Appendix A for further clarification of classification terminology. The figure shows various performance metrics as a function of either chl-a, C_{nap}, phycocyanin concentration (PC), and AOT at 550 nm, while the other three parameters are held constant to below some value noted in the figure caption. With increasing chl-a concentration (Fig. 9a), overall accuracy of the adjacency flag hovers between 60% - 70% until chl-a concentration reachs roughly 100 mg/m^3, in which accuracy begins to drop. The cause of this can be visualized through the other classification metrics. As chl-a increases above 10 mg/m^3, the accuracy becomes hindered by increased false identifications of adjacency (TPR and PPV). Accuracy still stays relatively high in these instances, because these incorrect classifications are offset by increased correct classifications of when no adjacency exists (TNR and NPV). With increasing C_{nap} (Fig. 9b), there is similarly a drop-off in accuracy. Once NAP concentrations reach around 3 mg/m^3, false identification of adjacency increases dramatically, lowering the accuracy of the flag to roughly 50% about a C_{nap} of 40 mg/m^3. Accuracy of the adjacency flag also drops quite dramatically with increasing

PC concentration to 50% from about 100 mg/m^3. Above this, only roughly 10% of cases which contain adjacency are correctly classified as adjacency contaminated spectra. Accuracy increases with increasing AOT until about a value of 0.5, in which correct identification of adjacency begins to decrease, keeping overall accuracy relatively constant at about 75% above 0.5. Global scores per cluster can be seen in Fig. 10. Cluster scores ranged from approximately 50% total correct classification to 70%. Clusters with high PC:Chl-a ratios or extreme C_{nap} cases generally performed worse due to very low scores for correctly classifying adjacency affected spectra.

These results suggest that, with the MPH adjacency flag, it becomes much harder to accurately classify a spectra as contaminated by adjacency with brighter waters, especially more productive waters. The adjacency included in this study is that of green vegetation which produces a highly elevated reflectance signal in the red and NIR. As water becomes more eutrophic, higher phytoplankton biomass similarly induces elevated signal in the red and NIR, and the MPH adjacency flag, which identifies adjacency if the max lambda is the 754 nm band, incorrectly classifies adjacency contaminated spectra more often. The adjacency flag attempts to reduce these incorrect classifications by also requiring the MPH value to be below 0.02, however, this threshold may not be very reliable when the signal is tainted by a turbid atmosphere.

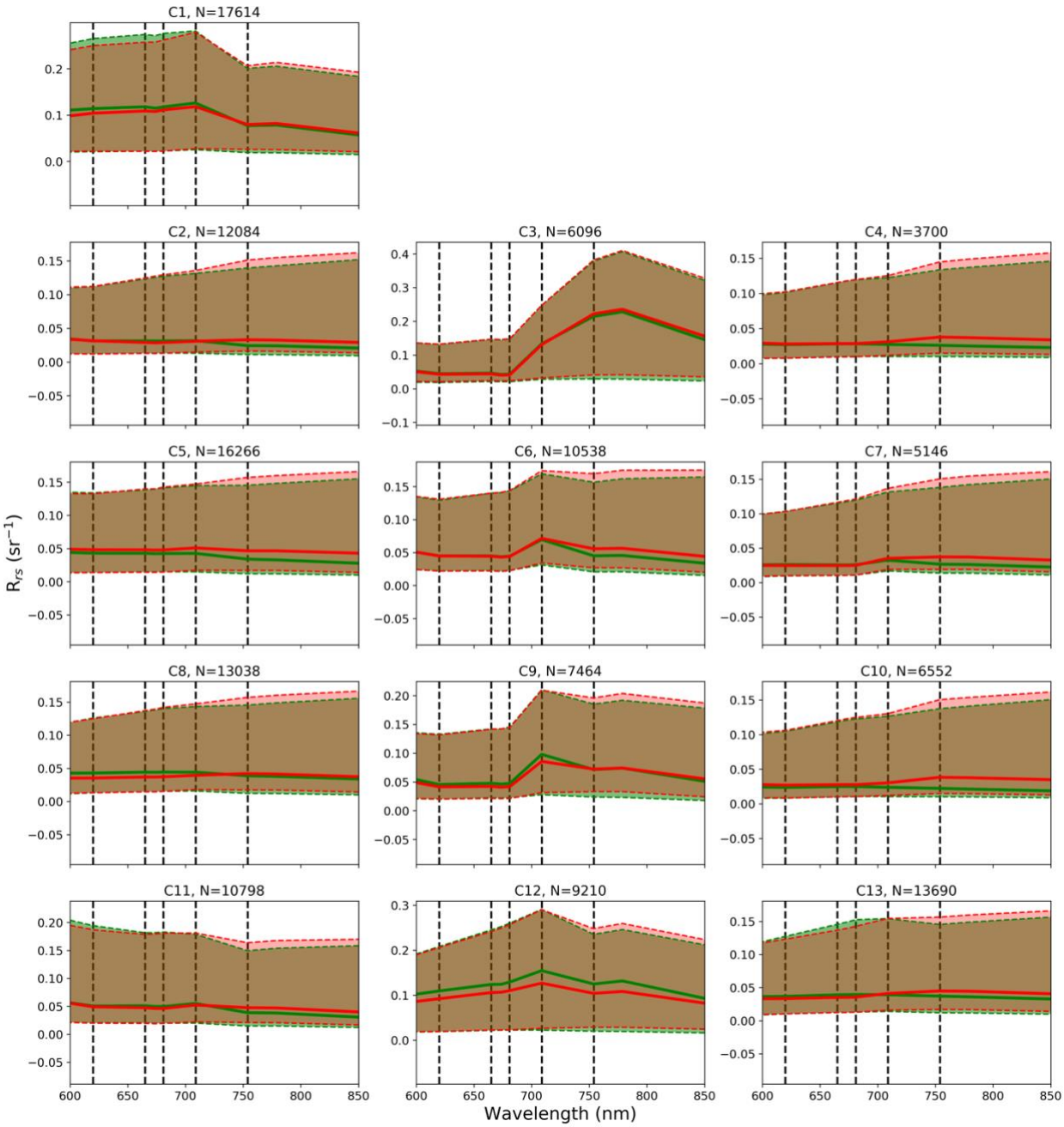

Fig. 8: Functional box plots of BRR spectra with and without adjacency for each OWT. The median BRR spectra are plotted as a solid line with shaded regions representing the IQR. Red are spectra with adjacency, green are spectra without adjacency.

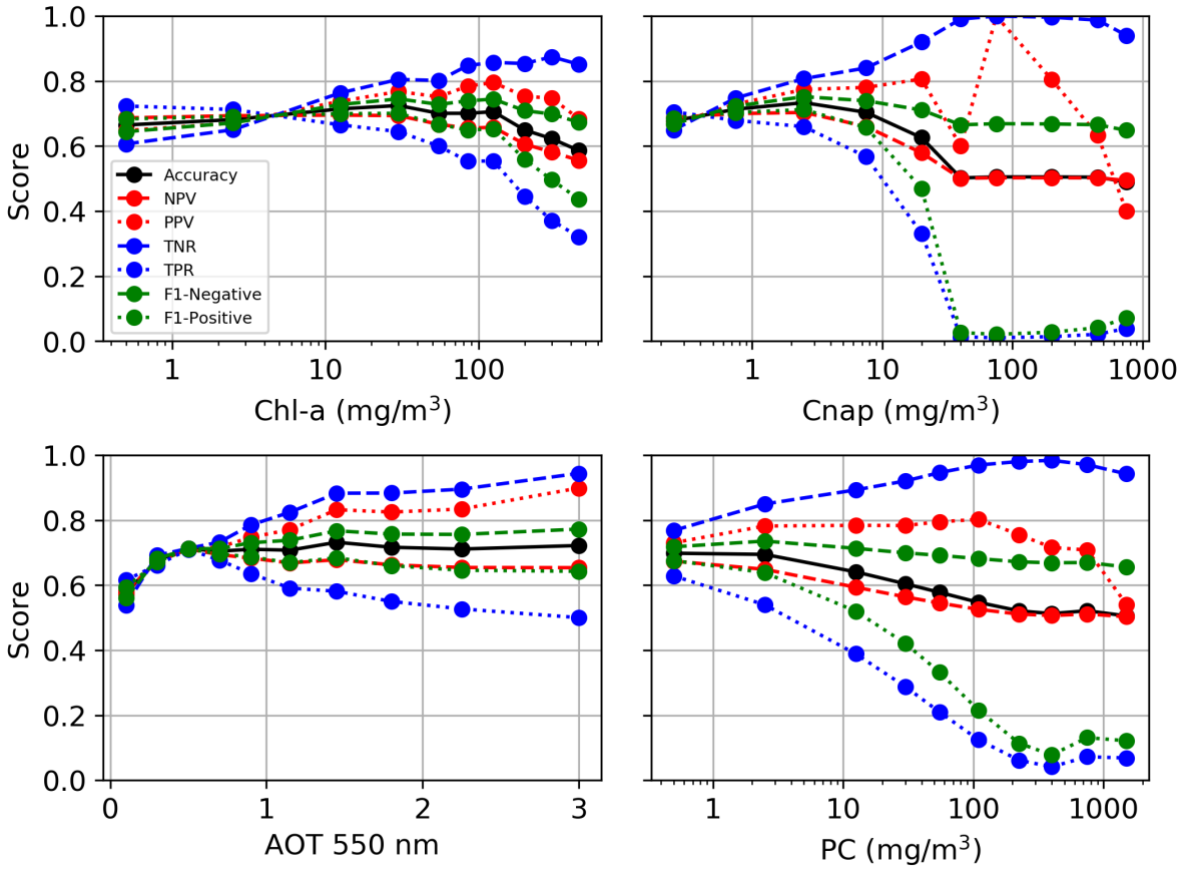

Fig. 9: Classification scores for identification of spectra with adjacency using the adjacency flag from MPH, as a function of increasing chl-a, C_{nap}, AOT at 550 nm , and PC. Overall classification accuracy is plotted in black. Definitions of other metrics can be found in Appendix B.

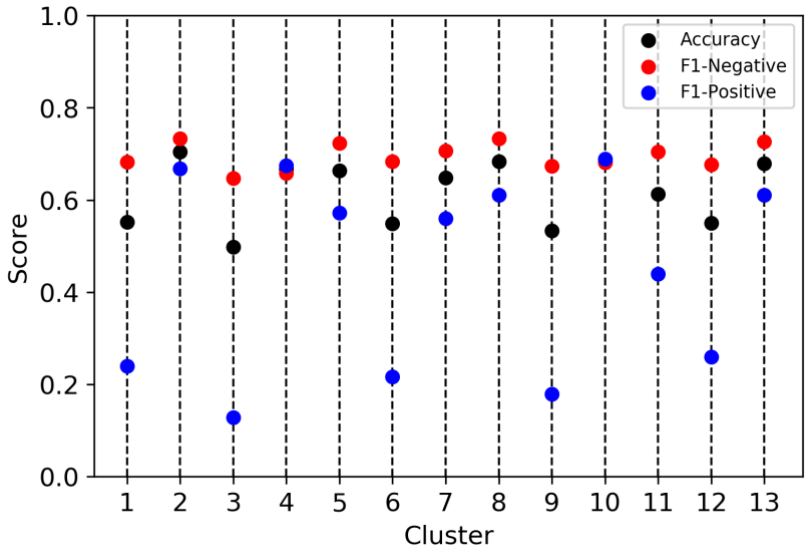

Fig. 10: Classification scores for identification of spectra with adjacency using the adjacency flag from MPH by OWT.

4.3.5 Cyanobacteria flag

In order to validate the MPH cyanobacteria flag against the synthetic data, it is important to investigate when the flagging classification considers a pixel "dominated" by cyanobacteria. The flagging procedure was first inspected on R_{rs} spectra, without interference from any atmosphere. In Fig. 11, overall accuracy of the cyanoFlag classification is plotted against chl-a concentration and C_{nap} for R_{rs} data. The admixture details the threshold for at which the minimum proportion of cyanobacteria compared to eukaryotic algae, would signify a pixel as a "True" cyanobacteria pixel (i.e. For an admixture of 0.4, all spectra built with an admix of at least 0.4 and above would be considered a spectra dominated by the cyanobacteria signal, while below 0.4 would be considered not dominated by cyanobacteria). At very low chl-a concentrations, accuracy scores remain quite high (> 80%) for all admixture thresholds. Once chl-a reaches about 3 mg/m^3, we begin to see large deviations in accuracy based on the threshold value. With increasing chl-a concentration, lower thresholds

experience worse classification accuracy. This makes sense in that when cyanobacteria contribution is low, the signal induced by the presence of cyanobacteria would also be low, making it more difficult to accurately predict the presence of cyanobacteria. At about 100 mg/m^3, the accuracy for the various thresholds begin to converge. With the exclusion of purely eukaryotic spectra (admix = 0), all threshold accuracy scores converge at around 80% above a chl-a concentration of roughly 200 mg/m^3. This also makes sense following that even at low fractions of cyanobacteria in the algal population, the strength of the backscattering signal from cyanobacteria is so much higher than their eukaryotic counterparts, that when biomass is high, it becomes easier to identify cyanobacteria.

Increasing the NAP concentration does not seem to affect intra-variability within a single admixture to a great extent. A slight decrease in accuracy appears around a NAP concentration of about 10 mg/m^3, although this is less pronounced at higher thresholds. Dramatic inter-variability between thresholds exists, with higher thresholds experiencing greater accuracy. This suggests that NAP concentration does not affect the cyanobacteria flag to a great extent.

When examining classification accuracy for BRR spectra (Fig. 12), accuracy at higher thresholds decrease faster with increasing chl-a when compared to R_{rs}, while accuracy at lower thresholds decrease less dramatically. There appears to be a pivotal point near a chlorophyll-a concentration of 10 mg/m^3, where accuracy at high thresholds begin to decrease quite dramatically, and accuracy at lower thresholds begin to increase. This can most likely be explained by the interplay between the chl-a fluorescence reflectance signal and the chl-a absorption peak between 675 nm and 685 nm. The cyano flag requires the SICF signal to be less than zero, which cyanobacteria satisfy from very low chl-a concentrations due to the lack of contributing chl-a fluorescence signal (Simis and Huot, 2012), while eukaryotic phytoplankton experience positive SICF. As chl-a concentration increases, the chl-a

absorption peak near 675 nm masks the reflectance peak induced by fluorescence, and the SICF signal begins diminishing, until eventually becoming less than zero even for eukaryotic phytoplankton. This is most likely causing misclassifications of cyanobacteria. The errors at BRR are more pronounced due to less sensitivity of the sensor to discriminate between these subtle differences. In Fig. 12a, the least amount of accuracy intra-variability as a function of chl-a occurs at an admixture of 0.6. Similar to R_{rs}, NAP concentration does not seem to affect the cyanobacteria flag much. Accuracy also only diminishes slightly with increasing AOT at 550 nm.

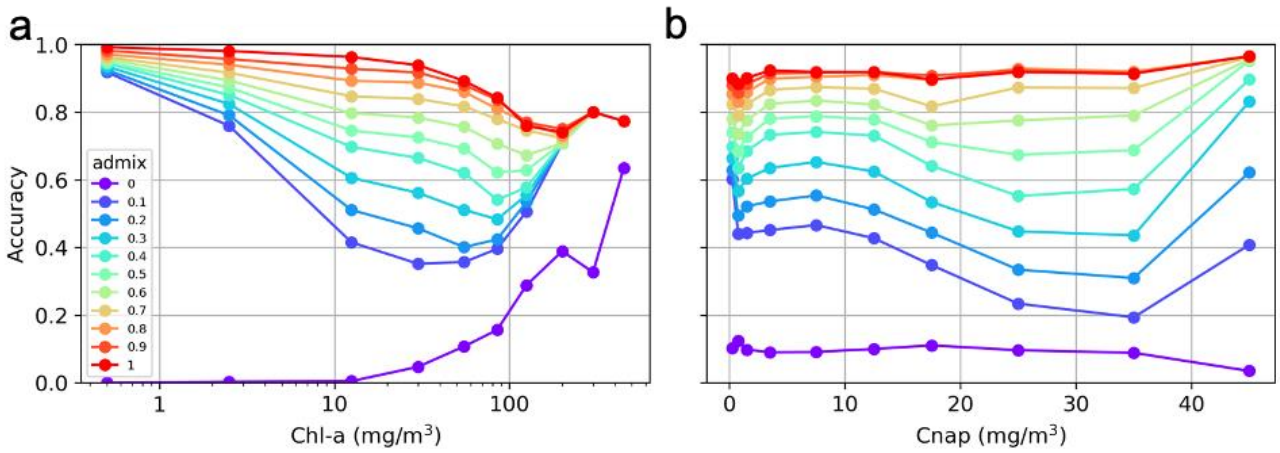

Fig. 11: Classification accuracy of the MPH cyanobacteria flag as a function of a) chl-a and b) NAP concentration. The different colored points represent the accuracy at different admixture thresholds for the relative contribution of cyanobacteria.

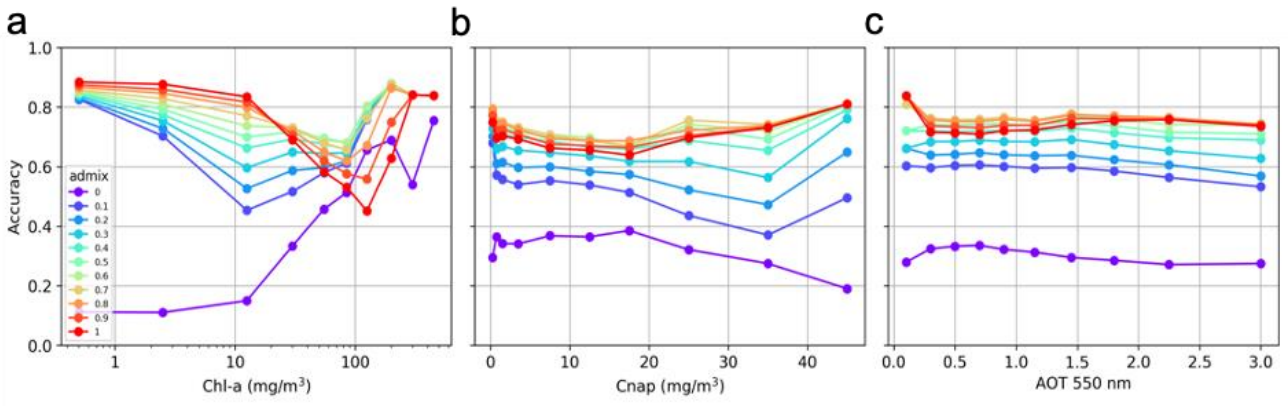

Fig. 12: As in Figure 11, except at BRR and including sensitivity to AOT 550 nm.

In Fig. 13, scores are now visualized as a function of the admixture threshold, thus the score is representative of entire ranges of optical constituents. There is a gradual increase in overall accuracy from about 50% to 90% as the threshold is set higher for R_{rs} spectra. For BRR, the accuracy is slightly higher at lower thresholds than R_{rs}, but plateaus at roughly 75% at an admixture of 0.6, after which there is a slight decrease. At an admixture of 0.6 for R_{rs}, the cyanoFlag accuracy reaches 80%. From Fig. 13a we can also see that the percentage of total cyanobacteria pixels correctly identified as cyanobacteria (TPR) ranges between 20 – 30% at lower thresholds while reaching 50 - 60% at higher thresholds, with accuracy values slightly higher in Fig. 13b at BRR. Contrary to this, at lower thresholds, there is an extremely high probability that spectra classified as cyanobacteria are actually cyanobacteria (PPV). In Fig. 13a, this probability starts to decrease rapidly at a threshold of about 0.6. Put simply, above an admixture threshold of 60% cyanobacteria, the MPH flag begins to incorrectly classify more spectra as cyanobacteria, when they are actually not. This interplay between TPR and PPV is most likely due to class imbalance in the dataset. When the admixture threshold is set higher, there would be much fewer spectra in the dataset which would be considered cyanobacteria dominant, thus probability of mis-classifying a spectra as cyanobacteria would increase (PPV),

however, percentage of correctly classified cyanobacteria spectra (TPR) would increase due to the overall smaller volume of cyanobacteria dominant spectra. In Fig. 13b there is a similar relationship, although with a more drastic decrease from low thresholds. Based on these results, we decided to use an admixture of 0.6 as the "real" threshold value to validate the MPH cyanobacteria flagging procedure.

