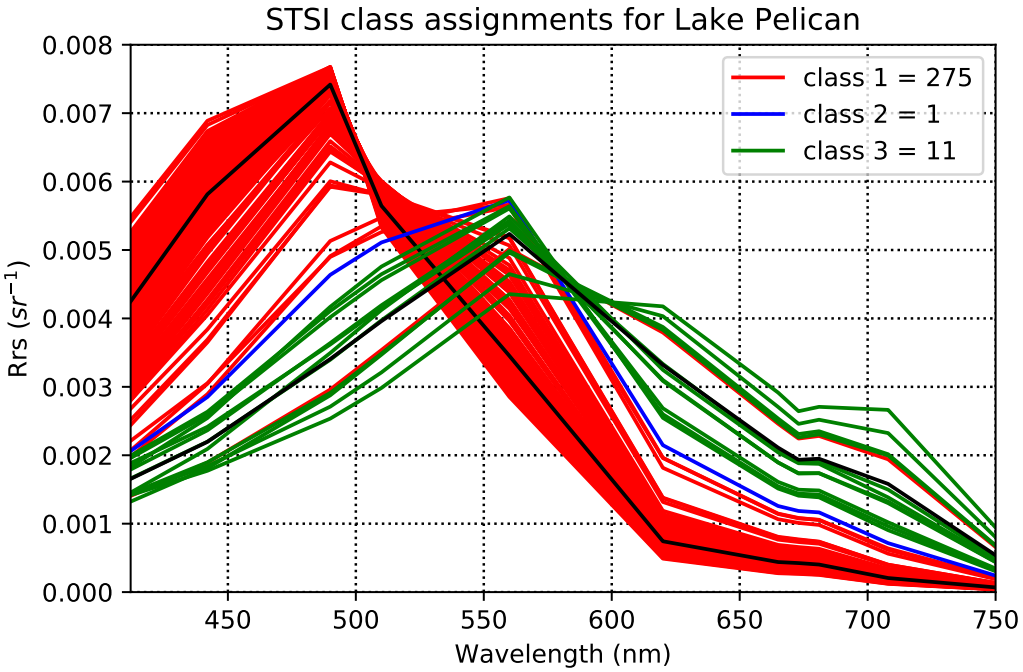
Geo-information Science and Remote Sensing

Thesis Report GIRS-2018-48

SPECTRAL TROPHIC STATE INDEX

A bio-optical algorithm to measure eutrophication in inland waters

Mortimer Werther



October 5, 2018

WAGENINGEN
UNIVERSITY & RESEARCH

Spectral Trophic State Index

A bio-optical algorithm to measure eutrophication in inland waters

Mortimer Werther

Registration number 91 09 15 941 130

Supervisors:

Jan Clevers
Kerstin Stelzer

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research,
The Netherlands.

October 5, 2018
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2018-48
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Acknowledgments

I thank my parents and my siblings, who gave me the strength, confidence and support to pursue the Master program at Wageningen University and undertake the research presented in this thesis.
The idea of the Spectral Trophic State Index emerged 2-3 years ago, in a conversation with Harald Krawczyk while being an intern at the German Aerospace Center (DLR) in Berlin. Back then the outline of the algorithm has been rather theoretical, but the fundamental question that remained until today is the same: can we classify a spectral library to find the trophic state classes? I learned a lot from Harald and he has been a great discussion partner throughout my undertaken work on the algorithm.
The overall development would not have been possible without financial support from Brockmann Consult (BC) during my studies. Not only that, but also the fruitful discussions and conversations with Carsten Brockmann and Kerstin Stelzer about the algorithm are highly appreciated. Special thanks to Kerstin, who not only supervised the whole development process, but also became a real mentor to guide my advancement in this sometimes very specific remote sensing field.
Besides the external supervision, I could not have asked for a better supervisor than Jan Clevers from WUR supporting the work with his large research expertise and helpful suggestions to improve my research methodological capabilities.

Abstract

A Trophic State Index (TSI) is a widely used classification system to measure eutrophication levels in surface waters. Besides nitrogen and phosphorous concentrations, chlorophyll-a (chl-a) can be used as the class indicator parameter. Traditionally, the TSI classes are estimated through in-situ samples. In recent times, the usage of remote sensors has evolved to complement in-situ measurement efforts, offering spatial and temporal revisit advantages. Hundreds of chl-a retrieval algorithms exist to calculate the desired chl-a concentrations from multi-spectral space born sensors. The most popular algorithms either use band-ratios or sophisticated inversion procedures, but they face several weaknesses as well as regional application limitations. In this study the Spectral Trophic State Index (STSI) algorithm has been developed. The aim of the STSI is to directly retrieve the trophic status and index classes without having to compute a chl-a concentration first. The core of the STSI algorithm is a gradient boosting machine classification algorithm (classifier) trained with synthetic remote-sensing reflectances (R_{rs}, sr^{-1}) that are the outcome of a hyperspectral forward simulation using the finite-element method. For every simulated R_{rs}, the chl-a concentration is known *a priori*, as it is one of the input parameters for the simulation procedure. Utilising the classifier, unseen reflectances from an ocean-colour sensor can then be classified into one of the pre-defined TSI classes to assign them a class label. As a result of the algorithm structure, chl-a retrievals using band-ratios or inversion procedures are circumvented. The STSI classifier can theoretically be applied to any ocean-colour sensor and its functionality has been exemplified using R_{rs} from Sentinel-3A OLCI. For this purpose, the two publicly available processors Idepix and C2RCC of ESA's Sentinel Application Platform (SNAP) were utilised. OLCI scenes over three different U.S. water bodies (Lake Pelican, Michigan and Jordan) were matched with in-situ measurements of in-water chl-a concentrations to validate the STSI class predictions. The results show that oligo- and eutrophic conditions were accurately predicted, with accuracy scores ranging from 0.66 to 1.00 for all in-situ stations in Lake Pelican and Michigan. The advantages of the proposed method are clearly determinable. The STSI algorithm does not require prior knowledge about local conditions or regional tuning as well as specific band combinations. Moreover, the design of the method proactively weakens the ocean-colour issue of spectral ambiguities. On the contrary, hypereutrophic conditions encountered in Lake Jordan are currently not well covered (accuracy of 0.28), revealing the limitations of the current STSI model. The extreme optical conditions faced in Lake Jordan indicate optical areas underrepresented in the simulated database. Optical water types (OWT) can help to constrain and validate the simulations. Coupled with a quality assurance system of calculated sensor R_{rs}, the STSI accuracies can be improved. A prototype of the STSI has been implemented in SNAP to enable the use as a eutrophication measurement tool for satellite imagery.

Keywords
Trophic State Index; Sentinel-3 OLCI; Hyperspectral remote-sensing reflectances; Gradient boosting machine; SNAP

Contents

List of Figures **V**

List of Tables **VI**

1 Introduction **1**
1.1 Thematic background . . . 1
1.2 Problem definition . . . 1
1.3 Research objectives and questions . . . 4

2 Spectral Trophic State Index framework **5**

3 Generation and processing of datasets **7**
3.1 Classifier training data . . . 7
3.2 Radiative transfer simulations . . . 8
3.2.1 Radiative transfer in atmosphere & water . . . 8
3.2.2 Simulation procedure . . . 9
3.2.3 Bio-optical model . . . 10
3.2.4 Calculation of remote-sensing reflectance database . . . 11
3.3 Trophic State Index labelling . . . 12
3.4 Spectral resampling . . . 15
3.5 Normalization . . . 17
3.6 Classifier test data . . . 17
3.6.1 Input datasets . . . 17
3.6.2 Idepix valid pixel classification . . . 18
3.6.3 Atmospheric correction . . . 19
3.6.4 Valid pixel selection . . . 20
3.6.5 Out-of-range check . . . 20
3.7 In-situ chl-a measurements . . . 20
3.8 Measurement selection for the validation . . . 21
3.9 Match-up generation . . . 22
3.10 Validation sites . . . 23
3.10.1 Lake Pelican . . . 24
3.10.2 Lake Jordan . . . 25
3.10.3 Lake Michigan . . . 27

4 Gradient boosting decision tree classification **29**
4.1 Supervised learning . . . 30
4.1.1 Loss function . . . 31
4.1.2 Parameters . . . 32
4.1.3 Learning algorithm . . . 33
4.1.4 Regularization . . . 33
4.1.5 Model complexity . . . 33
4.2 Tree-based methods . . . 34
4.2.1 STSI model structure . . . 35
4.2.2 Tree boosting . . . 36
4.3 STSI model hyperparameters . . . 38
4.4 Softmax prediction metric . . . 39

5 STSI operator **41**

6 Results **42**
6.1 Classification 42
6.1.1 Lake Pelican 42
6.1.2 Lake Jordan 43
6.1.3 Lake Michigan 45
6.2 Validation 47

7 Discussion **49**

8 Conclusion & outlook **52**

9 References **54**

List of Figures

1 STSI framework . . . 5
2 Processing scheme classifier training data . . . 7
3 Simulated remote-sensing reflectances . . . 16
4 OLCI remote-sensing reflectances processing chart . . . 18
5 Match-up locations of in-situ measurements and Sentinel-3 OLCI imagery . . . 23
6 Overview of the validation sites and selected lakes . . . 24
7 Lake Pelican OLCI imagery . . . 24
8 Lake Pelican original and normalized OLCI Rrs spectra . . . 25
9 Lake Jordan OLCI imagery . . . 26
10 Lake Jordan original and normalized OLCI Rrs spectra . . . 26
11 Lake Michigan, bay of Milwaukee OLCI imagery. . . . 27
12 Lake Michigan original and normalized OLCI Rrs spectra . . . 28
13 STSI classification model construction . . . 29
14 Tree partitions and CART . . . 35
15 Schematic of AdaBoost . . . 37
16 STSI SNAP Operator . . . 41
17 STSI class assignments for Lake Pelican . . . 42
18 Lake Pelican single STSI class plots . . . 43
19 Lake Pelican simulated vs. OLCI spectra . . . 43
20 STSI class assignments for Lake Jordan . . . 44
21 Lake Jordan single STSI class plots . . . 44
22 Lake Jordan simulated vs. OLCI spectra . . . 45
23 STSI class assignments for Lake Michigan . . . 46
24 Lake Michigan single STSI class plots . . . 46
25 Lake Michigan simulated vs. OLCI spectra . . . 46
26 Confusion matrices of the validation lakes . . . 48

List of Tables

1 Illumination geometries used in the FEMWAT simulation 9
2 Constituent scale of the bio-optical model 11
3 Simulated remote-sensing reflectances database statistics 13
4 Trophic State Index classes 14
5 Center wavelength and width of OLCI bands. 16
6 Probability vector of spectral class assignments. 40
7 In-situ measurement sites used in the validation 47

1 Introduction

1.1 Thematic background

Water eutrophication is a challenging environmental problem across the globe (Ayres et al., 1997; Yang et al., 2011). Since the 1970s, the worldwide severity of water eutrophication has increasingly gained attention of both the public and governments (Álvarez et al., 2017; Yang et al., 2008). The most predominant cause for eutrophication is the anthropogenic nutrient enrichment through the discharge of domestic wastes, non-point pollution from agricultural practices and urban development (Mainstone and Parr, 2002). Not only has increased nutrient loading severe impacts on the rich biological ecosystems found in aquatic environments, but it is also considered as a major threat to the health of coastal marine waters (Andersen et al., 2004). Once a water body is eutrophicated, it loses its primary functions and thus has consequences on wildlife communities, water supply, fishery and recreation activities (Wilkinson, 2017). To track changes in eutrophication levels of a water body, several trophic state indices have been formulated (Lambou et al., 1983). The most commonly used Trophic State Index (TSI) by Carlson (1977) has been adapted from several government institutions, namely the U.S. Environmental Protection Agency (USEPA), Brasil's São Paulo State Environment Company (CETESB) and the Chinese Environmental Protection Agency (CNEPA) (CNEPA, 2002; Novo et al., 2013; Rast and Lee, 1978).
Besides nitrogen and phosphorus, chlorophyll-a (chl-a) concentrations can be used to derive the classes of the TSI. Chl-a as the dominant active pigment in phytoplankton, including algae and bacteria, has a significantly positive relationship to phytoplankton biomass and is used globally as a simple proxy for phytoplankton in waters (Blanka, 1981). It is routinely measured by government agencies in water quality monitoring programs in the laboratory after extraction in an organic solvent or via direct measures based on fluorescence in vivo or in vitro (Gons et al., 2008; Schalles, 2006; Simis et al., 2007). In particular, chl-a is the vital indicator in monitoring programs to measure the impacts of eutrophication, ecological habitat status and health risks arising from harmful cyanobacteria and algal blooms (Matthews and Odermatt, 2015).
Conventional sampling methods to measure chl-a are labour-intensive and costly. Hence, they can not provide an extensive spatial and temporal coverage to capture the highly variable dynamics occurring in inland waters strongly influencing the concentrations of chl-a (Spyrakos et al., 2017). Fortunately, chl-a has a unique optical property measurable from remote instances allowing to estimate this parameter from remotely sensed data with wide spatial and temporal coverage.

1.2 Problem definition

To derive TSI classes from remote sensing imagery, current chl-a retrieval algorithms are utilised to calculate the concentrations of the parameter. The retrieved concentrations of chl-a are then assigned into one of the TSI classes (for application examples see: Novo et al. 2013; Papoutsa et al. 2014; Wang et al. 2005). For open ocean waters, algorithms for iteratively retrieving chl-a are well established using reflectance in the blue and green spectral regions, because the optical properties are generally controlled by phytoplankton and associated degradation products (Hu et al., 2012; Lee et al., 1996; O'Reilly et al., 1998). However, these algorithms fail when applied to inland waters because of highly differing, more complex optical properties (Dall'Olmo and Gitelson, 2005; Lavender et al., 2004; Le et al., 2011; Li et al., 2012; Sun et al., 2009). The reason for the inland water optical complexity is mainly due to the highly variable composition of the inherent optical properties (IOPs) (i.e. absorption, scattering and fluorescence) of chl-a, total suspended matter (TSM) and coloured dissolved organic matter (CDOM). The mass-specific relationship between chl-a and the other components can not be easily determined as in open ocean waters and is often lake specific (Loisel and Morel, 2001; Morel and Prieur, 1977;

Riddick et al., 2015; Sathyendranath et al., 1989). IOPs can not be determined by chl-a alone, but from the composition of all optically active substances leading to difficulties in precisely estimating chl-a with remote sensors (Gons et al., 2008; Le et al., 2009; Moses et al., 2012; Yacobi et al., 2011).

Two types of algorithms are most frequently used to retrieve chl-a concentrations: either based on a band-ratio or on numerical solutions that invert the water-leaving radiative spectrum. The most commonly applied are based on the difference or ratio between two bands derived from statistical relationships with the measured in-situ concentrations of chl-a (Dekker et al., 1991; Morel and Prieur, 1977; Stumpf and Tyler, 1988). To this category also algorithms belong that focus on the pursuit of particular signal attributes like the peak position and height of certain chl-a spectral features that seem to be unique, e.g. the chl-a fluorescence or cyanobacteria occurrences (Esaias et al., 1998; Matthews et al., 2012; Matthews and Odermatt, 2015). These algorithms are limited as they need well defined features in the water-leaving radiative spectra. This requirement is not always accomplished, especially in lakes with low chl-a concentrations and poor atmospheric corrections. Despite some of the limitations, these algorithms are usually comparatively easy to implement and provide robust concentration measures for regional lakes (Tyler et al., 2016). Other band-ratio approaches developed over the last decades are based on reflectances in the red and the near-infrared spectral regions. The results presented yield accurate chl-a estimates in local inland waters (Gilerson et al., 2010; Ruddick et al., 2001; Sun et al., 2009; Xu et al., 2010; Yacobi et al., 2011). Most of the parameters were obtained from relationships between data measured from remote instances and in-situ, both collected in a specific geographical or seasonal regime. These algorithms indeed are suitable for estimating chl-a of local waters or particular seasons. However, they require site-specific knowledge and training data. Hence, while the estimation of chl-a for one region is already a challenge, most of the specifically developed algorithms lack transferability to optically differing aquatic environments. The design of an algorithm for a local water type makes them less applicable to lakes outside of their training range and studies show that none of them is universally applicable (Moore et al., 2001; Kutser et al., 2001; Odermatt et al., 2012; Shi et al., 2013).

To overcome local confinement and to account for the prevailing contribution to the light field from inorganic and/or dissolved material, inversion algorithms were developed. These algorithms are strongly driven by an understanding of the relationships between the IOPs and the water-leaving signal through the use of physics-based bio-optical models. The inversion problem is examined as a two step process: first, the derivation of IOPs from measured radiance and second the biogeochemical parameters (such as chl-a) from the IOPs. Early approaches from Hoogenboom et al. (1998) used matrix inversion for retrieving chl-a and suspended matter. Another example is the adaptive implementation of the linear matrix inversion (LMI) method which iterates over a number of model parameter sets to account for the variability in the IOPs in a wide range of optically complex waters (Brando et al., 2012). Further, neural network inversion approaches have been optimised specifically for lakes or are partly transferable from the coastal zone setting to a range of inland waters (Brockmann et al., 2016; Doerffer and Schiller, 2007; Hieronymi et al., 2017). These models usually provide concentration normalized IOPs, so-called specific inherent optical properties (SIOP, i.e. absorption or scattering per unit mass). SIOP coefficient approximation is then required to convert the retrieved IOPs to the concentrations of water constituents like chl-a or TSM.

However, inversion-based algorithms face the major difficulty in solving the inverse problem of ocean-colour. The issues accompanying the inversion can be described using Equation 1 as an example of the relationships often used as the basis of many inversion algorithms:

$$R_{rs} = L_u/E_d = g\frac{b_{btot}}{a_{tot}}, \tag{1}$$

where R_{rs} is the sub-surface remote-sensing reflectance, defined by the ratio of absolute upwelling nadir radiance L_u (in $Wm^{-2}sr^{-1}$) to downwelling irradiance E_d (in Wm^{-2}) just beneath the water surface, b_{btot} is the total backscattering coefficient (in m^{-1}), a_{tot} is the total absorption coefficient (in m^{-1}) and g is a proportional factor (sr^{-1}). The difficulty originates from the fact that the relationship between the R_{rs} and the IOPs of each water component is not unique. IOPs have an additive property, meaning that several combinations of the water components IOPs can lead to the same, thus indistinguishable, reflectance spectrum making it difficult to ultimately retrieve precise concentrations of the in-water constituents. Further, a given value of the ratio b_{btot}/a_{tot} can be obtained from different values of b_{tot} and a_{tot}, resulting in a similar reflectance value as inferred by Equation 1. It is even in principle difficult to answer if two different sets of IOPs and boundary conditions can lead to the same radiance distribution. Because the solutions are not unique in practice, the inverse problem of ocean-colour is said to be ill-posed or ambiguous. Consequently, both the numerical part of the inversion (i.e. spectral optimization, linear and non-linear matrix algorithms) and the presence of ambiguities are sources of error in the measured IOPs necessary to retrieve the concentrations of the optically active constituents (Defoin-Platel and Chami, 2007). A further review about the numerical problems related to the spectral inversion algorithms is given in McCormick, 1992; Mobley, 1994.

To judge the performance of algorithms, it is necessary to distinguish between algorithmic performance bias originating from their retrieval techniques used (i.e. inversion or band-ratio) and the errors resulting from a failure of a proper atmospheric correction (AC) algorithm applied prior to the retrieval process to gain usable reflectances. Although the atmospheric correction is an external algorithm not linked to the retrieval, it is strongly influencing the final retrieval result.

The retrieval of the optically active substances in inland waters is a multifaceted task. As discussed, until now inland water remote sensing products are still facing issues. Nevertheless, this specific application field of ocean-colour remote sensing is rather young compared to the use of optical measurements from open oceans. Many new developments are emerging that show the potential for inland surface waters (Ogashawara et al., 2017; Schaeffer et al., 2013; Palmer et al., 2015).

From a methodological point of view, the purpose of chl-a retrieval algorithms is greater than retrieving eutrophication statements. Nonetheless, they are also currently the only method available to determine the trophic status of a surface water body using chlorophyll as the class indicator parameter. The herein developed Spectral Trophic State Index (STSI) provides an innovative alternative. To circumvent the aforementioned issues accompanying the retrieval process of chl-a, the proposed STSI algorithm avoids the retrieval of water constituent concentrations explicitly. The idea of the STSI algorithm is to view the TSI statements not as a retrieval, but rather as a classification problem. This is motivated by the circumstance that phytoplankton is just one of many lake parameters that have been used to biologically define the different eutrophication classes. The STSI inherently takes the perspective that chl-a is a linkage rather than the class defining parameter. The chl-a concentration ranges defined for the different trophic state classes are based on field knowledge that take into account the complex interactions in a lake environment, such as varying nutrient and oxygen levels, thermal stratification and light availability. The trophic state of a lake is defined by more than phytoplankton concentrations. Hence, it is simply necessary to retrieve the TSI classes, using the chl-a as the linkage, rather than as the class defining parameter. This is possible by the use of a supervised classification algorithm trained with synthetic simulated reflectances for which the TSI classes are known (based on the linkage parameter chl-a) *a priori*, as the chl-a concentration is an input into the simulation process already. The class boundaries to separate each class from another are already established for decades and the assumptions made have their reasoning mainly due to biological relationships found in lakes. The output of the STSI is a discrete TSI class for each reflectance of a satellite pixel. The aim is to provide a high-level information product

similar to the original TSI classes that even a non-remote sensing expert can interpret.

1.3 Research objectives and questions

Within this thesis several research questions are investigated:

1. Can the non-linear classes of the original TSI be adequately found using the STSI?
2. How well does the class assignment based on TSI chlorophyll-a concentrations work? And under which optical and eutrophic conditions?
3. How effectively deals the STSI with the issue of spectral ambiguities?
4. What are the limitations of the STSI and how reliable are the results?
5. Which future improvements can be identified?

Applying the STSI to the latest publicly available satellite imagery from the Sentinel-3A OLCI sensor will provide preliminary answers. Through the implementation of the STSI as a processor within the Sentinel Application Toolbox (SNAP), the tool might support lake managers and environmental policy makers in their use of remote sensing as a complementary source of information for surface water assessments.

2 Spectral Trophic State Index framework

Spectral remote-sensing reflectance (R_{rs}, sr^{-1}) is the fundamental key optical property for deriving ocean-colour retrievals such as chl-a (Toole et al., 2000). R_{rs} constitute the main spectral data source in the STSI, either used to train the classifier or to predict a TSI class on from a sensor.
The first part of the framework is about the simulated R_{rs}, calculated through a forward radiative transfer simulation using the finite-element method water-atmosphere transfer (FEMWAT) model (Bulgarelli et al., 1999). A dataset is then created using the simulated R_{rs} to train a modified supervised classification algorithm. This trained classifier is used to predict TSI classes on atmospherically corrected R_{rs} of a lake captured by an ocean-colour sensor. The result is the assignment of every R_{rs} to a previously defined trophic state class. Figure 1 shows the STSI algorithm structure and modules. These are interlinked through a framework that can be used to describe the manifold connections between them. The chain of classifier fitting, usage of satellite R_{rs} followed by the prediction and assignment to a TSI class builds the core of the STSI algorithm. Due to the modular structure of the framework it is possible to iterate and re-construct several of the entities providing the possibility to constantly update the algorithm.
The derived STSI includes several information features:

(a) **Probabilities**
The output values from the classifier are interpreted as a probability vector. This vector contains a membership probability of each spectrum belonging to an STSI class. This is the output of a function included in the supervised classifier's entity and discussed in section 4.4.

(b) **Class frequency**
Derived from the prediction of each pixel belonging to a STSI class, the class frequency can be used to elaborate whether a lake is eutrophically uniform or is actually consisting of multiple eutrophication classes. This is the relevant information product for lake managers, departments or policy makers.

(c) **Class of unknowns**
A class of unknowns as an additional class is created implementing an out-of-range check, including spectra that are unknown to the classifier and thus would most likely be falsely assigned to a class. This class of unkowns is not implemented in the first version of the STSI, but the general approach is outlined in section 3.6.5.

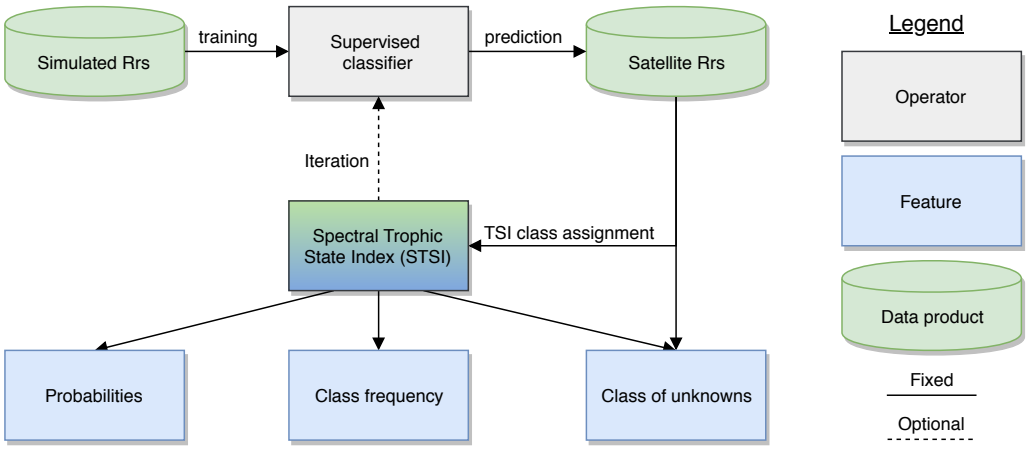


Figure 1: STSI framework. Simulated R_{rs} function as the training database for a classifer that is used to predict trophic state classes on satellite R_{rs}, resulting in multiple information features.

The STSI framework is applied to selected lakes and used to logically structure the thesis. First, all required data products within the STSI framework, classification and validation are described. Second, the classification assumptions and settings are envisioned that are used to predict the classes for the selected lakes. Third, the STSI results are validated using in-situ chl-a measurements being one of the first activities to proof the STSI concept. Finally, the results are presented and discussed.