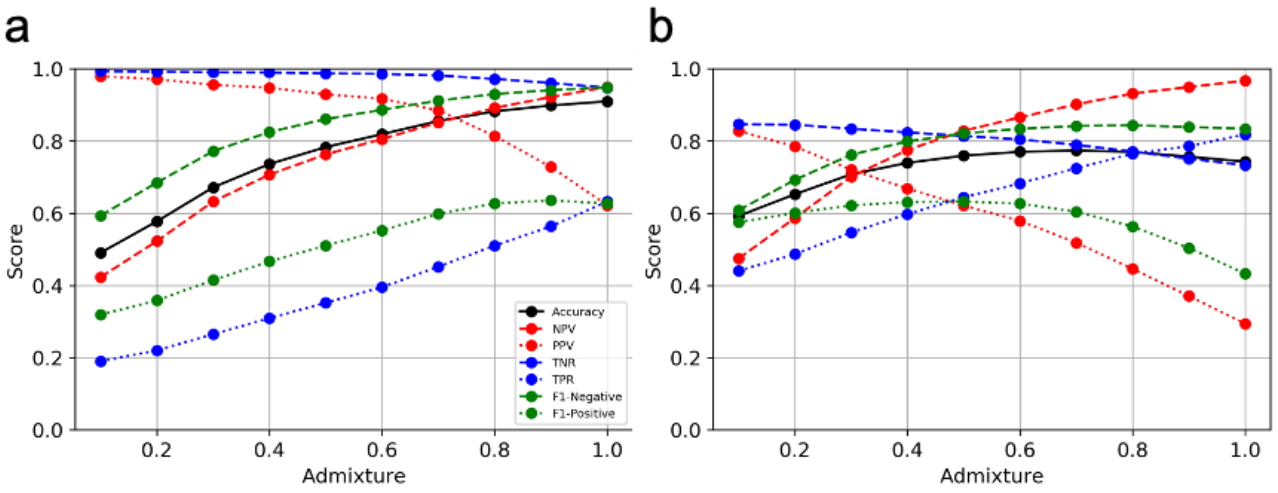

Fig. 13: Overall accuracy of the MPH cyanobacteria flag with increasing admixture threshold when applied to a) R_{rs} and b) BRR.

The cyanobacteria flag was also tested per OWT using an admixture threshold of 0.6 (Fig. 14). Considering the average accuracy for the total dataset was found to be just below 80% using a threshold of 0.6 (Fig. 13b), it follows that accuracies would hover between 70 – 90% amongst water types. Although, F1-positve scores, which represents an overall probability that the model is correctly classifying cyanobacteria dominant spectra as cyanobacteria, remain below 40% and reaching below 10% in some OWTs. The exception to this are clusters 3,6,9, and 11, which are all water types with high PC:Chl-a ratios. In these water types, the majority of the error instead stems from incorrectly classifying non-cyanobacteria spectra.

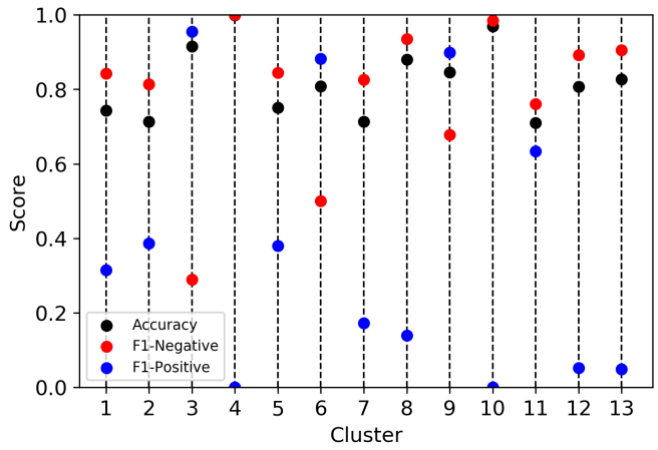

Fig. 14: : Classification accuracy of the MPH cyanobacteria flag using an admixture threshold of 0.6 for each OWT at BRR.

4.3.6 MPH chl-a estimation

The MPH model was recalibrated using the synthetic R_{rs} and BRR datasets with can be visualized in figure 16. New calibrations for eukaryotic and cyanobacteria dominated waters are provided, along with a single calibration which does not discriminate algal group. Calibrations were made only on data where C_{nap} was less than 20 mg/m^3, as greater than this causes dramatic errors in retrieval performance. A separate correction, or calibration, for waters with very high inorganic sediment loads will need to be investigated in the future. The new MPH model was trained via k-fold cross validation where 80% of the data was used to calibrate new regression coefficients, and 20% of the data was used to test predictive capability of the newly trained models. This was performed five times to reduce sampling bias and provide confidence intervals. Best fits for eukaryotic dominated spectra were found with a quadratic polynomial, while cyanobacteria dominated spectra and the single calibration were found with fourth and fifth order polynomials, respectively. Coefficients for the re-calibrated models can be found in Appendix B.

The MPH chl-a derived from the standard eukaryotic/cyanobacteria switching model and the single calibration are plotted against the synthetic chl-a in Figure 17 for both the R_{rs} and BRR datasets. Predictive statistics for the recalibrated models, tested on hold-out data from the synthetic dataset, are provided in figure 18 and broken down by OWT. It is clearly evident the drop in performance when applied to BRR and R_{rs} for OWT 4 and 10, which comprise of mostly oligotrophic to mesotrophic waters with low inorganic loads where surviving L_w at TOA is very low, rarely reaching above 10%. Considering the atmosphere comprises the vast majority of the signal for these waters, considerable error arises when using BRR data. Other OWTs with higher levels of water signal contribution at TOA see more similar statistics between BRR and R_{rs} results. The best overall results are seen for OWTs with relatively higher PC:Chl-a such as OWT 3, 6, 9, and 11. These OWTs generally see a MAPE between 20-50%. Standard RMSE will understandably be much higher for these OWTs due to the vastly greater concentrations of chl-a within these water types. There is not a consistent pattern in the results between the type of MPH calibration, whether based on a single or admixture based calibration. Although, there does appear to be slightly better results when using a single calibration at both BRR and R_{rs}. A table of overall predictive statistics can also be found in Appendix B.

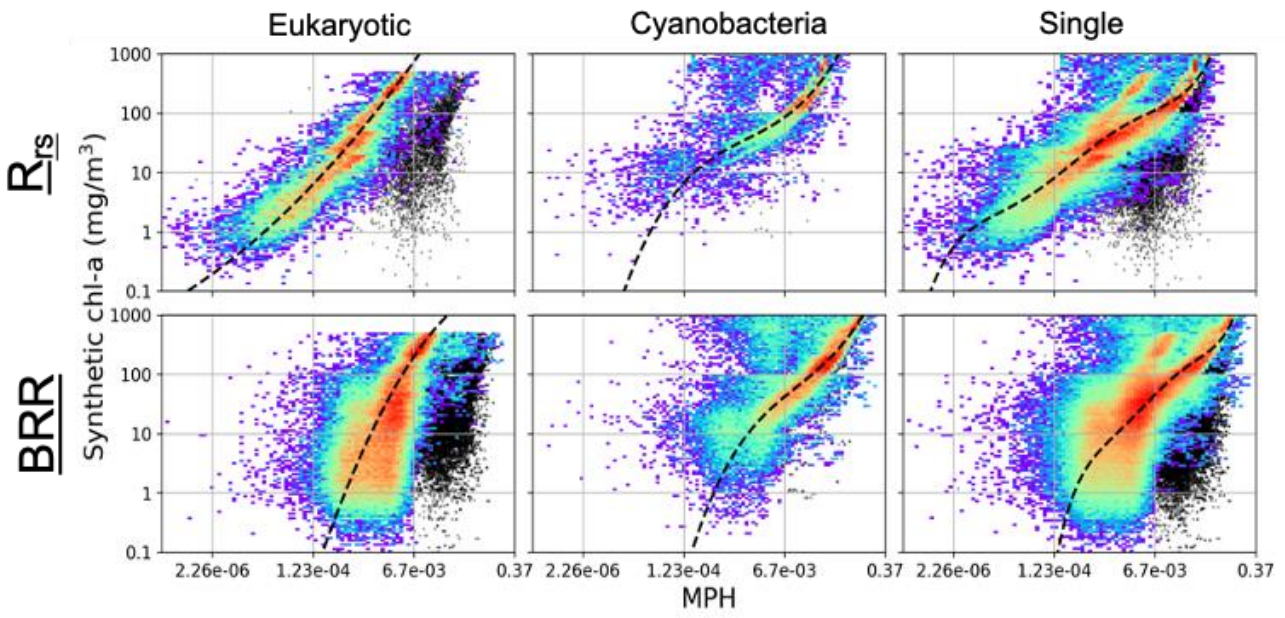

Fig. 16: Scatter plots comparing MPH values with synthetic chl-a concentrations. Calibration curves for eukaryotic and cyanobacteria dominated spectra, and a combine single calibration are plotted as a black dashed line. Scatter color represents point density, where warmer colors represents greater density of points. Black points represent instances where $C_{nap} > 20$ mg/m^3, and not considered for model calibration.

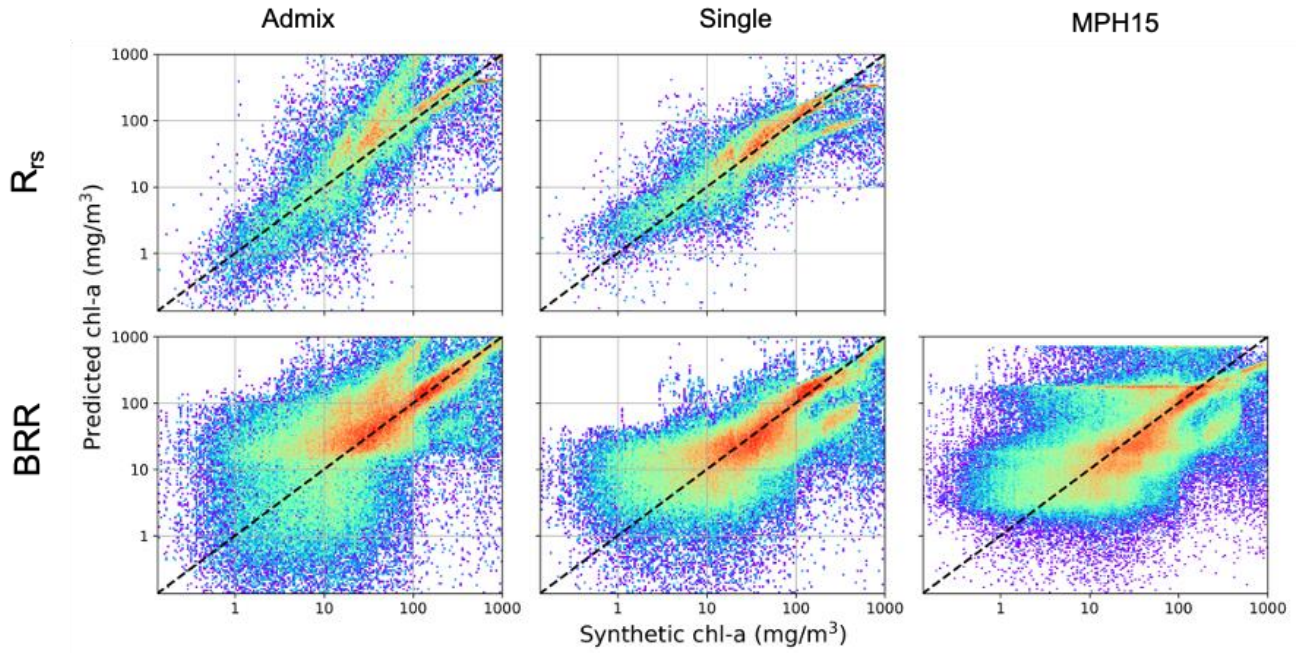

Fig. 17: Scatter plots comparing predicted and synthetic chl-a concentrations using an admixture based or single calibration for MPH, as well as the original MPH15 model. Black dotted line is 1:1 line, while scatter color is the same as in Fig. 16.

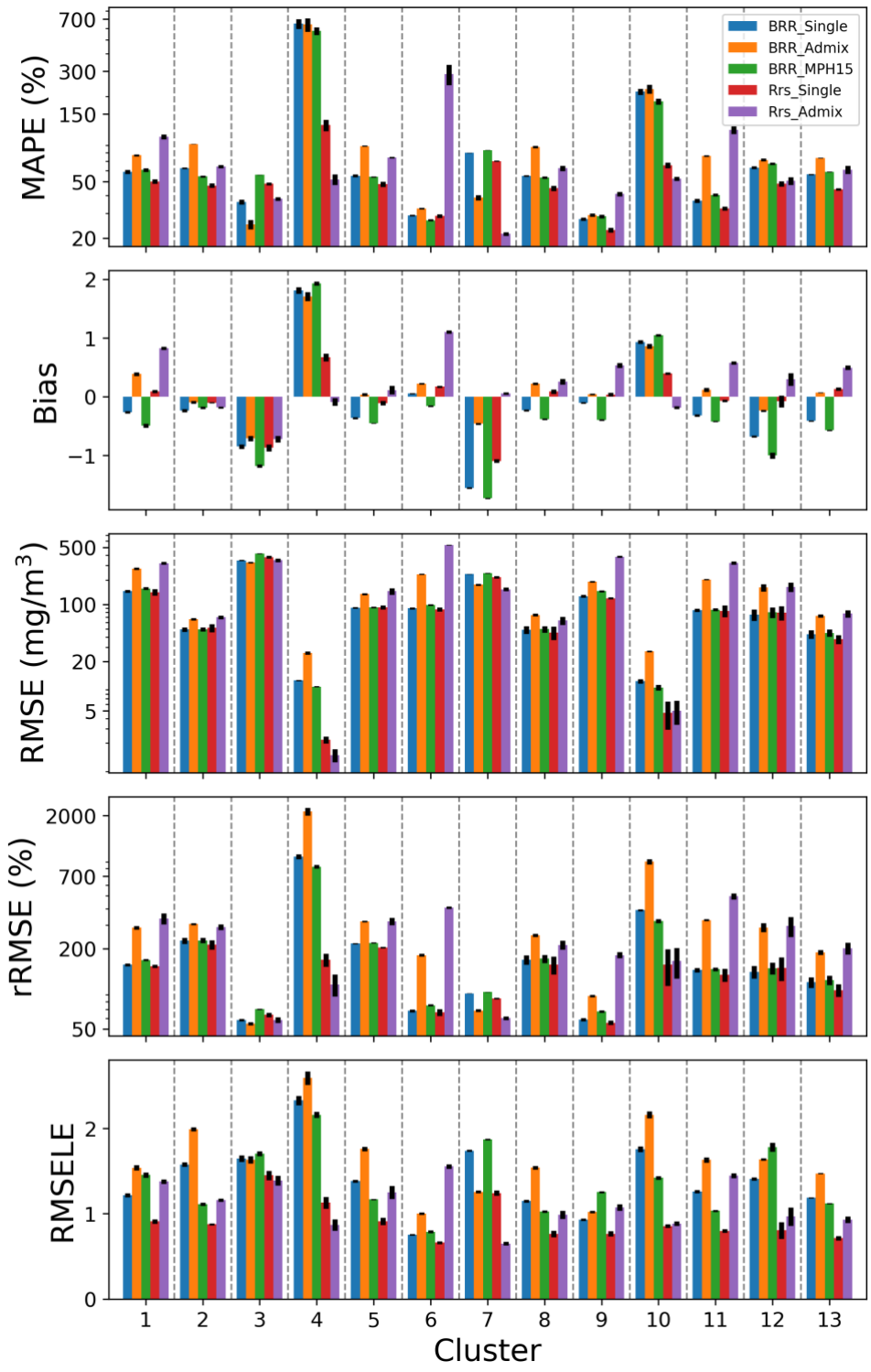

Fig. 18: Predictive statistics for MPH chl-a estimation represented by OWT.

4.4 Conclusion

This study further expanded the R_{rs} synthetic dataset from chapter 3 through the inclusion of a model synthetic atmosphere using MODTRAN radiative transfer code. The atmospheric modeling included dynamic ranges of pertinent atmospheric optical variables derived from the global NASA AERONET database, as well as inclusion of a green vegetation adjacency signal. The full dataset was used to perform an analysis of the fraction of surviving water-leaving signal at TOA. Productive water types with higher PC:Chl-a showed potential to reach surviving L_w fractions of 40% of the total signal when chl-a concentration was as low as 10 mg/m^3, with potential to reach upwards of 60% of the total signal. Highly scattering waters with elevated inorganic particle loads were also capable of reaching TOA L_w fractions of this magnitude. First order calculations of SNR on L_{tot} and L_w for specific OLCI bands on different water types showed that SNR for these water types to produce acceptable levels of error on retrieval products.

The synthetic dataset was also used to analyze various components of the MPH algorithm. An inspection of the peak height switching nature of the algorithm showed that at BRR, for common case 2 water types, the switch from a red/NIR peak height at the 681 nm band to the 709 nm band occurs at chl-a concentrations between 20 – 40 mg/m^3. Although, with elevated PC:Chl-a, this switch can occur at lower biomass. Elevated concentrations of NAP also make this switch hard to predict. An sensitivity analysis of the MPH adjacency flag demonstrated the difficulty for the flag to differentiate between elevated NIR signal due to high phytoplankton biomass or elevated NAP concentration, versus that of contaminate green vegetation signal. An analysis of the MPH cyanobacteria flag displayed overall accuracy of the flag ranging between 60% - 90% over all water types, with better performance in less productive waters where the SICF signal is distinctly different for eukaryotic and

cyanobacteria dominated spectra. Mis-classifications of cyanobacteria dominance tend to increase as biomass increases. Recalibration of the MPH model using the synthetic dataset and inspection by OWT demonstrated the dramatic increase in error for water types with low surviving L_w signal at TOA, although water types with much stronger L_w signal showed better performance using MPH. Using a single calibration for the MPH instead of the standard dual calibrations for cyanobacteria and eukaryotic waters appeared to give better results. Overall, as a simple and fast TOA based empirical model, the MPH algorithm produces satisfactory results for mesotrophic to hypertrophic waters, whereas darker waters with less signal strength have potential to produce considerable errors.

4.5 Appendix A

Further derivation of BRR is as follows:

$$\tau_r(\lambda) = P_{atm} / 1013.25 * 0.008569\lambda^4(1 + 0.0113\lambda^{-2} + 0.00013\lambda^{-4}) \quad (4.6)$$

$$P_r = \left(0.75 * (1 + cos^2_{-}\theta)\right) + \left(\rho_{sky}\theta_0 + \rho_{sky}\theta_v\right) + (0.75 * (1 + cos^2_{+}\theta)) \quad (4.7)$$

$$cos^2_{\mp}\theta = {}^{-}_{+}1 * cos\theta_0 * cos\theta_v - sin\theta_0 * sin\theta_v * \cos(|\phi_0 - \phi_v|) \quad (4.8)$$

$$\rho_{sky}\theta = 0.5 * \sin\left(\frac{\theta-\theta_t}{\theta+\theta_t}\right)^2 + \tan(\frac{\theta-\theta_t}{\theta+\theta_t})^2 \quad (4.9)$$

$$\theta_t = \text{asin}\,(1/1.34\sin(\theta)) \quad (4.10)$$

Performance metrics used based on a confusion matrix are as follows (in the context of this study):

Precision or Positive predicted value (PPV) = TP / TP + FP,

which calculates the probability that when the flag raises True, and the real result is True (e.g. Among pixels which were classified as cyanobacteria, the probability of it actually being cyanobacteria is 70%).

Negative predicted value (NPV) = TN / TN + FN,

which calculates the probability that when the flag raises False, and the real result is False (e.g. Among pixels which were classified as not cyanobacteria (Eukaryotes), the probability of it actually not being cyanobacteria is 60%).

Recall, sensitivity or true positive rate (TPR) = TP / TP + FN,

Calculates out of the proportion of pixels which are actually True, the probability that they are predicted True (e.g. The percentage of cyanobacteria pixels which were correctly classified as cyanobacteria was 80%).

Specificity or true negative rate (TNR) = TN / TN + FP

Calculates out of the proportion of pixels which are actually False, the probability they are predicted False (e.g. The percentage of non-cyanobactria pixels which were correctly classified as not cyanobacteria was 50%).

Overall Accuracy = TP + TN / TP + TN + FP + FN

Which simply calculates the how often the classifier is correct (True or False).

F1- positive = 2 x (PPV x TPR / PPV + TPR)

Which takes both PPV and TPR into account and is the harmonic mean between the two.

F1-negative = 2 x (NPV x TNR / NPV + TNR)

Which takes both NPV and TNR into account and is the harmonic mean between the two.

4.6 Appendix B

Table B1: Mph results

Dataset	***Data type***	***Calibration***	***RMSE (mg/m³)***	***RMSE (%)***	***RMSE (log)***	***Bias (log)***	***MAPE (%)***
Total	Rrs	Admix	202±9.7	152±12	1.34±0.012	0.65±0.01	82±0.9
		Single	201±8.1	153±11	1.42±0.015	0.73±0.018	82±1.8
	BRR	Admix	340±24	258±18	1.63±0.004	0.88±0.005	113±0.58
		Single	333±26	253±19	1.67±0.003	0.95±0.004	117±1.8
		MPH15	357±12	279±5	1.74±0.006	-0.12±0.004	68±0.32
chl-a < 1000	Rrs	Admix	109±4	130±5.8	1.34±0.008	0.65±0.01	81±0.44
		Single	111±4.1	135±6	1.46±0.015	0.76±0.01	84±1.8
	BRR	Admix	109±1	131±0.7	1.56±0.005	0.82±0.004	100±1
		Single	110±1.1	133±0.8	1.54±0.002	0.80±0.002	92±0.35

		MPH15	140±1.2	169±0.8	1.73±0.02	-0.08±0.002	67±0.4
chl-a < *300*	Rrs	Admix	51±0.4	100±1.4	1.25±0.01	0.58±0.01	65±1
		Single	50.5±1.47	99±2.7	1.31±0.007	0.59±0.013	65±1.2
	BRR	Admix	53±0.2	104±0.5	1.38±0.006	0.65±0.007	72±0.32
		Single	52±0.2	103±0.5	1.38±0.005	0.62±0.004	70±0.8
		MPH15	96±1.7	190±3.4	1.64±0.02	0.05±0.004	67±0.4

Table B2: calibration coefficients

Datatype	***Calibration***	***a0***	***a1***	***a2***	***a3***	***a4***	***a5***
Rrs	Single	24.996	13.653	3.724	0.5012	0.0319	0.000778
	Euk	14.35	1.75	0.04			
	Cyano	15.198	4.363	0.5947	0.0305		
BRR	Single	21.033	16.25	6.87	1.453	0.147	0.00575
	Euk	8.0	-0.358	-0.18			

Cyano	11.233	3.117	0.493	0.031	-0.00059

5

5 CHAPTER 5: THEORETICAL APPLICATION OF MACHINE LEARNING MODELS FOR GLOBAL WATER QUALITY RETRIEVAL USING EARTH OBSERVATION DATA

5.1 Introduction

The widespread increase of lake cyanoplankton blooms iscausing global eutrophication to intensify (Ho et al., 2019). The substantial increase in eutrophication can potentially increase methane from these systems by 30-90% over the next century, substantially contributing to global warming (Beaulieu et al., 2019). While previous chapters in this thesis have reviewed and explored the applicability of empirical or semi-analytical chl-a retrieval models, Machine learning (ML) and deep learning (DL) approaches are quickly becoming recognized as the new state-of-the-art in terms of classification and regression type problems, of which remote sensing is ideally suited (Ma et al., 2019, and references therein). The majority of ML and DL development and application has been within the terrestrial remote sensing community (Ghorbanzadeh et al., 2019; Maxwell et al., 2018; Ball et al., 2017; Li et al., 2018), although recent research suggests the benefit of ML and DL approaches for aquatic purposes (Pahlaven et al., 2020; Balasubramanian et al., 2020; Watanabe et al., 2020; Sagan et al., 2020; Peterson et al., 2020; Hafeez et al., 2019; Ruescas et al., 2018). While these studies generally found better performance over traditional empirical or semi-analytical methods, most note that the advanced models were trained on too few datapoints, and would greatly benefit from expanded datasets. DL architectures substantially benefit from greater volumes of high-quality training data. Vastly more coincident reflectance – biophysical parameter pairs, PC in particular, are required to train new and improved multi-parameter inversions for synoptic image analysis at global scales.

While Pahlevan et al., (2020) and Balasubramanian et al., (2020) presented highly convincing results for the transition to ML based models for aquatic particle retrievals using multi-spectral sensors, the authors note that adequate atmospheric correction of top of atmosphere (TOA) radiances to

bottom of atmosphere (BOA) reflectances continues to be one of the largest hurdles to robust, operational space-based water quality retrievals. Baseline type algorithms, which have proven to be robust estimators of trophic status, and relatively insensitive to poor atmospheric correction, have been utilized on partially corrected bottom-of-Rayleigh reflectance (BRR) in attempt to bypass the requirement for a full atmospheric correction (Matthews et al., 2012; Binding et al., 2011; Palmer et al., 2015c). This is indeed helpful for smaller water bodies where atmospheric correction induced uncertainty remains very high (Kravitz et al., 2020). Chapter 4 explored the feasibility of using TOA data over productive inland waters and found that situations involving elevated phytoplankton biomass, PC:Chl-a ratios, or inorganic particles, the fraction of surviving water-leaving radiance reaching TOA can be upwards of 50%, and even higher in more extreme cases. Thus, it follows that ML type models should also perform quite adequately when utilized on TOA data for inland water pixels. However, relatively few studies exist which investigate the actual fraction of the isolated water-leaving signal that reach the satellite sensor over productive inland water bodies.

This chapter aims to explore the potential for developing quick, robust multi-parameter aquatic retrieval models for both multi-spectral and hyper-spectral sensor specifications using a combined synthetic data and ML approach for productive inland waters. The development of the synthetic dataset is described in the previous chapters. Novel aspects of the synthetic dataset include: 1) physics-based, two-layered, size and type specific phytoplankton IOPs for mixed eukaryotic/cyanobacteria assemblages, 2) calculations of mixed assemblage chl-a fluorescence, 3) modeled PC concentration, 4) and paired sensor-specific TOA reflectances which include optically extreme cases and contribution of green vegetation adjacency. The objectives of this chapter are 1) to train and validate four current, and commonly used ML models for retrieval of common water quality parameters using synthetic BOA and TOA reflectances for several multi and hyperspectral

sensors, and 2) qualitatively asses the spatial integrity of models when applied to multi-spectral imagery. This chapter follows with a brief description of the ML models used, a description of the training and cross-validation of the models, an evaluation of the predictive capabilities of the models over multiple water types and sensor configurations, and assessment of the spatial integrity and product consistency between sensors when applied to real-world imagery. The description and development of the synthetic dataset can be found in chapters 3 and 4.

5.2 Methods

5.2.1 Machine learning models

5.2.1.1 K-nearest neighbors

The K-nearest neighbor (KNN) algorithm (Altman, 1992) is a non-parametric, lazy learning model, that can be used for regression and classification. The model is "lazy" in that all training data is used in the testing phase. This allows for faster training times, however slower and costlier testing and prediction. The core of the KNN model is based on identifying similarity between datapoints. This is done by calculating distance or proximity of all points to each other, and assuming similar data points are close to each other. The model is tuned by choosing the optimal number of K, which defines the number of training samples closest in distance to the new point, followed by a value prediction. How distance between points is calculated can also be defined. KNN has become popular for its simplicity and fast training with minimal tuning, however, predictions take much longer with increasing training data or number of features.

5.2.1.2 Random Forrest

The random forest (RF) algorithm (Ho, 1998; Breiman, 2001) is an extension of the decision tree model, which in simple terms, constructs a series of yes/no questions about the data until an answer is reached and can be used for classification or regression. RF is an ensemble method which instead builds tens to thousands of decision trees based on random sampling of training subsets and features, and averages (or majority voting for classification) all the results for a final product. There are a number of hyperparameters that can be tuned which generally differ in how the questions are formed and how deep the trees are. Training can be computationally expensive with extremely large datasets, however, prediction will be much faster than KNN.

5.2.1.3 XGBoost

The extreme gradiant boosting (XGBoost) framework (Chen and Guestrin, 2016) advances the random forest model by including gradient boosted decision trees. This ensemble method builds new, weak models sequentially by minimizing errors from previous models and increasing the influence of higher performing models (boosting), until no further model improvements can be made. Gradient boosting then uses the gradient descent algorithm to minimize the loss when adding new models. XGBoost runs exceptionally well on tabulated data for classification or regression purposes and has dominated data science competitions in recent years due to its efficiency and power.

5.2.1.4 Multi-layer Perceptron

The multi-layer perceptron (MLP) is a type of classical artificial neural network (ANN) that is capable of learning any non-linear mapping function and can be thought of as a universal approximation algorithm. The fundamental units MLPs are artificial neurons, each with their own weighting and

activation functions. The activation function maps the summed weighted inputs to the output of the neuron. Individual neurons can be merged into networks of neurons, generally in the form of a visible input layer and subsequent hidden layers which include the output layer. The activation function of the output layer constrains the model for the specific type of problem (i.e. regression or classification). With the advent of increasing computational resources, deep multi-layer networks composed of multiple layers of hundreds of neurons can now be constructed for highly complex problems.

5.2.2 Cross validation

5.2.2.1 Model inputs

The primary input to each model are the visible and near infrared channel TOA or BOA reflectances of the specific sensor. The modeled synthetic data were resolved to four multispectral sensor specifications: Sentinel 3 Ocean and Land Colour Imager (S3-OLCI), Sentinel 2 multi-spectral imager (S2-MSI) at the sensor's 60m, 20m, and 10m band, Landsat 8 operational land imager (L8-OLI), the moderate resolution imaging spectroradiometer (MODIS), and a hypothetical hyperspectral configuration based on the hyperspectral imager for the coastal ocean (HICO). The sixth configuration consists of the scores from the first ten Empirical Orthogonal Function (EOF) modes from a singular value decomposition (SVD) of HICO bands as a means of dimensionality reduction, as input to the model in replace of channel reflectances. A list of all channel configurations for model input can be found in Table 1 and visualized in Figure 1.

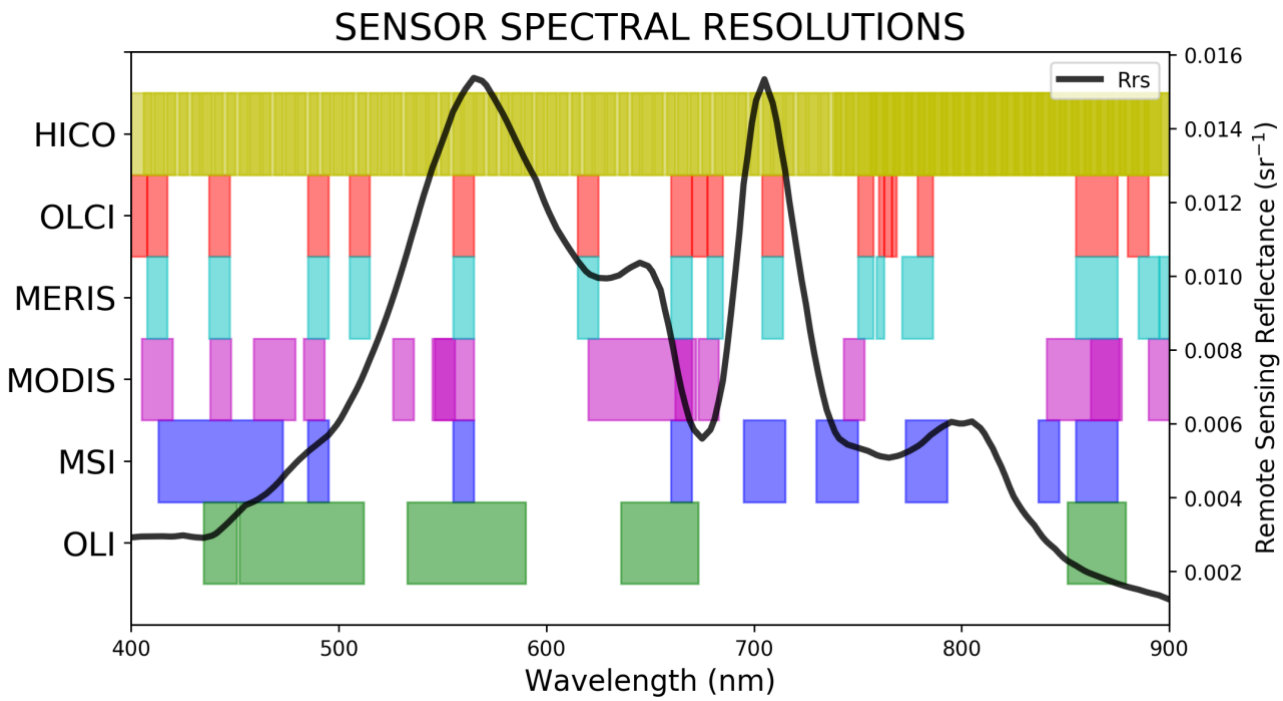

Figure 1: Spectral configurations for four multi-spectral instruments and one hyperspectral instrument, HICO.

Inputs to each model consist of three sets of features: 1) the visible and near infrared (NIR) bands of the specific sensor configuration, 2) the sun and sensor geometry if the model is applied to TOA reflectance, and 3) feature interactions. Feature tuning and extraction can have dramatic effects on resulting model errors or accuracies. Generally, interactions between variables can supplement the individual predictor variables to enhance the feature space to improve the predictive capability of the models. This has been confirmed for aquatic cases (Hafeeze et al., 2019; Ruescas et al., 2018), where including band interactions such as band ratios or line height models have improved model performance. A list of feature interactions used in model training can be found in Table 1. Model outputs are concentrations of chl-a, PC, and NAP in mg/m^3, as well as a_{phy} in m^{-1}, and the optical water type.

Table 1: Inputs for ML models. Inputs are the same for the four ML models used in the study, except for sun and sensor geometries which were only used on TOA models. References for certain feature interactions are located below. References from 1-13: (Gower et al., 2008, Hu et al., 2009, Dall'olmo et al., 2005, Mishra and Mishra, 2012, Gower et al., 1999, Moses et al., 2009, Qi et al., 2014, Matthews et al., 2012, Hunter et al., 2010, Mishra et al., 2013, Liu et al., 2017, Dekker, 1993, Shi et al., 2015)

Sensor	**Bands**	**Geometries (TOA only)**	**Feature Interactions**
L8 OLI	B1, B2, B3, B4, B5	OZA,OAA,SZA,SAA	B4/B3, B4/B2, B4/B1, B3/B2, B3/B1, B2/B1
S2 MSI 10m	B2, B3, B4, B8		B4/B3, B4/B2, B3/B2
S2 MSI 20m	B2, B3, B4, B5, B6, B7, B8, B8A		B5/B4, B5/B3, B5/B2, B4/B3, B4/B2, B3/B2, MCI[1], FAI[2], D3b[3], NDCI[4]
S2 MSI 60m	B1, B2, B3, B4, B5, B6, B7, B8, B8A		B5/B4, B5/B3, B5/B2, B4/B3, B4/B2, B3/B2, MCI, FAI, D3b, NDCI
S3 OLCI	Oa1, Oa2, Oa3, Oa4, Oa5, Oa6, Oa7, Oa8, Oa9, Oa10, Oa11,		FLH[5], MCI, FAI, M2b[6], D3B, NDCI, PCI[7], SIPF[8], H103b[9],

	Oa12, Oa16, Oa17, Oa18		M133b[10], L4b[11], D93[12]
MODIS	B1, B2, B3, B4, B8, B9, B10, B11, B12, B13, B14, B15, B16		FLH, SIPF, FAI, Shi15[13]
HICO	All bands 400-900nm		None
HICO SVD	EOF modes 1-10		None

The R_{rs} dataset contains roughly 70,000 samples, while the TOA reflectance dataset contains roughly 260,000 samples. For each dataset, models were evaluated using cross validation where the data was split into 80% for training and 20% for testing. This was done five times in order to avoid sampling bias. Performance metrics used in the evaluation consist of both linear and log-transformed root mean squared error (RMSE and RMSELE, respectively), relative RMSE (rRMSE), bias, and median absolute percent error (MAPE).

5.2.2.2 Hyper-parameter tuning

In order to obtain results of the highest fidelity possible, ML models require optimization of their respective hyper-parameters before evaluation. The hyper-parameters govern the training process itself and define the model architecture and structure. These parameters are not updated during the learning process and are used to configure the model in various ways. Hyperparameters are not model parameters and cannot be directly trained from the data. In this study, hyper-parameter

tuning was accomplished using grid search, which builds a single model for each possible combination for the range of all possible values for hyperparameters, evaluates each model, and selects the architecture with the lowest mean squared error (MSE) for regression models, or accuracy for classification models. Computational requirements for extensive hyperparameter tuning can be very high, especially when dealing with more complex or deep models. An exhaustive grid search for each model was beyond the scope of this research, however, a brief hyperparameter tuning exercise was performed to optimize each of the models most sensitive hyperparameters. Final model hyperparameters may be found in Table 2.

Table 2: Final hyperparameters used for model training

Model	**Hyperparameter**	**Value**
Random Forest	Decision trees	150
	Samples per leaf node	2
K-nearest neighbors	Leaf size	30
	Minkowski power	2nd
	Neighbors	6
XGBoost	Decision trees	500
	Max depth	6
	Gamma	0.4
	Objective	Reg/squared error
MLP TOA	Layers/Neurons	5 layers:[500,300,200,100,output]
	Activation	ReLu

	Optimizer	Adam
	Loss	MAE
MLP BOA	Layers/Neurons	4 layers:[500,200,50,output]
	Activation	ReLu
	Optimizer	Adam
	Loss	MAE

5.3 Results

5.3.1 Model performance

5.3.1.1 Overall model performance

Evaluation of overall model performance applied to TOA or BOA spectral data, per sensor, can be found in Fig. 2 for retrieval of chl-a, PC, and NAP concentrations, and $a_{aphy}(440)$. The MLP generally outperforms the other ML models in almost every case in terms of MAPE and RMSELE when evaluated against the entire dataset, while KNR generally performs the worst. Although retrievals using R_{rs} performed significantly better than when applied to TOA reflectances in every case, MAPE remained below 30% when applied to both TOA and BOA reflectance for Chl-a, PC, and $a_{phy}(440)$ retrievals for top-performing sensors. RMSELE was found to be less than one for Chl-a and $a_{phy}(440)$ retrievals while ranging between one and one and a half for PC and NAP. S3-OLCI displays considerably better retrieval performance of PC than other multispectral sensors, in line with HICO retrieval performance. The 620 nm band of OLCI, which aligns with the maximum absorption peak of PC, appears to provide a significant advantage over other multispectral sensors for the

quantification of cyanobacteria. Considering the variability of these products within the synthetic dataset, the MLP shows promising predictive capabilities at all trophic states.

5.3.1.2 Performance by sensor

Similar performance relationships were found for each product based on different sensor configurations. Generally, performance showed slight improvement with sensor configurations incorporating more channels. This is more drastic when applied to TOA reflectance, although at BOA, these differences are not significant in most cases when comparing extents of standard deviations from cross validation. The most dramatic discrepancy between TOA and BOA performance occurs when applied to L8 sensor configuration, while differences between the three S2-MSI configurations appear to be quite minimal. Results based on medium resolution configurations also compare reasonably well with model performance from the HICO hyperspectral configuration. Models were also applied to the scores of the first ten modes from SVD of hyperspectral HICO data to compare performance after a dimensionality reduction. Performance was significantly worse in all cases when compared to models applied to the full HICO hyperspectral configuration as well as other medium resolution configurations. Models applied to L8 specifications were worst performing overall relative to the other sensors, however, still provide very promising prediction performance for aquatic retrieval products compared to previous methodologies.

5.3.1.3 Performance by OWT

Considering the wide range of concentrations and IOPs of aquatic products found in productive waters (Spyrakos et al., 2018), model performance is also evaluated by OWT. Fig. 3 displays the performance of the MLP algorithm separated by OWT for Chl-a, PC, NAP, and $a_{phy}(440)$. Fig. 6 only

displays results for the S2-20m and L8 sensor configurations to illustrate general relationships. All predictive statistics for every model, sensor, product, and OWT can be found in supplementary material. The S2-20m configuration, which includes three more channels in the NIR, shows overall better performance in pigment retrievals for more productive water types, while L8 shows better pigment retrieval performance in more oligotrophic cases. This is likely attributed to the inclusion of a band centered at 440 nm for L8, which is only included in the 60m sensor configuration for S2. The retrieval of NAP incurs the highest errors of the four products for both sensors, especially in OWTs with lower mineral content. Largest discrepancies between TOA and BOA product retrievals, understandably occur in cases where surviving L_w signal at TOA is lower, such as in low bulk scattering OWTs 4, 7, and 10.