3 Generation and processing of datasets

Several datasets are used in the STSI framework and the validation. The STSI classifier relies on simulated R_{rs}, thus the radiative transfer simulations are described initially. The description is followed by the supervised process to gain R_{rs} from lakes captured by the Sentinel-3 OLCI sensor (entity "Satellite R_{rs}" in Fig. 1). Lastly, the in-situ water constituent measurement data used to validate the STSI predictions are presented.

3.1 Classifier training data

A major entity within the overall STSI framework (see Fig. 1) are the simulated remote-sensing reflectances. These constitute the input training space for the classification algorithm (entity "Supervised classifier" in Fig. 1). Several processing steps are required to deduce high quality input training data, so that it can be readily used as the training database for the classifier. Within the framework to process the simulated R_{rs}, several entities are independent of each other providing the possibility to integrate changes and improvements without impacts on other entities. Figure 2 shows the entire flow chart of the processing steps to obtain the classifier training dataset. The first part consists of the simulated radiances required to gain the R_{rs} above the water surface. To simulate the

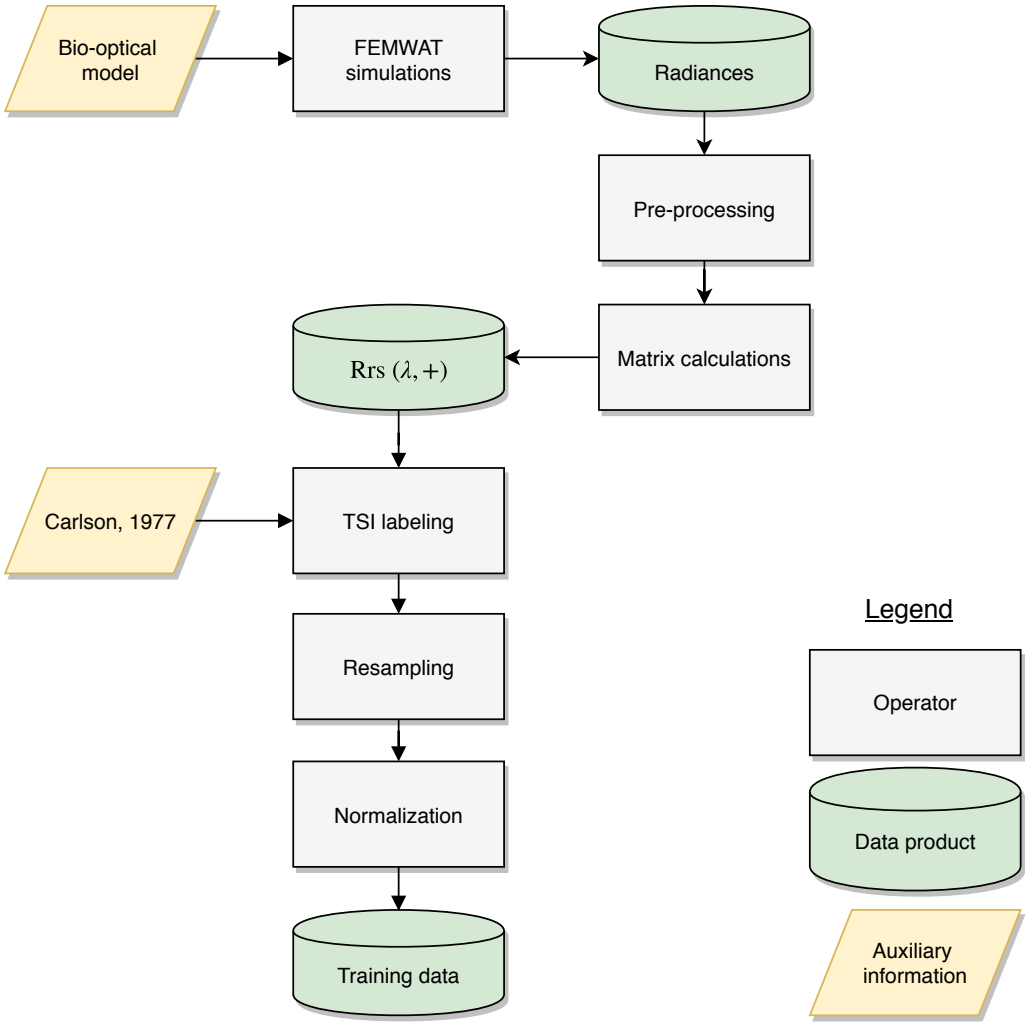


Figure 2: Processing chart of the classifier training data. The resulting fluxes from the FEMWAT simulation necessary to derive R_{rs} are pre-processed and subsequently used to calculate R_{rs}. Those spectra are then labelled based on the TSI from Carlson (1977), spectrally resampled to the target sensor's spectral response function, normalized and correctly formatted to finally constitute the training dataset for the classification algorithm.

optical complexity found in lakes, a bio-optical model has been created. These and other settings of the model are then input parameters to the radiative transfer simulations using the finite-element method (Bulgarelli et al., 1999). The radiances calculated using the FEM model are processed to bring them into format so that the required matrix algebra can be applied to derive R_{rs} (section 3.2.4).
The aim of the STSI is to assign R_{rs} satellite spectra into one of the pre-defined TSI classes based on non-linearly increasing chl-a concentrations. Because the chl-a concentrations of the simulated R_{rs} are known, a supervised classifier has been chosen. Prior to the classification process all R_{rs} spectra are labelled based on their chlorophyll-a concentration that has been used to simulate them (section 3.3). The class assignment is then based on the TSI classes from Carlson (1977). Additionally it is necessary to spectrally resample the labelled database to match the spectral response functions (SRF) of the target sensor, in this case Sentinel-3 OLCI. An algorithm has been developed enabling the resampling to basically any target sensor SRF (section 3.4). Further, the resulting spectra are normalized to base the classification on spectral shapes, rather than their amplitude (section 3.5).

3.2 Radiative transfer simulations

Numerical simulations of irradiance and radiance distributions, in water and atmosphere, have become increasingly important in optical oceanographic remote sensing applications during the last decades (Bulgarelli and Doyle, 2004). The retrieval of the only indirectly measureable R_{rs} from ocean-colour missions strongly relies on the capability of the accurate simulation of the radiative transfer processes in the coupled atmosphere-water system (AWS). It is necessary to solve the radiative transfer equation (RTE) correctly and model the propagation media appropriately. The desired, synthetic dataset requires the availability of (1) an accurate and comprehensive inland water bio-optical model that contains simultaneous measurements of the atmospheric and water IOPs (i.e. absorption and scattering coefficients and scattering phase functions), radiometric quantities (e.g. radiances and irradiances) and environmental parameters (e.g. surface roughness, wind speed) and (2) accurate codes for solving the RTE. The simulations were executed by Harald Krawczyk and Helge Witt from the German Aerospace Center (DLR) within the Photogrammetry and Image Analysis department in Berlin. The code is not publicly available.
The following sections contain the assumptions made using a radiative transfer model (RTM) with a coupled atmosphere-water system by applying a bio-optical model based on IOPs.

3.2.1 Radiative transfer in atmosphere & water

A formulation of the terms describing the radiative transfer in scattering and absorbing media is given by the classic RTE (Bukata, 1995; Gege, 2017; Jerlov, 1976):

$$\frac{\mathrm{d}L(\lambda, z, \theta, \varphi)}{\mathrm{d}r} = -c(\lambda, z)L(\lambda, z, \theta, \varphi) + L^*(\lambda, z, \theta, \varphi) + L^*_S(\lambda, \lambda', z, \theta, \varphi). \tag{2}$$

The RTE describes the change of radiance $\mathrm{d}L$ of a light beam of wavelength λ travelling distance $\mathrm{d}r$ in a medium at depth z in the direction (θ, φ). The first term $-c(\lambda, z)L(\lambda, z, \theta, \varphi)$ is the loss by attenuation, the second term $L^*(\lambda, z, \theta, \varphi)$ is the gain by elastic scattering and the third term $L^*_S(\lambda, \lambda', z, \theta, \varphi)$ is the gain by luminescence. For further adaptations (e.g. in the case of a plane parallel medium) and the derivation of the volume scattering function β, the interested reader is referred to Bukata 1995; Gege 2017.
For this thesis it is relevant to note that the RTE describes the propagation of radiation in turbid media (i.e. atmosphere, water) phenomenologically (Chandrasekhar, 1960; Mobley, 1994; Rybicki, 1996). Basically, the RTE is a differential-integral equation which has no analytical solution for realistic medium configurations and boundary conditions. Thus, it is necessary to adapt the solving approach to the illumination geometry, the shape, consistency and composition of the considered medium, in this case inland waters. While the

mathematical formulation of the medium water is already a challenge, also the illumination geometry is. Unlike in meteorological optics where the incident radiation is given by the Sun position, for inland waters the illumination is also influenced by the sky and possible clouds that illuminate the surface of the water hemispherically. Additionally, the incident radiation field under water has to be modelled, which is difficult due to the dynamics of the water surface (e.g. roughness, foam, waves, ripples) that refracts incoming radiation. In practice, the atmospheric conditions and water surface geometry are not known, thus approximate solutions of the RTE are used. Required are coupled models of the atmosphere, water surface and water body. Computational efficient numerical solutions are utilised to apply the RTE theory. A non-complete list of several methods to solve the RTE is given in Zhai et al. (2009). Solving the RTE in an AWS is of particular interest to simulate reflectances comparable to those measured with ocean-colour sensors. The numerical simulations were performed using the finite-element method water atmosphere transfer (FEMWAT) model solving the RTE for an atmosphere-inland water model (Kisselev et al., 1994, 1995; Bulgarelli et al., 1999). The FEM model has been extensively benchmarked with other popular radiative transfer codes that solve the RTE (Bulgarelli and Zibordi, 2018). Additionally, the code has been used to perform radiative transfer simulations in realistic cases (Bulgarelli, Mélin and Zibordi, 2003; Bulgarelli, Zibordi and Berthon, 2003; Zibordi and Bulgarelli, 2007).

3.2.2 Simulation procedure

FEWMAT simulates the illumination of the atmosphere by incident radiation from the Sun. It is modeled as an incoming parallel flux on the top boundary of the atmosphere in the ultraviolet, visible and near-infrared regions of the electromagnetic spectrum. The physical processes included in the algorithm are multiple scattering, bottom boundary bi-directional reflectivity, refraction and reflection at the interface between media with different refractive properties (i.e. atmosphere-water system). The air-water interface is assumed to be flat (Bulgarelli, Mélin and Zibordi, 2003; Bulgarelli and Doyle, 2004). Each propagation medium (atmosphere, water) can be divided into plane-parallel layers of uniform optical properties. For this study specifically, the atmospheric influence on the simulations has been excluded (see "Atmosphere" below for more details). For a mathematical formulation of the FEM numerical code the reader is referred to Bulgarelli et al. (1999); Bulgarelli and Doyle (2004). The output of FEWMWAT is the angular distribution of the diffuse radiance and direct and diffuse irradiance. A standard simulation model was designed for the inland water propagation medium with the elements presented in the next sections.

Spectral and angular resolution

The simulations are performed to reproduce the bands in the visible light range of any ocean-colour sensor, therefore the according wavelength range λ 400 - 1025 nm has been selected (in 5 nm steps). The simulations were azimuthally resolved and carried out for several Sun and viewing zenith angles (Table 1).

Table 1: Illumination geometries used in the FEMWAT simulation

Geometry	**Values**
Sun zenith angle (°)	20; 30; 40; 50; 60; 70; 80
View zenith angle (°)	0; 10; 20; 30; 40
Azimuthal difference (°)	0; 30; 60; 90; 120; 150; 180

Atmosphere

Within FEM, the atmosphere can be divided into plane-parallel layers resolving the aerosol, gas molecules and ozone vertical distributions. For the STSI, the database is not used as a look-up table (LUT) for an atmospheric algorithm, but to serve as the training database for the classifier that is then used to predict the class assignment of atmospherically corrected reflectances retrieved from lakes. To compare the simulated and at-sensor radiative quantities, the simulation of the FEM reflectance values has been carried out without atmospheric disturbances, thus representing values directly above the water level surface (also called bottom of atmosphere reflectances (BOA)). They are comparable to atmospherically corrected R_{rs} from OLCI.

Water surface and body

The water is assumed to be infinitely deep with a refractive index (n = 1.340) and the surface is assumed to be foam free and flat. The latter assumption is supported by initial considerations made in the development process of the FEMWAT model. In-situ measurements were taken with a wind speed lower than threshold U_B (Beaufort velocity) at which white-caps first appear. Monahan and O'Muircheartaigh (1986) state that U_B strongly depends on various meteorological and oceanographic conditions that influence the parameters. They related U_B (ms^{-1}) to the seawater surface temperature $T_w(°C)$ with

$$U_B = 3.36 \times 10^{-0.00309 T_w}. \tag{3}$$

This assumption of a flat water surface (i.e. no surface roughness) could induce significant differences in the comparison between simulated and sensor measured radiometric quantities (Mobley, 1994; Ronald et al., 2001). They are expected to increase with wind speed and Sun zenith angle. Other, unaccounted factors occurring in lake environments might further influence the conditions.

Sea water absorption and scattering

The spectral absorption and scattering coefficients (m^{-1}) for pure seawater are given by

$$a_w = 0.0257, \quad b_w = 0.00149. \tag{4}$$

The absorption and scattering of pure water are also influenced by salinity and temperature in the UV and NIR, and thus could be modelled as a linear expansion with coefficients for both (Simis et al., 2017). For this version of the STSI, the in Equation 4 provided static coefficients were used. FEMWAT is limited in its simulation capabilities as it does not account for inelastic scattering, other sources of light within the water body and it makes the generic assumption that inland water is freshwater with salinities close to zero.

3.2.3 Bio-optical model

The bio-optical model implemented is originally based on the three-component model of ocean-colour for coastal waters proposed by Prieur and Sathyendranath (1981), but extended to account for a higher constituent concentration range usually found in inland waters. The IOPs (i.e. absorption and backscatter) of the model are described separately and partitioned into the contributions from each optically active constituent.
Besides the absorption by sea water (a_w), three groups of substances can be considered responsible for significant modifications of the total absorption coefficient $a_{tot}(\lambda)$, describing the absorbing properties of water: phytoplankton $a_{ph}(\lambda)$, modelled using chl-a specific $a^*_{ph,\mathrm{chl}}(\lambda)$, 'non-algal' particles (NAP), i.e. the difference between total particulate and phytoplankton pigment absorption, modelled using TSM specific $a^*_{\mathrm{NAP,TSM}}(\lambda)$ and coloured dissolved organic matter a_{CDOM} modelled using $a_{\mathrm{CDOM}}(440)$ and CDOM specific value S = 0.014, the spectral slope of $a_{\mathrm{CDOM}}(\lambda)$. The absorption related to each

of the constituents chl-a, TSM and CDOM is modelled as a product of its concentration and the corresponding constituent specific absorption coefficients (SIOPs) a^*_{ph}, a^*_{NAP}, a^*_{CDOM} (m^{-1} per unit concentration). The concentrations are given in Table 2 and the SIOPs are taken from Prieur and Sathyendranath (1981). To describe the (back-) scattering properties, the total backscatter coefficient $b_{btot}(\lambda)$ consists of the coefficient for the backscatter of sea water $b_w(\lambda)$ and the backscatter of phytoplankton $a_\phi(\lambda)$, modelled as $0.0002142 \times \mathrm{chl}^{0.63}$. The scatter and backscatter of suspended matter $b_{b,\mathrm{NAP}}(\lambda)$ were modelled using TSM specific $b^*_{\mathrm{NAP,TSM}}(\lambda)$ accounting for all 'non-algal' particles that scatter. CDOM is generally assumed to be non-scattering (Riddick et al., 2015). The backscattering-to-scatter ratio has been assumed static (0.016) and non-seasonal specific. Several phase functions are required for water, phytoplankton and TSM. The scattering phase function of sea water is the Rayleigh phase function given in Bulgarelli et al. (1999), where the total phase function is modelled as a function of depth. For phytoplankton, the particulate phase function is the Petzold volume scattering function [$p_p(\cos\theta)$] as given in Petzold (1992) and for TSM it is the Two-term-Henyey-Greenstein (TTHG) phase function (Haltrin, 2002; Henyey and Greenstein, 1941).
By adopting concentration-specific IOPs from the literature it is assumed that these measured SIOPs are valid over the whole concentration range and all combinations of the constituents for which they were modelled. This is a simplification considered to be acceptable for the purpose of evaluating initial STSI prototype algorithm performances, but should not be considered strictly representative of variations in SIOPs that will occur in nature over such a wide concentration range.
The combined in-water and geometric permutations result in 125.000 R_{rs} spectra.

Table 2: Constituent scale of the bio-optical model

Constituent	**Concentration scale**	**Units**
Chlorophyll-a (chl-a)	0., 1., 1.1, 1.2, 1.3, 1.5, 1.6, 1.8, 2., 2.2, 2.4, 2.6, 2.9, 3.2, 3.5, 3.8, 4.2, 4.6, 5.1, 5.6, 6.2, 6.8, 7.5, 8.3, 9.1, 10., 11., 12.1, 13.3, 14.7, 16.2, 17.8, 19.6, 21.5, 23.7, 26.1, 28.7, 31.6 34.8, 38.3, 42.2, 46.4, 51.1, 56.2, 61.9, 68.1, 75., 82.5, 90.9, 100	mg m^{-3}
Coloured dissolved organic matter (CDOM)	0., 0.01, 0.0112, 0.0125, 0.0139, 0.0156, 0.0174, 0.0194, 0.0217, 0.0242, 0.027, 0.0302, 0.0337, 0.0376, 0.042, 0.0469, 0.0524, 0.0585, 0.0653, 0.0729, 0.0814, 0.0909, 0.1016, 0.1134, 0.1266, 0.1414, 0.1579, 0.1764, 0.1969, 0.2199, 0.2456, 0.2742, 0.3063, 0.342, 0.3819, 0.4265, 0.4762, 0.5318, 0.5939, 0.6632, 0.7406, 0.827, 0.9236, 1.0313, 1.1517, 1.2861, 1.4362, 1.6038, 1.791, 2.	m^{-1}
Total suspended matter (TSM)	0., 0.1, 0.115, 0.133, 0.154, 0.178, 0.205, 0.237, 0.274, 0.316, 0.365, 0.422, 0.487, 0.562, 0.649, 0.75, 0.866, 1., 1.155, 1.334, 1.54, 1.778, 2.054, 2.371, 2.738, 3.162, 3.652, 4.217, 4.87, 5.623, 6.494, 7.499, 8.66, 10., 11.548, 13.335, 15.399, 17.783, 20.535, 23.714, 27.384, 31.623, 36.517, 42.17, 48.697, 56.234, 64.938, 74.989, 86.596, 100.	g m^{-1}

3.2.4 Calculation of remote-sensing reflectance database

The simulated remote-sensing reflectances R_{rs} are not a direct model output, but have to be derived from up- and downwelling radiances and irradiances measured just above the water surface. The spectral irradiance reflectance (or irradiance ratio), $R(z,\lambda)$, is defined as the ratio of spectral upwelling to downwelling plane irradiances (Mobley, 1999):

$$R(z,\lambda) = \frac{E_u(z,\lambda)}{E_d(z,\lambda)}. \tag{5}$$

$R(z, \lambda)$ is the measure of how much radiance travelling in all downward directions is reflected upward back into any direction, that then can be measured by a cosine collector as existent in an ocean-colour sensor. The depth z is assumed to be in the air just above the water surface, but can theoretically be also any depth within the water column. FEMWAT outputs the diffuse and direct irradiance as E_d, whereas the diffuse part is neglectable, because the simulation is without atmospheric influences.
The spectral remote-sensing reflectance R_{rs} of interest is defined as

$$R_{rs}(\theta_s, \phi, \lambda) = \frac{L_w(0^+, \theta_s, \phi, \lambda)}{E_d(0^+, \lambda)} \quad (sr^{-1}), \tag{6}$$

for all viewing directions ϕ and Sun zenith angle θ_s, where the depth argument (0^+) indicates that R_{rs} is evaluated using the water-leaving radiance L_w and E_d in the air, just above the water surface (Lee et al., 1999). Generally speaking, the remote-sensing reflectance is a measure of how much of the downwelling radiance incident onto the water surface in any direction is eventually returned through the surface into a small solid angle $\Delta\Omega$ centered on a particular direction (θ, ϕ). These radiances strongly depend on wavelength. L_w can not be measured directly and the output of FEMWAT is thus the absolute upwelling radiance L_u (times π to account for the small solid angle $\Delta\Omega$) above the surface, being the sum of the water-leaving radiance L_w and the downward Sun and sky radiance that is reflected upward by the sea surface L_{surf} (Lee et al., 1999).
The FEMWAT radiances are subsequently used to compute the $R_{rs}(\theta_s, \theta_v, \phi)$ for all viewing directions θ_v, ϕ for the given Sun zenith angle θ_s and the IOPs as well as the boundary conditions (e.g. wind speed):

$$R_{rs}(\theta_s, \theta_v, \phi, \lambda) = \frac{L_u(0^+, \theta_s, \theta_v, \phi)}{E_d(0^+, \theta_s)} \quad (sr^{-1}). \tag{7}$$

Therefore the derived FEMWAT R_{rs} also incorporate BRDF effects for various viewing directions (see Fan et al. (2016); Lee et al. (2011) for a background on this topic).
It is to note that - either simulated or sensor-retrieved - R_{rs} are imperfect light field measurements. In inversion algorithms these measures are used to retrieve as much information as possible about the water constituents used to derive a TSI class. However, R_{rs} is far off from being a measure of the full radiance distribution and the measurements we do have may contain substantial errors due to (but not limited to) inaccurate simulation procedures, poor atmospheric corrections and inaccurate radiometer calibration. Moreover, using the R_{rs} to retrieve the in-water concentrations, it should be expected *a priori* that it is not possible to recover a full set of water IOPs, and that even what is recovered may contain large errors. Avoiding to retrieve the constituent concentrations from R_{rs} to derive chl-a concentration values is one of the key considerations that led to the development of the STSI.

3.3 Trophic State Index labelling

Resulting from the forward simulation that makes use of a bio-optical model for inland waters, for every simulated R_{rs} the chlorophyll-a concentration is known *a priori*. The TSI classification system by Carlson (1977) defines the trophic classes based on non-linearly increasing chlorophyll-a concentrations (see Table 4). Using the chlorophyll-a ranges of the TSI, every simulated spectrum in the database has been assigned a class label (1-4). Basic statistics describing the distribution of the spectra are given in Table 3. The TSM and CDOM median and mean values are similar for all classes, as the simulation primarily varies TSM and CDOM within smaller concentration changes in the TSI defined chl-a ranges. Consequently, the whole range of TSM and CDOM concentrations is reached in every TSI class (based on chl-a) resulting in the same statistical values. The overall value range is 0 to 100 chl-a (mg m^{-3}), defining the lower and upper concentrations limits of

the STSI (see Table 2). This database is then appropriately formatted and resampled (see next section).
An advantage following this approach is that the original TSI classification itself defines ranges for every class (e.g. oligotrophic 0 - 2.6 mg m^{-3}), allowing for more spectral flexibility inherently including different concentration combinations. Moreover, it weakens the issue to perfectly retrieve a chl-a concentration as opted for by the classic retrieval algorithms: a range itself does not require one specific concentration, but allows several concentrations to count as a correct class assignment. Furthermore, this partly has an influence on the previously described problem of spectral ambiguities (several different combinations of water constituent concentrations resulting in the same spectrum) that might not be as predominant using the STSI classification. Classifying spectra based on their initially defined chl-a values takes into account several constituent combinations already, that might create the same spectrum, but are still correctly classified. Ultimately, a class range takes into account higher dynamics of the water constituent concentrations and thus reflects the natural intermediate states of trophic states more realistically.

Table 3: Simulated remote-sensing reflectances database statistics

Statistical features					**Units**
Class	1	2	3	4	
Chl-a median	1.55	4.60	21.50	78.75	mg m^{-3}
Chl-a mean	1.55	4.8	25.36	79.73	mg m^{-3}
TSM median	2.95	2.95	2.95	2.95	g m^{-1}
TSM mean	14.90	14.90	14.90	14.90	g m^{-1}
CDOM median	0.134	0.134	0.134	0.134	m^{-1}
CDOM mean	0.38	0.38	0.38	0.38	m^{-1}
Number of reflectances	30.000	27.500	52.500	15.000	

Table 4: Trophic State Index classes based on Carlson, 1977

TSI index	CHL (ug/L)	Attributes	Water supply	Fisheries & recreation
<30	<0.95	**Oligotrophy:** Clear water, oxygen throughout the year in the hypolimnion.	Water may be suitable for an unfiltered supply.	Salmonid fisheries dominate.
30 - 40	0.95 - 2.6	Hypolimnia of shallower lakes may become anoxic.		Salmonid fisheries in deep lakes only.
40 - 50	2.6 - 7.3	**Mesotrophy:** Water moderately clear, increasing probability of hypolimnetic anoxia during summer.	Iron, manganese, taste and odor issues worsen. Raw water turbidity requires filtration.	Hypolimnetic anoxia results loss of salmnoids. Walleye may predominate.
50 - 60	7.3 - 20	**Eutrophy:** Anoxia hypolimnia, macrophyte problems possible.	Episodes of severe taste and odor possible.	Warm-water fisheries only. Bass may dominate.
60 - 70	20 - 56	Blue-green algae dominate, algal scums and macrophyte problems.		Nuisance macrophytes, algal scums, and low transparency may discourage swimming and boating.
70 - 80	56 - 155	**Hypereutrophy:** Light productivity is limited. Dense algae and macrophytes		
>80	>155	Algal scums, few macrophytes		Rough fish dominate, summer fish kills are possible.