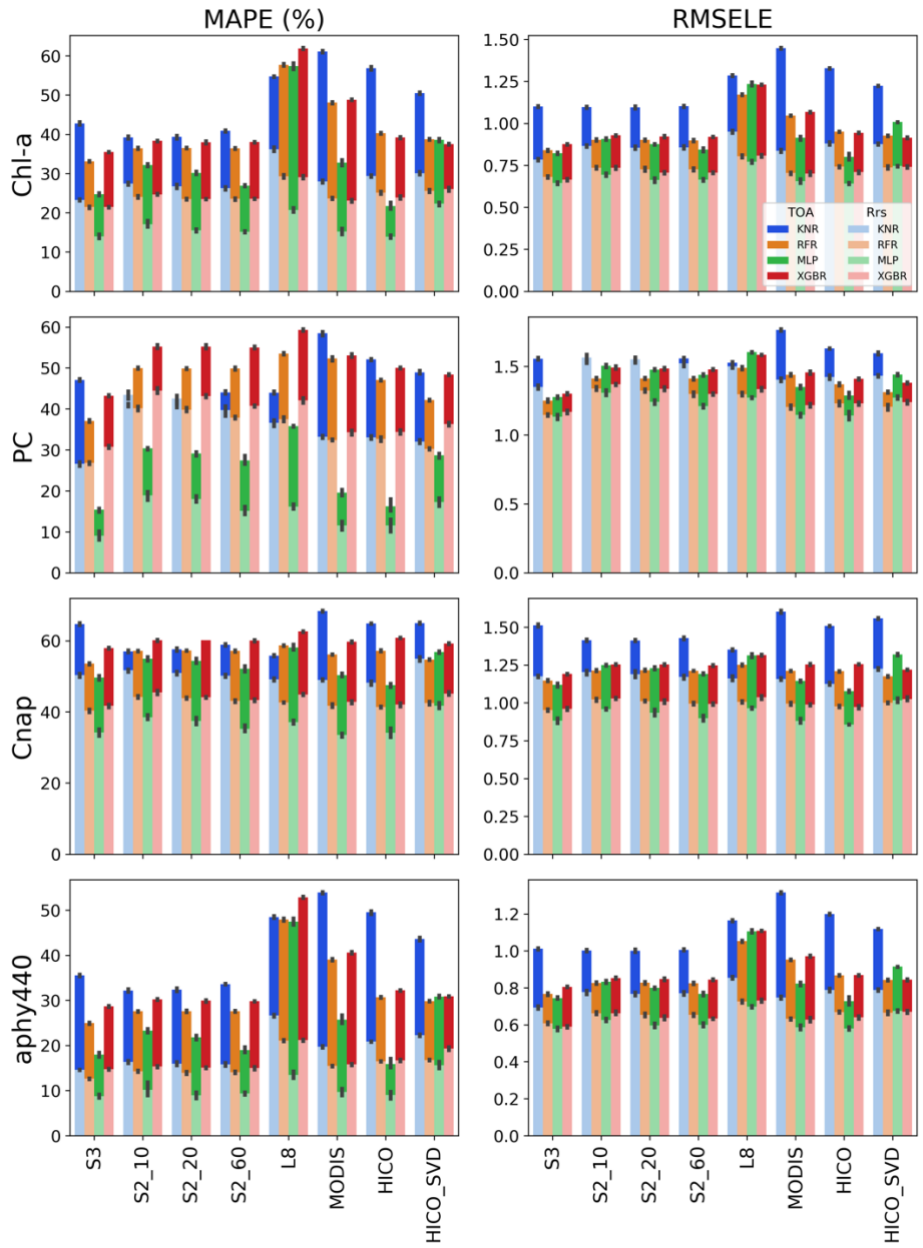

Figure 2: Overall model performance based on MAPE and RMSELE for each sensor configuration at both TOA and BOA. Additional statistics can be found in supplementary material.

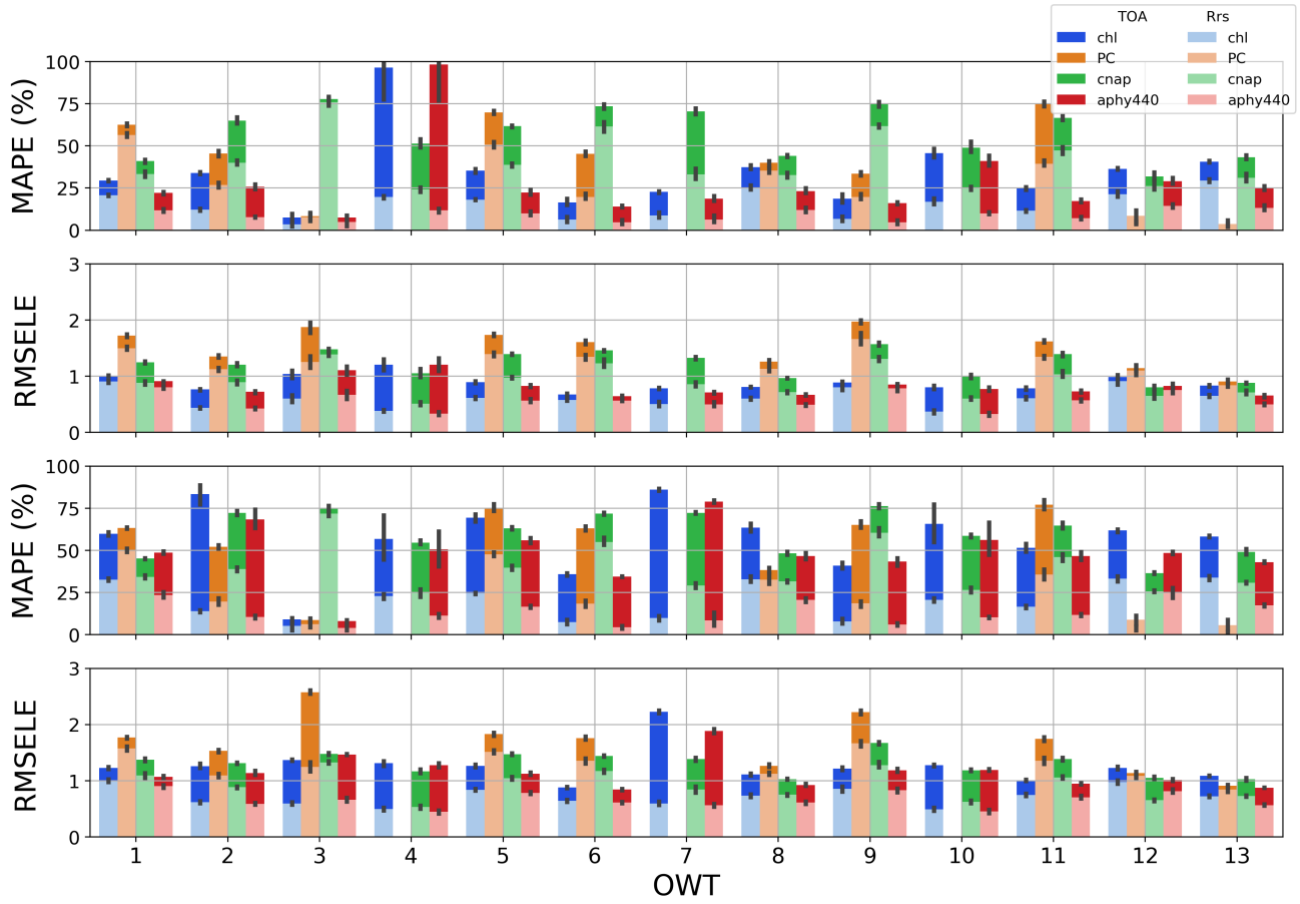

Figure 3: MLP performance by OWT based on MAPE and RMSELE at both TOA and BOA for the S2 20m channel (upper two plots), and L8 (lower two plots).

5.3.1.4 OWT classification

General relationships of the OWT classification results can be seen in Figure 4, using S2 and HICO as an example. Displayed are the heatmaps for overall accuracy of the classification which provides the percentage of the overall correctly predicted classes. In this instance, the threshold is the probability output from the classifier model that a certain spectra belongs to a class, in which above a probability of 0.5, a spectra would be assigned to that class. Accuracy using the S2 10m configuration at R_{rs} was found to be 84%, which was unexpectedly high considering only four vis/NIR bands are used. However, when classifying spectra using TOA reflectance, accuracy drops by about 50% using S2 10m

data. Accuracy increases to 57% and 62% for S2 in 60m configuration and HICO using TOA reflectance, respectively. Overall trends in the results, which include those not shown, display a general increase in accuracy when more spectral bands are used, while high accuracies persist above 80% in all spectral configurations when used on R_{rs} data.

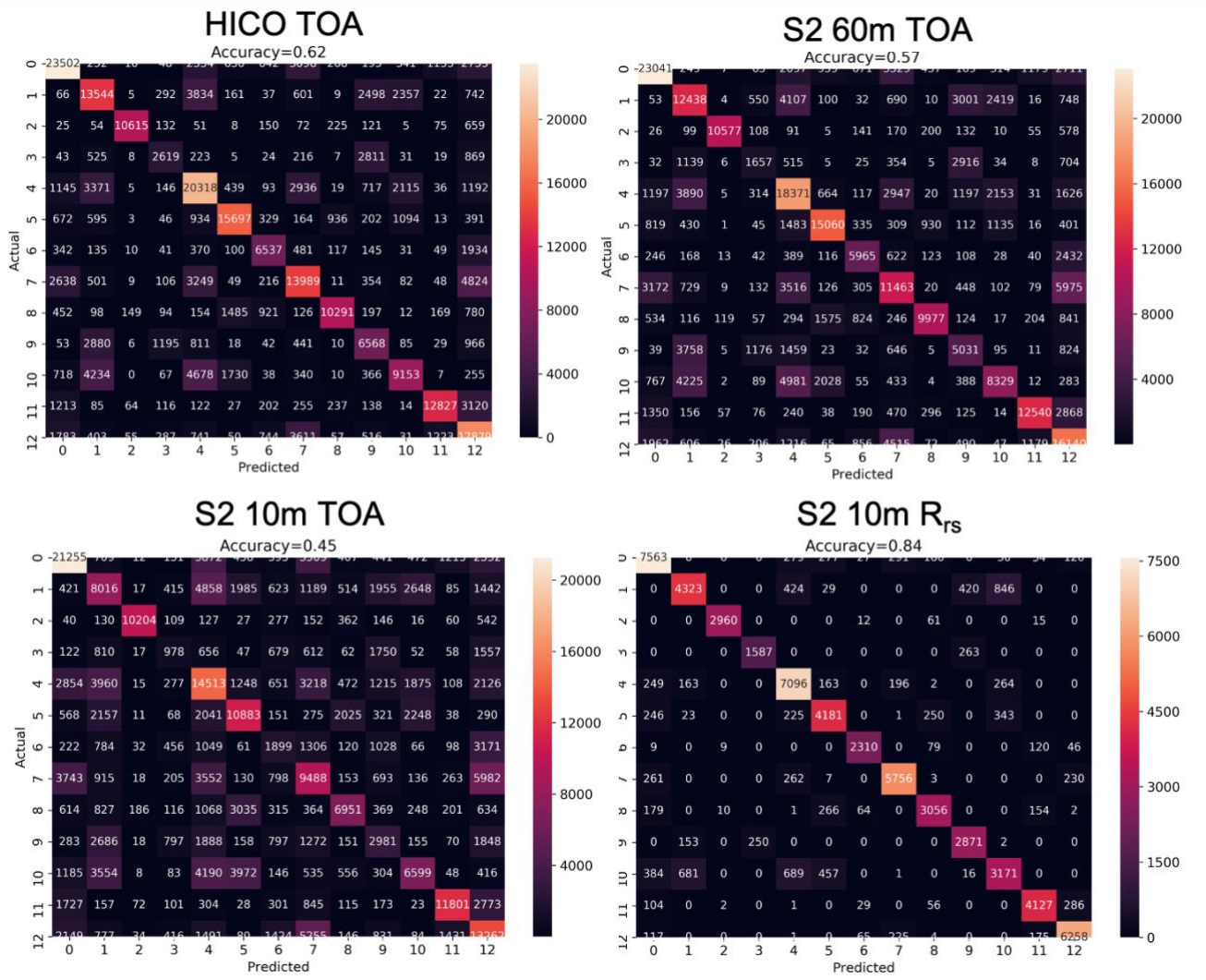

Figure 4: Heatmaps displaying number of correct OWT classifications and overall accuracy.

5.3.2 Qualitative evaluation using EO data

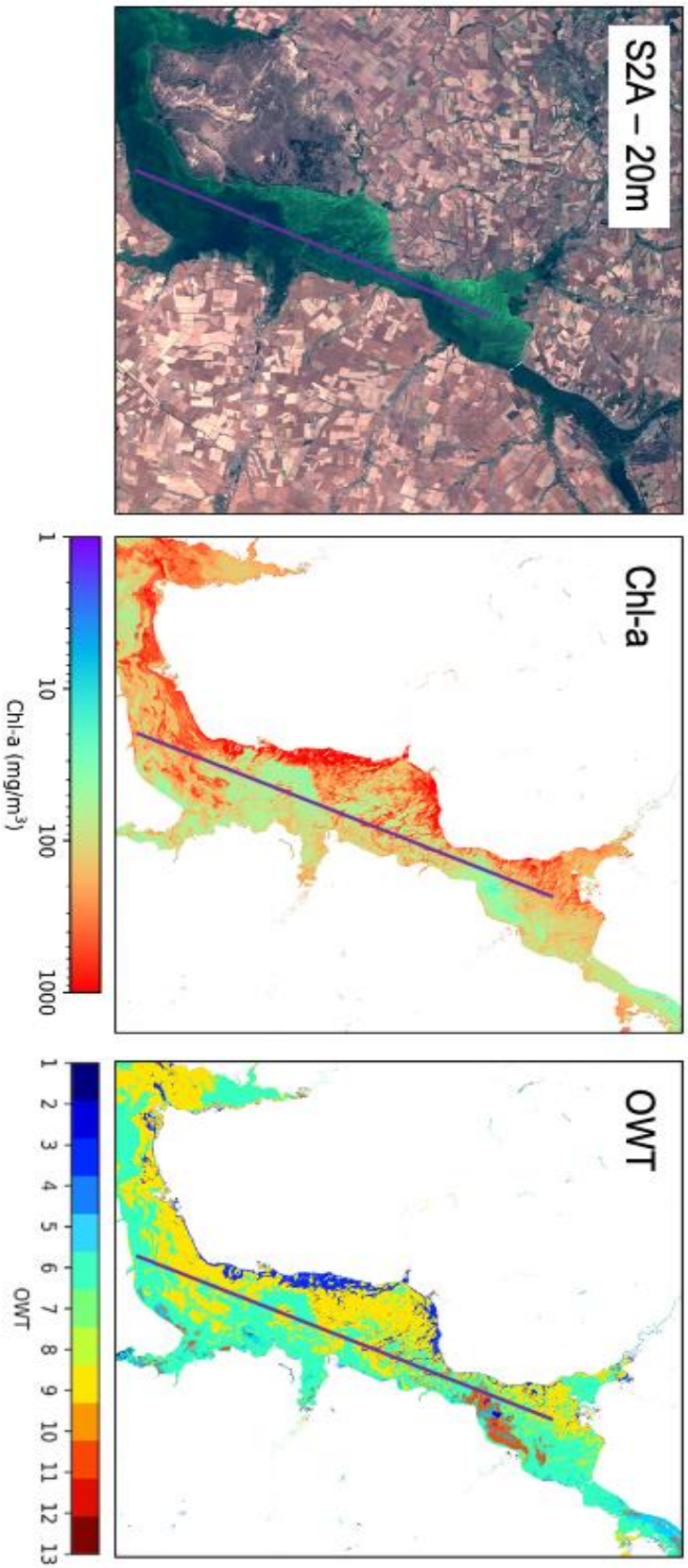

Figure 5: S2A MSI RGB and MLP derived chl-a concentration and OWT from TOA reflectance for Tsimlyansk reservoir, Russia on September 8, 2018. The purple line indicates the transect taken for Fig. 6.

To assess the spatial integrity of retrieval products as well as test cross-sensor consistency, a qualitative examination of productive freshwater scenes was undertaken. Fig. 5 displays a S2-MSI scene of an intense cyanobacteria bloom at Tsimlyansk reservoir in Southern Russia taken on September 8, 2018. The chl-a and OWT products produced from the MLP using TOA reflectances are displayed along with an associated RGB image. Although no validation points are present, the ranges of chl-a concentration conform to expert knowledge of situations of intense cyanobacteria surface blooms, where chl-a concentration is capable of reaching into the thousands mg/m^3. The model also produced OWTs which would be expected given these scenarios. OWT 3 (blue) is produced for areas of highly concentrated surface blooms, which coincides with extremely high chl-a concentrations and the intense green color from the RGB. OWT 9 (yellow) is then displayed representing more subsurface cyanobacteria blooms, of still quite highly elevated chl-a, while the majority of remaining pixels are represented by OWT 6 (cyan) for even milder subsurface cyanobacteria blooms. This can also be visualized in the RGB as fading of the intensity of the green color, where the absorption of the water becomes stronger again due to less phytoplankton biomass. Atmospheric correction over intense bloom waters such as these are error prone and can lead to large uncertainties in retrieval products (chapter 2). Strong water-leaving signal at TOA allows for very reasonable product estimates, with no atmospheric correction involved.

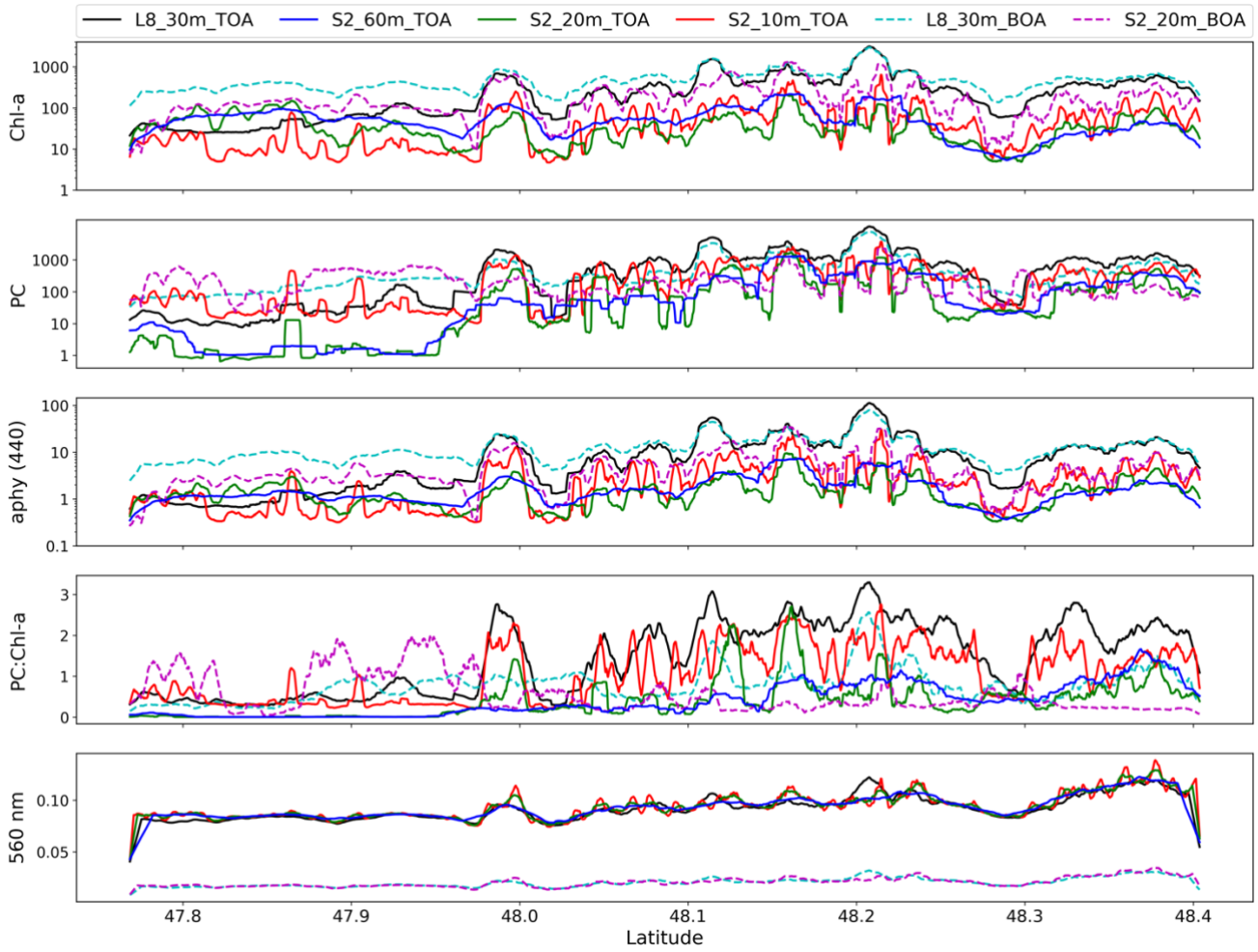

Figure 6: Transect of MLP derived products from Tsimlyansk reservoir, Russia on September 8, 2018 (purple line in Fig. 5).

A near coincident L8 scene was also captured for the same day at Tsimlyansk reservoir. As a means to assess cross-sensor product consistency, a transect was taken from North to South through the center of the main basin of the reservoir (purple line in Fig. 5). Figure 6 displays the product retrievals along the transect for L8 at standard 30m resolution and S2 in its 10m, 20m, and 60m resolution sensor configurations. Product retrieval from atmospherically corrected reflectance for L8 and S2 at 20m are also presented, produced from R_{rs} using the ACOLITE dark spectrum fitting scheme (Vanhellemont, 2019). Chl-a, PC, the absorption due to phytoplankton at 440 nm, and the PC:Chl-a

ratio are displayed along with the 560 nm band of both sensors. The transect demonstrates the spatially dynamic nature of these water types where swirls of changing surface and subsurface blooms create dramatic swings in pigment concentrations over small spatial scales. A general smoothing along the transect is evident when comparing larger to smaller spatial resolutions, however, the multiple sensor configurations follow similar patterns and relationships. L8 is generally over-estimated when compared to the other configurations, with both TOA and BOA products. BOA products also tend to over-estimate relative to TOA products, more than likely from atmospheric under-correction from ACOLITE. However, while overestimated, it is still promising to see similar patterns in BOA products, as TOA. Ranges of $a_{phy}(440)$ are similar to those found in literature for such water types, and conform to similar relationships with chl-a as found in nature (Matthews et al., 2013). Some sizeable discrepancies exist for the PC:chl-a ratio, with S2-10m and L8 producing generally higher ratios than their higher spectral resolution counterparts. Although, ranges of calculated PC:chl-a are also similar to those found in nature, where intense cyanobacteria blooms generally exhibit a ratio > 1. The transect of the 560 nm band is also plotted for each configuration as a means to compare to product variation and sensor consistency and validates the smoothing occurring in the products at different spatial resolutions.

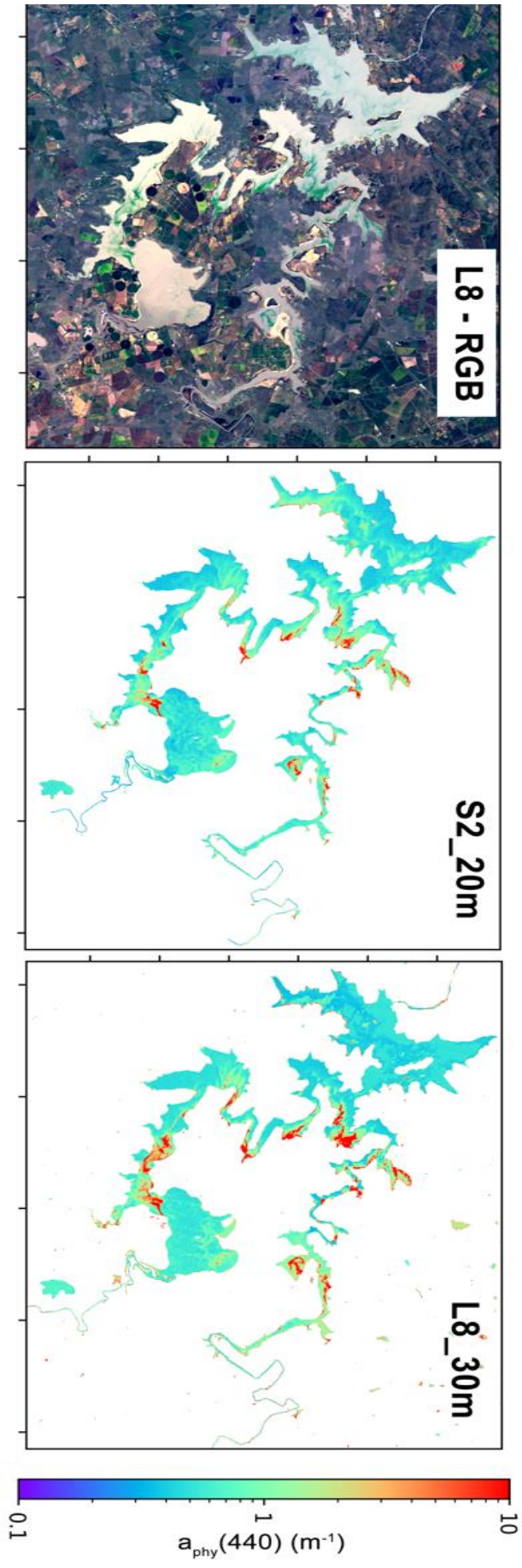
L8 - RGB
S2_20m
L8_30m
0.1
1
10
$a_{phy}(440)$ (m^{-1})

Figure 7: MLP derived $a_{phy}(440)$ for S2-MSI at 20m resolution and for L8 for Vaal dam, South Africa on April 6, 2020.

Similar cross sensor consistency between S2 and L8 can be found in Fig. 7 of $a_{phy}(440)$ over Vaal Dam in South Africa using TOA reflectance. This reservoir is a major water supply for the densely populated Gauteng Province, and regularly exhibits intense cyanobacteria surface blooms. This reservoir has proven to be particularly problematic for chl-a retrievals due to the spatially dynamic nature of surface blooms combined with relatively high mineral loads (chapter 2). The two products are surprisingly consistent given that S2 at 20m resolution utilizes three more spectral bands throughout the NIR. The L8 scene was taken roughly 45 minutes after S2, however, the spread and intensity of the bloom can already be visualized in the productive regions, which was corroborated by comparing the two RGB scenes (only L8 shown).

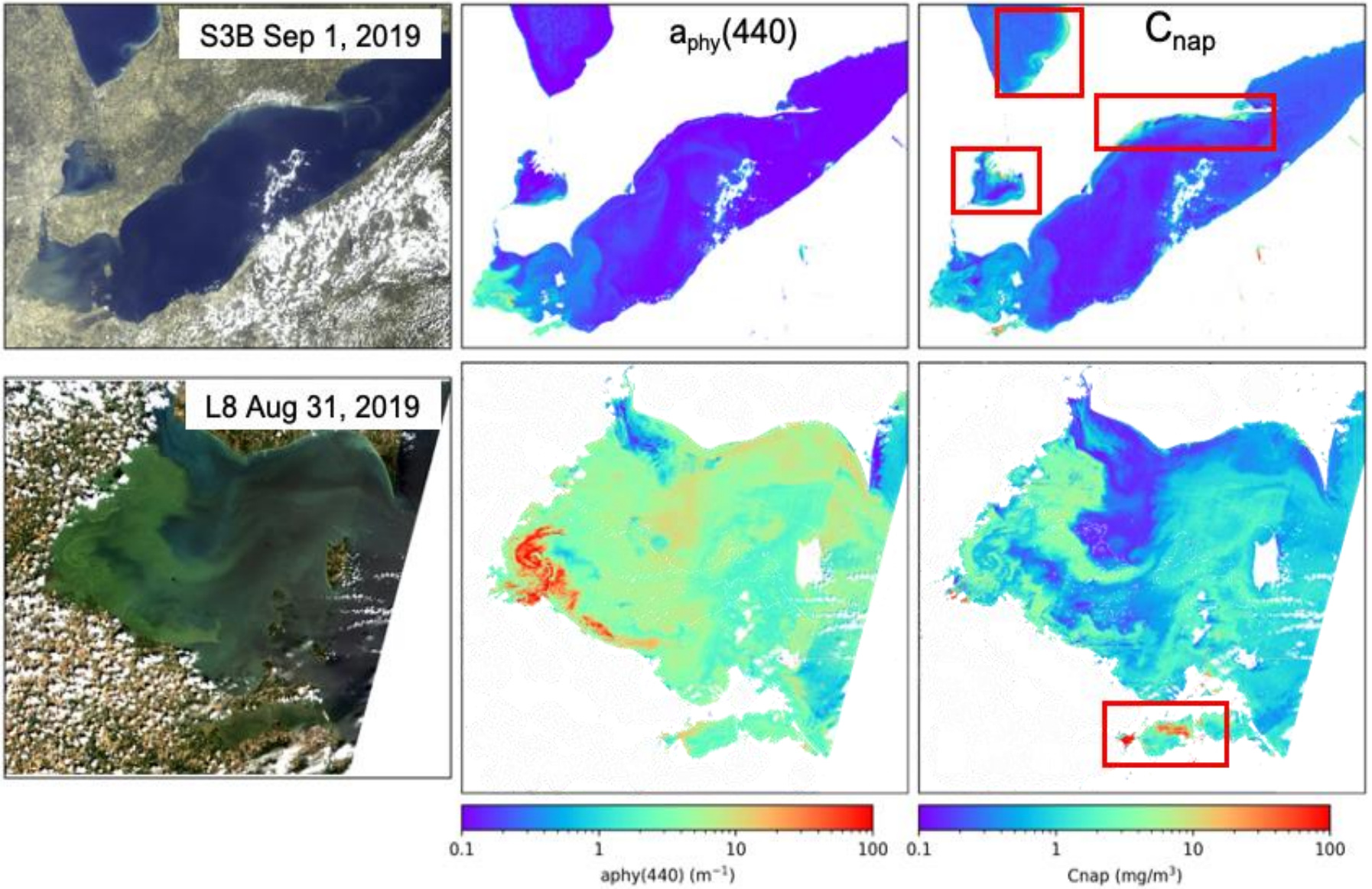
S3B Sep 1, 2019
$a_{phy}(440)$
C_{nap}
L8 Aug 31, 2019
0.1
1
10
100
aphy(440) (m^{-1})
0.1
1
10
100
Cnap (mg/m^3)

Figure 8: MLP derived $a_{phy}(440)$ and C_{nap} from TOA reflectance from S3B and L8 during an intense cyanobacteria bloom at Lake Erie, USA. Regions outlined in red indicate areas of elevated mineral concentrations.

A fairly common issue relating to product retrievals in turbid and productive waters is the general overestimation of phytoplankton biomass in the presence of higher inorganic sediment loads (Zeng et al., 2019, Heironymi et al., 2017). Fig. 8 displays a cyanobacteria bloom in the western part of Lake Erie on August 31, 2019, captured by L8, and then S3B on the following day. The associated map of phytoplankton absorption at 440 nm, produced with the MLP at TOA, produces a realistic gradient of $a_{phy}(440)$ from the intense surface swirls of cyanobacteria at the far western edge, to the more oligotrophic waters coming down from the northern lakes. Certain areas are highlighted in the adjacent map of non-algal particles, depicting regions of elevated mineral content that are reasonably accounted for by the MLP model, and sufficiently decoupled from phytoplankton absorption. Fig. 9 illustrates another spatially dynamic scene caught by S2 over Roodeplaat dam in South Africa. A few near-coincident validation points were obtained of chl-a from a fieldwork campaign from Kravitz et al., (2020). Exceptional consistency is evident between the 60m and 10m MLP products using TOA reflectance despite all nine visible and NIR channels being used at 60m, while only four channels are utilized at 10m. Validation points also align with very well with the chl-a product.

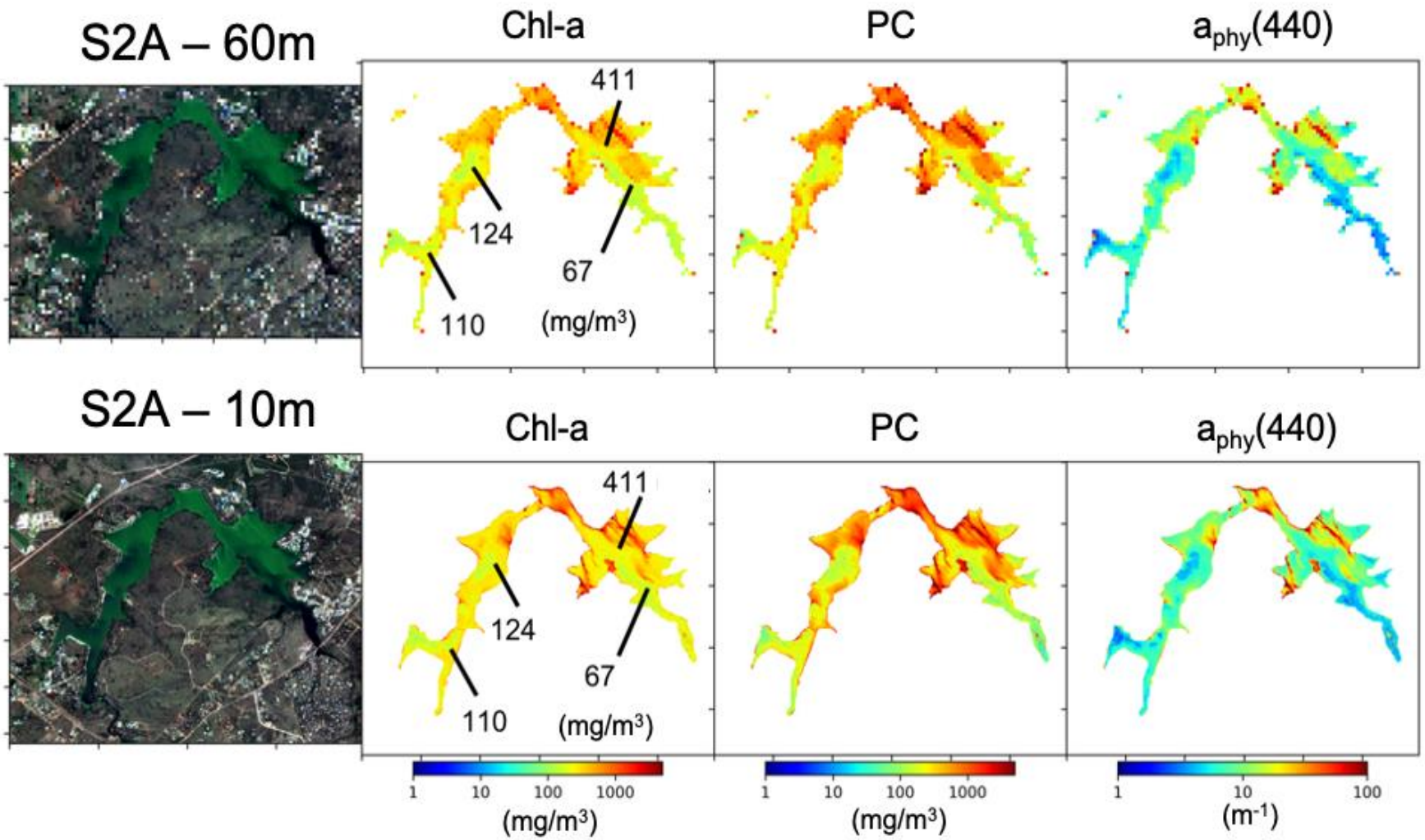
S2A – 60m
Chl-a
PC
$a_{phy}(440)$
411
124
67
110
(mg/m³)
S2A – 10m
Chl-a
PC
$a_{phy}(440)$
411
124
67
110
(mg/m³)
1
10
100
1000
(mg/m³)
1
10
100
1000
(mg/m³)
1
10
100
(m^{-1})

Figure 9: MLP derived products from S2A at 60m and 10m channel configuration over Roodeplaat dam, South Africa on March 23, 2017.

5.4 Discussion

5.4.1 Machine learning models

Four out-of-the-box ML models were trained using synthetic data and implemented on EO data all within the Python programming language. It is important to note that the aim of this study was not to produce an optimal, finalized retrieval model for operational use, but rather explore the capability of a range of well documented ML models to make adequate predictions of water quality variables, trained from synthetic data. ML has shown to be an extremely powerful tool which is now more accessible, and easier to implement than ever before. The models used in this manuscript were trained with minimal hyperparameter tuning, as conducting an exhaustive parameter tuning exercise for every trained model explored in this study would be very computational expensive. All of the analysis conducted for this research was performed on a personal laptop with 16 GB of RAM. This study confirmed other reports of ANNs outperforming other "shallow" ML models such as decision trees or support vector machines (SVM) (Hafeez et al., 2019; Peterson et al., 2018). Other ML techniques utilized in recent aquatic work such as feature fusion (Peterson et al., 2019) was also implemented to a degree in this study. Multiple "feature interactions" in the form of band ratios or line height models were included in model training along with sensor visible and NIR bands. Ruescas et al., (2018) found increasing model performance by including more feature interactions for a ML model for CDOM retrieval. Although the results are not shown here, a subset of ML models in this work were trained with and without the inclusion of feature interactions with significant increase in performance when included and thus was decided to include them for all models.

Pahlaven et al., (2020) and Balasubramanian et al., (2020) found that a mixture density network (MDN), which is essentially an ANN, except the final layer is mapped to a mixture of distributions, produced extremely robust results for chl-a and suspended solid material. MDNs would theoretically be the most optimal choice, as you can still design a highly efficient deep neural network (DNN), however, address the signal ambiguity problem of optical remote sensing through the addition of a mixture of parametrized Gaussians. This approach was also attempted in this study, however, was taking considerably longer for training and cross validation, and producing roughly similar results as the MLP model, thus discarded for the purposes of this research. Future work, with access to higher computational resources, would include training of deeper neural networks and the inclusion of mixture distributions. Shallow ML models such as Random Forest and XGBoost still provide highly adequate results, although require much less parameterization and computational resources. It is also important to note that these models were both trained and validated only using the synthetic database. Future work will entail validating products against available in-situ data. It is expected that performance will drop somewhat when validated against fieldwork data due to spatial inconsistencies and uncertainty due to fieldwork methods.

5.4.2 Product integrity and consistency

Pahlevan et al., (2020) notes atmospheric correction to still be one of the major challenges for operational inland and coastal remote sensing. This work explored the capability of product retrievals from TOA reflectances. Chapter 4 illustrated how water types for turbid or productive inland waters have substantially higher percentages of surviving water-leaving radiances reaching the satellite sensor than oligotrophic waters, or highly absorbing waters dominated by dissolved organic matter. The separation of MLP performance by OWT in Fig. 3 confirmed water types with stronger scattering signal have smaller discrepancy between TOA and BOA retrieval results. An OWT based framework

could be used to run atmospheric correction only on oligotrophic pixels where AC processors are more ideally suited, with product retrievals made from TOA in more productive water types. Due to uncertainties still inherent in current AC processors, especially for smaller water bodies, product maps in this manuscript are mostly made from TOA data. Although, promising AC processors have been developed using a combined synthetic data/NN approach (Fan et al., 2017), and the dataset developed in this research could indeed be used to train an AC.

The adjacency effect (AE), where strong spatial heterogeneity from surrounding terrestrial sources contaminates the water signal, has the potential to induce considerable errors in retrieved products (Bulgarelli et al., 2017). Incorporating contamination by green terrestrial vegetation at TOA in the synthetic modeling was done in attempt to mitigate this issue. Theoretically, with enough training data, a properly optimized deep ANN should then be able to account for this. Fig. 10 shows a S2 scene over a small dam in South Africa which was found to be affected by considerable adjacency in chapter 2. The figure presents the predicted contribution of adjacency to water pixels based on a simple random forest model trained using the synthetic dataset. The plot produces realistic gradation of increasing adjacency contribution towards the edges of the dam in the darker waters, as well as in instances near bright surface cyanobacteria blooms. Areas of intense algal surface bloom would be less affected by green vegetation adjacency since they exhibit similar reflectance patterns in the red and NIR, and would themselves be potentially contaminating nearby “less bright” water pixels. The fact that the model demonstrates acceptable relationships of the adjacency effect, gives confidence that other retrieval products would be inherently corrected for this effect. Future work would incorporate more sources of adjacency, and could include other

sources of signal contamination such as sun glint.

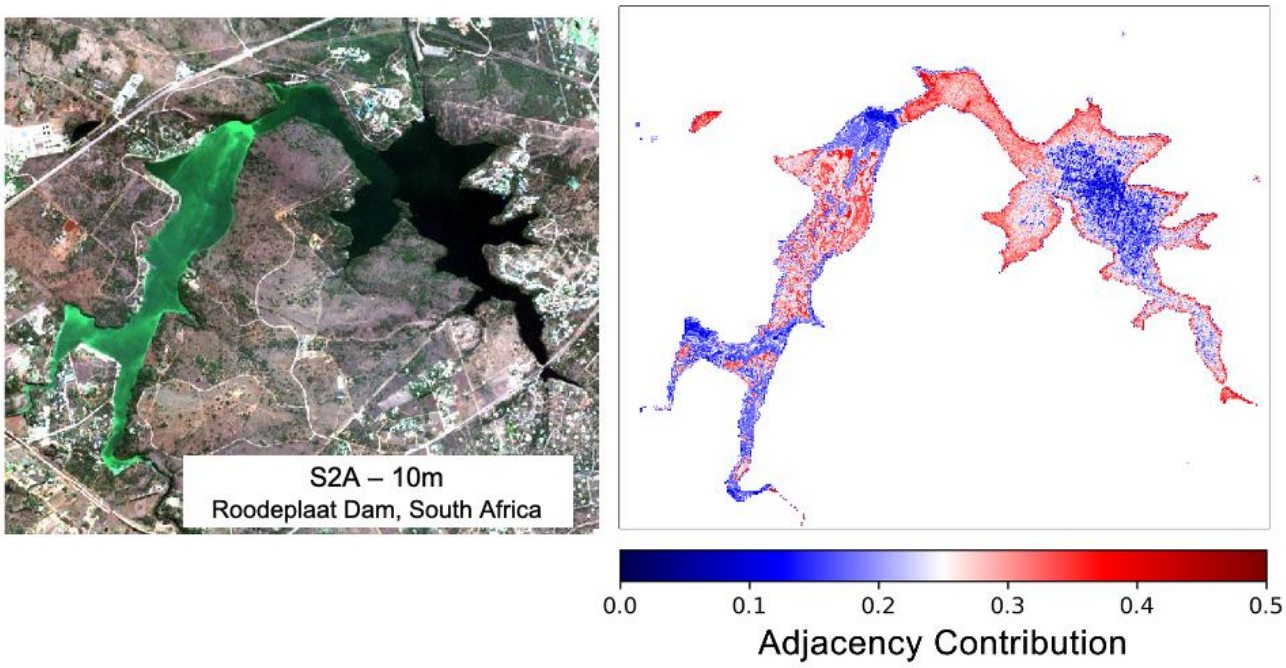

Figure 10: Percent adjacency contribution derived using RF regression over Roodeplaat dam, South Africa using TOA reflectance data.