3.4 Spectral resampling

In the STSI framework, the simulated R_{rs} represent the training database for the classifier. The aim is to classify the OLCI sensor reflectances into one of the classes defined in the training database. For this classification, a necessary requirement is a common band discretization. The simulated R_{rs} spectra are of full width half maximum (FWHM) 5 nm hyperspectral resolution simulated over the entire visible to near-infrared part of the electromagnetic spectrum (400 - 1020 nm). The spectral resolution is higher than any ocean-colour sensor currently available that the STSI classification can be applied to.
Multi-spectral remote sensors represent spectral signals of viewed surfaces in discrete wavelengths or spectral channels. Every spectral channel is defined through a wavelength dependent course of their relative sensibility, often also called spectral response function (SRF). The simulated, hyperspectral database has to be convolved to the OLCI SRF to compute band-integrated R_{rs} for each OLCI band i. This process is also called spectral resampling and can be divided into two steps (Pahlevan et al., 2017; Witt, 1998).
The first step is to calculate the representative center wavelength of every band (see Table 5):

$$\lambda_i = \frac{\int_{\lambda_1}^{\lambda_2} \lambda \phi_i(\lambda) E_0(\lambda) \mathrm{d}\lambda}{\int_{\lambda_1}^{\lambda_2} \phi_i(\lambda) E_0(\lambda) \mathrm{d}\lambda}, \tag{8}$$

where

λ	is the wavelength,
λ_i	is the center wavelength in the i-th spectral band,
ϕ_i	is the spectral response function in the i-th spectral band,
E_0	is the solar irradiance and
λ_1, λ_2	are the boundary wavelengths of the considered spectral range.

For public sensors, these values are usually also distributed by the respective space agencies. It is then necessary to calculate the values of the spectral albedo for the bands:

$$\overline{R}(\lambda_i) = \frac{\int_{\lambda_1}^{\lambda_2} R(\lambda) \phi_i(\lambda) \mathrm{d}\lambda}{\int_{\lambda_1}^{\lambda_2} \phi_i(\lambda) \mathrm{d}\lambda}, \tag{9}$$

where

R	is the spectral albedo,
$\overline{R}(\lambda_i)$	is the mean spectral albedo in the i-th spectral band.

The mean values of the spectral albedo in the spectral bands are used to represent the spectral albedo of a sensor. This is often also called spectral signature of the viewed surface. The range and values of the spectral response function for each of the spectral bands can be publicly downloaded from ESA. A comparison of the original and resampled simulated spectra is illustrated in Fig. 3.

Band 1 (400 nm) has been omitted for several reasons. First, this wavelength is mostly used for atmospheric correction (AC) purposes and hence strongly influenced by the quality of an AC algorithm (see section 3.6.3). Second, additionally to the already existent difficulties arising from AC procedures, OLCI calibration in shorter wavelengths on Sentinel-3A is off compared to MERIS, increasing the uncertainty of quality derived reflectances. It has been reported that calculated R_{rs} using the C2RCC processor can result in values too high in the blue part of the spectrum (Toming et al., 2017). This is not only related to C2RCC, as the problems with atmospheric correction procedures are generally the highest in the blue part of the spectrum. Even if another AC module would have been used, the AC process will remain difficult anyway, as atmospheric and sun glint effects are the highest in the blue spectral range (Ligi et al., 2017). ESA has ongoing activities to improve the OLCI calibration in these shorter wavelength ranges and for Sentinel-3B the calibration is in-line with their sensor requirements.

Table 5: Center wavelength and width of OLCI bands. Yellow band is new on OLCI compared to MERIS, red band has been omitted in this study.

Center (λ_i)	**Band width (nm)**
400	10
412	10
442.5	10
490	10
510	10
560	10
620	10
665	10
673.75	7.5
681	7.5
708.75	10
753.75	7.5

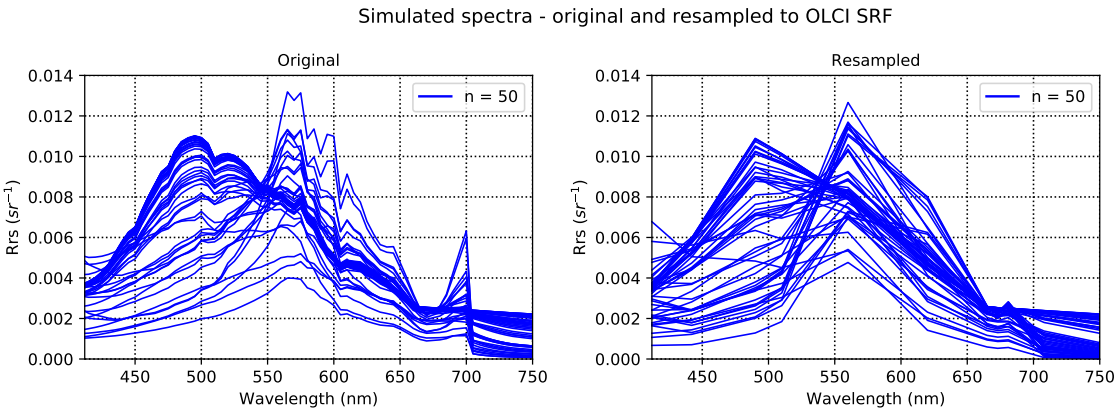


Figure 3: Random subset (n = 50) of the simulated spectra. Left shows the original R_{rs}, right the resampled version.

3.5 Normalization

In the past, optical classification studies aimed at classifying reflectance spectra based on raw R_{rs} (Lee et al., 2010b; Mélin et al., 2011; Moore et al., 2001, 2009). However, the latest studies shifted towards a prior normalization of the reflectances (Mélin and Vantrepotte, 2015; Xi et al., 2015, 2017). It has been recognized that all optically significant constituents have an effect on both amplitude and shape of R_{rs} and that the relationship between the two is most-likely not unique (Loisel and Morel, 2001; Sathyendranath et al., 1989). The variance shifts in non-normalized R_{rs} are associated with back-scattering and/or to the concentration of (mainly non-chlorophyllus) particles, whereas absorption by phytoplankton cells, CDOM or detritus predominantly have an impact on the spectral shape (Mélin and Vantrepotte, 2015). As a consequence, when the R_{rs} spectra are normalized, a preference is given to the latter in the classification process. This avoids results of class assignments to be based on a gradient of scattering particle concentrations. For this application it is highly preferable, as the assignment of the spectra to the pre-defined class ranges is then based on the spectral shapes, primarily influenced by chlorophyll and CDOM concentrations allowing for a more precise class assignment and differentiation between the classes.
In order to reduce the variance of the reflectances and to focus on the reflectance spectral shape, each R_{rs} spectrum was normalized prior to the classification. Using trapezoidal integration every spectrum was normalized by its integrated value (i.e. the surface below the spectrum) between λ_1 and λ_2, based on the following formula (for each wavelength (λ)):

$$r_n(\lambda) = \frac{R_{rs}}{\int_{\lambda_1}^{\lambda_2} R_{rs}(\lambda)\mathrm{d}\lambda} \tag{10}$$

3.6 Classifier test data

As introduced in the description of the STSI framework, quality ocean-colour reflectances are a key item of interest. To end up with TSI classes for every pixel of an inland water, a combination of several processing steps is required. For the thesis, several of the processing entities are exercised manually (supervised) using SNAP that has evolved from its predecessor BEAM (Fomferra and Brockmann, 2005). A future objective is to test the processing chain on a large amount of tiles. The final processing chain will be deployed on the data processing cluster Calvalus developed and run by Brockmann Consult (BC) (Fomferra et al., 2012). Figure 4 shows the entities of the processing chain that are applied to all test sites.

3.6.1 Input datasets

The OLCI sensor has been selected to test the supervised processing chain. Other optical ocean-colour sensors such as MERIS, SeaWiFis, VIIRS or MODIS can theoretically also be utilised with the STSI, only requiring a resampling of the simulated database to match their SRF. It is to note that sensor specific band placements play a role, as the resampling is strongly affected by the band positions. Further tests on these differing sensors would be required.
Throughout the processing of the OLCI scenes two different levels occur: level 1 (L1) consisting of top-of-atmosphere (TOA) signals and level 2 (L2) atmospherically corrected products including derived geophysical quantities (such as the trophic classes for each pixel). Level 3 (L3) products representing spatio- and temporally aggregated data are of long-term interest to showcase trophic state changes overtime, but are not exercised in this thesis. OLCI FRS L1 images were used as the input to the processing chain.

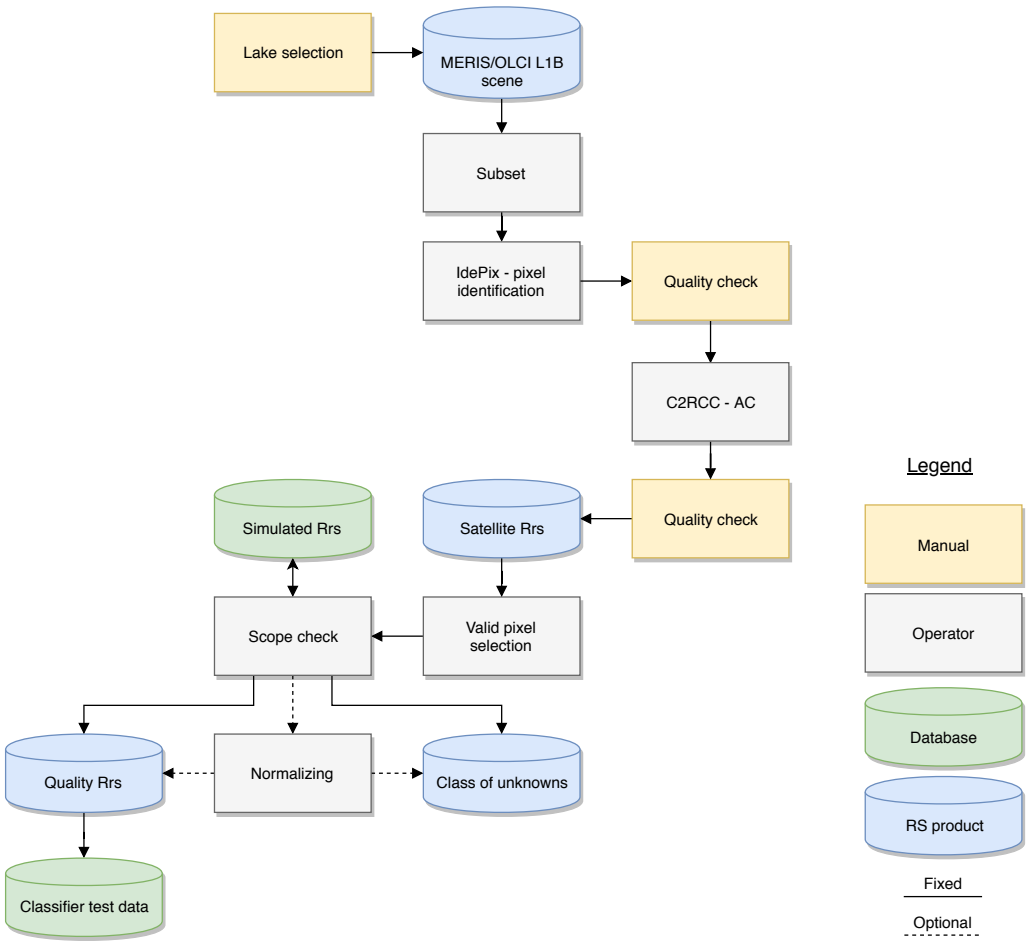


Figure 4: Ocean-colour sensor processing scheme used to gain remote-sensing reflectances to predict the TSI classes on. First, the selected lakes are processed with the Idepix tool for a scene classification to flag invalid pixels. Second, an atmospheric correction processor (C2RCC) is applied to correct for atmospheric disturbances. Both processing steps are manually reviewed. The two processing steps enable a valid pixel selection. Subsequently, the spectra can then be used in a range check to test whether a satellite spectrum is within the range of the simulated R_{rs} values. In a future version of the STSI, the spectrum is then normalized and included in the pool of spectra the classifier is applied to. Else it is assigned to a class of unknowns as an additional information feature.

As apparent from Figure 4, the first entity in this chain is the lake selection that includes a user's area of interest (AOI), captured on an OLCI scene. These lakes and test sites are described in section 3.10. The selected water bodies show varying trophic states, they differ in their ecological statuses and are large enough to enable quantitative tests. For the test chain, a subset of the full scene has been selected, mainly covering the relevant lake parts. To further avoid erroneous pixels, only visibly cloud-free scenes at L1 RGB level were selected.

3.6.2 Idepix valid pixel classification

An essential step to correctly classify the pixels into one of the TSI classes is the identification of pure water pixels. Pixel contributions of land surfaces can strongly affect the resulting class assignment. The simulated R_{rs} only account for pure water R_{rs}, hence it is even more crucial to only classify pure water pixels. For this procedure the Idepix algorithm has been selected, as it is an open source SNAP processor and performs the identification of land, clouds, cloud shadows, cloud buffers, snow/ice, Sun glint and ambiguous mixed pixel areas (Danne, 2016). Idepix is based on molecular scattering corrected bottom-of-

Rayleigh reflectances (BRR) and therefore avoids an error-prone aerosol correction over optically complex waters. BRR already includes the partial correction of atmospheric effects, as it represents reflectance at an hypothetical boundary between infinitesimal-sized aerosol- and gaseous air layers above. When background reflectances for the estimation of aerosol optical thickness (AOT) are highly uncertain, it is the preferred signal. Therefore BRR intermediate products are also used for the identification of non-optically deep water areas such as shallow areas in close proximity to land (Odermatt et al., 2018; Santer et al., 1999). The identification of shallow areas are not implemented in the Idepix version used to generate the valid pixel expression. With the application of Idepix to a L1B OLCI scene the first part of the valid pixel selection (section 3.6.4) becomes available. Manually quality checking the results and flags further improves this part of the valid pixel selection process.

3.6.3 Atmospheric correction

The signal received by the Sentinel-3 OLCI sensor contains an optical return from the atmosphere. One of the most crucial steps therefore is to correct the total radiative signal for atmospheric noise and disturbances (IOCCG, 2000; Saythendranath, 1986). In an AWS, the TOA radiance $L_{\mathrm{TOA}}(\lambda)$ is calculated from its distinctive physical contributions:

$$L_{\mathrm{TOA}}(\lambda) = L_r(\lambda) + L_a(\lambda) + L_{ra}(\lambda) + T_u(\lambda)L_{wc} + T_u(\lambda)L_g + T_u(\lambda)T_d(\lambda)\cos\theta_s^+[L_W(\lambda)]_N, \tag{11}$$

$$L_{\mathrm{path}}(\lambda) = L_r(\lambda) + L_a(\lambda) + L_{ra}(\lambda), \tag{12}$$

where

$L_{\mathrm{TOA}}(\lambda)$	is the top-of-atmosphere (TOA) radiance,
L_r	is molecular scattered radiance,
L_{ra}	is combined molecular aerosol scattered radiance,
L_{wc}	is reflectance due to white caps on the water surface,
L_g	is the specular reflection of direct sunlight at the water surface (sun glint),
T_d	denotes the downwelling and
T_u	the upwelling diffuse atmospheric transmittances from TOA to the target and back,
$[L_W(\lambda)]_N$	is the radiance L_w normalized to nadir direction,
$L_{\mathrm{path}}(\lambda)$	defines the TOA atmospheric path radiance (including both contributions from atmospheric scattering and surface reflection).

Atmospheric correction schemes generally try to estimate L_{path} and remove it from L_{TOA} (IOCCG, 2010).
For the AC in this framework, the neural network of the Case 2 Regional Coast Colour (C2RCC) processor is used on the L1C (Idepix) product to derive remote-sensing reflectances (Brockmann et al., 2016). The Case 2 Regional (C2R) part originates from a processor that originally has been developed by Doerffer and Schiller, 2007. Through the CoastColour (CC) project improvements were implemented. The C2RCC processor relies on large databases of simulated water-leaving reflectances and related TOA radiances. Neural networks were trained in order to perform the inversion of spectra for the atmospheric correction, thus the determination of the water-leaving reflectances from TOA radiances as well as the retrieval of IOPs for the respective water body. The IOP retrieval is not of interest for the STSI, but the AC has been specifically designed for complex water types and thus can be utilised. Similar to Idepix, it is made available as an open source

processor in SNAP. Within the C2RCC processor settings it is possible to automatically compute the AC reflectances as remote-sensing reflectances. Again, the processing result is manually checked using the processor inherent flags.
It is important to note that the atmospheric correction process has a strong influence on the derived spectral shapes of the R_{rs} possibly leading to erroneous STSI class assignments. As described in the spectral resampling section (3.4), calibration issues on Sentinel-3A exist that might influence the retrieved R_{rs} values.

3.6.4 Valid pixel selection

Combining the flags from both Idepix and C2RCC, the quality filtered subset can then be used to manually create a polygon area of the lake areas deemed as pure water pixels. The raw R_{rs} of a scene have to be normalized (with the same method described in section 3.5) to constitute the test dataset for the classifier to enable predictions of the TSI classes. Integrating over a spectrum changes its scaling, making it necessary to apply the normalization procedure to both datasets. Otherwise similar spectral shapes have an offset leading to erroneous class assignments.

3.6.5 Out-of-range check

All bands on OLCI include for every AC corrected subset of the respective lake a reflectance value for every valid pixel. This reflectance value can be checked with previously computed *min* and *max* values of the reflectance values per band of the simulated R_{rs}. The out-of-range check is performed on the non-normalised version of the R_{rs}. The comparison has the aim to check whether or not the value of each band of an OLCI scene is within the value range of the simulated database. If this is not the case, they are assigned to a 5th class, the class of unknowns. The absolute min/max values of the simulated database are computed prior to this range check. The range check is not yet implemented in the SNAP algorithm and operator (Chapter 5).

3.7 In-situ chl-a measurements

Publicly available in-situ data are often difficult to find due the obstructive combination of license terms and the dedicated use for a user's specific purpose. Regarding water constituent data, the U.S. Environmental Protection Agency (EPA) offers freely available water constituent in-situ data sampled from water bodies throughout the entire USA. The availability is based on the Clean Water Act (CWA) which is a federal legislation that established the basic structure for regulating quality standards for surface waters in the USA. The main aim of the legislation is to restore and maintain the chemical, physical and biological integrity of the U.S. waters (USEPA, 2002). In addition, all navigable water bodies in the USA are protected by the CWA from 1988. This federal mandate authorizes states, tribes and U.S. territories, with guidance and oversight from the EPA, to develop and implement water quality standards. They include designated use cases, defined as the services that a water body provides, e.g. drinking water, aquatic life, harvestable species and recreation. These standards are applicable within state waters, defined as $<$ 3 nautical miles from shore. Therefore, a majority of water quality management decisions address near shore coastal waters, estuaries, lakes, reservoirs, rivers and streams where applicable water quality regulation can be implemented (Schaeffer et al., 2013). Both the EPA and the environmental protection departments at the state level recognize that water resources can not be managed without monitoring. At the federal level, section 305(b) of the CWA directs each state to (1) prepare and submit a report every two years that includes a description of water quality of all of its navigable surface waters to the EPA and (2) protect balanced indigenous populations (FDEP, 2015). Those reports, referred to as 305(b) reports, describe surface and groundwater as well as trends of water quality

and major impacts on them (FDEP, 2012).
Being forced to report periodically on the condition of the nation's waters, the departments transfer large amounts of water quality data about the majority of the monitored lakes to a database management system run by the EPA and United States Geological Survey (USGS). Usually, most of the monitoring is performed through routine samples taken from observation stations or by state-specific projects to carry out water quality monitoring for specific events. Further, in 2007 and 2012 two National Lake Assessments (NLA) were realized from the EPA to overcome different approaches of collecting and evaluating data that varies from state to state, making it difficult to compare the information across states, on a nationwide basis, or over time (USEPA, 2009, 2012). Simultaneously, every two years the departments are in charge to hand in a report that lists the general water quality of the lakes (referred to as the 305 (b) report) in their state. The gathered in-situ water quality data is available through various databases online (e.g. STORET, Water Quality Portal) and can be utilized by any public user.
A dataset has been compiled including chl-a measurements from 2002 - 2018 for all inland water bodies that are resolvable from MERIS and OLCI. In this thesis the focus is on OLCI, thus a subset of the actual database is used (01/2016 - 04/2018). The raw in-situ measurements require several processing steps to be compared to the STSI class predictions. These pre-processing steps are described in the following section.

3.8 Measurement selection for the validation

Several processing steps have been exercised prior to the usage of the in-situ data for the validation of the STSI predictions:

1 Valid measurement location
The position of the measurement stations used for the validation need to fulfil certain criteria in order to guarantee a good comparability with the satellite data. They need to be taken in optically deep water ($>$ 15 m) to avoid influences of the bottom albedo on the retrieved reflectances utilised for the STSI from the same location (Albert and Mobley, 2003). In addition, it is to recognise that eventual perturbations from the nearby land areas influence the optical signals retrieved. TOA radiance contamination between neighbouring surfaces with different reflectances is usually called adjacency effect (AE) (Bulgarelli and Zibordi, 2018). To neglect reflectances from the surrounding land and bottom, a filter grid has been designed.
First, a land/water mask for the U.S. has been created. Second, a 3x3 matrix grid has been designed and added as a layer. The size of each cell within the 3x3 matrix is 300m, thus matching the spatial resolution of a pixel from OLCI and MERIS. Third, the in-situ chl-a measurements have been spatially queried with the 3x3 grid to discard every in-situ location that is not surrounded by at least 8 complete neighbouring pixels. This procedure excludes all shoreline pixels. The assumption is made that in lakes 300 meters from shoreline the pixels only include optically deep water and that AE influences do not occur due to the distance. In combination with the selected lakes, all exceeding the optically deep water minimum depth, these are reasonable assumptions.

2 Quality filtering of the measurements
The measurements provided by the EPA databases are frequently gathered from different environmental state departments, entities, private persons, universities, environmental programmes and require an extensive quality filtering to ensure only reliable measurements are used in the validation.

a Chlorophyll
Several types of chlorophyll were included and deselected, as not all are compa-

rable to chl-a (i.e. chlorophyll-b, c) used in the bio-optical model to generate the simulated R_{rs}.

b Analytical measurement identifier (IDs)
Plenty of different analytical IDs provide information about the respective measurement method used to extract the chl-a concentrations. No single document exists listing all methods used, but a database covers most that can be queried online: `https://www.nemi.gov/home/`. The measurements were filtered for IDs for which clear protocols exist.

c Unit harmonisations
Measurement units have been unified to ug/l.

d Measurement depth
Only surface measurements were used, evaluated for each lake individually. Underwater radiometric in-situ measurements were not available measuring the diffuse attenuation. Under ideal conditions, for which the incident light is provided by the Sun, the various radiances and irradiances all decrease approximately exponentially with depth, when they are far enough below the surface to be free of boundary effects. Radiometric measurements were not available to measure the available light influencing the measurement of R_{rs} from the sensor to precisely account for each possible (or maximum) depth per measurement.

e Time zones
The U.S. mainland has several timezones, but the measurement time of the European sensors OLCI and MERIS is provided in GMT/UTC. Consequently, all U.S. dates have been converted.

For most of the chl-a measurements standard fluorometric methods were used. Yet, several studies report that fluorometric analyses underestimate chl-a values (Kumari, 2005; Pinckney et al., 1994). They recommend to use the more precise high performance liquid chromatography (HLPC) methods. However, HLPC samples are more costly and take longer to process per sample than fluorometric ones. Basically all of the measurements in the validation database are not based on HLPC methods. Consequently, an uncertainty might also be introduced using these in-situ measurements to compare the STSI predictions to. It would actually require the extraction techniques and analytical procedures to be water type dependent, which is completely out of scope regarding the nature and purpose of this publicly available database.

3.9 Match-up generation

The in-situ measurements were used on Calvalus (BC's processing infrastructure) to generate match-ups with Sentinel-3A OLCI imagery. In-situ data has been acquired from January 2016 onwards, while the first usable OLCI product is from the 23rd of November. Therefore the time frame of possible match-ups is restricted to 23/11/2016 until 01/04/2018 (end of in-situ data collection). During the winter time in most of the U.S. states in-situ measurements were not collected, reducing the availability of match-ups effectively to the year 2017 (see Figure 5).
A time range of +- 5 hours between the actual in-situ measurement and the satellite overpass has been chosen. This relatively strict time constraint ensures a considerably stable water column at the match-up location. An identical measurement time between sensor and in-situ is nearly impossible to find in the database, as the original measurement purpose has not been to use them for satellite validations. A macropixel size of 3x3 has been set.

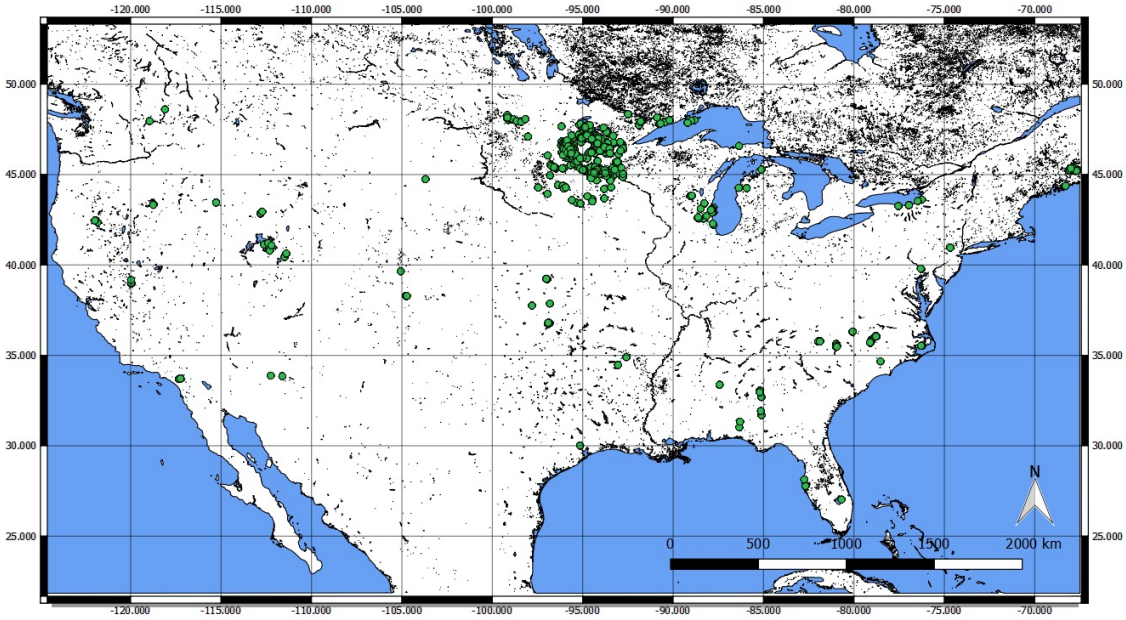


Figure 5: Match-up locations of in-situ measurements and Sentinel-3 OLCI imagery between 23/11/2016 and 01/04/2018. The map does not include the valid pixel expression applied, as a single flag can change the overall availability of possible match-up locations.