The water bodies examined in this manuscript have the potential to experience high spatial and temporally dynamic blooms. Sensor requirements for operational monitoring of these waters are recommended to be <60m spatial resolution and daily to tri-weekly revisit times (Muller-Karger et al., 2018; Hestir et al., 2015, Mouw et al., 2015). The transect in Figs. 5 and 6 demonstrates the fine scale spatial distributions of cyanobacteria blooms. MLP products at different spatial resolutions demonstrate how spatial smoothing from just 10-60m can cause significant differences in product retrievals, which was corroborated by inspecting the 560 nm band of the various sensor configurations. Extreme temporal dynamics can additionally be visualized in Fig. 7 where a difference in scene acquisition time of just 45 minutes was able to capture the start of a spreading bloom. Likewise, the Lake Erie scene in Fig. 8 captured a diminishing bloom, albeit at substantially different

spatial resolutions. For such dynamic waters, it is apparent how cross-mission capability could provide substantial benefit for operational monitoring of harmful algal blooms.

5.4.3 Outlook

While this study acts more as a proof-of-concept rather than finalized products, the results suggest the capability of using a synthetic dataset and ML approach to develop operational global freshwater monitoring products. Expansion of the synthetic dataset by incorporating more diverse phytoplankton IOPs and other sources of signal contamination would be the logical next step. While the amount of synthetic data developed in this research (~260,000 TOA spectra) is still quite small with respect to current advancements in Big Data analytics, access to high powered computing resources would allow the development of extremely large synthetic datasets into the tens and hundreds of millions of data points for which to train advanced deep learning networks. Validation of models using global in-situ datasets would then be the final step to compare product outputs trained from synthetic data to outputs trained on in-situ data as in Pahlaven et al., (2020) and Balasubramanian et al., (2020). Although, it is very promising to see model performance described in this research relating very well to results in the aforementioned studies. Further research would also include parameterized sensitivity studies identifying most optimal spectral and radiometric resolutions that ML can exploit. This study suggests that both L8, and S2 at its various sensor configurations, contain enough spectral information to produce reasonable estimates of various aquatic products for productive water bodies. Highly consistent product outputs were found between S2 at 60m and 10m resolutions (Fig. 9), considering the benefit of five additional NIR spectral bands at 60m. This could have major implications on future sensor design, where more resources can be instead invested in increasing SNR or spatial resolutions of sensors, while spectral resolution can remain fairly low, at least for the water types depicted in this research. The research suggests that

relevant bands for assessing wide ranging trophic levels should at least include a short wavelength blue band around 440 nm as in L8 for more oligotrophic instances and highly absorbing scenarios, a band around 620 nm to aid in cyanobacteria detection and quantification and a band in the red edge around 710 nm to capture the phytoplankton scattering peak.

5.5 Conclusion

Four types of current ML architectures were tested and trained using the synthetic data with an ANN providing the most optimistic results for multi-parameter retrieval from different sensors. Application to EO imagery provided realistic gradients of concentrations and phytoplankton absorption for wide ranging trophic scenarios for small inland water bodies using TOA reflectance data. While this research only examined a subset of possible products, which already has never been done at such fine spatial scales in such dynamic water bodies, more complex NNs will allow for the derivation of additional functional types such as algal size, cyanobacteria type classification, surface scum classification, aquatic macrophyte classification, absorption and scattering properties for all optical constituents, as well as deriving relevant atmospheric parameters such as aerosol optical thickness. With a combination of current and past sensor spatial resolutions ranging from 10 m to 4 km scales, a synergistic evaluation of water constituents, with known uncertainties by OWT, will allow for an unprecedented global snapshot of fine scale ecological dynamics of coastal and inland waters. This synthetic dataset acts as the first step towards this goal. It is by no means, a fully comprehensive inclusion of all possible natural values and relationships found in inland waters, however, works as a proof-of-concept to show the capability of these techniques to create accurate simulations of real-world aquatic environments.

6

6 SUMMARY AND CONCLUSIONS

This body of work contributes advancements towards the field of inland water remote sensing. The application of Earth Observation imagery to retrieve high fidelity water quality products from global inland water bodies has long been a goal for environmental remote sensing scientists. Significant progress has been made over the last two decades, but substantial limitations still exist preventing operational inland water retrievals. This thesis outlines and demonstrates many of these limitations as well as proposes a solution to the many problems facing inland water remote sensing.

This thesis will be summarized by re-addressing the aims laid out within the introduction:

Aim 1: Chapter 2 provided one of the first comprehensive validations of common atmospheric corrections and chl-a retrieval algorithms for Sentinel-3 OLCI for small inland water bodies. This chapter illustrated the difficulty of producing reliable remote sensing reflectances from state-of-the-art atmospheric corrections over productive inland waters. The optical complexity of both the water and atmosphere, combined with significant adjacency effects from bright adjacent land pixels, contributes to substantial errors in derived radiometric products and subsequent chl-a retrievals. Simple, empirically calibrated spectral derivative models, such as the MPH, produced the best results when utilized on partially corrected top-of-atmosphere reflectances. These types of models are relatively less sensitive to poor atmospheric correction or other sources of signal contamination. Through atmospheric modeling, significant signal contamination was quantified in the red and NIR spectrum for one of the small water bodies under investigation. OLCI proved to be able to produce adequate estimates of trophic status, as long as end-users understand the limitations involved, and aware of the significant errors that was shown to be induced at higher observation zenith angles.

Aim 2: Chapter 3 described a novel synthetic dataset of remote sensing reflectance built through radiative transfer modeling. This dataset included bio-optical relationships based on our most current

understanding of inland water optics, and associated optical constituents. The most significant contribution of this dataset stems from its attention to mixed cyanobacteria assemblages. Previous work has identified the vast differences in inherent optical properties of bloom forming cyanobacteria and eukaryotic phytoplankton, creating substantial differences in water-leaving reflectance depending on the assemblage mixture. The phytoplankton component within the dataset is better optimized to productive inland waters through the use of a two-layered equivalent algal populations model as well as more accurate estimates of chl-a fluorescence amplitude and shape. The dataset also addressed the scarcity of paired PC concentration data with radiometric data, which has critically hindered the progression of PC retrieval models. The final dataset was scrutinized against radiometry and optics of natural waters and was found to include highly realistic ranges and relationships of water constituents and optical characteristics. The derivation of 13 distinct optical water types which conform to previous research will allow more sophisticated and in-depth analysis of radiometric relationships in future research.

Aim 3: The utility of a simple empirically calibrated global MPH model was discussed in chapter 4 using the synthetic dataset. This chapter extended the aquatic synthetic dataset developed in chapter 3 by modeling above surface reflectances to at-sensor radiances resolved to multiple multispectral and hyperspectral sensor configurations. Only Sentinel-3 OLCI was considered in this chapter as the only current operational sensor capable of utilizing the MPH algorithm. The atmospheric modeling also allowed a first order examination of typical surviving water-leaving radiance at top-of-atmosphere. It was found that water types composed of elevated phytoplankton biomass, PC:chl-a ratios, or inorganic particles contributed much higher fractions of water-leaving signal at TOA, capable of reaching 40-60% of the total at-sensor radiance signal on average. First order estimates of typical SNR ranges for OLCI for the OWTs defined in chapter 3 were also calculated, which suggested

the majority of inland water types produce SNR values above the thresholds described in previous research as the required minimum. The sensitivity analysis of the MPH algorithm and its associated flags also suggested the difficulty of simple indices to produce reliable classifications of adjacency and cyanobacteria dominated spectra. While accuracy of cyanobacteria detection ranged between 60 – 90%, significant error still occurs in consistently predicting when cyanobacteria dominate in more productive waters. A recalibration of the MPH model using the synthetic dataset illustrated the error inherent in more oligotrophic or highly absorbing water types, where the fraction of the surviving water-leavingsignal is very low with respect to the total radiance signal at TOA.

Aim 4: The potential for development of cross-sensor, multi-parameter retrieval models for global inland water quality monitoring was illustrated in chapter 5. Four current and commonly used machine learning models were trained for the retrieval of multiple water quality parameters using the combined aquatic and atmospheric synthetic datasets developed in chapters 3 and 4. These models showed very promising performance when validated against hold-out data from the synthetic dataset with an artificial neural network providing the most reliable and consistent results amongst different sensor configurations. Application of the models to R_{rs} data still produced better results than when applied to TOA reflectance data, however, predictive statistics still suggested comparable good performance when applied at TOA. The analysis also demonstrated how similar performance can be obtained between sensor configurations incorporating vastly fewer spectral bands versus configurations with a larger number of bands. Application to satellite imagery of multiple multispectral sensors displayed adequate product consistency amongst varying sensor spectral, radiometric, and spatial resolutions, with results conforming to expert knowledge of typical natural conditions. The models more than adequately produced realistic gradients of chl-a, PC, and NAP concentrations as well as phytoplankton absorption at 440 nm. The potential for an inherent

adjacency correction due to the inclusion of green adjacency contamination in the synthetic dataset was also demonstrated.

This thesis provides a novel contributions towards realizing global inland water quality retrieval. Although, there is still substantial work to be done to bring operational products to fruition. The synthetic dataset and machine learning approach to water quality retrieval developed in this work acts as a proof-of-concept to demonstrate our current capacity to produce naturally realistic synthetic data to train sufficiently reliable out-of-the-box machine learning algorithms. Future work would be to greatly expand the synthetic dataset using high powered computing. The EAP phytoplankton model would also be used to model different phytoplankton functional types and various types of harmful algae. Other sources of signal contamination such as other types of adjacency, sun glint, and cloud cover/shadow could also be incorporated into the dataset. While minimal hyperparameter tuning was performed in this study, high powered computing would allow for the training of more advanced machine learning models which could potentially provide even better performance.

7 REFERENCES

Alikas, K., Kangro, K., & Reinart, A. (2010). Detecting cyanobacterial blooms in large North European lakes using the Maximum Chlorophyll Index. *Oceanologia*, *52*(2), 237-257.

Alikas, K., & Reinart, A. (2007). Validation of the MERIS products on large European lakes: Peipsi, Vänern and Vättern. In *European Large Lakes Ecosystem changes and their ecological and socioeconomic impacts* (pp. 161-168). Springer, Dordrecht.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.

Antoine, D., d'Ortenzio, F., Hooker, S. B., Bécu, G., Gentili, B., Tailliez, D., & Scott, A. J. (2008). Assessment of uncertainty in the ocean reflectance determined by three satellite ocean color sensors (MERIS, SeaWiFS and MODIS-A) at an offshore site in the Mediterranean Sea (BOUSSOLE project). *Journal of Geophysical Research: Oceans*, *113*(C7).

Arabi, B., Salama, M., Wernand, M. R., & Verhoef, W. (2016). MOD2SEA: a coupled atmosphere-hydro-optical model for the retrieval of chlorophyll-a from remote sensing observations in complex turbid waters. *Remote sensing*, *8*(9), 722.

Babin, M., Morel, A., & Gentili, B. (1996). Remote sensing of sea surface sun-induced chlorophyll

fluorescence: consequences of natural variations in the optical characteristics of phytoplankton and the quantum yield of chlorophyll a fluorescence. *International Journal of Remote Sensing*, *17*(12), 2417-2448.

Balasubramanian, S. V., Pahlevan, N., Smith, B., Binding, C., Schalles, J., Loisel, H., ... & Bunkei, M. (2020). Robust algorithm for estimating total suspended solids (TSS) in inland and nearshore coastal waters. *Remote Sensing of Environment*, 111768.

Ball, J. E., Anderson, D. T., & Chan, C. S. (2017). Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *Journal of Applied Remote Sensing*, *11*(4), 042609.

Bassani, C., Manzo, C., Braga, F., Bresciani, M., Giardino, C., & Alberotanza, L. (2015). The impact of the microphysical properties of aerosol on the atmospheric correction of hyperspectral data in coastal waters. *Atmospheric Measurement Techniques*, *8*(3), 1593-1604.

Beaulieu, J. J., DelSontro, T., & Downing, J. A. (2019). Eutrophication will increase methane emissions from lakes and impoundments during the 21st century. *Nature communications*, *10*(1), 1-5.

Behrenfeld, M. J., Westberry, T. K., Boss, E. S., O'Malley, R. T., Siegel, D. A., Wiggert, J. D., ... & Moore, J. K. (2009). Satellite-detected fluorescence reveals global physiology of ocean phytoplankton. *Biogeosciences*, *6*(5), 779.

Bélanger, S., Ehn, J. K., & Babin, M. (2007). Impact of sea ice on the retrieval of water-leaving reflectance, chlorophyll a concentration and inherent optical properties from satellite ocean color data. *Remote Sensing of Environment*, *111*(1), 51-68.

Bennett, E. M., Carpenter, S. R., & Caraco, N. F. (2001). Human impact on erodable phosphorus and eutrophication: a global perspective: increasing accumulation of phosphorus in soil threatens rivers, lakes, and coastal oceans with eutrophication. *BioScience*, *51*(3), 227-234.

Bernard, S., Probyn, T. A., & Quirantes, A. (2009). Simulating the optical properties of phytoplankton cells using a two-layered spherical geometry. *Biogeosciences Discussions*, *6*(1).

Bi, S., Li, Y., Wang, Q., Lyu, H., Liu, G., Zheng, Z., ... & Miao, S. (2018). Inland Water Atmospheric Correction Based on Turbidity Classification Using OLCI and SLSTR Synergistic Observations. *Remote Sensing*, *10*(7), 1002.

Bidigare, R. R., Prezelin, B. B., & Smith, R. C. (1992). Bio-optical models and the problems of scaling. In *Primary productivity and biogeochemical cycles in the sea* (pp. 175-212). Springer, Boston, MA.

Binding, C. E., Greenberg, T. A., & Bukata, R. P. (2011). Time series analysis of algal blooms in Lake of the Woods using the MERIS maximum chlorophyll index. *Journal of Plankton Research*, *33*(12), 1847-1852.

Binding, C. E., Greenberg, T. A., McCullough, G., Watson, S. B., & Page, E. (2018). An analysis of

satellite-derived chlorophyll and algal bloom indices on Lake Winnipeg. *Journal of Great Lakes Research*, *44*(3), 436-446.

Bláha, L., Babica, P., & Maršálek, B. (2009). Toxins produced in cyanobacterial water blooms-toxicity and risks. *Interdisciplinary toxicology*, *2*(2), 36-41.

Blix, K., Pálffy, K., R Tóth, V., & Eltoft, T. (2018). Remote Sensing of Water Quality Parameters over Lake Balaton by Using Sentinel-3 OLCI. *Water*, *10*(10), 1428.

Blondeau-Patissier, D., Gower, J. F., Dekker, A. G., Phinn, S. R., & Brando, V. E. (2014). A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Progress in oceanography*, *123*, 123-144.

Bidigare, R. R., Marra, J., Dickey, T. D., Iturriaga, R., Baker, K. S., Smith, R. C., & Pak, H. (1990). Evidence for phytoplankton succession and chromatic adaptation in the Sargasso Sea during spring 1985. *Marine Ecology Progress Series*, 113-122.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brewin, R. J., Tilstone, G. H., Jackson, T., Cain, T., Miller, P. I., Lange, P. K., ... & Airs, R. L. (2017). Modelling size-fractionated primary production in the Atlantic Ocean from remote sensing. *Progress in Oceanography*, *158*, 130-149.

Bricaud, A., Roesler, C., & Zaneveld, J. R. V. (1995). In situ methods for measuring the inherent

optical properties of ocean waters. *Limnology and Oceanography*, *40*(2), 393-410.

Brockmann, C., Doerffer, R., Peters, M., Kerstin, S., Embacher, S., & Ruescas, A. (2016, August). Evolution of the C2RCC neural network for Sentinel 2 and 3 for the retrieval of ocean colour products in normal and extreme optically complex waters. In *Living Planet Symposium* (Vol. 740, p. 54).

Bukata, R. P. (1995). The Effects of Chlorophyll, Suspended Mineral, and Dissolved Organic Carbon on Volume Reflectance. *Optical Properties and Remote Sensing of Inland and Coastal Waters*, 135-166.

Bulgarelli, B., & Zibordi, G. (2018). On the detectability of adjacency effects in ocean color remote sensing of mid-latitude coastal environments by SeaWiFS, MODIS-A, MERIS, OLCI, OLI and MSI. *Remote sensing of environment*, *209*, 423-438.

Bulgarelli, B., Kiselev, V., & Zibordi, G. (2014). Simulation and analysis of adjacency effects in coastal waters: a case study. *Applied optics*, *53*(8), 1523-1545.

Bulgarelli, B., Kiselev, V., & Zibordi, G. (2017). Adjacency effects in satellite radiometric products from coastal waters: a theoretical analysis for the northern Adriatic Sea. *Applied optics*, *56*(4), 854-869.

Byrne, M., Hill, M., Robertson, M., King, A., Katembo, N., Wilson, J., ... & Jadhav, A. (2010). Integrated management of water hyacinth in South Africa. *WRC Report*, (454/10).

Campbell, D., Eriksson, M. J., Öquist, G., Gustafsson, P., & Clarke, A. K. (1998). The cyanobacterium Synechococcus resists UV-B by exchanging photosystem II reaction-center D1 proteins. *Proceedings of the National Academy of Sciences*, *95*(1), 364-369.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Cilliers, C. J. (1991). Biological control of water hyacinth, Eichhornia crassipes (Pontederiaceae), in South Africa. *Agriculture, ecosystems & environment*, *37*(1-3), 207-217.

Coetzee, J. A., & Hill, M. P. (2012). The role of eutrophication in the biological control of water hyacinth, Eichhornia crassipes, in South Africa. *BioControl*, *57*(2), 247-261.

Crichton, M., Hutchings, L., Lamont, T., & Jarre, A. (2013). From physics to phytoplankton: prediction of dominant cell size in St Helena Bay in the Southern Benguela. *Journal of plankton research*, *35*(3), 526-541.

Dall'Olmo, G., & Gitelson, A. A. (2005). Effect of bio-optical parameter variability on the remote estimation of chlorophyll-a concentration in turbid productive waters: experimental results. *Applied optics*, *44*(3), 412-422.

De Keukelaere, L., Sterckx, S., Adriaensen, S., Knaeps, E., Reusen, I., Giardino, C., ... & Vaiciute,

D. (2018). Atmospheric correction of Landsat-8/OLI and Sentinel-2/MSI data using iCOR algorithm: validation for coastal and inland waters. *European Journal of Remote Sensing*, *51*(1), 525-542.

Dekker, A. G. (1993). Detection of optical water quality parameters for eutrophic waters by high resolution remote sensing.

Delgado, A. L., Pratolongo, P. D., Gossn, J. I., Dogliotti, A. I., Arena, M., Villagran, D., & Severini, M. F. (2018, October). Evaluation of derived total suspended matter products from Ocean and Land Colour Instrument Imagery (OLCI) in the inner and mid-shelf of Buenos Aires Province (Argentina). In *Extended Abstract submitted to the XXIV Ocean Optics Conference, Dubrovnik, Croatia*.

Doerffer, R., & Schiller, H. (2007). The MERIS Case 2 water algorithm. *International Journal of Remote Sensing*, *28*(3-4), 517-535.

Doerffer, R., & Schiller, H. (2008). MERIS lake water algorithm for BEAM—MERIS algorithm theoretical basis document. V1.0, 10 June 2008. Geesthacht, Germany: GKSS Research Center.

Donlon, C., Berruti, B., Buongiorno, A., Ferreira, M. H., Féménias, P., Frerick, J., ... & Nieke, J. (2012). The global monitoring for environment and security (GMES) sentinel-3 mission. *Remote Sensing of Environment*, *120*, 37-57.

Doxaran, D., Cherukuru, R. N., & Lavender, S. J. (2004). Estimation of surface reflection effects on upwelling radiance field measurements in turbid waters. *Journal of Optics A: Pure and Applied Optics*, *6*(7), 690.

Evers-King, H., Bernard, S., Lain, L. R., & Probyn, T. A. (2014). Sensitivity in reflectance attributed to phytoplankton cell size: forward and inverse modelling approaches. *Optics express*, *22*(10), 11536-11551.

Falconer, I. R., & Humpage, A. R. (2006). Cyanobacterial (blue-green algal) toxins in water supplies: Cylindrospermopsins. *Environmental Toxicology: An International Journal*, *21*(4), 299-304.

Fan, Y., Li, W., Gatebe, C. K., Jamet, C., Zibordi, G., Schroeder, T., & Stamnes, K. (2017). Atmospheric correction over coastal waters using multilayer neural networks. *Remote Sensing of Environment*, *199*, 218-240.

Fischer, J., & Kronfeld, U. (1990). Sun-stimulated chlorophyll fluorescence 1: Influence of oceanic properties. *Remote Sensing*, *11*(12), 2125-2147.

Forsberg, C. (1998). Which policies can stop large scale eutrophication?. *Water Science and Technology*, *37*(3), 193-200.

Gallegos, C. L., Correll, D. L., & Pierce, J. W. (1990). Modeling spectral diffuse attenuation, absorption, and scattering coefficients in a turbid estuary. *Limnology and Oceanography*, *35*(7), 1486-1502.

Ganf, G. G., Oliver, R. L., & Walsby, A. E. (1989). Optical properties of gas-vacuolate cells and colonies of Microcystis in relation to light attenuation in a turbid, stratified reservoir (Mount Bold Reservoir, South Australia). *Marine and Freshwater Research*, *40*(6), 595-611.