3.10 Validation sites

The validation lakes and all information products related were selected before the classification can be applied. In addition, they are relevant for the water constituent validation part of the framework. Lake sites were selected following several criteria:

1. **Lake selection**

   - The selected lakes are from different ecological environments to account for varying optical variability and eutrophication levels. The seasonality is an important circumstance to consider, as trophic conditions greatly change during the course of a year.
   - To compare the results, also prior studies are included in the analyses in the ecological state of the lake. They can not be compared to the STSI directly, but should provide an ecological impression about the lake with specific environmental conditions related to the aquatic environment, e.g. occurring algal blooms.

2. **In-situ data**

   - In-situ chl-a measurements have to be existent for the same date as an OLCI intake and should cover a large spatial extent. This is secured through the match-up process described in the previous section.

3. **L1 products**

   - OLCI scenes from the match-up process need to be manually reviewed and checked using the flags of Idepix and C2RCC. Single flags can have a major influence on the remaining match-ups, thus it is crucial to review this process.
   - Although IdePix is a pixel classification tool that includes the classification of all sorts of clouds, the scenes should also visually be free of clouds and other meteorological disturbances such as haze or Sun glint.

Considering these requirements for a lake to be used as a valid match-up location, three lakes (amongst many others) were chosen after the match-up processing and manual review: Lake Pelican in Minnesota, Lake Jordan in North Carolina and Lake Michigan in Wisconsin (see Figure 6).

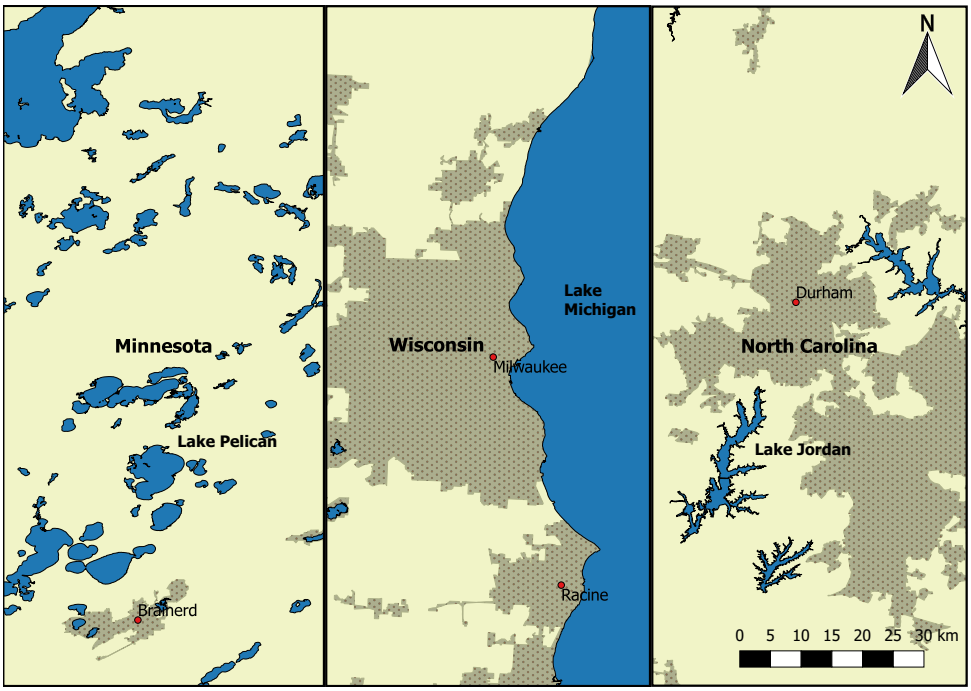


Figure 6: Overview of the validation sites and selected lakes. Left: Lake Pelican in Minnesota, middle: Lake Michigan and the Bay of Milwaukee in Wisconsin, right: Lake Jordan in North Carolina.

3.10.1 Lake Pelican

Lake Pelican is located in Crow Wing County in the U.S. State Minnesota. Most of the lakes in this area are important protected habitats that also partly serve as recreational

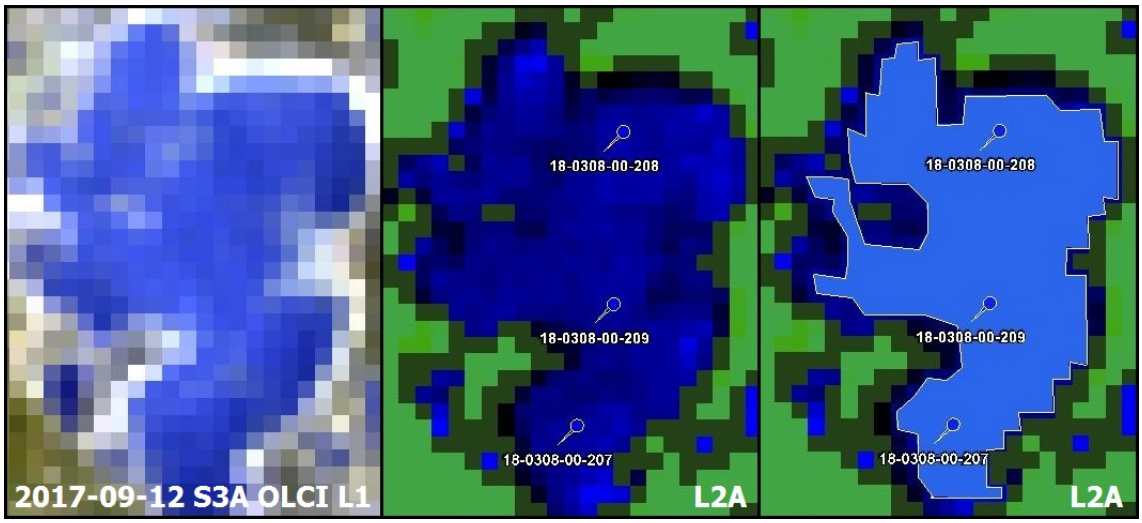


Figure 7: Lake Pelican OLCI imagery. Left image shows the L1 product as obtained from ESA, middle the image after processing it with Idepix and C2RCC to derive the required R_{rs}. The pins depict the locations of available in-situ measurements. The right image displays a polygon of the same product indicating the R_{rs} used in the STSI classification.

and aesthetic resources (Olmanson et al., 2008). In terms of size, Lake Pelican covers 33.86 km^2 placing it in the upper 5% of lakes in Minnesota. The maximum depth is reported to be over 100 feet at the center of the lake and at least 30 feet in proximity to shore (MPCA, 2018). Derived from historical data (2008 - 2017), the mean TSI (using chl-a as the indicator) is within the oligotrophic-mesotrophic range (MPCA, 2018). Three in-situ measurements were taken across the lake on the 12th of September 2017, under cloud free conditions enabling to use this lake to compare it with the STSI classification (see Figure 7). Not all areas of the lake are usable as the North-West area has an island (Gooseberry Island) that is frequently used in summer to beach boats. Due to its uneven coastline, patchiness and possible radiative distortion through other adjacency effects, Lake Pelican can be considered as a boundary case for any of the applied algorithms or multi-spectral ocean-colour imagery in general. Figure 8 displays the R_{rs} (287 considered valid water pixels) from the right image of Figure 7 that are subsequently used in the STSI classification. The gradient in reflectivity (left image) will most likely be caused by small amounts of scattering particles for which the normalization originally has been designed.

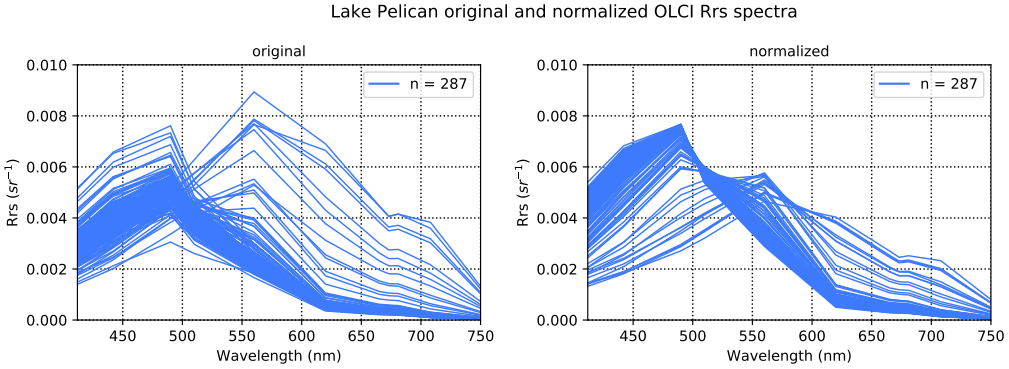


Figure 8: Lake Pelican original (left) and normalized (right) R_{rs} spectra used in the STSI classification.

3.10.2 Lake Jordan

Lake Jordan (officially B. Everett Jordan Lake) is an artificial reservoir in close proximity to the cities Cary and Durham in North Carolina. It spans a size of 56.41 km^2 and was developed as part of a flood control project in the late 1940's. Lake Jordan serves as a drinking-water supply for the towns of Cary, Apex and Morrisville and is listed as impaired (based on the CWA regulations) due to nutrient over-enrichment and occasionally experienced algal blooms. Nine different locations all over the lake were available to use for a match-up with OLCI, however after the manual review step inspecting the OLCI L1 radiances, the retrieved reflectances and the valid pixel selection, several stations were discarded due to poor or insufficient quality. Seven measurements remained to estimate the in-situ TSI from the lake. While Lake Pelican is a boundary case for current ocean-colour methods, Lake Jordan is difficult due to multiple layers interfering with each other making this lake an extremely challenging environment to test the STSI in (see Figure 9). The lake itself has eutrophic-hypereutrophic conditions (based on historical and current in-situ data used for this study). The C2RCC atmospheric correction has been reported to not provide realistic reflectances in cyanobacterial bloom situations, while at the same time it is uncertain for this lake if surface scums or subsurface blooms exist during the capture of OLCI (Toming et al., 2017). Unrealistic reflectances will cause the STSI classification to fail, as it is highly dependent on the training reflectances. Even if the reflectances with C2RCC are retrieved correctly, the classification might still fail due to insufficient training reflectances not including reflectances covering this water type. Further, adjacency effects constitute a layer of uncertainty on the retrieved reflectances, predominantly because of

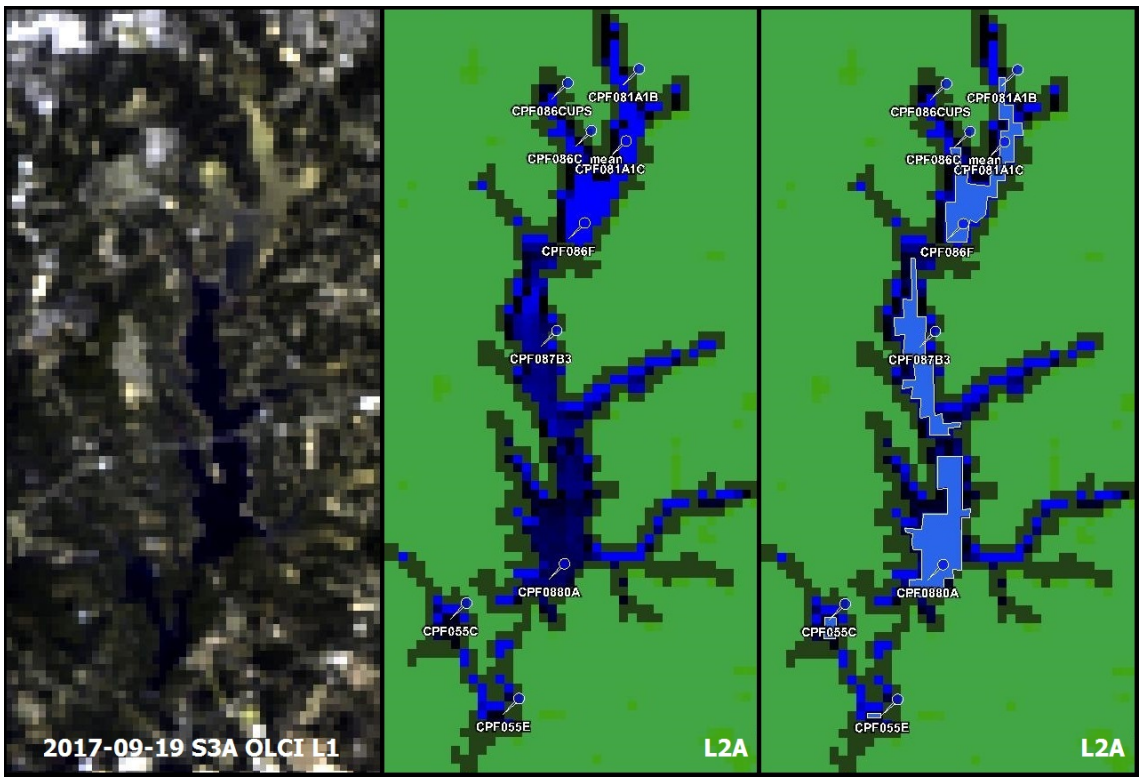


Figure 9: Lake Jordan OLCI imagery. Left image shows the L1 product as obtained from ESA, middle the image after processing it with Idepix and C2RCC to derive the required R_{rs}. The pins depict the locations of available in-situ measurements (not all were used). The right image displays a polygon of the same product that was used to derive the valid water pixel R_{rs} for the STSI classification.

the inconsistent shape and close proximity to recreational areas. The influence should be actively weakened through the imposed 3x3 grid to only select in-situ stations surrounded by at least one water pixel and the precise coastline flag of Idepix. Still, AE might influence the retrieved reflectances. Figure 10 displays the R_{rs} of the 180 considered valid water pixels that are used in the STSI classification. A strong gradient in scattering particles is imminent on the left image containing the original R_{rs} from the L2A OLCI product. In comparison to Lake Pelican this scattering gradient might not only be caused by suspended material (TSM), but also due to high concentrations of phytoplankton or submerged algal. Whether or not these reflectances are realistic can currently not be judged without reference measurements or a validation methodology.

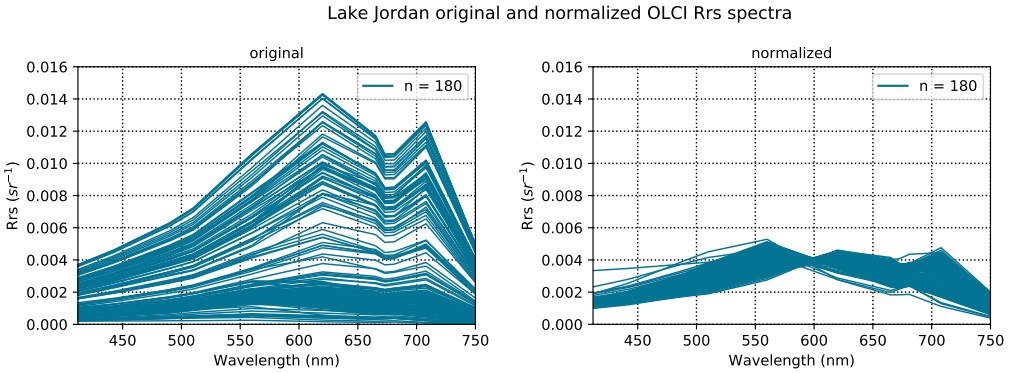


Figure 10: Lake Jordan original (left) and normalized (right) R_{rs} spectra used in the STSI classification.

3.10.3 Lake Michigan

Lake Michigan is one of the largest Great Lakes water basins. Besides other functionalities it serves as a drinking water reservoir for cities such as Chicago, Milwaukee, Waukegan and Green Bay. The study area is part of the southern basin whose limnological dynamics strongly differ from the central and northern basins of the lake (Yousef et al., 2014). Spring blooms can occur and in the absence of winter ice cover a counter-clockwise gyre can form contributing to these blooms (Chen et al., 2004). Horizontal and vertical transfers last until mid-May or early June when a thermal bar forms that limits spatial transports (Kerfoot et al., 2008, 2010).
On the 18th of July 2017 ten chl-a measurements were taken close to the harbour and bay of the city Milwaukee (Figure 11). Similar to Lake Jordan, not all locations of the in-situ measurements could be utilised. They were either too close to shore showing clear land influences or the measurements were taken in depths too large to compare them with OLCI reflectances. For the validation, eight measurements were used.
The trophic status of the bay of Milwaukee rapidly changes with distance to shore. Currents around the harbour area exist that capture phosphorous-rich discharges and re-suspended near-shore sediments, transporting the nutrient-rich waters into offshore regions. Therefore gradients in chl-a concentrations as well as spatial differences are expected in this region of interest. The present dataset shows three trophic states for this region, most of it being oligotrophic (offshore) and meso- or eutrophic closer to shore.
The OLCI scene is cloud free and 1270 R_{rs} were used in the classification. Opposite to Lake Jordan it is expected that the calculated C2RCC R_{rs} are realistic. However, also here AE can play a role as the closeness to the harbour and coastline can influence the retrieval.

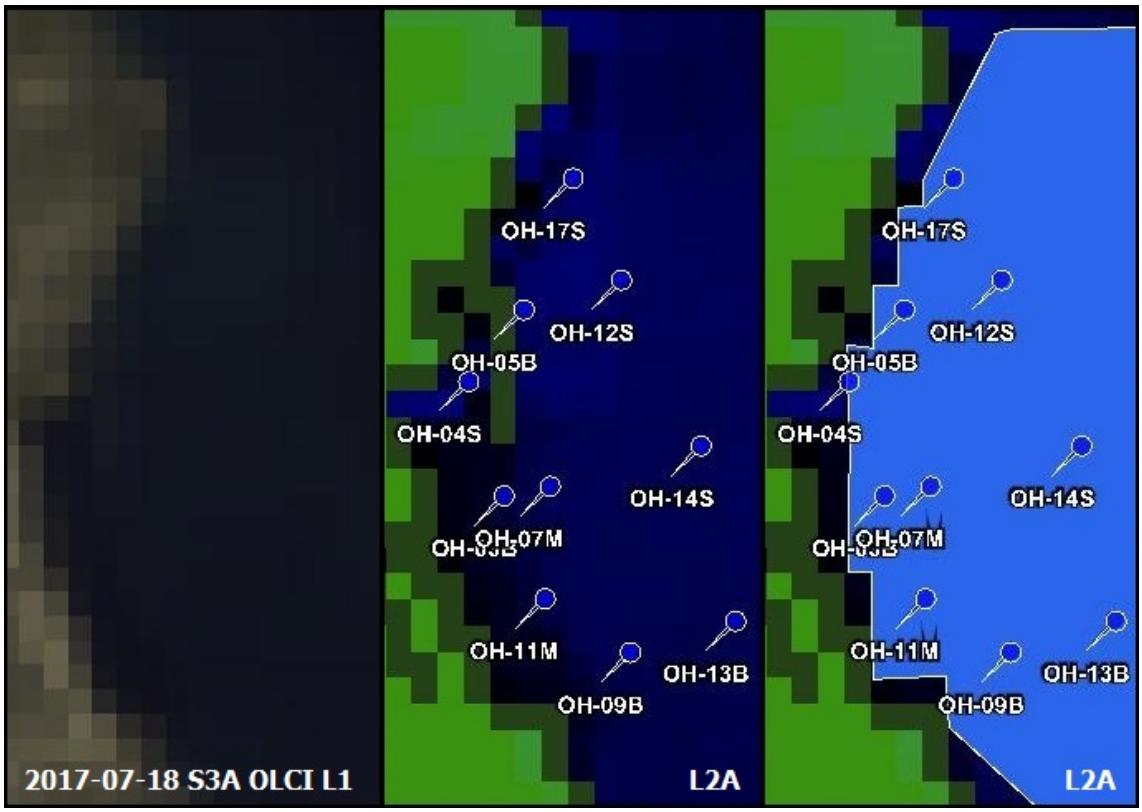


Figure 11: Lake Michigan, bay of Milwaukee OLCI imagery. Left image shows the L1 product as obtained from ESA, middle the image after processing it with Idepix and C2RCC to derive the required R_{rs}. The pins depict the locations of available in-situ measurements (not all were used). The right image displays a polygon of the same product that was used to derive the valid water pixel R_{rs} for the STSI classification.

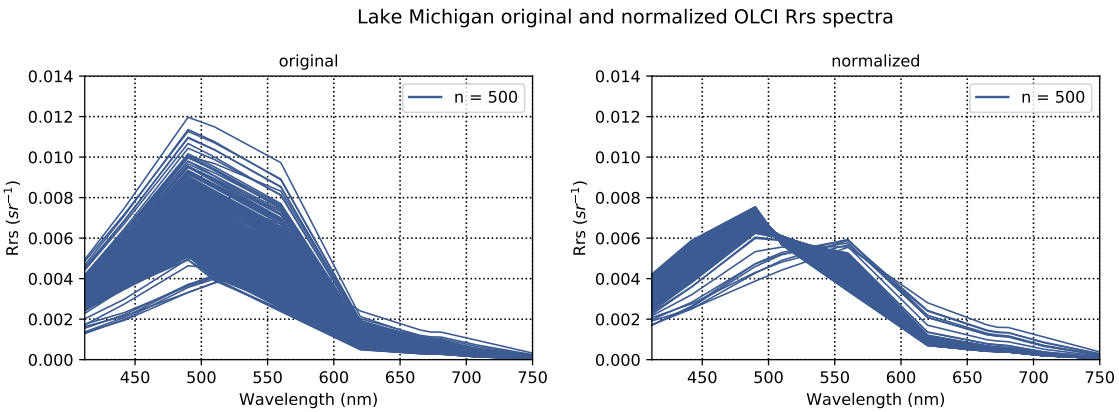


Figure 12: Lake Michigan original (left) and normalized (right) R_{rs} spectra used in the STSI classification.

As for the two other lakes, the flagging of coastline and non-water pixels has been reviewed manually. Figure 12 shows a large subset (500 reflectances) of the valid water pixels covering most of the optical gradient found in this area.

4 Gradient boosting decision tree classification

Over the last two decades, most optical classification studies were concerned with pattern finding in reflectance data to foster the development and selection of bio-optical retrieval algorithms (Eleveld et al., 2017; Gonçalves-Araujo et al., 2018; Jackson et al., 2017; Lubac and Loisel, 2007; Moore et al., 2014; Wang et al., 2010). Originating from the aim to find new patterns in the reflectances, the majority of studies made use of unsupervised learning techniques. Mostly hierarchical and fuzzy clustering clustering techniques were used to find the patterns in the studied datasets (Shi et al., 2013; Moore et al., 2001; Vilas et al., 2011).
Within the STSI framework, the aim is to predict for every valid pixel (and the derived reflectance) a trophic state class. Opposed to the mentioned optical classification studies, unsupervised techniques can not be utilised. Instead, the selected model has to be based on the process of finding patterns that generalize well to unseen (or unobserved) measurements. If an algorithm is able to find patterns that generalize well, accurate predictions are possible. This is the goal in supervised learning, which is about the relationship between a response variable y_i and a set of predictor variables x_i. In statistical learning, this process is called *predictive modelling*. Applied to the STSI goal, the predicted classes are the outcome of a supervised classification algorithm trained with the simulated R_{rs} derived in section 3.2. After the model selection process the resulting classification model (classifier)

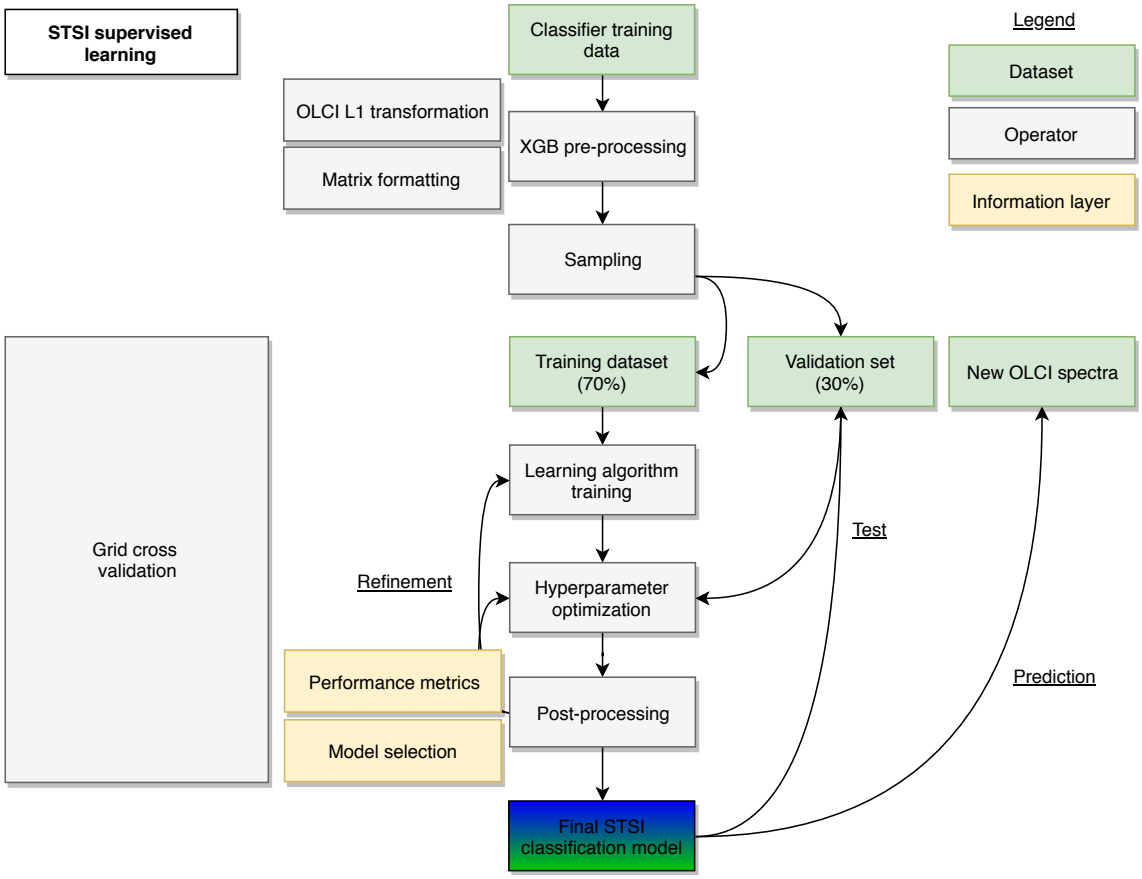


Figure 13: STSI classification construction using the XGBoost library. The derived training data (section 3.1) are brought into the correct XGBoost matrix format and are then split into a train set (70%) and validation set (30%). To evaluate which hyperparameter combination is most suited, grid cross-validation is applied to the training data. After an initial model construction, the predictive capabilities are tested on the validation set allowing for refinement of the model. The model with the lowest test error is then exported and used to predict the TSI classes on OLCI R_{rs}.

is then applied to unseen OLCI R_{rs} derived from the chain of algorithms explained in section 3.6 to finally predict for every valid reflectance a TSI class. Figure 13 displays a schematic of the steps exercised to derive the STSI classification model. The classification algorithm utilised is based on gradient boosting, introduced by Friedman (2001). Gradient boosting has been specifically successful when applied to tree models, in which it fits additive tree models. For this combination specifically, the methods were given names such as Gradient Boosting Machine (GBM) or Gradient Boosted Decision Trees (GBRT) (Friedman, 2002). Moreover, gradient tree boosting has proven itself to be highly suited for all kinds of regression and classification problems. Extreme Gradient Boosting (XGBoost), proposed by Chen and Guestrin (2016), is a recent variation of Friedman's GBM, now with contributions from many developers. It belongs to a broader collection of tools under the umbrella of the Distributed Machine Learning Community (DMLC). Following, XGBoost is an open access software library implementing gradient boosting machines. XGBoost is one of the most popular data science methods used for predictive modelling, but it has not been used in any ocean-colour related optical classification study. The classification model of the STSI is constructed using the XGBoost library. Nevertheless, this thesis is not about advancing the XGBoost library further. The library basically combines the benefits of three machine learning methods:

1. Trees as base learners

2. Boosting to improve predictive capabilities (applied to trees, also named Tree boosting)

3. Numerical optimization via gradient descent

The focus for this thesis is on those parts of the XGBoost framework that provide significant insight into the algorithm implementation used to construct the STSI model. Therefore, relevant parts include the tree-method utilised to classify the spectra into the classes and boosting, a technique to improve the prediction accuracy of the tree. Details about gradient descent are omitted, as essentially the whole purpose of gradient descent algorithms is to minimize the loss function used, hence to numerically optimize the classification procedure to make it computationally feasible even on large datasets. The only requirement to use gradient descent as the algorithm to minimize loss is that the loss function itself has to be derivable. To reason about the STSI classifier, it is required to review the relevant basics of supervised learning, as they are the building blocks the STSI machine learning model is based on.

4.1 Supervised learning

The general model in supervised learning refers to the mathematical structure of how to make a prediction of the response variable y_i using a set of covariates $X = (x_1, ..., x_p)$. At hand are the simulated data

$$D = \{(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)\}, \tag{13}$$

consisting of 125.000 simulated R_{rs}. The response y_i is also referred to as dependent variable. In the STSI case, y_i can only take a finite number of classes C, i.e. the TSI classes previously assigned (4), thus the STSI is a classification task. The covariates $X = (x_1, ..., x_p)$ are also referred to as the predictors, the attributes, the features, explanatory variables, the independent variables or the input variables.
Several approaches exist that can be used to construct a predictive model. The approach followed herein is applying the framework of statistical learning theory, providing the theoretical basis for modern machine learning algorithms (von Luxburg and Schölkopf, 2011).

The STSI model built is consisting of predictive functions designed to make accurate predictions on the new, unseen OLCI R_{rs}. To make predictions based on the input variables x_i, a function should be learned (also called hypothesis) such as

$$f : x \rightarrow y, \tag{14}$$

that maps every input x_i to a corresponding prediction y_i. Hence, for a given x, the prediction

$$\hat{y} = f(x), \tag{15}$$

can be made. This model can also be referred to as prediction function, decision function or a decision rule. Let y^x denote the set of all functions mapping from x to y. The problem of estimating a model $\hat{f}$ thus can be viewed as a problem of selecting a function $\hat{f}$ from the set y^x, based on the available data. Consequently, predictive modelling can also be viewed as a problem of function estimation (Vapnik, 1999).

4.1.1 Loss function

A small amount of theory is necessary to provide a framework for developing models such as those discussed informally so far. Generally, a function $f(x)$ is searched for to predict y_i given values of the given input. In a machine learning model this requires a loss function to penalize errors in prediction. The generic loss function

$$L : Y \rightarrow \mathbb{R}_+, \tag{16}$$

gives a quantitative measure of the loss resulting from a prediction when the true result ends up being y. The prediction accuracy of the function defined is measured using a loss function. In other words, the training loss measures how predictive the model is (James et al., 2014).
For the STSI classification, the multi-class logarithmic loss (log loss) is implemented. Log loss quantifies the accuracy of the classifier by penalising false classifications. Minimising the log loss is equivalent to maximising the accuracy of the classifier.
The reasoning to choose log loss as the STSI loss function originates from the aim of the overall STSI framework. Instead of receiving a clear TSI classification label for a reflectance directly, the interest is to receive the probabilities of a spectrum belonging to each class (that sum up to 1). Log loss includes the idea of probabilistic prediction confidence when predicting a label. It is the cross entropy between the distribution of the true labels and the predictions made. In other words, cross entropy incorporates the entropy (the measure of the randomness / unpredictability in the processed sensor R_{rs}), plus the extra uncertainty originating from having two distributions (cross), i.e. the simulated reflectances being a different distribution than the true distribution (the actual sensor reflectances). Simplified, log loss is a measure to gauge (or rate) the noise that accompanies the process of using a predictor as opposed to the true labels (Bishop, 2006; Nielsen, 2016).
Class probabilities allow for interference into the decision making to control the process. Application examples and benefits are discussed in Chapter 7. In order to calculate log loss, the classifier has to assign a probability that a spectrum belongs to each class rather than simply yielding the most likely class. Log loss is mathematically defined as

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \log p_{ij}, \tag{17}$$

where N is the number of samples (125.000), M is the number of possible labels (4), y_{ij} is a binary indicator of whether or not label j is the correct classification for spectrum i, and p_{ij} is the model probability of assigning the class label j to spectrum i (Bishop, 2006). The perfect classifier would have a log loss of zero. Minimising this function is the

objective, thus it is also referred to as objective function. In the XGBoost framework used to build the STSI model, parameters θ and a regularization term $\Omega(\theta)$ are also added to this objective function, discussed in sections 4.1.4 and 4.3.

4.1.2 Parameters

To exemplify how the learning algorithm selects its prediction function with required parameters, the theory of risk minimization is shortly introduced. Assuming that a joint probability distribution $\mathbb{P}_{Y,X}$ exists over X and Y, then the training set of the algorithm consists of m instances $(x_1, y_1), ..., (x_m, y_m)$ drawn from $\mathbb{P}_{Y,X}$. As defined, any loss function measures the accuracy of a prediction after the outcome is observed. However, at the time the prediction is made, the true outcome is still unknown and the loss assigned can consequently be viewed as a random variable $L(Y, \hat{y})$ (Murphy, 2012). The risk associated with the hypothesis function $f(x)$ can be defined as the expectation of the loss function:

$$R(\hat{y}) = E[L(Y, \hat{y})]. \tag{18}$$

The risk is important, because the goal of a learning algorithm is to find the optimal hypothesis (f) among a set of functions for which the risk $R(\hat{y})$ is minimal. The risk $R(\hat{y})$ can not be directly computed, because the distribution $\mathbb{P}_{Y,X}$, i.e. the related instances of X and Y, are unknown to the learning algorithm (Tewari and Bartlett, 2014). Therefore an approximation is computed, called empirical risk, by averaging the loss function on the training set:

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)). \tag{19}$$

The empirical risk minimization principle (ERM) by Vapnik (1999) states that the learning algorithm has to choose a hypothesis $\hat{f}$ minimizing the empirical risk:

$$\hat{f} = \arg\min \hat{R}(f). \tag{20}$$

The model defined by ERM is the empirical risk minimizer $\hat{f}$. It is an empirical approximation of a defined target function that seeks to reduce the risk. ERM is a criterion that is optimised to select the function $\hat{f}$ from a set of functions F, whereby the selection is of major importance. A naive approach would be to allow any function in Equation 14 to be a solution. However, this naive approach would make the function estimation an ill-posed problem, as it would be an attempt to estimate an infinite number of parameters using only a finite data set (Evgeniou et al., 2000; Rakotomamonjy and Canu, 2005). In practice, this problem is solved restricting F to be a subset of the total function space, also known as the hypothesis space. Restricting the function space essentially defines a class of models, thus F is referred to as a model class. One of the most popular model classes is the class of linear models, where the prediction is given by a linear combination of weighted input features (Nielsen, 2016):

$$F = \{f : f(\hat{y}) = \sum_{j} \theta_j x_j\}. \tag{21}$$

The prediction value $\hat{y}$ can have different interpretations. A popular example is the logistic transformation, but many more exist (Evgeniou et al., 2000). In the STSI classifier, the prediction value is interpreted as the softmax value (described in section 4.4).
Most important to recognise is that this model simplifies the function estimation to a problem of estimating a parameter θ or a full vector of several parameters $\theta = (\theta_0, \theta_1, ..., \theta_p)$. These parameters are the undetermined part required to learn from the data.

4.1.3 Learning algorithm

Combining the model class with the ERM principle reduces the learning problem of a supervised classifier to an optimization problem to estimate the optimal parameters θ. The optimization problem has two aspects. First, the model class is a set of functions that all are considered to be candidate solutions, while the ERM is the criterion to select a function from this set of functions. This approach defines the statistical aspect of the problem. Second, computationally the problem is to solve the optimization problem defined by the ERM. This then is the job of the learning algorithm, which essentially is an optimization algorithm itself. In the XGBoost implementation, this is the part of the gradient descent algorithm implemented to numerically optimize the loss incurred, hence to solve the ERM optimization problem.
More generally, any learning algorithm takes the data set D as input and outputs the fitted/trained model $\hat{f}$. As introduced, the function estimation of the model classes have some parameters θ that the learning algorithm then iteratively adjusts to fit the data. Estimating the parameters θ is sufficient to estimate the model

$$\hat{f}(x) = f(x, \hat{\theta}). \tag{22}$$

Different model classes and loss functions chosen lead to different optimization problems. Parts of the model class used in XGBoost belong to the class of continuous optimization problems, i.e. the objective function is continuous with respect to the parameter θ. Many methods exist for continuous numerical optimization problems, further examples can be found in Nocedal and Wright (2006). For the XGBoost classifier the method is gradient descent. As mentioned initially, the details of the gradient decent algorithm are omitted, but the interesting reader is referred to Friedman (2001, 2002); Hastie et al. (2009).

4.1.4 Regularization

There is more to machine learning than optimizing an objective (loss) function. Training the STSI model from the simulated R_{rs} has the purpose of trying to prepare it for the unseen reflectances from an OLCI scene. The preparation of the model does not take into account the added complexity of the unseen OLCI spectra. Thus, a general training loop that tries to minimize the objective (loss) function defined in Equation 17 may overfit the OLCI spectra. A model might be capable of capturing the structure of the simulated R_{rs}, however due to the already present complexity of capturing these structures, the model probably has no flexibility left to incorporate new structures inherent to the unseen OLCI spectra. This process of essentially fitting the model too well on the training data is called overfitting. A way to combat overfitting is through a regularization function $\Omega(\theta)$. Regularization of tree models is achieved by constraining or penalizing the complexity of the tree (see section 4.3).

4.1.5 Model complexity

The goal of a learning method is to be able to generalize, thus perform well on unseen observations from $\mathbb{P}_{Y,X}$. Therefore the objective is a model $\hat{f}$ with as low true risk $R(\hat{f})$ as possible. For a well-performing model it is necessary to use a flexible model class capable of fitting all the relevant structures existent in the spectra. However, when the model class is too flexible, it ends up precisely fitting the structure of the training data, in this case the simulated R_{rs}. On the other hand, if the model class is not flexible enough, it will not be able to fit the relevant structure, thus called underfitting. The tradeoff in selecting an appropriate model complexity is referred to as bias-variance tradeoff (for more details see Tewari and Bartlett (2014)). For the STSI, the approach is used splitting the simulated and OLCI R_{rs} into the training and test set. The STSI model is fit by minimizing the empirical risk on the training set, while the generalization error is measured by calculating

the empirical risk on a validation set. For this, 10 k-fold grid cross-validation is used, combining all hyperparameters (see section 4.3) with each other to ultimately select the best parameter settings. More details about model complexity and cross-validation procedures can be found in Allen 1974; Hastie et al. 2009; James et al. 2014. The hyperparameter settings can subsequently be selected providing the model on the training data that is expected to perform best on the unseen test data. After testing the initial model on the validation set, iterations of the model fitting procedure are manually possible, i.e. to include empirical knowledge gained by reviewing the initial test results and incorporating it into the XGBoost model. This procedure is a semi-supervised form of finding the best XGBoost model that is then exported and implemented into the SNAP operator described in Chapter 5.

4.2 Tree-based methods

From several model classes included in XGBoost, trees in combination with boosting are used to create the model of the STSI. First, the tree-based method used in XGBoost is explained, which then expanded to tree ensembles is combined with the boosting technique explained in section 4.2.2. Trees partition the feature space X into a set of rectangles T and then fit a simple model (called constant c herein) in each one. The node at the top of the tree can be called root node, this node then has branches below it. Nodes that have branches below them are internal nodes or splits. The lowest nodes at the bottom of a tree are called terminal nodes or leaves. Tree models usually exhibit limited predictive utility. However, there are certain improvements available, like the combination of trees in bagged trees (Breiman, 1996), Random Forests (Breiman, 2001) or making use of boosting algorithms. The combination of standard tree model classes with these algorithms results in high predictive capabilities. To illustrate tree methods, a simple regression problem with continuous response Y and inputs X_1 and X_2 is assumed. The top left panel of Figure 14 shows a partition of the feature space by lines that are parallel to the coordinate axes. In each partition element the response Y can be modelled with a different constant c. This would result in a simple description like $X_1 = c$, but for some of the resulting regions this is complicated to describe. Therefore recursive binary partitions are used. The top right panel of Figure 14 shows these partitions, as also used in classification trees. First, the space is split into two regions and the response is modelled by the mean of Y in each region. The variable and split-point are chosen for the best fit. Then one or both of these regions are split again into two more regions. This process is repeated until a defined stopping criterion is reached, e.g. a previously defined maximum number of terminal nodes. Restricting the view again on the top right panel of Figure 14, a first split is done at $X_1 = t_1$. Then the region $X_1 \leq t_1$ is split at $X_2 = t_2$ and the region $X_1 > t_1$ is split at $X_1 = t_3$. Lastly, the region $X_1 > t_3$ is split at $X_2 = t_4$. The results of this binary recursive splitting procedure is a partition into the five regions $R_1, R_2, ..., R_5$. The corresponding model predicts Y with a constant c_m in region R_m, that is,

$$\hat{f}(X) = \sum_{m=1}^{5} c_m I\{(X_1, X_2) \in R_m\}. \tag{23}$$

The same model can be represented by the binary tree in the bottom left panel of Figure 14. In this case, the measurements that would satisfy the condition at each junction are then assigned to the left branch, others to the right branch. The terminal nodes or leaves of the tree correspond to the regions $R_1, R_2, ..., R_T$. The bottom right panel of Figure 14 is a perspective plot of the regression surface from the model used to describe tree methods in its simplest form. With more than two inputs (like in the real STSI case), partitions like that in the top right panel of Figure 14 are difficult to draw, but the binary tree representations just described work in the same way.

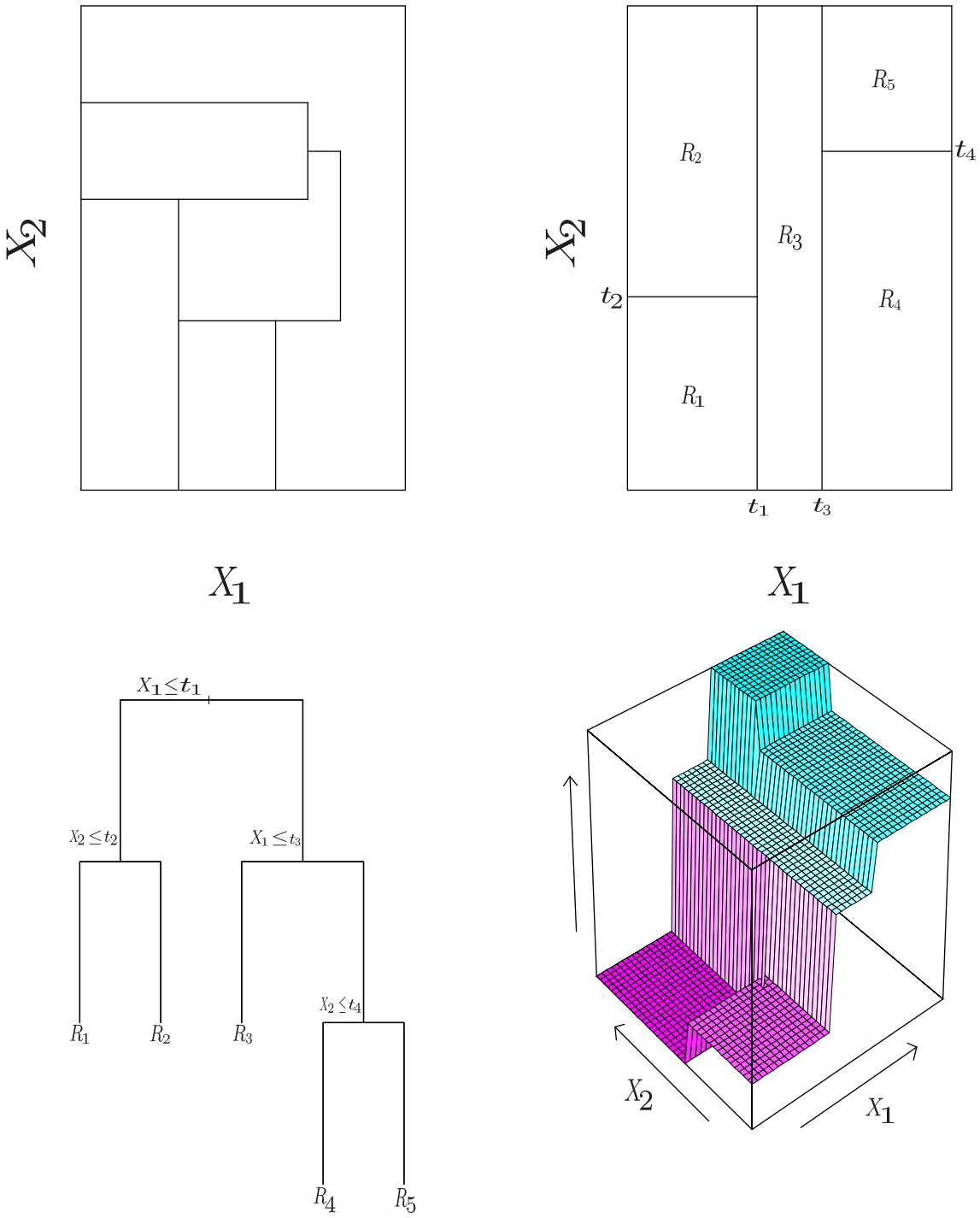


Figure 14: Top left panel shows a partition that can not be obtained from recursive binary splitting as opposite to the top right panel showing a partition of a two-dimensional feature space by recursive binary splitting as used in CART (explanation in the next section) and XGBoost. Bottom left shows the tree corresponding to the partition in the top right panel and a perspective plot of the prediction surface appears in the bottom right panel (James et al., 2014).

4.2.1 STSI model structure

For explanatory purposes, the learning part of the classification tree is considered without a regularization function. In the actual STSI model several regularization parameters $\Omega(\theta)$ are used, as described in section 4.3. The STSI spectra used to grow the tree consist of $p_1, p_2, ..., p_n$ inputs and $y_i = y_1, y_2, ..., y_4$ qualitative responses (one of k values for the TSI classes), for each of the N spectra (125.000): that is, (x_i, y_i) for $i = 1, 2, ..., N$

with $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$. The algorithm needs to automatically decide on the splitting variables and split points and also what structure the tree should have. Suppose the first partition is into M regions $R_1, R_2, ..., R_T$ and the modelled response has a constant c_m in each region, the tree model f can be written as:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_T). \tag{24}$$

One can see that the best $\hat{c}_m$ is just the average of y_i in region R_m

$$\hat{c}_m = \text{avg}(y_i | x_i \in R_m), \tag{25}$$

implying that the class predicted for each OLCI spectrum originates from the *most commonly occurring* class of training spectra in the region to which an OLCI spectrum has been assigned to. Mathematically, in a node m, representing region R_m with N_m spectra, let

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k), \tag{26}$$

be the proportion of class k spectra in node m, where I is an indicator function, indicating spectrum y_i is a member of class k. In node m, the spectra are then classified to class $k(m)$ based on $\arg\max_k \hat{p}_{mk}$, the majority class k in node m (James et al., 2014). Finding the best partition in terms of minimizing the objective function is typically computationally infeasible, i.e. learning the structure of the tree is hard. In fact, the problem is NP-complete (Hyafil and Rivest, 1976). Consequently the problem is simplified by instead computing an approximate solution. For this, XGBoost makes use of a greedy learning algorithm, so-called CART (Classification and Regression Trees) (Breiman et al., 1984). CART grows a tree greedily in a top-down fashion using binary splits, starting with the only root node (see also Figure 14). Then every split parallel to the coordinate axes is considered and the split minimizing the defined objective function is chosen. Next, a certain split parallel to the coordinate axes within each of the current regions is considered (Hastie et al., 2009). The mathematical details of greedily selecting the splits which minimize the empirical risk or training error are skipped here, as they do not provide additional insight necessary to explain the STSI model.

4.2.2 Tree boosting

Boosting is the approach used in XGBoost to improve the predictions resulting from a decision tree. It is not only restricted to tree methods, but useful in combination. The described boosting procedure is not identical to the actual boosting algorithm implemented in XGBoost, but similar in its approach and hence used to envision the idea and its benefits to be used in combination with trees. For a mathematical derivation of the actual boosting procedure (so-called Newton Boosting) the interested reader is referred to Nielsen (2016). The idea of boosting is that unlike fitting a single large decision tree to the data, which is equal to fitting the data hard and thus potentially overfitting, a boosting approach instead learns slowly. Boosting grows the tree sequentially: each new tree is grown using information from previously grown trees, hence it combines many simple models (Schapire and Freund, 2012). Boosting can be well-described using the most popular boosting algorithm called "AdaBoost"(Freund and Schapire, 1996, 1997). The algorithm is explained herein to clarify the idea behind boosting in a simplified manner.
Considering a multi-class problem like in the STSI framework, with the output variable coded as $Y \in \{1, 4\}$. Given a vector of predictor variables x_i, a classifier $G(X)$ produces a

prediction taking one of the values of y_i. The error rate on the training sample is

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} I(y_1 \neq G(x_i)), \tag{27}$$

and the expected error rate on the test set is $E_{XY}I(Y \neq G(X))$. A weak classification model (or classifier $G(X)$) is one whose error rate is only slightly better than random guessing. The purpose of the boosting technique is to sequentially apply the weak classification algorithm to repeatedly modified versions of the data, thereby producing a sequence of weak classifiers $G_m(x), m = 1, 2, ..., M$. The predictions of all of these weak classifiers are then combined through a weighted majority vote to produce a final prediction:

$$G(x) = \text{sign}\left[\sum\nolimits_{m=1}^{M} \alpha_m G_m(x)\right]. \tag{28}$$

Here $\alpha_1, \alpha_2, ..., \alpha_m$ are the weight coefficients computed by the boosting algorithm that essentially weight the contribution of each respective $G_m(x)$ to the final prediction (see Figure 15). Effectively, they try to give higher influence to the more accurate classifiers in the sequence. Figure 15 shows the schematic of the AdaBoost procedure. Linking the AdaBoost algorithm to the STSI, the application of the weak classifiers $G_m(x)$ to modified versions of the spectral dataset at each boosting step is possible due to the application of weight $w_1, w_2, ..., w_N$ to each of the training spectra $(x_i, y_i), i = 1, 2, ..., N$. The first step simply trains the classifier on the spectra in the usual manner, i.e. the weights are set to $w_i = 1/N$. Then, for each successive iteration $m = 2, 3, ..., M$ the weights of the spectra are individually modified and the classification algorithm is re-applied to the weighted spectra. At step m, those spectra misclassified by the classifier $G_m - 1(x)$ induced at the previous step have their weights increased, whereas the weights are decreased for those that were correctly classified. Proceeding with increasing iterations, spectra that are difficult to correctly classify receive ever-increasing influence. Thereby each classifier is forced to

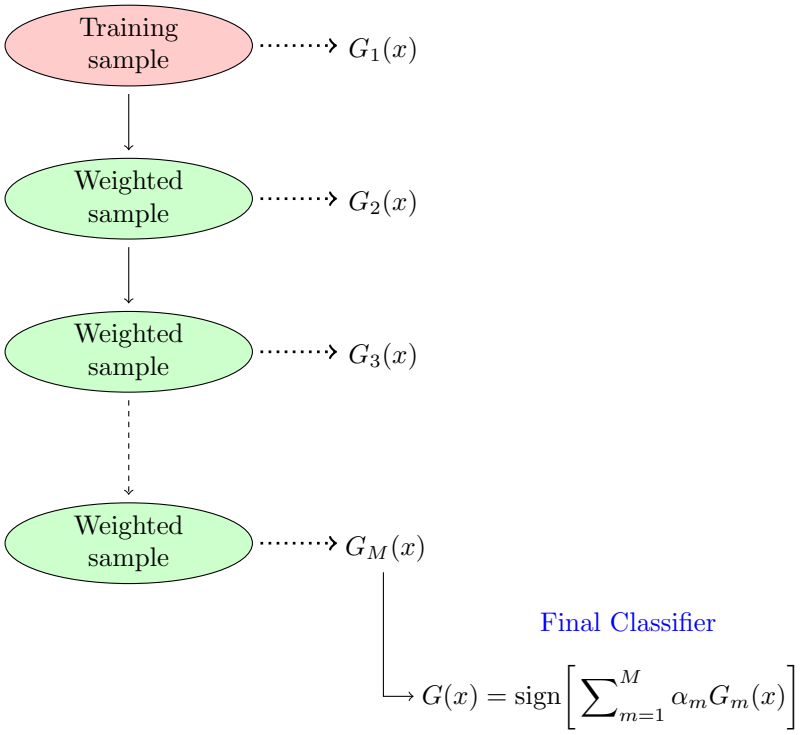


Figure 15: Schematic of AdaBoost. Classifiers are trained on weighted versions of the dataset and then combined to produce a final prediction. Adapted from James et al. (2014).