Garver, S. A., & Siegel, D. A. (1997). Inherent optical property inversion of ocean color spectra and its biogeochemical interpretation: 1. Time series from the Sargasso Sea. *Journal of Geophysical Research: Oceans*, *102*(C8), 18607-18625.

Gholizadeh, M. H., Melesse, A. M., & Reddi, L. (2016). A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors*, *16*(8), 1298.

Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D., & Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing*, *11*(2), 196.

Giardino, C., Brando, V. E., Dekker, A. G., Strömbeck, N., & Candiani, G. (2007). Assessment of water quality in Lake Garda (Italy) using Hyperion. *Remote Sensing of Environment*, *109*(2), 183-195.

Giardino, C., Bresciani, M., Cazzaniga, I., Schenk, K., Rieger, P., Braga, F., ... & Brando, V. E. (2014). Evaluation of multi-resolution satellite sensors for assessing water quality and bottom depth of Lake Garda. *Sensors*, *14*(12), 24116-24131.

Giardino, C., Bresciani, M., Pilkaityte, R., Bartoli, M., & Razinkovas, A. (2010). In situ

measurements and satellite remote sensing of case 2 waters: first results from the Curonian Lagoon. *Oceanologia*, *52*(2), 197-210.

Gilerson, A. A., Gitelson, A. A., Zhou, J., Gurlin, D., Moses, W., Ioannou, I., & Ahmed, S. A.

(2010). Algorithms for remote estimation of chlorophyll-a in coastal and inland waters using red and near infrared bands. *Optics Express*, *18*(23), 24109-24125.

Gilerson, A., Zhou, J., Hlaing, S., Ioannou, I., Amin, R., Gross, B., ... & Ahmed, S. (2007, October).

Fluorescence contribution to reflectance spectra for a variety of coastal waters. In *Coastal Ocean Remote Sensing* (Vol. 6680, p. 66800C). International Society for Optics and Photonics.

Gilerson, A., Zhou, J., Hlaing, S., Ioannou, I., Gross, B., Moshary, F., & Ahmed, S. (2008).

Fluorescence component in the reflectance spectra from coastal waters. II. Performance of retrieval algorithms. *Optics express*, *16*(4), 2446-2460.

Gitelson, A. (1992). The peak near 700 nm on radiance spectra of algae and water: relationships

of its magnitude and position with chlorophyll concentration. *International Journal of Remote Sensing*, *13*(17), 3367-3373.

Gitelson, A. A., Dall'Olmo, G., Moses, W., Rundquist, D. C., Barrow, T., Fisher, T. R., ... & Holz, J.

(2008). A simple semi-analytical model for remote estimation of chlorophyll-a in turbid waters: Validation. *Remote Sensing of Environment*, *112*(9), 3582-3593.

Gitelson, A. A., Gao, B. C., Li, R. R., Berdnikov, S., & Saprygin, V. (2011). Estimation of

chlorophyll-a concentration in productive turbid waters using a Hyperspectral Imager for the Coastal Ocean—the Azov Sea case study. *Environmental Research Letters*, *6*(2), 024023.

Gitelson, A. A., Gurlin, D., Moses, W. J., & Barrow, T. (2009). A bio-optical algorithm for the remote estimation of the chlorophyll-a concentration in case 2 waters. *Environmental Research Letters*, *4*(4), 045003.

Gons, H. J. (1999). Optical teledetection of chlorophyll a in turbid inland waters. Environmental Science & Technology, 33(7), 1127–1132.

Gons, H. J., Auer, M. T., & Effler, S. W. (2008). MERIS satellite chlorophyll mapping of oligotrophic and eutrophic waters in the Laurentian Great Lakes. *Remote Sensing of Environment*, *112*(11), 4098-4106.

Gordon, H. R., Brown, O. B., & Jacobs, M. M. (1975). Computed relationships between the inherent and apparent optical properties of a flat homogeneous ocean. *Applied optics*, *14*(2), 417-427.

Gordon, H. R., & Ding, K. (1992). Self-shading of in-water optical instruments. *Limnology and Oceanography*, *37*(3), 491-500.

Gordon, H. R., & Wang, M. (1994). Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with SeaWiFS: a preliminary algorithm. *Applied optics*, *33*(3), 443-452.

Gossn, J. I., Ruddick, K. G., & Dogliotti, A. I. (2019). Atmospheric Correction of OLCI Imagery over Extremely Turbid Waters Based on the Red, NIR and 1016 nm Bands and a New Baseline Residual Technique. *Remote Sensing*, *11*(3), 220.

Govindjee, G. (2004). Chlorophyll a fluorescence: a bit of basics and history. *Chlorophyll a fluorescence: a signature of photosynthesis Springer, Dordrecht*, 1-42.

Gower, J. F. R., & Borstad, G. (1981). Use of the in vivo fluorescence line at 685 nm for remote sensing surveys of surface chlorophyll a. In *Oceanography from space* (pp. 329-338). Springer, Boston, MA.

Gower, J. F. R., Doerffer, R., & Borstad, G. A. (1999). Interpretation of the 685nm peak in water-leaving radiance spectra in terms of fluorescence, absorption and scattering, and its observation by MERIS. *International Journal of Remote Sensing*, *20*(9), 1771-1786.

Gower, J., King, S., Borstad, G., & Brown, L. (2005). Detection of intense plankton blooms using the 709 nm band of the MERIS imaging spectrometer. *International Journal of Remote Sensing*, *26*(9), 2005-2012.

Gower, J., King, S., & Goncalves, P. (2008). Global monitoring of plankton blooms using MERIS MCI. *International Journal of Remote Sensing*, *29*(21), 6209-6216.

Greene, R. M., Geider, R. J., Kolber, Z., & Falkowski, P. G. (1992). Iron-induced changes in light

harvesting and photochemical energy conversion processes in eukaryotic marine algae. *Plant Physiology*, *100*(2), 565-575.

Grossman, A. R., Schaefer, M. R., Chiang, G. G., & Collier, J. L. (1993). Environmental effects on the light-harvesting complex of cyanobacteria. *Journal of bacteriology*, *175*(3), 575.

Guanter, L., Alonso, L., & Moreno, J. (2005). A method for the surface reflectance retrieval from PROBA/CHRIS data over land: Application to ESA SPARC campaigns. *IEEE Transactions on Geoscience and Remote Sensing*, *43*(12), 2908-2917.

Guanter, L., Estellés, V., & Moreno, J. (2007). Spectral calibration and atmospheric correction of ultra-fine spectral and spatial resolution remote sensing data. Application to CASI-1500 data. *Remote Sensing of Environment*, *109*(1), 54-65.

Guanter, L., Ruiz-Verdú, A., Odermatt, D., Giardino, C., Simis, S., Estellés, V., ... & Moreno, J. (2010). Atmospheric correction of ENVISAT/MERIS data over inland waters: Validation for European lakes. *Remote Sensing of Environment*, *114*(3), 467-480.

Gurlin, D., Gitelson, A. A., & Moses, W. J. (2011). Remote estimation of chl-a concentration in turbid productive waters—Return to a simple two-band NIR-red model?. *Remote Sensing of Environment*, *115*(12), 3479-3490.

Hafeez, S., Wong, M. S., Ho, H. C., Nazeer, M., Nichol, J., Abbas, S., ... & Pun, L. (2019). Comparison

of machine learning algorithms for retrieval of water quality indicators in case-II waters: a case study of Hong Kong. *Remote sensing*, *11*(6), 617.

Harding, W. R. (2015). Living with eutrophication in South Africa: a review of realities and challenges. *Transactions of the Royal Society of South Africa*, *70*(2), 155-171.

Hestir, E. L., Brando, V. E., Bresciani, M., Giardino, C., Matta, E., Villa, P., & Dekker, A. G. (2015). Measuring freshwater aquatic ecosystems: The need for a hyperspectral global mapping satellite mission. *Remote Sensing of Environment*, *167*, 181-195.

Hieronymi, M., Müller, D., & Doerffer, R. (2017). The OLCI Neural Network Swarm (ONNS): A bio-geo-optical algorithm for open ocean and coastal waters. *Frontiers in Marine Science*, *4*, 140.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832-844.

Ho, J. C., Michalak, A. M., Pahlevan, N. (2019) Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature*. *Nature*, 2019; DOI: 10.1038/s41586-019-1648-7

Holben, B. N., Eck, T. F., & Fraser, R. S. (1991). Temporal and spatial variability of aerosol optical depth in the Sahel region in relation to vegetation remote sensing. *International Journal of Remote Sensing*, *12*(6), 1147-1163.

Hu, C. (2009). A novel ocean color index to detect floating algae in the global oceans. *Remote Sensing of Environment*, *113*(10), 2118-2129.

Hu, C., Chen, Z., Clayton, T. D., Swarzenski, P., Brock, J. C., & Muller–Karger, F. E. (2004). Assessment of estuarine water-quality indicators using MODIS medium-resolution bands: Initial results from Tampa Bay, FL. *Remote Sensing of Environment*, *93*(3), 423-441.

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

Hudnell, H. K. (2010). The state of US freshwater harmful algal blooms assessments, policy and legislation. *Toxicon*, *55*(5), 1024-1034.

Hunter, P. D., Tyler, A. N., Carvalho, L., Codd, G. A., & Maberly, S. C. (2010). Hyperspectral remote sensing of cyanobacterial pigments as indicators for cell populations and toxins in eutrophic lakes. *Remote Sensing of Environment*, *114*(11), 2705-2718.

Huot, Y., Brown, C. A., & Cullen, J. J. (2005). New algorithms for MODIS sun-induced chlorophyll fluorescence and a comparison with present data products. *Limnology and Oceanography: Methods*, *3*(2), 108-130.

Huot, Y., Brown, C. A., & Cullen, J. J. (2007). Retrieval of phytoplankton biomass from simultaneous

inversion of reflectance, the diffuse attenuation coefficient, and Sun-induced fluorescence in coastal waters. *Journal of Geophysical Research: Oceans*, *112*(C6).

James, C., Fisher, J., Russell, V., Collings, S. & Moss, B. (2005) Nitrate availability and hydrophyte species richness in shallow lakes. *Freshwater Biology*, 50, 1049–1063.

Johnsen, G., & Sakshaug, E. (2007). Biooptical characteristics of PSII and PSI in 33 species (13 pigment groups) of marine phytoplankton, and the relevance for pulse-amplitude-modulated and fast-repetition-rate fluorometry 1. *Journal of Phycology*, *43*(6), 1236-1251.

Jorge, D. S., Barbosa, C. C., De Carvalho, L. A., Affonso, A. G., Lobo, F. D. L., & Novo, E. M. D. M. (2017). Snr (signal-to-noise ratio) impact on water constituent retrieval from simulated images of optically complex amazon lakes. *Remote Sensing*, *9*(7), 644.

Jupp, D. L., Kirk, J. T., & Harris, G. P. (1994). Detection, identification and mapping of cyanobacteria—using remote sensing to measure the optical quality of turbid inland waters. *Marine and Freshwater Research*, *45*(5), 801-828.

Kotchenova, S. Y., Vermote, E. F., Matarrese, R., & Klemm Jr, F. J. (2006). Validation of a vector version of the 6S radiative transfer code for atmospheric correction of satellite data. Part I: Path radiance. *Applied optics*, *45*(26), 6762-6774.

Kravitz, J., Matthews, M., Bernard, S., & Griffith, D. (2020). Application of Sentinel 3 OLCI for chl-

a retrieval over small inland water targets: Successes and challenges. *Remote Sensing of Environment*, *237*, 111562.

Kudela, R. M., Hooker, S. B., Houskeeper, H. F., & McPherson, M. (2019). The Influence of Signal to Noise Ratio of Legacy Airborne and Satellite Sensors for Simulating Next-Generation Coastal and Inland Water Products. *Remote Sensing*, *11*(18), 2071.

Kumar, K. R., Sivakumar, V., Yin, Y., Reddy, R. R., Kang, N., Diao, Y., ... & Yu, X. (2014). Long-term (2003–2013) climatological trends and variations in aerosol optical parameters retrieved from MODIS over three stations in South Africa. *Atmospheric environment*, *95*, 400-408.

Kutser, T. (2004). Quantitative detection of chlorophyll in cyanobacterial blooms by satellite remote sensing. *Limnology and Oceanography*, *49*(6), 2179-2189.

Kutser, T., Metsamaa, L., & Dekker, A. G. (2008). Influence of the vertical distribution of cyanobacteria in the water column on the remote sensing signal. *Estuarine, Coastal and Shelf Science*, *78*(4), 649-654.

Kutser, T., Soomets, T., Toming, K., Uiboupin, R., Arikas, A., Vahter, K., & Paavel, B. (2018, June). Assessing the Baltic Sea Water Quality with Sentinel-3 OLCI Imagery. In *2018 IEEE/OES Baltic International Symposium (BALTIC)* (pp. 1-6). IEEE.

Lain, L. R., & Bernard, S. (2018). The fundamental contribution of phytoplankton spectral scattering

to ocean colour: implications for satellite detection of phytoplankton community structure. *Applied Sciences*, *8*(12), 2681.

Lain, L. R., Bernard, S., & Matthews, M. W. (2016). Biophysical modelling of phytoplankton communities from first principles using two-layered spheres: Equivalent Algal Populations (EAP) model: erratum. *Optics Express*, *24*(24), 27423-27424.

Lavender, S., Doxaran, D., & Nagur Cherukura, R. C. (2005, June). High spatial resolution remote sensing of the Plymouth coastal waters. In *ESA Special Publication* (Vol. 593).

Lee, Z. P. (2003). Models, parameters, and approaches that used to generate wide range of absorptionand backscattering spectra. *Ocean Color Algorithm Working Group, IOCCG*.Lee, Z. (2006). Remote sensing of inherent optical properties: fundamentals, tests of algorithms, and applications.Lee, Z., Carder, K. L., & Arnone, R. A. (2002). Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep waters. *Applied optics*, *41*(27), 5755-5772.

Li, Y., Zhang, H., Xue, X., Jiang, Y., & Shen, Q. (2018). Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(6), e1264.

Li, L., Li, L., & Song, K. (2015). Remote sensing of freshwater cyanobacteria: An extended IOP

Inversion Model of Inland Waters (IIMIW) for partitioning absorption coefficient and estimating phycocyanin. *Remote Sensing of Environment*, *157*, 9-23.

Ligi, M., Kutser, T., Kallio, K., Attila, J., Koponen, S., Paavel, B., ... & Reinart, A. (2017). Testing the performance of empirical remote sensing algorithms in the Baltic Sea waters with modelled and in situ reflectance data. *Oceanologia*, *59*(1), 57-68.

Liu, G., Simis, S. G., Li, L., Wang, Q., Li, Y., Song, K., ... & Shi, K. (2017). A four-band semi-analytical model for estimating phycocyanin in inland waters from simulated MERIS and OLCI data. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(3), 1374-1385.

Lubac, B., & Loisel, H. (2007). Variability and classification of remote sensing reflectance spectra in the eastern English Channel and southern North Sea. *Remote Sensing of Environment*, *110*(1), 45-58.

Lunetta, R. S., Schaeffer, B. A., Stumpf, R. P., Keith, D., Jacobs, S. A., & Murphy, M. S. (2015). Evaluation of cyanobacteria cell count detection derived from MERIS imagery across the eastern USA. *Remote Sensing of Environment*, *157*, 24-34.

Martins, V. S., Barbosa, C. C. F., de Carvalho, L. A. S., Jorge, D. S. F., Lobo, F. D. L., & Novo, E. M. L. D. M. (2017). Assessment of Atmospheric Correction Methods for Sentinel-2 MSI Images Applied to Amazon Floodplain Lakes. *Remote Sensing*, *9*(4), 322.

Matsushita, B., Yang, W., Yu, G., Oyama, Y., Yoshimura, K., & Fukushima, T. (2015). A hybrid

algorithm for estimating the chlorophyll-a concentration across different trophic states in Asian inland waters. *ISPRS journal of photogrammetry and remote sensing*, *102*, 28-37.

Matthews, M. W. (2011). A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters. *International Journal of Remote Sensing*, *32*(21), 6855-6899.

Matthews, M. W. (2014). Eutrophication and cyanobacterial blooms in South African inland waters: 10years of MERIS observations. *Remote Sensing of Environment*, *155*, 161-177.

Matthews, M., & Bernard, S. (2013). Characterizing the absorption properties for remote sensing of three small optically-diverse South African reservoirs. *Remote Sensing*, *5*(9), 4370-4404.

Matthews, M. W., Bernard, S., & Robertson, L. (2012). An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters. *Remote Sensing of Environment*, *124*, 637–652. doi:10.1016/j.rse.2012.05.032

Matthews, M. W., Bernard, S., & Winter, K. (2010). Remote sensing of cyanobacteria-dominant algal blooms and water quality parameters in Zeekoevlei, a small hypertrophic lake, using MERIS. *Remote Sensing of Environment*, *114*(9), 2070-2087.

Matthews, M. W., & Odermatt, D. (2015). Improved algorithm for routine monitoring of

cyanobacteria and eutrophication in inland and near-coastal waters. *Remote Sensing of Environment*, *156*, 374-382.

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, *39*(9), 2784-2817.

Metsamaa, L., Kutser, T., & Strömbeck, N. (2006). Recognising cyanobacterial blooms based on their optical signature: a modelling study. *Boreal Environment Research*, *11*(6), 493-506.

Mishra, S., & Mishra, D. R. (2012). Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. *Remote Sensing of Environment*, *117*, 394-406.

Mobley, C. D. (1999). Estimation of the remote-sensing reflectance from above-surface measurements. *Applied Optics*, *38*(36), 7442-7455.

Mobley, C. D. (2015). Polarized reflectance and transmittance properties of windblown sea surfaces. *Applied optics*, *54*(15), 4828-4849.

Mobley, C. D., Sundman, L. K., & Boss, E. (2002). Phase function effects on oceanic light fields. *Applied optics*, *41*(6), 1035-1050.

Moses, W. J., Bowles, J. H., & Corson, M. R. (2015). Expected improvements in the quantitative

remote sensing of optically complex waters with the use of an optically fast hyperspectral spectrometer—A modeling study. *Sensors*, *15*(3), 6152-6173.

Moore, T. S., Campbell, J. W., & Feng, H. (2001). A fuzzy logic classification scheme for selecting and blending satellite ocean color algorithms. *IEEE Transactions on Geoscience and Remote sensing*, *39*(8), 1764-1776.

Moore, T. S., Campbell, J. W., & Dowell, M. D. (2009). A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product. *Remote Sensing of Environment*, *113*(11), 2424-2430.

Moore, T. S., Dowell, M. D., Bradt, S., & Verdu, A. R. (2014). An optical water type framework for selecting and blending retrievals from bio-optical algorithms in lakes and coastal waters. *Remote sensing of environment*, *143*, 97-111.

Morel, A., & Bricaud, A. (1981). Theoretical results concerning light absorption in a discrete medium, and application to specific absorption of phytoplankton. *Deep Sea Research Part A. Oceanographic Research Papers*, *28*(11), 1375-1393.

Morel, A., & Prieur, L. (1977). Analysis of variations in ocean color 1. *Limnology and oceanography*, *22*(4), 709-722.

Moses, W. J., Gitelson, A. A., Berdnikov, S., & Povazhnyy, V. (2009a). Estimation of chlorophyll-a

concentration in case II waters using MODIS and MERIS data—successes and challenges. *Environmental Research Letters*, *4*(4), 045005.

Moses, W. J., Gitelson, A. A., Berdnikov, S., & Povazhnyy, V. (2009b). Satellite estimation of chlorophyll-a concentration using the red and NIR bands of MERIS—The Azov sea case study. *IEEE Geoscience and Remote Sensing Letters*, *6*(4), 845-849.

Moses, W. J., Gitelson, A. A., Berdnikov, S., Saprygin, V., & Povazhnyi, V. (2012). Operational MERIS-based NIR-red algorithms for estimating chlorophyll-a concentrations in coastal waters—The Azov Sea case study. *Remote Sensing of Environment*, *121*, 118-124.

Mouw, C. B., Greb, S., Aurin, D., DiGiacomo, P. M., Lee, Z., Twardowski, M., ... & Moses, W. (2015). Aquatic color radiometry remote sensing of coastal and inland waters: Challenges and recommendations for future satellite missions. *Remote sensing of environment*, *160*, 15-30.

Mueller, J. L. (2003). *Ocean optics protocols for satellite ocean color sensor validation, revision 4: radiometric measurements and data analysis protocols*(Vol. 3). Goddard Space Flight Center.

Muller-Karger, F. E., Hestir, E., Ade, C., Turpie, K., Roberts, D. A., Siegel, D., ... & Morgan, F. (2018). Satellite sensor requirements for monitoring essential biodiversity variables of coastal ecosystems. *Ecological applications*, *28*(3), 749-760.

Neil, C., Spyrakos, E., Hunter, P. D., & Tyler, A. N. (2019). A global approach for chlorophyll-a

retrieval across optically complex inland waters based on optical water types. *Remote Sensing of Environment*, *229*, 159-178.