Algorithm 1 AdaBoost

1. Initialize the spectrum weights $w_i = 1/N, i = 1, 2, ..., N$
2. For $m = 1$ to M:
   (*a*) Fit a classifier $G_m(x)$ to the training data using weights w_i,
   (*b*) Compute
   $$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i},$$
   (*c*) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$,
   (*d*) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))], i = 1, 2, ..., N$.
3. Output $G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$

concentrate on those training spectra that are missed by previous ones in the sequence (Hastie et al., 2009; James et al., 2014).
Algorithm 1 provides the details of the AdaBoost procedure. A description of the algorithm is as follows: at the beginning, the current classifier $G_m(x)$ is induced on the weighted spectra at line 2a. The resulting weighted error rate is computed at line 2b. Line 2c calculates the weight α_m given to $G_m(x)$ in producing the final classifier $G(x)$ (line 3). Lastly, the individual weights of each of the spectra are updated for the next iteration at line 2d. Spectra misclassified by $G_m(x)$ have their weights scaled by a factor $\exp(\alpha_m)$, increasing their relative influence for inducing the next classifier $G_{m+1}(x)$ in the sequence. The key of boosting lies in Equation 28. Defined in other words, boosting is a way of fitting an additive expansion in a set of elementary "basis" functions. Here the basis functions are the individual classifiers $G_m(x) \in \{1, 4\}$. More generally, basis function expansions take the form

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \theta_m) \tag{29}$$

where $\beta_m, m = 1, 2, ..., M$ are the expansion coefficients and $b(x; \theta) \in \mathbb{R}$ are the simple functions (or constants in trees) of the set of parameter functions x (instead of a variable, also called multivariate argument), characterized by a set of parameters θ. For trees θ can parametrize the nodes, e.g. the split decisions and split points at the internal nodes, and the predictions made at the terminal nodes. These models are fit as usual by minimizing the loss function averaged over the training data. For many loss functions this requires computationally intensive numerical optimization techniques. XGBoost makes use of the gradient descent algorithm. As mentioned already, details of the gradient descent algorithm are not further elaborated in this study.

4.3 STSI model hyperparameters

The STSI model is a highly non-linear version including the objective (loss) function to minimise a set of several additional parameters to either construct the model and/or regularize the model complexity to prevent overfitting on the unseen OLCI R_{rs}, that is,

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{30}$$

where the loss function L is defined as the log loss explained in section 4.1.1 and the regularization Ω can be embedded using several methods to specifically constrain the model complexity of trees, e.g. the depth of the tree, the number of terminal nodes in a tree, the

size of local neighbourhoods or the relative difference in the constants $c_{m1}, ..., c_{mN}$. All of them allow for more theoretical considerations, but the focus is on the applied methods of penalizing the STSI model complexity. More parameters θ are actually used for the STSI, so-called hyperparameters. Hyperparameters are a set of parameters that are used in the training process to manually tune the model before the training process starts, e.g. they influence the rate at which the tree is learnt or the number of estimators used to build the tree. Those parameters essentially set the boundary conditions of the gradient boosting machine to learn from the data and to construct an initial model. It can be evaluated using the log loss and other metrics such as plotting the feature importance or the tree structure itself. As there is an arbitrary high number of hyperparameters possible (depending on the amount of parameters used in combination with each other) grid cross-validation is a useful method to find the best hyperparameter combination on the training data. This method allows to define a list of settings for each hyperparameter and then combines all hyperparameter settings with each other. Every combination is evaluated on a hold-out set of the training data and the final output is the best combination of these iterations. This output can then be manually tuned testing the grid cross-validation model on the initially hold-out validation set (30% of the same data) for which it has to perform well. Following is the final list of the additional hyperparameters θ that were used to build the STSI model for the classification of all three lakes. The model is the same for all validation lakes and includes no regional tuning:

- Column sample by tree = 0.3
- Gamma rate = 0.0
- Learning rate = 0.13
- Maximum depth of the tree = 2
- Minimum child weight = 2
- Number of estimators = 3000
- Amount of subsampling = 0.05

4.4 Softmax prediction metric

The XGBoost framework enables to use different evaluation metrics for the predicted output values of the applied STSI model that an OLCI spectrum i belongs to class k. The prediction values for the STSI are interpreted using the softmax function $\sigma(y)$, enabling to interpret the class assignments as probabilities (de Brébisson and Vincent, 2015; Duan et al., 2003). Before explaining the benefits of this interpretation, consider that the STSI model is not dealing with regression, meaning that the interest is not to have one output value that takes either 0 or 1. The classification task is about the output layer of the STSI model, consisting of 4 classes. Each spectrum will belong to a class and an output value representing the probability of the spectrum belonging to a class is of interest. This is primarily motivated by the circumstance that the finally assigned class label is then the one of the highest weight of the probability vector for each spectrum. An example of the probability vector can be seen in Table 6, of spectra belonging to a class for Lake Jordan. In mathematical terms, the STSI classification model with C (4) classes generates $y \in \mathbb{R}^4$, a vector of C scores. These scores are arbitrary real numbers in the range from 0 to 1. To go from these scores $y \in \mathbb{R}^4$ to probability estimates $p \in \mathbb{R}^4$ for a spectrum, exponentiation is used. The softmax function $\sigma(y)$ takes the model output as a vector $y \in \mathbb{R}^4$ with 4 values, exponentiates each value and then normalizes them by dividing through the sum of all exponentiated values:

$$\sigma(y)_i = \frac{\exp y_i}{\sum_{c=1}^{C} \exp y_c}, \tag{31}$$

where $i, c \in \{1, ..., C\}$ is the range over the classes, and y_i, y_c refer to class probabilities and scores for a single spectrum. As described, the vector of C scores add up to 1 for each spectrum. The final target class is the one that has the highest probability, thus the highest weight of the classification model and its prediction capability. Opposed to an outcome consisting of only a number between 1-4 for each prediction, probabilities provide informational insight into the decision making. This has two benefits: first, when provided as an additional information band in a L2 STSI product, the probability vector given for each pixel allows for a user-based judgement, whether the assigned probability that a pixel/spectrum belongs to a class is high enough to use this pixel for a valid TSI statement. This provides transparency about the final class assignments and enables user interpretation of the assigned class. Second, the probabilities can be used to constrain the class assignments using an independent decision tree for further analyses, i.e. if the classifier is indecisive, a spectrum can be assigned to the two classes having the highest probabilities.
Nevertheless, the softmax interpretation of the output does not allow to interpret the probabilities as uncertainties, as they sum up to 1 for each spectrum and thus are not an appropriate measurement of the uncertainty propagating through the model included in the assigned probabilities that a spectrum belongs to a class. In other words, the probabilities do not showcase how certain the classifier is in assigning the relative class probability. Rather they are simply the probabilities of the trained model (with its assumptions made and included uncertainty in the overall datasets used) that each pixel/spectrum belongs to a class. The probabilities allow for an assessment of the model itself and for further model refinement and building considerations.

Table 6: Probability vector of spectral class assignments. The columns Pc1,..., Pc4 contain the probability of a spectrum belonging to class k.

Spectrum	**Pc1**	**Pc2**	**Pc3**	**Pc4**	**Final class assignment**
1	0.244617	0.175018	0.263105	0.317261	4
2	0.254129	0.181823	0.273336	0.290712	4
3	0.267271	0.191227	0.287472	0.254031	3

5 STSI operator

The structure of the software implementation of the STSI framework is split into two parts. First, the STSI model resulting from the supervised learning process described in chapter 4 is exported and made available to the processor. It can be loaded as an external input (see left box "External Input" in Figure 16) simply specifying a path to the model. This implementation flexibility allows for updates to the model without having to change the core part of the operator (see right box "STSI SNAP Operator"). In this work C2RCC has been used, but also other algorithms are possible for retrieving R_{rs}, as the user has the option to define a list of the available band names. The core part essentially requires a L2A processed OLCI product with available R_{rs} bands from 412 nm to 753.75 (same wavelength range as defined in Table 5). These bands are then processed using the trapezoidal integration normalization as described in section 3.5. The resulting spectra are converted into XGBoost matrix format, enabling the usage of the XGBoost prediction function to assign every pixel of the OLCI product a TSI class. The result of the class assignments are added to the original L2A product as a new band.
The implementation has been written in Python, requiring the SNAP toolbox to be configured to use the SNAP Python API (snappy). Furthermore, the code is hosted under the General Public License (GNU) v3.0 allowing for distribution, modification, private and commercial use of the software when license and copyrights are granted. The source code is hosted on `https://github.com/bcdev/stsi` and will be made publicly available once version 1.0 is ready to be released. A plug-in for SNAP can also be received from the author.

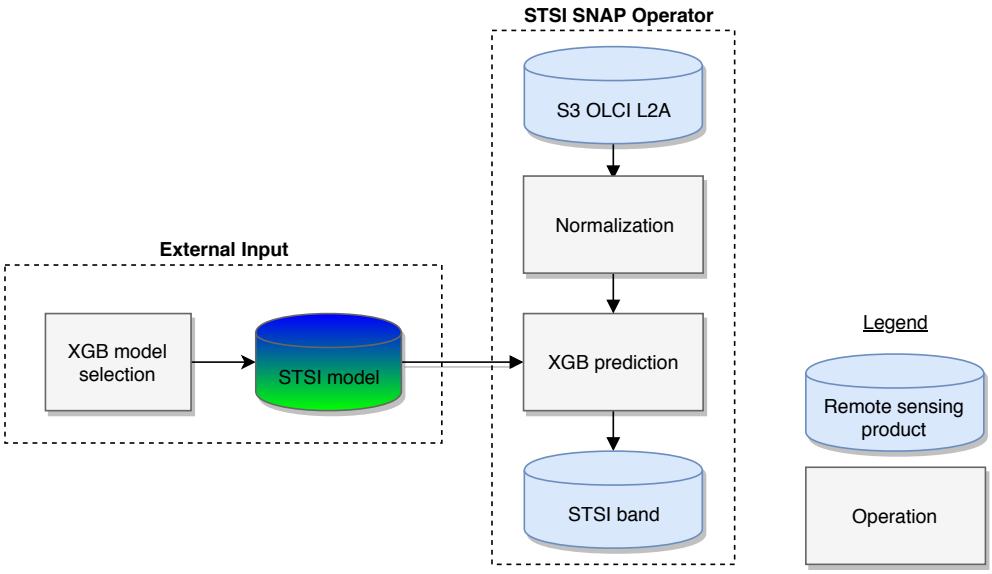


Figure 16: Schematic of the STSI SNAP operator implementation to produce L2 STSI OLCI products of surface water bodies.

6 Results

The prototype of the STSI has been finished and it can be used on inland water bodies for which R_{rs} of OLCI can be qualitatively derived. To evaluate the findings, the result section is split into two. First, the class assignments of the classifier are presented and second the class assignments are compared to the TSI values for the respective locations allowing for a judgement about the accuracy of the methodology.

6.1 Classification

The results for each lake were achieved using the STSI framework outlined in Chapter 2 that has been described thoroughly in the preceding sections. The retrieved OLCI reflectances provide a holistic view on the lakes' optical variability enabling to use the STSI to provide a TSI class for every valid pixel.

6.1.1 Lake Pelican

Of the total 287 valid R_{rs}, 275 have been assigned to the oligotrophic, 1 to the mesotrophic and 11 to the eutrophic TSI classes (Figure 17). To begin with, the retrieved R_{rs} display realistic spectral patterns, supporting the assumption that C2RCC provides accurate R_{rs} in low biomass waters. In addition, the valid pixel selection worked well as none of the reflectances accounts for a spectral shape showing typical structures originating from land influences. Lake Pelican is a challenging environment for ocean-colour remote sensing, but the results document the overall robustness of the methodological approach to retrieve the necessary reflectances. Regarding the class assignments, the class differentiation between the three eutrophic classes is clearly recognizable (see Figure 18). However, a slightly worse spectral distinction between the oligotrophic and mesotrophic classes (1 and 2) is apparent (wavelength region from 500 - 560 nm). While nearly all of those spectra are assigned to class 1, several spectra could also be class 2 (see Figure 17). In general, spectra seem to be well distinguished between the oligotrophic and eutrophic classes. Mostly only spectra of class 3 show an increase in reflectivity in the red part of the spectrum around the red-edge between 670 - 720 nm. Figure 19 shows the comparison of the simulated R_{rs} versus the R_{rs} of Lake Pelican classes 1 and 3, respectively. While for class 3 the OLCI spectra are approximately covered by the simulated spectra, the spectral shapes of class 1 (especially between 400 - 490 nm) show higher reflectance values outranging the simulated R_{rs}.

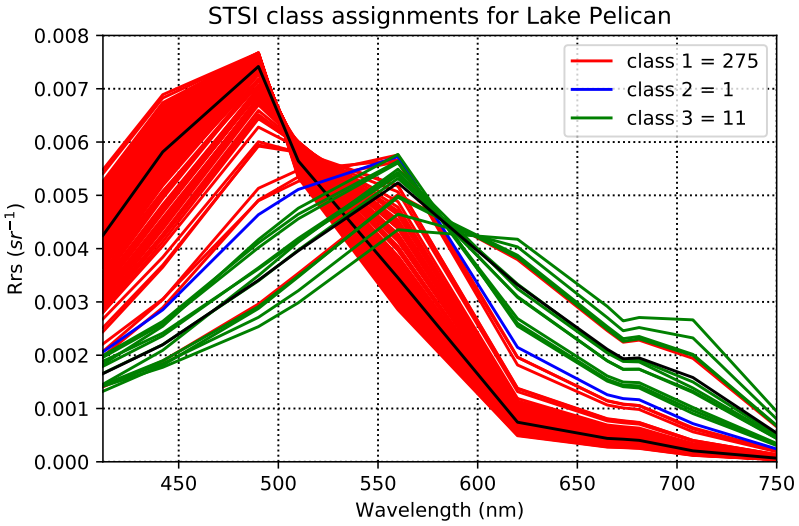


Figure 17: STSI class assignments for Lake Pelican. Normalized R_{rs} (n = 287).

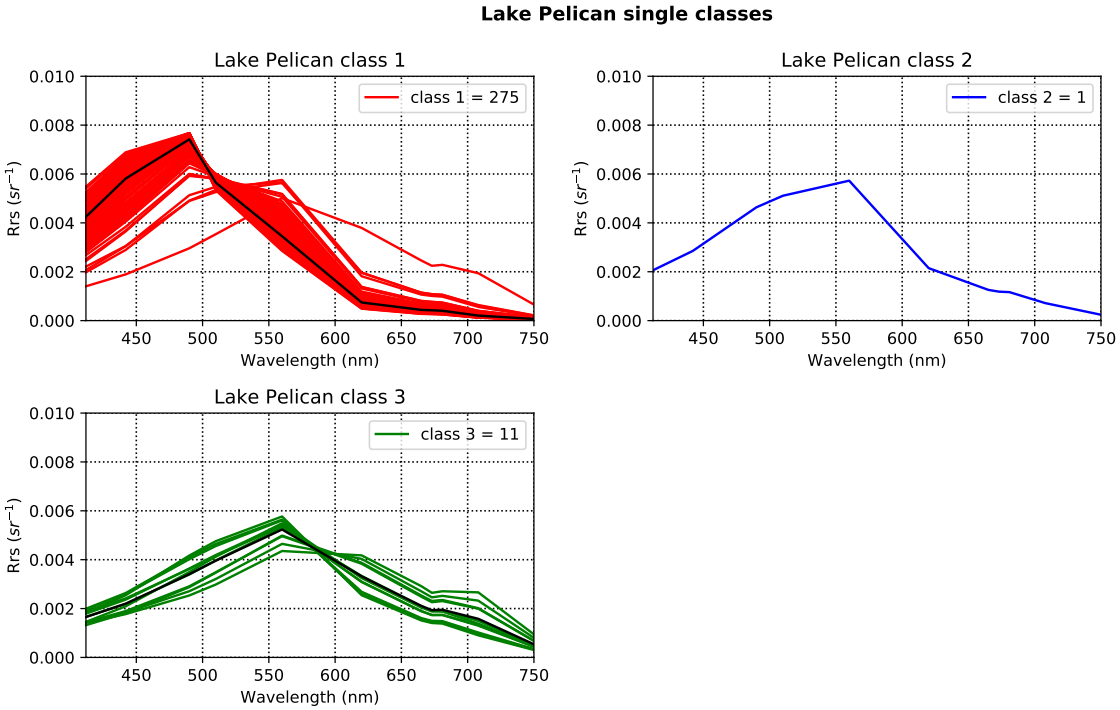


Figure 18: Lake Pelican single STSI class plots.

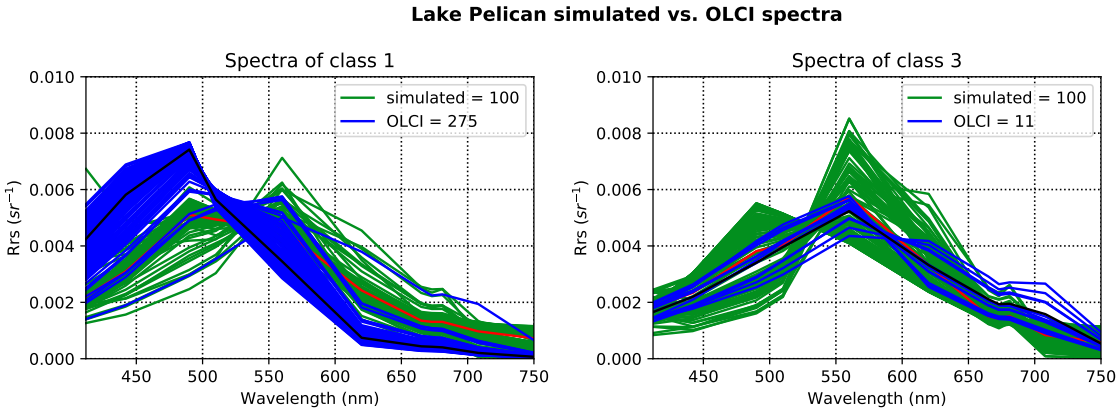


Figure 19: Lake Pelican simulated R_{rs} (green) vs STSI assignments of the OLCI R_{rs} (blue). Both reflectance datasets are normalized.

6.1.2 Lake Jordan

Lake Jordan has been the most challenging environment to retrieve quality reflectances from. Besides the difficult valid pixel selection, the AC module of C2RCC might not provide realistic reflectances for cyanobacterial bloom situations (as described in section 3.10.2). Hypereutrophic conditions are dominating Lake Jordan on the day of sampling, possibly including cyanobacteria events. Due to the lack of in-situ data covering this water type, it is not possible to judge whether or not algal blooms exist during the OLCI acquisition. An error might be included in the C2RCC produced reflectances, but necessary radiometric validation data are currently lacking to confirm this hypothesis. Nevertheless, it is to assume that not all areas of the lake are affected by cyanobacteria and that consequently most of the spectra represent valid reflectances. The STSI method assigns 19 reflectances to the oligotrophic class, 112 to the eutrophic and 49 to hypereutrophic classes (see Figure 20). All of the reflectances are similar to another in magnitude, with differences only in shape requiring a high sensibility from the classifier to recognise class defining common features.

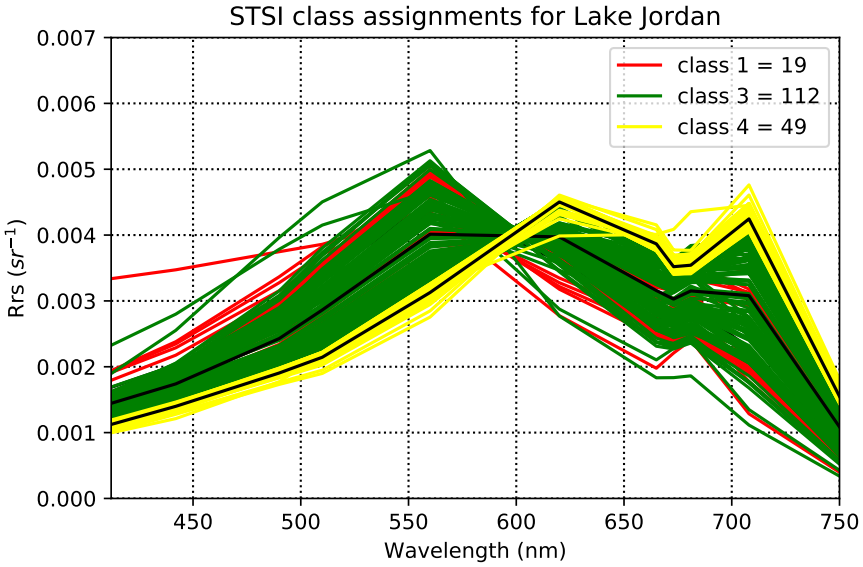


Figure 20: STSI class assignments for Lake Jordan. Normalized R_{rs} (n = 180).

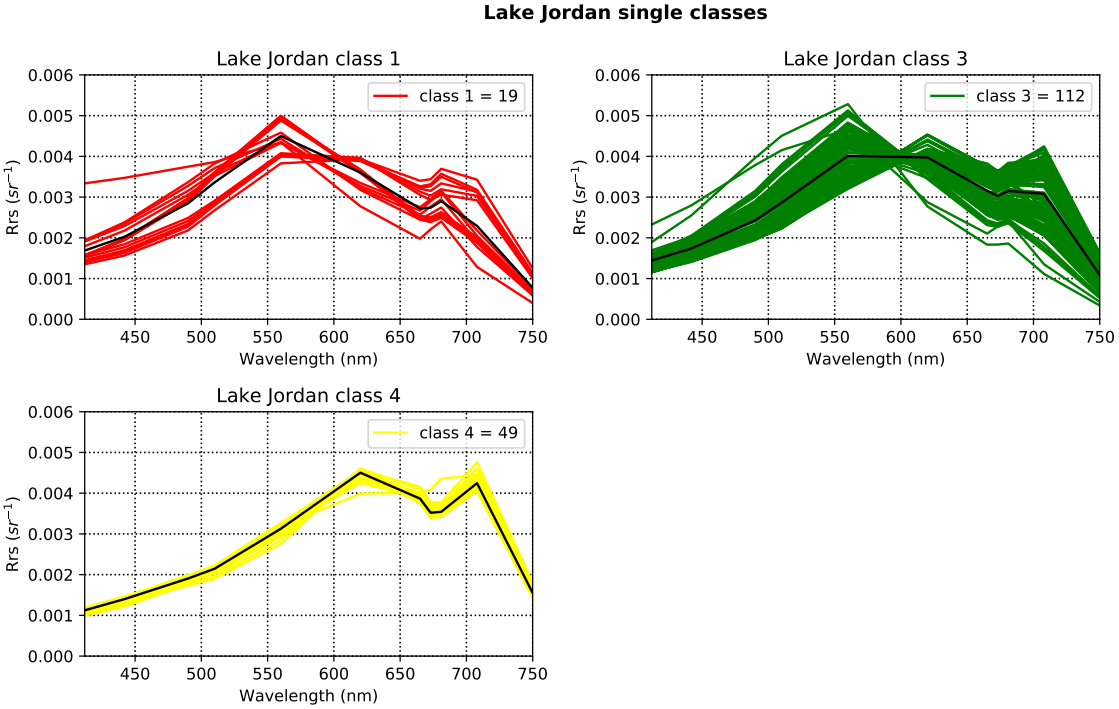


Figure 21: Lake Jordan single STSI class plots.

The mean reflectance vectors depicted in Figure 21 show the clear spectral separation of the three classes providing insight into the decision making of the classifier. The spectral shapes of the classes are all unique and separable from each other, but the class assignment itself does not provide enough information to resonate about the STSI class predictions for Lake Jordan. Figure 22 provides an additional view on the class assignments. For all three classes basically none of the spectral shapes is entirely covered in the simulated database providing new insight into the limitations of the database. Although 19 spectra were assigned to the oligotrophic clear water class, the spectral shapes of Lake Jordan do not indicate clear oligotrophic water states anywhere across the lake where R_{rs} were used in the classification. Class 3 is the closest regarding the spectral similarity, but then again the spectra of class 4 are not covered by the simulated R_{rs} at all. The class assignments for

Lake Jordan indicate a degree of arbitrariness whether or not a spectrum is class 1,2, 3 or 4 as those spectral shapes are not covered by the training database the STSI classifier bases its class prediction on. This in turn indicates that STSI predictions on similar trophic conditions might produce completely different results while they could at the same time be entirely wrong.

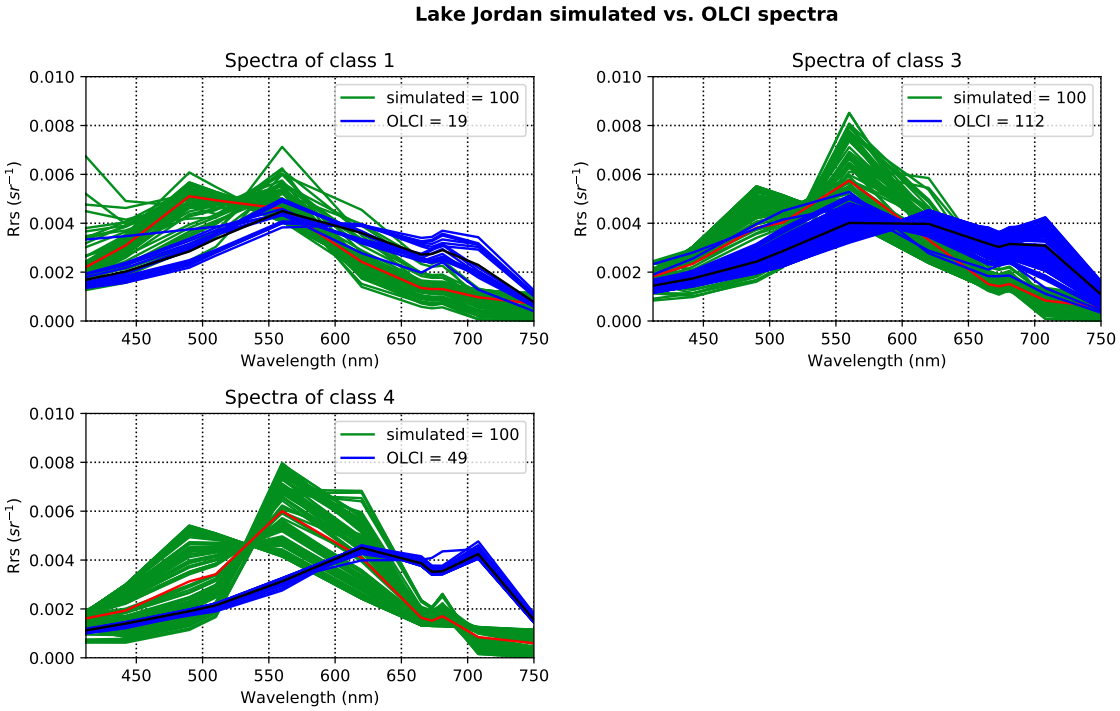


Figure 22: Lake Jordan simulated R_{rs} (green) vs STSI assignments of the OLCI R_{rs} (blue). Both reflectance datasets are normalized.

6.1.3 Lake Michigan

Comparable to Lake Pelican, the C2RCC calculated reflectances for Lake Michigan are realistic both in shape and magnitude. The narrow degree to correctly separate between valid and invalid pixels has been accurately managed using the Idepix tool.
Nearly all of the R_{rs} from Lake Michigan are assigned to the oligotrophic class 1, except for 9 spectra being assigned to class 3, indicating eutrophic conditions (see Figure 23). Associated therewith, also the shapes strongly differ from each other (see Figure 24). The polygon of usable R_{rs} is greater than the area of the bay where the in-situ measurements were sampled. The large gradient in trophic conditions can be explained with the close proximity to the harbour of Milwaukee accounting for the main source of nutrient intake causing the eutrophic conditions. The nutrient rich waters disperse with greater distance to shore leading to an oligotrophic state of the main lake basin. The main basin of Lake Michigan is extraordinary large for an inland water body and plays a major role in the distribution of nutrients. Another interesting feature is that none of the spectra has been assigned to class 2, while also the in-situ values do not indicate mesotrophic conditions (see next section). Especially the mean reflectance vectors of the OLCI spectra are close to the simulated dataset, being the reason for the class label assignments (see Figure 25). Between the two encountered trophic states the intermediate mesotrophic conditions might still exist in an outer part of the bay, but they are most likely simply not covered by the in-situ sampling during this day in the studied area.

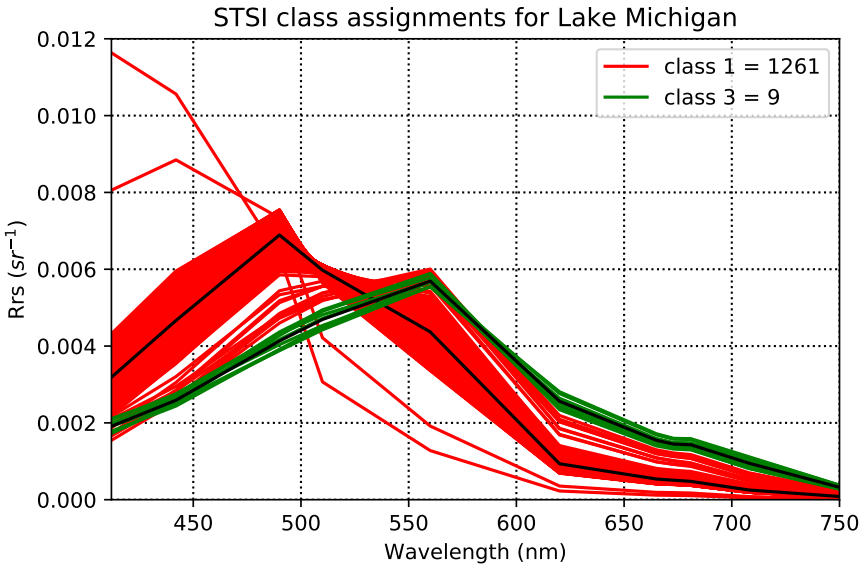


Figure 23: STSI class assignments for Michigan. Normalized R_{rs} (n = 1270).

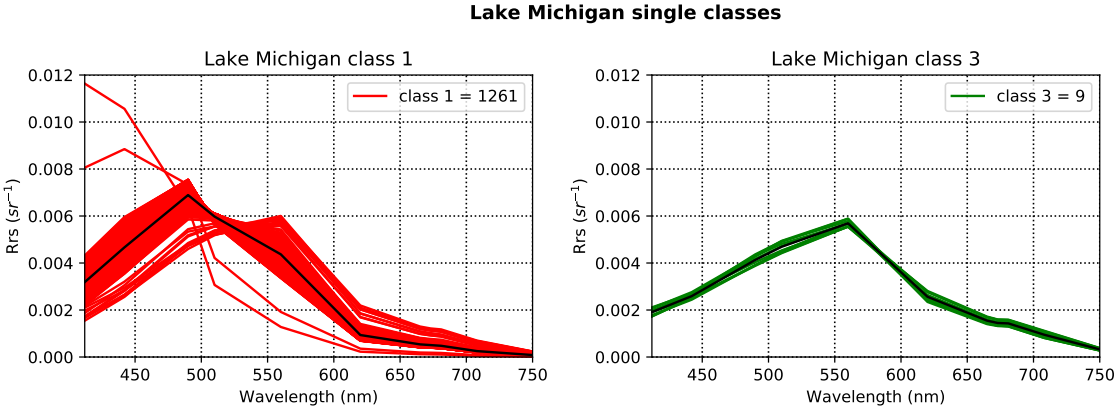


Figure 24: Lake Michigan single STSI class plots.

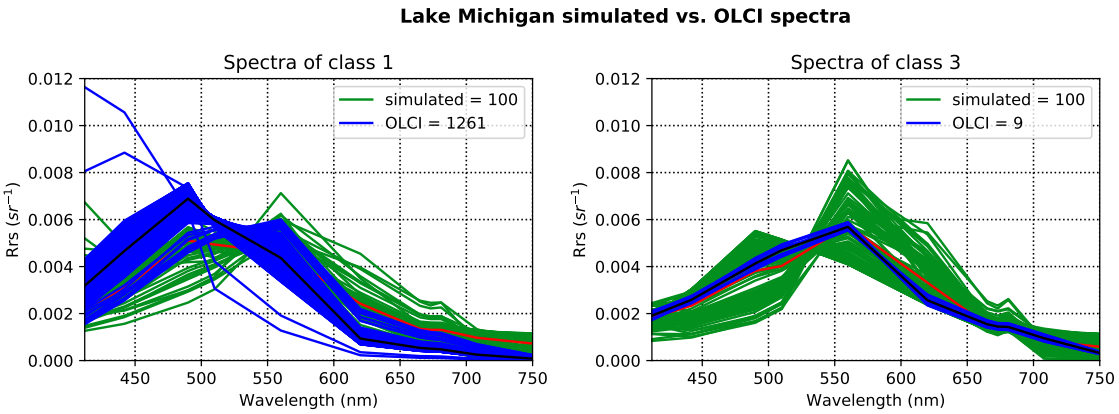


Figure 25: Lake Michigan simulated R_{rs} (green) vs STSI assignments of the OLCI R_{rs} (blue). Both reflectance datasets are normalized.

6.2 Validation

The STSI spectral class assignments presented in the previous section were validated using the prepared chl-a in-situ measurements of the locations in the lakes. Using the class defining chl-a concentrations of Table 4, the TSI class labels were derived for all of the in-situ stations (see Table 7). Further information about the stations can be found in section 3.10.
Together the lakes cover all trophic states, but mesotrophic conditions are underrepresented in the dataset. Eutrophic conditions are available, but more measurements of this trophic state are required to test the STSI. The varying trophic states are the result of the large range of chl-a in-situ concentrations (0.18 - 88 ug/l). The measurement maximum depths range from surface (Pelican, Jordan) to slightly more deep water (Michigan). Although radiometric attenuation measurements were not available for the lakes, the in-situ measurements for Lake Michigan were used despite partly large measurement depths. The lake is generally oligotrophic allowing for the assumption that light water penetration can reach larger depths than the standard 0 - 5 m range of usually estimated maximum measurement depth. For the same location, the resulting TSI class labels of the in-situ measurements were then compared against the assigned TSI labels from the STSI method. The results are displayed using a confusion matrix (see Figure 26).
The overall prediction accuracies strongly differ between the lakes. For Lake Pelican the accuracy is 0.66 (2/3), as one out of three stations has been misclassified (oligotrophic instead of mesotrophic). The spectral class plot (Figure 17) also indicates 11 spectra to be eutrophic, but in-situ data is not available to validate these predictions. The STSI method has the lowest prediction accuracy for Lake Jordan (0.28) with only two correct predictions (2/7). All in-situ stations indicate hypereutrophic conditions throughout this lake, while most of the spectra are assigned to the eutrophic class 3. The STSI class predictions are

Table 7: In-situ measurement sites used for the validation of the STSI class assignments. The TSI classes are derived from the class defining chl-a concentration values listed in Table 4.

Lake name	**Station name**	**Chl-a conc.**	**Unit**	**Meas. Depth (m)**	**TSI class**
Pelican	18-0308-00-208	3.0	ug/l	0 - 2	Mesotrophic
Pelican	18-0308-00-209	2.0	ug/l	0 - 2	Oligotrophic
Pelican	18-0308-00-207	2.0	ug/l	0 - 2	Oligotrophic
Jordan	CPF081A1B	82.0	ug/l	0.2	Hypereutrophic
Jordan	CPF086C	60.0	ug/l	0.2	Hypereutrophic
Jordan	CPF086F	78.0	ug/l	0.8	Hypereutrophic
Jordan	CPF087B3	72.0	ug/l	1.0	Hypereutrophic
Jordan	CPF0880A	62.0	ug/l	1.0	Hypereutrophic
Jordan	CPF055C	88.0	ug/l	1.2	Hypereutrophic
Jordan	CPF055E	72.0	ug/l	1.4	Hypereutrophic
Michigan	OH-17S	0.18	mg/m^3	1.0	Oligotrophic
Michigan	OH-12S	0.2	mg/m^3	1.0	Oligotrophic
Michigan	OH-05B	0.78	mg/m^3	9.1	Oligotrophic
Michigan	OH-14S	0.28	mg/m^3	1.0	Oligotrophic
Michigan	OH-07M	0.46	mg/m^3	5.8	Oligotrophic
Michigan	OH-03B	8.2	mg/m^3	8.2	Eutrophic
Michigan	OH-11M	36.0	mg/m^3	4.8	Eutrophic
Michigan	OH-13B	0.66	mg/m^3	14.3	Oligotrophic

not off by far from the true TSI label (class 4). However, 19 spectra were assigned to the oligotrophic class not even slightly reflecting the actual trophic conditions occurring in the lake. As described in the previous sections, the spectral class assignments currently are not reliable and precise for optically extreme conditions like those encountered in Lake Jordan. Opposite to the low accuracies and instability of results for Lake Jordan, the prediction of Lake Michigan is extremely accurate (1.00). All TSI classes were correctly predicted (8/8), even those measured in larger depths that are both oligo- and eutrophic. The spectral plot of Lake Michigan (see Figure 23) in the previous section provides the necessary insight: remarkable is the separation of spectra that show similar reflectance values around 560 nm, but differences in the wavelength range before (450 - 550 nm) that ultimately led to difference class assignments. This confirms that clearly separable spectra will be assigned to different classes, if only the mean reflectance vectors of the simulated database accurately represent these spectral shapes.
The overall prediction accuracy of the STSI classifier is 0.66, but it is questionable how reliable this value is regarding the poor performance of the current model for hypereutrophic lakes.

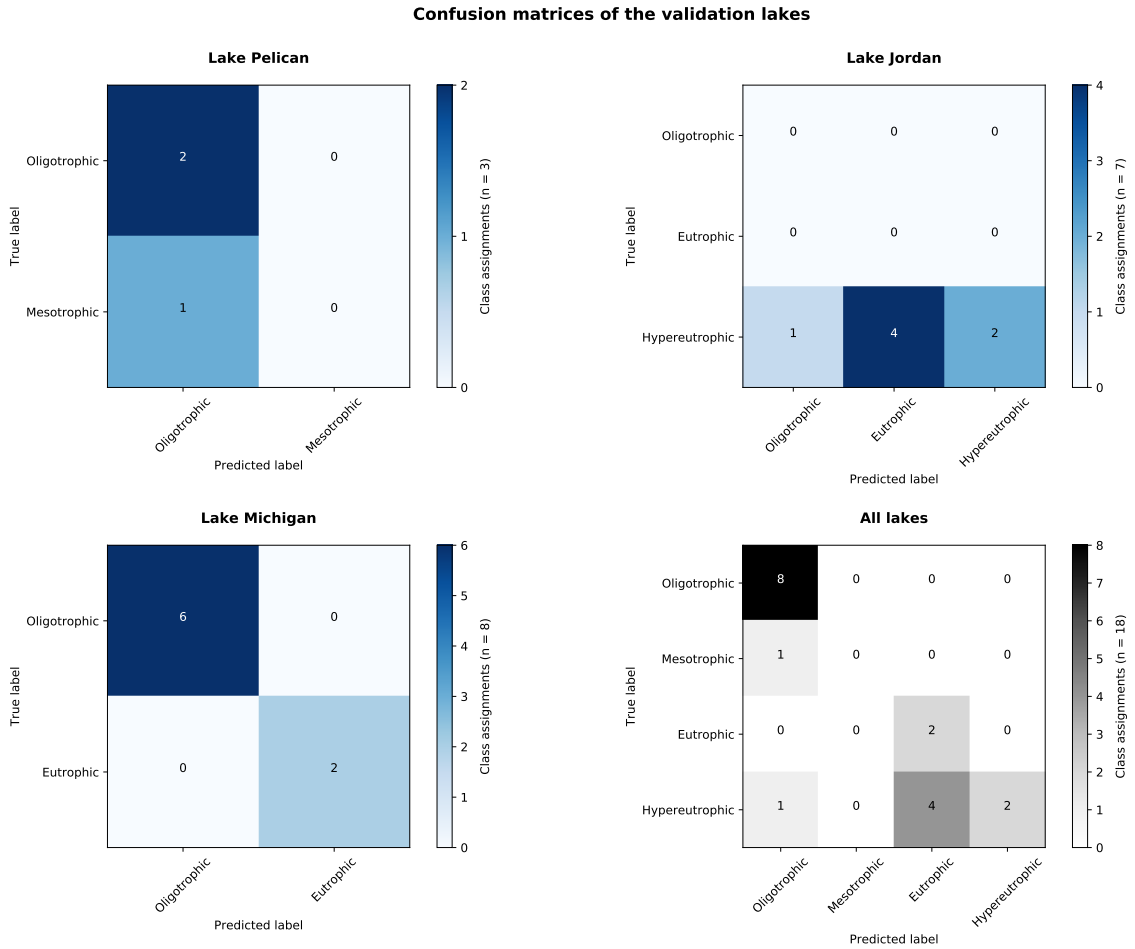


Figure 26: Confusion matrices of the three validation lakes. The diagonal elements represent the number of matches for which the predicted STSI label is equal to the true TSI in-situ label. Off-diagonal elements are those that are misclassified by the STSI model. Higher values in the diagonal elements of the confusion matrix indicate many correct predictions.

7 Discussion

The STSI algorithm depends on quality assured R_{rs}. The methodology outlined in section 3.6 using Idepix and C2RCC represents a robust approach to derive the required reflectances. While Idepix performed well for all of the analysed lakes, the AC module of C2RCC faces limitations that can currently not be investigated due to missing in-situ reflectance measurements. Upcoming updates to the neural networks utilised in the atmospheric correction module will certainly improve the weaknesses in the addressed extreme optical and trophic lake conditions. Nevertheless, the retrieved sensor R_{rs} are currently not quality assured. Therefore it is highly desirable to implement a quality assurance system of the calculated R_{rs} before a reflectance is classified into one of the TSI classes. In particular, the methodological approach designed by Wei et al. (2016) seems promising: every R_{rs} is compared to quality controlled hyperspectral in-situ measured R_{rs} enabling to assure that artefacts are excluded and only realistic reflectances are considered in the classification. The method has been successfully used in recent studies (Shang et al., 2017; Zheng and DiGiacomo, 2017). Topically related, the herein presented, but not implemented, out-of-range check (see section 3.6.5) has a closely related intention. The aim is to discard reflectances that are not covered by the training database to only base the classification on those that are known to the trained classifier to ultimately improve the STSI class assignments. Theoretically this also excludes different reflectances produced by the AC module (compared to the simulated spectra). While this approach might work for conventional retrieval-based methods like inversion algorithms, it is questionable if the approach for the STSI is appropriate. To illustrate this, it is useful to compare the use of reflectances in an inversion algorithm and the STSI method. Inversion-based algorithms can only correctly invert a reflectance if it is covered by the IOP combination defined in its LUT. Logically, all unknown or out-of-range (OOR) reflectances are most likely inverted erroneously, as the mathematical method (e.g. linear matrix inversion) utilized can not retrieve an IOP combination that is non-existent in its LUT. On the contrary, the STSI method is not based on an inversion, but rather classifies a signal. The IOP combination that led to the reflectance shape is not relevant for the assignment process, hence a non-existent spectrum in the training database can still be assigned to a correct class. This phenomenon inherent to a classifier is completely separating the STSI method from conventional approaches trying to retrieve a concentration from a radiative signal. An example are the reflectances of class 1 from Lake Pelican. Their shape is not identically covered by the simulation, but they are still closest to those found in the training class 1 (see Figure 19) and hence receive the related class label. In return this indicates that a simulated trophic class does not have to perfectly hold the optical diversity existing for the range of chlorophyll concentrations defining a trophic class, as long as the major optical water types of a class are covered by the training database. Nonetheless, spectra that strongly differ in shape to those existent in the training classes are often assigned to a wrong class, as the predicted class most likely only remotely matches the spectral pattern of the encountered reflectances (e.g. classes 3 and 4 of Lake Jordan) resulting in low predictive accuracies. Logically, improving the simulated database or the atmospheric correction algorithms seem to be the best options to improve the prediction capabilities of the classifier.

Furthermore, spectral shapes that are close to either of the classes might be assigned falsely if the simulated spectra do not cover enough spectral diversity. While spectral diversity is a key element for correct class assignments, the separation of boundary trophic and optical cases needs to be viewed in a greater multi-dimensional context, that not only the STSI but also the classic TSI faces. To illustrate this, a chlorophyll concentration of e.g. 3.0 mg/m^3 (e.g. station '18-0308-00-20' of Lake Pelican) is a typical example faced by anyone using the definite TSI classification system. Whether or not a chl-a concentration of 3.0 mg/m^3 is meso- or still oligotrophic is even with in-situ analyses difficult to judge. The uncertainties of the measurement and analysis procedures make this concentration a class candidate for both trophic states. In-situ approaches solve this issue using different

methods and by the sampling of additional parameters, e.g. with more accurate methods (HLPC), chemical analyses of the nutrient-levels like nitrogen and phosphorous, strongly correlated parameters such as transparency or by explicitly defining the uncertainty range included in their analysis. For the STSI classification these measurements are not available, hence crucial information is lacking that would clarify the class assignment. Nevertheless, the STSI has to correctly deal with these boundary cases.
Going back to the initially defined research objectives, the question for the STSI to answer is: can trophic states be optically separated using chl-a concentrations as the separating parameter. In this context, a TSI chl-a boundary concentration like 3.0 mg/m^3 entails two issues: First, the classifier's internal classification process and second the spectral ambiguity of the target spectrum. For the first issue, the current STSI methodology offers two approaches. The spectral class assignments allow to resonate about the class assignment sensitivity of the XGBoost classier that can be controlled and adapted using several hyperparameters such as the learning rate, maximum depth of the tree or the minimum child weight. Adjusting the classifier to distinguish better between spectral deviations from the mean reflectance vectors will lead to a better separation of boundary conditions, i.e. the one encountered in Lake Pelican whose trophic conditions vary between high oligotrophic and low mesotrophic conditions. The spectral shapes differ only slightly, but this small spectral difference can be captured by a well-discriminating classifier. Further, a feature of the current STSI method is the interpretation of the classification output. It is interpreted using the softmax function providing probabilities of the class affinities for each spectrum. The probabilities display the class affiliation of a spectrum that is either clear (i.e. high probabilities for a class) or poor (similar probabilities for all classes). For a well-trained classifier, a spectrum with high class assignment probabilities (e.g. those of Lake Michigan) constitutes no class separation issue (disregarding the spectral ambiguity issue here), while those with similar probabilities (Lake Jordan) are difficult for the classifier to assign to a class. As mentioned, an improvement of the simulated database would lower the chances of spectra encountered whose class affinity is poor, but these boundary cases will eventually occur anyway. The probabilities allow to use several thresholds instead of simply assigning the spectrum to a class with the highest probabilities (arg max). For similar probabilities a threshold can be introduced that requires a class affinity to be at least 10% larger than the second highest probability (a selected threshold value has to be tested before its usage). All spectra not meeting the threshold can be assigned to the two classes with the highest probabilities, indicating the spectrum is a boundary case and that the STSI method can not confidently separate between the classes. In a future STSI version the probabilities can also be used in a hierarchical tree manner. For example, only spectra with a class affinity of e.g. 80% are directly assigned the respective class label. Otherwise the class affinities are checked for similarity and a decision is made whether the spectrum is assigned to the boundary class or another class type. Several cases exist that could be iteratively considered. The softmax interpretation of the output allows for several controlled class assignments that would improve the decision making process and altogether make the STSI a better and more transparent method.
While the internal class assignment process faces boundary cases, the spectrum itself might be ambiguous and hence a boundary case for the classifier. Spectral ambiguity has been shortly thematised in the first chapter and remains an issue for basically all ocean-colour algorithms. The STSI is not independent of spectral ambiguity, but effectively weakens the issue. The presented normalization procedure stresses the spectral shapes that are primarily influenced by absorbing materials like phytoplankton and CDOM and therefore weakens magnitude changes caused by scattering particles. The procedure certainly showed its value for all of the lakes, as the spectral ambiguity became less pronounced benefiting the classification to base the assignments on the absorbing materials. In the blue part of the spectrum, most conventional algorithms try to retrieve low chlorophyll-a concentrations from the absorption coefficients of phytoplankton. Ambiguity exists herein, as CDOM ef-

fectively can make up to 80% of the total absorption (*atot* in Equation 1), distorting the retrieval procedure that leads to an overestimation of the chl-a concentration (IOCCG, 2015). This retrieval issue is one of many the STSI avoids, as long as the spectrum is covered in the database. Lastly, the TSI itself weakens the impact of ambiguity naturally, as it accounts for a range of valid chlorophyll concentrations. The class ranges already include room for several different spectral compositions creating the same indistinguishable spectrum that is known to be an ill-posed problem for inversion algorithms. As the results of the study indicate, most spectra that strongly differ from another also fall into separate TSI classes. Only boundary concentrations of the TSI classes (the mentioned examples) seem to pose an issue, but they might not only be optically ambiguous, but also biologically. Still, the amount of test sites needs to be increased to be able to judge the influence of spectral ambiguity on the classification. An increase in test sites would be accompanied by the use of more in-situ stations providing more statistically profound insight into the accuracies of the STSI. However, unsolved is the issue of uncertainties inherent to the STSI predictions caused by the measured in-situ data. Errors in the measured in-situ samples add up to the uncertainty in the STSI caused by sensor and atmospheric constraints and the simulation (including the bio-optical model). Ideally an in-situ dataset is used for the validation that has been compiled for ocean-colour remote sensing purposes such as SeaBASS from NASA's Ocean Biology Processing Group (OBFG) or UK's LIMNADES originating from the GloboLakes project. While currently not possible, the use of suited validation measurements would enable to separate between the uncertainty inherent to the STSI algorithm and the validation data, both influencing the accuracies of the final STSI class predictions.