O'neil, J. M., Davis, T. W., Burford, M. A., & Gobler, C. J. (2012). The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. *Harmful algae*, *14*, 313-334.

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

Oberholster, P. J., & Ashton, P. J. (2008). State of the nation report: An overview of the current status of water quality and eutrophication in South African rivers and reservoirs. *Parliamentary Grant Deliverable. Pretoria: Council for Scientific and Industrial Research (CSIR)*.

Odermatt, D., Giardino, C., & Heege, T. (2010). Chlorophyll retrieval with MERIS Case-2-Regional in perialpine lakes. *Remote Sensing of Environment*, *114*(3), 607-617.

Odermatt, D., Gitelson, A., Brando, V. E., & Schaepman, M. (2012). Review of constituent retrieval in optically deep and complex waters from satellite imagery. *Remote sensing of environment*, *118*, 116-126.

Odermatt, D., Kiselev, V., Heege, T., Kneubühler, M., & Itten, K. I. (2008, September). Adjacency

effect considerations and air/water constituent retrieval for Lake Constance. In *Proceedings of the 2nd MERIS/(A) ATSR user workshop. Frascati, Italy* (Vol. 1).

Ogashawara, I. (2020). Determination of Phycocyanin from Space—A Bibliometric Analysis. *Remote Sensing*, *12*(3), 567.

Organelli, E., Claustre, H., Bricaud, A., Barbieux, M., Uitz, J., D'Ortenzio, F., & Dall'Olmo, G. (2017). Bio-optical anomalies in the world's oceans: An investigation on the diffuse attenuation coefficients for downward irradiance derived from Biogeochemical Argo float measurements. *Journal of Geophysical Research: Oceans*, *122*(5), 3543-3564.

Padedda, B. M., Sechi, N., Lai, G. G., Mariani, M. A., Pulina, S., Sarria, M., ... & Lugliè, A. (2017). Consequences of eutrophication in the management of water resources in Mediterranean reservoirs: A case study of Lake Cedrino (Sardinia, Italy). *Global Ecology and Conservation*, *12*, 21-35.

Pahlevan, N., Smith, B., Schalles, J., Binding, C., Cao, Z., Ma, R., ... & Matsushita, B. (2020). Seamless retrievals of chlorophyll-a from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in inland and coastal waters: A machine-learning approach. *Remote Sensing of Environment*, 111604.

Palmer, S. C., Hunter, P. D., Lankester, T., Hubbard, S., Spyrakos, E., Tyler, A. N., ... & Tóth, V. R. (2015a). Validation of Envisat MERIS algorithms for chlorophyll retrieval in a large, turbid and optically-complex shallow lake. *Remote Sensing of Environment*, *157*, 158-169.

Palmer, S. C., Kutser, T., & Hunter, P. D. (2015b). Remote sensing of inland waters: Challenges,

progress and future directions, *Remote Sensing of Environment, 157, pp. 1-8.*

Palmer, S. C., Odermatt, D., Hunter, P. D., Brockmann, C., Presing, M., Balzter, H., & Tóth, V. R. (2015c). Satellite remote sensing of phytoplankton phenology in Lake Balaton using 10 years of MERIS observations. *Remote Sensing of Environment*, *158*, 441-452.

Palmer, K. F., & Williams, D. (1974). Optical properties of water in the near infrared. *JOSA*, *64*(8), 1107-1110.

Peterson, K. T., Sagan, V., Sidike, P., Cox, A. L., & Martinez, M. (2018). Suspended sediment concentration estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers using an extreme learning machine. *Remote Sensing*, *10*(10), 1503.

Peterson, K. T., Sagan, V., & Sloan, J. J. (2020). Deep learning-based water quality estimation and anomaly detection using Landsat-8/Sentinel-2 virtual constellation and cloud computing. *GIScience & Remote Sensing*, *57*(4), 510-525.

Peterson, K. T., Sagan, V., Sidike, P., Hasenmueller, E. A., Sloan, J. J., & Knouft, J. H. (2019). Machine Learning-Based Ensemble Prediction of Water-Quality Variables Using Feature-Level and Decision-Level Fusion with Proximal Remote Sensing. *Photogrammetric Engineering & Remote Sensing*, *85*(4), 269-280.

Pieterse, A. J. H., & Rohrbeck, M. A. (1990). Dominant phytoplankters and environmental variables in Roodeplaat Dam, Pretoria, South Africa. *Water SA*, *16*(4), 211-218.

Pitarch, J., Ruiz-Verdú, A., Sendra, M. D., & Santoleri, R. (2017). Evaluation and reformulation of the maximum peak height algorithm (MPH) and application in a hypertrophic lagoon. *Journal of Geophysical Research: Oceans*, *122*(2), 1206-1221.

Prieur, L., & Sathyendranath, S. (1981). An optical classification of coastal and oceanic waters based on the specific spectral absorption curves of phytoplankton pigments, dissolved organic matter, and other particulate materials 1. *Limnology and Oceanography*, *26*(4), 671-689.

Qi, L., Hu, C., Duan, H., Cannizzaro, J., & Ma, R. (2014). A novel MERIS algorithm to derive cyanobacterial phycocyanin pigment concentrations in a eutrophic lake: Theoretical basis and practical considerations. *Remote sensing of environment*, *154*, 298-317.

Qi, L., Lee, Z., Hu, C., & Wang, M. (2017). Requirement of minimal signal-to-noise ratios of ocean color sensors and uncertainties of ocean color products. *Journal of Geophysical Research: Oceans*, *122*(3), 2595-2611.

Quirantes, A., & Bernard, S. (2004). Light scattering by marine algae: two-layer spherical and nonspherical models. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *89*(1), 311-321.

Ramsay, J. O., & Silverman, B. (2006). Functional data analysis. Hoboken.

Randolph, K., Wilson, J., Tedesco, L., Li, L., Pascual, D. L., & Soyeux, E. (2008). Hyperspectral

remote sensing of cyanobacteria in turbid productive water using optically significant pigments, chlorophyll a and phycocyanin. *Remote Sensing of Environment*, *112*(11), 4009-4019.

Richter, R., Bachmann, M., Dorigo, W., & Muller, A. (2006). Influence of the adjacency effect on ground reflectance measurements. *IEEE Geoscience and Remote Sensing Letters*, *3*(4), 565-569.

Robarts, R. D., & Zohary, T. (1987). Temperature effects on photosynthetic capacity, respiration, and growth rates of bloom-forming cyanobacteria. *New Zealand Journal of Marine and Freshwater Research*, *21*(3), 391-399.

Roesler, C. S., & Perry, M. J. (1995). In situ phytoplankton absorption, fluorescence emission, and particulate backscattering spectra determined from reflectance. *Journal of Geophysical Research: Oceans*, *100*(C7), 13279-13294.

Ruddick, K. G., De Cauwer, V., Park, Y. J., & Moore, G. (2006). Seaborne measurements of near infrared water-leaving reflectance: The similarity spectrum for turbid waters. *Limnology and Oceanography*, *51*(2), 1167-1179.

Ruddick, K. G., Gons, H. J., Rijkeboer, M., & Tilstone, G. (2001). Optical remote sensing of chlorophyll a in case 2 waters by use of an adaptive two-band algorithm with optimal error properties. *Applied optics*, *40*(21), 3575-3585.

Ruescas, A. B., Hieronymi, M., Mateo-Garcia, G., Koponen, S., Kallio, K., & Camps-Valls, G.

(2018). Machine learning regression approaches for colored dissolved organic matter (CDOM) retrieval with S2-MSI and S3-OLCI simulated data. *Remote Sensing*, *10*(5), 786.

Sagan, V., Peterson, K. T., Maimaitijiang, M., Sidike, P., Sloan, J., Greeling, B. A., ... & Adams, C. (2020). Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computing. *Earth-Science Reviews*, 103187.

Santer, R., & Schmechtig, C. (2000). Adjacency effects on water surfaces: primary scattering approximation and sensitivity study. *Applied Optics*, *39*(3), 361-375.

Sarada, R. M. G. P., Pillai, M. G., & Ravishankar, G. A. (1999). Phycocyanin from Spirulina sp: influence of processing of biomass on phycocyanin yield, analysis of efficacy of extraction methods and stability studies on phycocyanin. *Process biochemistry*, *34*(8), 795-801.

Schalles, J. F. (2006). Optical remote sensing techniques to estimate phytoplankton chlorophyll a concentrations in coastal. In *Remote sensing of aquatic coastal ecosystem processes* (pp. 27-79). Springer, Dordrecht.

Schindler, D. W. (2012). The dilemma of controlling cultural eutrophication of lakes. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1746), 4322-4333.

Scott, W. E., Ashton, P. J., & Steyn, D. J. (1979). *Chemical control of the water hyacinth on Hartbeespoort Dam*. Water Research Commission.

Scott, W. E., Ashton, P. J., Walmsley, R. D., & Seaman, M. T. (1980). Hartbeespoort Dam: a case study of a hypertrophic, warm, monomictic impoundment. In *Hypertrophic ecosystems* (pp. 317-322). Springer Netherlands.

Schalles, J. F., Gitelson, A. A., Yacobi, Y. Z., & Kroenke, A. E. (1998). Estimation of chlorophyll a from time series measurements of high spectral resolution reflectance in an eutrophic lake. *Journal of Phycology*, *34*(2), 383-390.

Sei, A. (2007). Analysis of adjacency effects for two Lambertian half-spaces. *International Journal of Remote Sensing*, *28*(8), 1873-1890.

Seppälä, J., Ylöstalo, P., Kaitala, S., Hällfors, S., Raateoja, M., & Maunula, P. (2007). Ship-of-opportunity based phycocyanin fluorescence monitoring of the filamentous cyanobacteria bloom dynamics in the Baltic Sea. *Estuarine, Coastal and Shelf Science*, *73*(3-4), 489-500.

Shen, M., Duan, H., Cao, Z., Xue, K., Loiselle, S., & Yesou, H. (2017). Determination of the downwelling diffuse attenuation coefficient of lake water with the Sentinel-3A OLCI. *Remote Sensing*, *9*(12), 1246.

Shi, K., Zhang, Y., Zhu, G., Liu, X., Zhou, Y., Xu, H., ... & Li, Y. (2015). Long-term remote monitoring of total suspended matter concentration in Lake Taihu using 250 m MODIS-Aqua data. *Remote Sensing of Environment*, *164*, 43-56.

Simis, S. G., Huot, Y., Babin, M., Seppälä, J., & Metsamaa, L. (2012). Optimization of variable

fluorescence measurements of phytoplankton communities with cyanobacteria. *Photosynthesis research*, *112*(1), 13-30.

Simis, S. G. H., Peters, S. W. M., & Gons, H. J. (2005). Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water. *Limnology and Oceanography*, *50*(1), 237–245.

Smith, V. H. (2003). Eutrophication of freshwater and coastal marine ecosystems a global problem. *Environmental Science and Pollution Research*, *10*(2), 126-139.

Smith, V. H., Dodds, W. K., Havens, K. E., Engstrom, D. R., Paerl, H. W., Moss, B., & Likens, G. E. (2014). Comment: Cultural eutrophication of natural lakes in the United States is real and widespread. *Limnology and Oceanography*, *59*(6), 2217-2225.

Smith, M. E., Lain, L. R., & Bernard, S. (2018). An optimized Chlorophyll a switching algorithm for MERIS and OLCI in phytoplankton-dominated waters. *Remote sensing of environment*, *215*, 217-227.

Spyrakos, E., O'Donnell, R., Hunter, P. D., Miller, C., Scott, M., Simis, S. G., ... & Bresciani, M. (2018). Optical types of inland and coastal waters. *Limnology and Oceanography*, *63*(2), 846-870.

Steinmetz, F., Deschamps, P. Y., & Ramon, D. (2011). Atmospheric correction in presence of sun glint: application to MERIS. *Optics express*, *19*(10), 9783-9800.

Sterckx, S., Knaeps, E., Adriaensen, S., Reusen, I., Keukelaere, L. D., & Hunter, P. (2015a).

OPERA: An atmospheric correction for land and water. In *Proceedings of the ESA Sentinel-3 for Science Workshop, Venice, Italy* (pp. 2-5).

Sterckx, S., Knaeps, S., Kratzer, S., & Ruddick, K. (2015b). SIMilarity Environment Correction (SIMEC) applied to MERIS data over inland and coastal waters. *Remote Sensing of Environment*, *157*, 96-110.

Stumpf, R. P., Davis, T. W., Wynne, T. T., Graham, J. L., Loftin, K. A., Johengen, T. H., ... & Burtner, A. (2016). Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful Algae*, *54*, 160-173.

Sukenik, A., Hadas, O., Kaplan, A., & Quesada, A. (2012). Invasion of Nostocales (cyanobacteria) to subtropical and temperate freshwater lakes–physiological, regional, and global driving forces. *Frontiers in microbiology*, *3*, 86.

Sun, Y., & Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, *20*(2), 316-334.

Thirion, C. (2000). A new biomonitoring protocol to determine the ecological health of impoundments using artificial substrates. *Southern African Journal of Aquatic Sciences*, *25*(1), 123-133.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via

the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.

Toerien, D. F., Hyman, K. L., & Bruwer, M. J. (1975). A preliminary trophic status classification of some South African impoundments. *Water SA*, *1*(1), 15-23.

Toming, K., Kutser, T., Uiboupin, R., Arikas, A., Vahter, K., & Paavel, B. (2017). Mapping water quality parameters with sentinel-3 ocean and land colour instrument imagery in the Baltic Sea. *Remote Sensing*, *9*(10), 1070.

Twardowski, M. S., Boss, E., Macdonald, J. B., Pegau, W. S., Barnard, A. H., & Zaneveld, J. R. V. (2001). A model for estimating bulk refractive index from the optical backscattering ratio and the implications for understanding particle composition in case I and case II waters. *Journal of Geophysical Research: Oceans*, *106*(C7), 14129-14142.

Vaillancourt, R. D., Brown, C. W., Guillard, R. R., & Balch, W. M. (2004). Light backscattering properties of marine phytoplankton: relationships to cell size, chemical composition and taxonomy. *Journal of plankton research*, *26*(2), 191-212.

Vanhellemont, Q., & Ruddick, K. (2015). Advantages of high quality SWIR bands for ocean colour processing: Examples from Landsat-8. *Remote Sensing of Environment*, *161*, 89-106.

Vanhellemont, Q. (2019). Adaptation of the dark spectrum fitting atmospheric correction for aquatic

applications of the Landsat and Sentinel-2 archives. *Remote Sensing of Environment*, *225*, 175-192.

Van Ginkel, C. E. (2011). Eutrophication: Present reality and future challenges for South Africa. *Water SA*, *37*(5), 693-702.

Van Ginkel, C. E. (2012). Algae, phytoplankton and eutrophication research and management in South Africa: past, present and future. *African journal of aquatic science*, *37*(1), 17-25.

Van Ginkel, C. E., & Silberbauer, M. J. (2007). Temporal trends in total phosphorus, temperature, oxygen, chlorophyll a and phytoplankton populations in Hartbeespoort Dam and Roodeplaat Dam, South Africa, between 1980 and 2000. *African Journal of Aquatic Science*, *32*(1), 63-70.

Van Wyk, E., & Van Wilgen, B. W. (2002). The cost of water hyacinth control in South Africa: a case study of three options. *African Journal of Aquatic Science*, *27*(2), 141-149.

Vantrepotte, V., Loisel, H., Dessailly, D., & Mériaux, X. (2012). Optical classification of contrasted coastal waters. *Remote Sensing of Environment*, *123*, 306-323.

Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016). Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sensing of Environment*, *185*, 46-56.

Vermote, E. F., Tanré, D., Deuze, J. L., Herman, M., & Morcette, J. J. (1997). Second simulation

of the satellite signal in the solar spectrum, 6S: An overview. *IEEE transactions on geoscience and remote sensing*, *35*(3), 675-686.

Viskari, P. J., & Colyer, C. L. (2003). Rapid extraction of phycobiliproteins from cultured cyanobacteria samples. *Analytical biochemistry*, *319*(2), 263-271.

Walsby, A. E., Hayes, P. K., & Boje, R. (1995). The gas vesicles, buoyancy and vertical distribution of cyanobacteria in the Baltic Sea. *European Journal of Phycology*, *30*(2), 87-94.

Wang, J., & Christopher, S. A. (2003). Intercomparison between satellite-derived aerosol optical thickness and PM2. 5 mass: implications for air quality studies. *Geophysical research letters*, *30*(21).

Wang, M., & Gordon, H. R. (2018). Sensor performance requirements for atmospheric correction of satellite ocean color remote sensing. *Optics express*, *26*(6), 7390-7403.

Wang, M., & Shi, W. (2005). Estimation of ocean contribution at the MODIS near-infrared wavelengths along the east coast of the US: Two case studies. *Geophysical research letters*, *32*(13).

Wang, M., & Shi, W. (2007). The NIR-SWIR combined atmospheric correction approach for MODIS ocean color data processing. *Optics Express*, *15*(24), 15722-15733.

Walmsley, R. D., & Toerien, D. F. (1978). The chemical composition of the waters flowing into

Roodeplaat Dam. *Water S. A.*, *4*(4), 192-202.

Walmsley, R. D., Toerien, D. F., & Steÿn, D. J. (1978). An introduction to the limnology of Roodeplaat Dam. *Journal of the Limnological Society of Southern Africa*, *4*(1), 35-52.

Watanabe, F. S. Y., Alcântara, E., & Stech, J. L. (2018). High performance of chlorophyll-a prediction algorithms based on simulated OLCI Sentinel-3A bands in cyanobacteria-dominated inland waters. *Advances in Space Research*, *62*(2), 265-273.

Whitmire, A. L., Boss, E., Cowles, T. J., & Pegau, W. S. (2007). Spectral variability of the particulate backscattering ratio. *Optics express*, *15*(11), 7019-7031.

Wilson, R. T. (2013). Py6S: A Python interface to the 6S radiative transfer model. *Computers & Geosciences*, *51*(2), 166.

Wynne, T. T., Stumpf, R. P., Tomlinson, M. C., & Dyble, J. (2010). Characterizing a cyanobacterial bloom in western Lake Erie using satellite imagery and meteorological data. *Limnology and Oceanography*, *55*(5), 2025-2036.

Wynne, T. T., Stumpf, R. P., Tomlinson, M. C., Warner, R. A., Tester, P. A., Dyble, J., & Fahnenstiel, G. L. (2008). Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes. *International Journal of Remote Sensing*, *29*(12), 3665-3672.

Xi, H., Hieronymi, M., Röttgers, R., Krasemann, H., & Qiu, Z. (2015). Hyperspectral differentiation

of phytoplankton taxonomic groups: a comparison between using remote sensing reflectance and absorption spectra. *Remote Sensing*, *7*(11), 14781-14805.

Xue, K., Ma, R., Wang, D., & Shen, M. (2019). Optical Classification of the Remote Sensing Reflectance and Its Application in Deriving the Specific Phytoplankton Absorption in Optically Complex Lakes. *Remote Sensing*, *11*(2), 184.

Yacobi, Y. Z., Gitelson, A., & Mayo, M. (1995). Remote sensing of chlorophyll in Lake Kinneret using highspectral-resolution radiometer and Landsat TM: spectral features of reflectance and algorithm development. *Journal of Plankton Research*, *17*(11), 2155-2173.

Yacobi, Y. Z., Moses, W. J., Kaganovsky, S., Sulimani, B., Leavitt, B. C., & Gitelson, A. A. (2011). NIR-red reflectance-based algorithms for chlorophyll-a estimation in mesotrophic inland and coastal waters: Lake Kinneret case study. *Water research*, *45*(7), 2428-2436.

Yu, X., Shi, C., Ma, J., Zhu, B., Li, M., Wang, J., ... & Kang, N. (2013). Aerosol optical properties during firework, biomass burning and dust episodes in Beijing. *Atmospheric environment,* 81, 475-484.

Zeng, C., & Binding, C. (2019). The effect of mineral sediments on satellite chlorophyll-a retrievals from line-height algorithms using red and near-infrared bands. *Remote Sensing*, *11*(19), 2306.

Zhou, W., Wang, G., Sun, Z., Cao, W., Xu, Z., Hu, S., & Zhao, J. (2012). Variations in the optical

scattering properties of phytoplankton cultures. *Optics express*, *20*(10), 11189-11206.

Zhu, Y., Chen, X. B., Wang, K. B., Li, Y. X., Bai, K. Z., Kuang, T. Y., & Ji, H. B. (2007). A simple method for extracting C-phycocyanin from Spirulina platensis using Klebsiella pneumoniae. *Applied microbiology and biotechnology*, *74*(1), 244-248.

Zohary, T. (1985). Hyperscums of the cyanobacterium Microcystis aeruginosa in a hypertrophic lake (Hartbeespoort Dam, South Africa). *Journal of Plankton Research*, *7*(3), 399-409.

Xue, K., Ma, R., Wang, D., & Shen, M. (2019). Optical classification of the remote sensing reflectance and its application in deriving the specific phytoplankton absorption in optically complex lakes. *Remote Sensing*, *11*(2), 184.