The STSI classification is based on a hyperspectral simulation. Due to its hyperspectral nature, the spectral resampling can be applied to other ocean-colour sensors like MERIS, MODIS, SeaWiFS and VIIRS enabling new application possibilities. If the method is proven to work on all of these sensors, it would be possible to fill gaps in long-term European (MERIS and OLCI) monitoring efforts. Adjoining is the question, if the method is limited to ocean-colour sensors only. The main argument for a multi-sensor applicability would be that the STSI does not require a band composition of the usual ocean-colour multi-spectral sensors like MERIS or OLCI, as it is not based on band-ratios requiring certain band positions. As long as the spectral shapes of atmospherically corrected reflectances are comparable to those in the simulated database, basically any sensor can be utilised making the STSI a multi-sensor algorithm. Examples are Sentinel-2 MSI (S2-MSI) and Landsat-8 (L8). Nevertheless, correct band positions to detect various chlorophyll features in water are necessary to distinguish between the spectra. This circumstance probably excludes some, if not all land sensors. An example is the missing red-edge band on L8. Consequently, the possibility to differentiate between chl-a and CDOM is missing, as both are already difficult to distinguish in blue and green areas of the VIS wavelength range. Certainly this would lead to errors in the STSI classification, because CDOM dominated waters would then be assigned to TSI classes with high chl-a concentrations. Adjoining is the issue of atmospheric correction schemes posing a large source of uncertainty that need to be evaluated for each sensor individually, as the procedures are often sensor specific.

8 Conclusion & outlook

The presented STSI methodology is capable of accurately predicting the non-linear TSI classes for the oligo- eutrophic lakes Pelican and Michigan, while at the same time the hypereutrophic conditions of Lake Jordan were not well covered. Using the STSI methodology, the validation results show that it is possible to use a simulated dataset to train a classifier and predict a trophic state class using only chl-a concentrations as the separating class parameter. Neither has the algorithm been specifically trained with local datasets nor do the results indicate a regional limitation usually being the limiting factor for band-ratio algorithms. The STSI classifier is a generic model not relying on local tuning.
While the three validation lakes were strongly differing from each other, the STSI needs to be applied to a larger time frame and in-situ dataset that covers more meso- and eutrophic conditions to draw more general conclusions about its functionality and limitations. Despite the high prediction accuracies for the two lakes, at the lake level difficulties were encountered that required an in-depth investigation of the validation datasets, the algorithms used to derive valid water pixels as well as the atmospheric correction procedure. The STSI strongly depends on correctly derived R_{rs} and the envisaged quality assurance system of Wei et al. (2016) can assure the STSI methods also works with other valid pixel and atmospheric correction algorithms.
Even though the issue of spectral ambiguities also exists for the STSI, the influence is effectively weakened and several issues faced by conventional inversion algorithms can be practically circumvented. More specifically, the conclusion can be drawn that clear distinguishable optical patterns of R_{rs} between the trophic state classes exist. However, the large optical variety of water types in each TSI class are currently only partly contained in the simulated database. Adjoining is the circumstance that adaptation and evolution of the STSI classification model is necessary to incorporate more optical complexity and heterogeneity. The current classification model is too strict and does not greatly treat boundary cases. Both, the improvement of the simulated database and an adaptation of the classifier settings are closely linked.
In the last years, optical water types (OWT) became widely popular in the field of inland water remote sensing to characterize the optical diversity found in these optically complex environments (Mélin and Vantrepotte, 2015; Moore et al., 2014; Spyrakos et al., 2017). OWTs of inland waters can help to improve the class assignment process of the STSI in several ways. First, the trophic state classes inherently include all OWTs naturally: even though most of the OWT clusters are not based on dominating chlorophyll-a concentrations, but the contribution of all optically active constituents, all clusters are existent throughout the TSI classes. Thereby, whether or not a OWT is chl-a dominated, they must be dealt with in the STSI classification. To use the gained knowledge about OWTs, either the STSI simulated database is constrained using measured (S)IOP values of OWTs to generate realistic spectra for each class or the database is validated against the spectra dominating the OWT classes to identify spectral gaps in the simulations. Each approach needs further evaluation but would lead to an improvement of the simulated database and consequently the classification. Several OWT classes include high chl-a concentrations and/or cyanobacteria abundance (Spyrakos et al., 2017). This optical variability is certainly underrepresented in the current STSI simulated database. At present, the bio-optical model that constitutes FEMWAT can not be changed, thus replacing it will be one of the key future activities. Because of the modular nature of the STSI framework, it can be replaced by a publicly available database that covers the most trophic extreme cases encountered in Lake Jordan. One example is the database originating from ESA's Case2Extreme project that is based on Hydrolight simulations for several different extreme water types.
The improvement of the STSI prototype will lead to a more precise STSI model. Three main fields can be distinguished in which updates would have a large impact on the quality of the predictions: a simulated database that accounts for higher chlorophyll-a concentra-

tions and that contains more optical diversity, a R_{rs} quality assurance system to only use valid reflectances in the classification and the use of optical water type frameworks to constrain, validate and enhance the simulations overall. Furthermore, a comparison between valid atmospherically corrected and the simulated reflectances would enable to learn about the discrepancies between the spectra and investigate the causes. The comparison can be enhanced applying different AC algorithms to get a good match between the sensor and simulated spectra.
The architecture of the STSI would enable to use the method on other ocean-colour, but possibly also sensors primarily designed for land usage (e.g. S2-MSI, L8). The multi-sensor application would enable to monitor longer periods of time, as trophic state changes are important habitat indicators over long-term scales. Together with future hyperspectral sensors like the German EnMAP and U.S. PACE missions, the STSI can prove to be a valuable tool to track and monitor eutrophication changes in optically complex water environments.

9 References

Albert, A. and Mobley, C. (2003). An analytical model for subsurface irradiance and remote sensing reflectance in deep and shallow case-2 waters, *Opt. Express* **11**(22): 2873–2890.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction, *Technometrics* **16**(1): 125–127.

Álvarez, X., Valero, E., Santos, R., Varandas, S., Fernandes, L. S. and Pacheco, F. (2017). Anthropogenic nutrients and eutrophication in multiple land use watersheds: Best management practices and policies for the protection of water resources, *Land Use Policy* **69**: 1 – 11.

Andersen, J. H., Conley, D. J. and Hedal, S. (2004). Palaeoecology, reference conditions and classification of ecological status: the eu water framework directive in practice, *Marine Pollution Bulletin* **49**(4): 283 – 290.

Ayres, W., Busia, A., Dinar, A., Hirji, R., Lintner, S., McCalla, A. and Robelus, R. (1997). Integrated Lake and Reservoir Management. World Bank Approach and Experience, *Technical report*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg.

Blanka, D. (1981). Relationship between chlorophyll-a concentration and phytoplankton biomass in several reservoirs in czechoslovakia, *Internationale Revue der gesamten Hydrobiologie und Hydrographie* **66**(2): 153–169.

Brando, V. E., Dekker, A. G., Park, Y. J. and Schroeder, T. (2012). Adaptive semi-analytical inversion of ocean color radiometry in optically complex waters, *Appl. Opt.* **51**(15): 2808–2833.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**(2): 123–140.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, CA.

Brockmann, C., Doerffer, R., Peters, M., Kerstin, S., Embacher, S. and Ruescas, A. (2016). Evolution of the c2rcc neural network for sentinel 2 and 3 for the retrieval of ocean color products in normal and extreme optically complex waters, *Living Planet Symposium*, Vol. 740 of *ESA Special Publication*, p. 54.

Bukata, R. P. (1995). *Optical properties and remote sensing of inland and coastal waters*, Boca Raton, Fla. : CRC Press.

Bulgarelli, B. and Doyle, J. P. (2004). Comparison between numerical models for radiative transfer simulation in the atmosphere–ocean system, *Journal of Quantitative Spectroscopy and Radiative Transfer* **86**(3): 315 – 334.

Bulgarelli, B., Kisselev, V. B. and Roberti, L. (1999). Radiative transfer in the atmosphere–ocean system: the finite-element method, *Appl. Opt.* **38**(9): 1530–1542.

Bulgarelli, B., Mélin, F. and Zibordi, G. (2003). Seawifs-derived products in the baltic sea: performance analysis of a simple atmospheric correction algorithm, *Oceanologia* **45**(4): 655–677.

Bulgarelli, B. and 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., Zibordi, G. and Berthon, J.-F. (2003). Measured and modeled radiometric quantities in coastal waters: toward a closure, *Appl. Opt.* **42**(27): 5365–5381.

Carlson, R. E. (1977). A trophic state index for lakes, *Limnology and Oceanography* **22**(2): 361–369.

Chandrasekhar, S. (1960). *Radiative transfer*, Vol. 76, Dover Publications.

Chen, C., Wang, L., Ji, R., Budd, J. W., Schwab, D. J., Beletsky, D., Fahnenstiel, G. L., Vanderploeg, H., Eadie, B. and Cotner, J. (2004). Impacts of suspended sediment on the ecosystem in lake michigan: A comparison between the 1998 and 1999 plume events, *Journal of Geophysical Research: Oceans* **109**(C10): 1–18.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *CoRR* **abs/1603.02754**.

CNEPA (2002). *Environmental Quality Standard for Surface Water*, Chinese Environmental Protection Agency. GB 3838-88.

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

Danne, O. (2016). Beam-idepix operator, Brockmann Consult, Geesthacht, Germany. **URL:** *https://github.com/bcdev/beam-idepix*

de Brébisson, A. and Vincent, P. (2015). An exploration of softmax alternatives belonging to the spherical loss family, *CoRR* **abs/1511.05042**.

Defoin-Platel, M. and Chami, M. (2007). How ambiguous is the inverse problem of ocean color in coastal waters?, *Journal of Geophysical Research: Oceans* **112**(C3): 1–16.

Dekker, A. G., Malthus, T. J. and Seyhan, E. (1991). Quantitative modeling of inland water quality for high-resolution mss systems, *IEEE Transactions on Geoscience and Remote Sensing* **29**(1): 89–95.

Doerffer, R. and Schiller, H. (2007). The meris case 2 water algorithm, *International Journal of Remote Sensing* **28**(3-4): 517–535.

Duan, K., Keerthi, S. S., Chu, W., Shevade, S. K. and Poo, A. N. (2003). Multi-category classification by soft-max combination of binary classifiers, *in* T. Windeatt and F. Roli (eds), *Multiple Classifier Systems*, Springer, Berlin, Heidelberg, pp. 125–134.

Eleveld, M. A., Ruescas, A. B., Hommersom, A., Moore, T. S., Peters, S. W. M. and Brockmann, C. (2017). An optical classification tool for global lake waters, *Remote Sensing* **9**(5).

Esaias, W. E., Abbott, M. R., Barton, I., Brown, O. B., Campbell, J. W., Carder, K. L., Clark, D. K., Evans, R. H., Hoge, F. E., Gordon, H. R., Balch, W. M., Letelier, R. and Minnett, P. J. (1998). An overview of modis capabilities for ocean science observations, *IEEE Transactions on Geoscience and Remote Sensing* **36**(4): 1250–1265.

Evgeniou, T., Pontil, M. and Poggio, T. (2000). Statistical learning theory: A primer, *International Journal of Computer Vision* **38**(1): 9–13.

Fan, Y., Li, W., Voss, K. J., Gatebe, C. K. and Stamnes, K. (2016). Neural network method to correct bidirectional effects in water-leaving radiance, *Appl. Opt.* **55**(1): 10–21.

FDEP (2012). *Integrated Water Quality Assessment for Florida: 2012 Sections 303 (d), 305 (b), and 314 Report and Listing Update*, Florida Department of Environmental Protection. Division of Environmental Assessment and Restoration. Tallahassee.

FDEP (2015). *Florida Watershed Monitoring Status and Trend Program Design Document*, Florida Department of Environmental Protection. EPA 62-40.540.

Fomferra, N. and Brockmann, C. (2005). Beam - the envisat meris and aatsr toolbox, *MERIS AATSR Workshop*.

Fomferra, N., Böttcher, M., Zühlke, M., Brockmann, C. and Kwiatkowska, E. (2012). Calvalus: Full-mission eo cal/val, processing and exploitation services, *2012 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5278–5281.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm, *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 148–156.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* **55**(1): 119 – 139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine., *Ann. Statist.* **29**(5): 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting, *Computational Statistics and Data Analysis* **38**(4): 367 – 378.

Gege, P. (2017). Chapter 2 - radiative transfer theory for inland waters, *in* D. R. Mishra, I. Ogashawara and A. A. Gitelson (eds), *Bio-optical Modeling and Remote Sensing of Inland Waters*, Elsevier, pp. 25 – 67.

Gilerson, A. A., Gitelson, A. A., Zhou, J., Gurlin, D., Moses, W., Ioannou, I. and Ahmed, S. A. (2010). Algorithms for remote estimation of chlorophyll-a in coastal and inland waters using red and near infrared bands, *Opt. Express* **18**(23): 24109–24125.

Gonçalves-Araujo, R., Rabe, B., Peeken, I. and Bracher, A. (2018). High colored dissolved organic matter (cdom) absorption in surface waters of the central-eastern arctic ocean: Implications for biogeochemistry and ocean color algorithms, *PLOS ONE* **13**(1): 1–27.

Gons, H. J., Auer, M. T. and 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.

Haltrin, V. I. (2002). One-parameter two-term henyey-greenstein phase function for light scattering in seawater, *Appl. Opt.* **41**(6): 1022–1028.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*, 2 edn, Springer.

Henyey, L. G. and Greenstein, J. L. (1941). Diffuse radiation in the Galaxy, *Astrophys. J.* **93**: 70–83.

Hieronymi, M., Müller, D. and 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.

Hoogenboom, H., Dekker, A. and de Haan, J. (1998). Retrieval of chlorophyll and suspended matter from imaging spectrometry data by matrix inversion, *Canadian Journal of Remote Sensing* **24**(2): 144–152.

Hu, C., Lee, Z. and Franz, B. (2012). Chlorophyll aalgorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference, *Journal of Geophysical Research: Oceans* **117**(C1): n/a–n/a. C01011.

Hyafil, L. and Rivest, R. L. (1976). Constructing optimal binary decision trees is np-complete, *Information Processing Letters* **5**(1): 15 – 17.

IOCCG (2000). *Remote Sensing of Ocean Colour in Coastal, and Other Optically-Complex, Waters*, Vol. No. 3 of *Reports of the International Ocean Colour Coordinating Group*, IOCCG, Dartmouth, Canada.

IOCCG (2010). *Atmospheric Correction for Remotely-Sensed Ocean-Colour Products*, Vol. No. 10 of *Reports of the International Ocean Colour Coordinating Group*, IOCCG, Dartmouth, Canada.

Jackson, T., Sathyendranath, S. and Mélin, F. (2017). An improved optical classification scheme for the ocean colour essential climate variable and its applications, *Remote Sensing of Environment* **203**: 152 – 161.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*, Springer Publishing Company, Incorporated.

Jerlov, N. (1976). Chapter 7 theory of radiative transfer in the sea, *Marine Optics*, Vol. 14 of *Elsevier Oceanography Series*, Elsevier, pp. 83 – 100.

Kerfoot, W. C., Budd, J. W., Green, S. A., Cotner, J. B., Biddanda, B. A., Schwab, D. J. and Vanderploeg, H. A. (2008). Doughnut in the desert: Late-winter production pulse in southern lake michigan, *Limnology and Oceanography* **53**(2): 589–604.

Kerfoot, W. C., Yousef, F., Green, S. A., Budd, J. W., Schwab, D. J. and Vanderploeg, H. A. (2010). Approaching storm: Disappearing winter bloom in lake michigan, *Journal of Great Lakes Research* **36**: 30 – 41.

Kisselev, V. B., Roberti, L. and Perona, G. (1995). Finite-element algorithm for radiative transfer in vertically inhomogeneous media: numerical scheme and applications, *Appl. Opt.* **34**(36): 8460–8471.

Kisselev, V., Roberti, L. and Perona, G. (1994). An application of the finite element method to the solution of the radiative transfer equation, *Journal of Quantitative Spectroscopy and Radiative Transfer* **51**(4): 603 – 614.

Kumari, B. (2005). Comparison of high performance liquid chromatography and fluorometric ocean colour pigments, *Journal of the Indian Society of Remote Sensing* **33**(4): 541–546.

Kutser, T., Herlevi, A., Kallio, K. and Arst, H. (2001). A hyperspectral model for interpretation of passive optical remote sensing data from turbid lakes, *Science of The Total Environment* **268**(1): 47 – 58.

Lambou, V., Taylor, W., Hern, S. and Williams, L. (1983). Comparisons of trophic state measurements, *Water Research* **17**(11): 1619 – 1626.

Lavender, S. J., Pinkerton, M. H., Froidefond, J.-M., Morales, J., Aiken, J. and Moore, G. F. (2004). Seawifs validation in european coastal waters using optical and biogeochemical measurements, *International Journal of Remote Sensing* **25**(7-8): 1481–1488.

Le, C., Li, Y., Zha, Y., Sun, D., Huang, C. and Lu, H. (2009). A four-band semi-analytical model for estimating chlorophyll a in highly turbid lakes: The case of taihu lake, china, *Remote Sensing of Environment* **113**(6): 1175 – 1182.

Le, C., Li, Y., Zha, Y., Sun, D., Huang, C. and Zhang, H. (2011). Remote estimation of chlorophyll a in optically complex waters based on optical classification, *Remote Sensing of Environment* **115**(2): 725 – 737.

Lee, Z., Ahn, Y.-H., Mobley, C. and Arnone, R. (2010b). Removal of surface-reflected light for the measurement of remote-sensing reflectance from an above-surface platform, *Opt. Express* **18**(25): 26313–26324.

Lee, Z., Carder, K. L., Mobley, C. D., Steward, R. G. and Patch, J. S. (1999). Hyperspectral remote sensing for shallow waters: 2. deriving bottom depths and water properties by optimization, *Appl. Opt.* **38**(18): 3831–3843.

Lee, Z. P., Carder, K. L., Peacock, T. G., Davis, C. O. and Mueller, J. L. (1996). Method to derive ocean absorption coefficients from remote-sensing reflectance, *Appl. Opt.* **35**(3): 453–462.

Lee, Z. P., Du, K., Voss, K. J., Zibordi, G., Lubac, B., Arnone, R. and Weidemann, A. (2011). An inherent-optical-property-centered approach to correct the angular effects in water-leaving radiance, *Appl. Opt.* **50**(19): 3155–3167.

Li, Y., Wang, Q., Wu, C., Zhao, S., Xu, X., Wang, Y. and Huang, C. (2012). Estimation of chlorophyll a concentration using nir/red bands of meris and classification procedure in inland turbid water, *IEEE Transactions on Geoscience and Remote Sensing* **50**(3): 988–997.

Ligi, M., Kutser, T., Kallio, K., Attila, J., Koponen, S., Paavel, B., Soomets, T. and 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.

Loisel, H. and Morel, A. (2001). Non-isotropy of the upward radiance field in typical coastal (case 2) waters, *International Journal of Remote Sensing* **22**(2-3): 275–295.

Lubac, B. and 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.

Mainstone, C. P. and Parr, W. (2002). Phosphorus in rivers — ecology and management, *Science of The Total Environment* **282-283**: 25 – 47.

Matthews, M. W., Bernard, S. and 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.

Matthews, M. W. and 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.

McCormick, N. J. (1992). Inverse radiative transfer problems: A review, *Nuclear Science and Engineering* **112**(3): 185–198.

Mélin, F. and Vantrepotte, V. (2015). How optically diverse is the coastal ocean?, *Remote Sensing of Environment* **160**: 235 – 251.

Mélin, F., Vantrepotte, V., Clerici, M., D'Alimonte, D., Zibordi, G., Berthon, J.-F. and Canuti, E. (2011). Multi-sensor satellite time series of optical properties and chlorophyll-a concentration in the adriatic sea, *Progress in Oceanography* **91**(3): 229 – 244.

Mobley, C. D. (1994). *Light and water : radiative transfer in natural waters*, San Diego : Academic Press.

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

Monahan, E. C. and O'Muircheartaigh, I. G. (1986). Whitecaps and the passive remote sensing of the ocean surface, *International Journal of Remote Sensing* **7**(5): 627–642.

Moore, T. S., Campbell, J. W. and 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., Campbell, J. W. and 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., Dowell, M. D., Bradt, S. and 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. and Prieur, L. (1977). Analysis of variations in ocean color, *Limnology and Oceanography* **22**(4): 709–722.

Moses, W. J., Gitelson, A. A., Berdnikov, S., Saprygin, V. and 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.

MPCA (2018). Minnesota pollution control agency surface water data. Assessed: 2018-09-04.
URL: *https://www.pca.state.mn.us/surface-water-data*

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*, Cambridge, MA.

Nielsen, D. (2016). *Tree boosting with xgboost - why does xgboost win "every" machine learning competition?*, Master's thesis, Norwegian University of Science and Technology, Norway.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*, 2nd edn, Springer, New York.

Novo, E. M. L. A. d. M., Londe, L. d. R., Barbosa, C., Araujo, C. A. S. d. and Renna, C. D. (2013). Proposal for a remote sensing trophic state index based upon Thematic Mapper/Landsat images, *Revista Ambiente & Agua* **8**: 65 – 82.

Odermatt, D., Danne, O., Philipson, P. and Brockmann, C. (2018). Diversity ii water quality parameters for 300 lakes worldwide from envisat (2002–2012), *Earth System Science Data Discussions* **2018**: 1–26.

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

Ogashawara, I., Mishra, D. R. and Gitelson, A. A. (2017). Chapter 1 - remote sensing of inland waters: Background and current state-of-the-art, *in* D. R. Mishra, I. Ogashawara and A. A. Gitelson (eds), *Bio-optical Modeling and Remote Sensing of Inland Waters*, Elsevier, pp. 1 – 24.

Olmanson, L. G., Bauer, M. E. and Brezonik, P. L. (2008). A 20-year landsat water clarity census of minnesota's 10,000 lakes, *Remote Sensing of Environment* **112**(11): 4086 – 4097.

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

Pahlevan, N., Smith, B., Binding, C. and O'Donnell, D. M. (2017). Spectral band adjustments for remote sensing reflectance spectra in coastal/inland waters, *Opt. Express* **25**(23): 28650–28667.

Palmer, S. C., Kutser, T. and Hunter, P. D. (2015). Remote sensing of inland waters: Challenges, progress and future directions, *Remote Sensing of Environment* **157**: 1 – 8. Special Issue: Remote Sensing of Inland Waters.

Papoutsa, C., Akylas, E. and Hadjimitsis, D. (2014). Trophic state index derivation through the remote sensing of case-2 water bodies in the mediterranean region, *Central European Journal of Geosciences* **6**(1): 67–78.

Petzold, T. J. (1992). *Volume scattering functions for selected ocean waters*, UC San Diego: Scripps Institution of Oceanography.

Pinckney, J., Papa, R. and Zingmark, R. (1994). Comparison of high-performance liquid chromatographic, spectrophotometric, and fluorometric methods for determining chlorophyll a concentrations in estaurine sediments, *Journal of Microbiological Methods* **19**(1): 59 – 66.

Prieur, L. and 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, *Limnology and Oceanography* **26**(4): 671–689.

Rakotomamonjy, A. and Canu, S. (2005). Frames, reproducing kernels, regularization and learning, *J. Mach. Learn. Res.* **6**: 1485–1515.

Rast, W. and Lee, G. F. (1978). *Summary Analysis Of The North American (US Portion) OCED Eutrophication Project: Nutrient Loading - Lake Response Relationships And Trophic State Indices*, U.S. Environmental Protection Agency. EPA;600/3-78-008.

Riddick, C. A. L., Hunter, P. D., Tyler, A. N., Martinez-Vicente, V., Horváth, H., Kovács, A. W., Vörös, L., Preston, T. and Présing, M. (2015). Spatial variability of absorption coefficients over a biogeochemical gradient in a large and optically complex shallow lake, *Journal of Geophysical Research: Oceans* **120**(10): 7040–7066.

Ronald, J., Zaneveld, V., Boss, E. and Hwang, P. A. (2001). The influence of coherent waves on the remotely sensed reflectance, *Opt. Express* **9**(6): 260–266.

Ruddick, K. G., Gons, H. J., Rijkeboer, M. and Tilstone, G. (2001). Optical remote sensing of chlorophyll a in case 2 waters by use of an adaptivetwo-band algorithm with optimal error properties, *Appl. Opt.* **40**(21): 3575–3585.

Rybicki, G. B. (1996). Radiative transfer, *Journal of Astrophysics and Astronomy* **17**(3): 95–112.

Santer, R., Carrere, V., Dubuisson, P. and Roger, J. C. (1999). Atmospheric correction over land for meris, *International Journal of Remote Sensing* **20**(9): 1819–1840.

Sathyendranath, S., Prieur, L. and Morel, A. (1989). A three-component model of ocean colour and its application to remote sensing of phytoplankton pigments in coastal waters, *International Journal of Remote Sensing* **10**(8): 1373–1394.

Saythendranath, S. (1986). Remote sensing of phytoplankton: A review, with special reference to picoplankton, **214**: 561–583.

Schaeffer, B. A., Schaeffer, K. G., Keith, D., Lunetta, R. S., Conmy, R. and Gould, R. W. (2013). Barriers to adopting satellite remote sensing for water quality management, *International Journal of Remote Sensing* **34**(21): 7534–7544.

Schalles, J. F. (2006). *Optical remote sensing techniques to estimate phytoplankton chlorophyll a concentrations in coastal waters with varying suspended matter and cdom concentrations*, Vol. 9 of *Remote Sensing and Digital Image Processing*, Springer International Publishing, pp. 27–79.

Schapire, R. E. and Freund, Y. (2012). *Boosting: Foundations and Algorithms*, The MIT Press.

Shang, S., Lee, Z., Lin, G., Hu, C., Shi, L., Zhang, Y., Li, X., Wu, J. and Yan, J. (2017). Sensing an intense phytoplankton bloom in the western taiwan strait from radiometric measurements on a uav, *Remote Sensing of Environment* **198**: 85 – 94.

Shi, K., Li, Y., Li, L., Lu, H., Song, K., Liu, Z., Xu, Y. and Li, Z. (2013). Remote chlorophyll-a estimates for inland waters based on a cluster-based classification, *Science of The Total Environment* **444**: 1 – 15.

Simis, S. G. H., Ylöstalo, P., Kallio, K. Y., Spilling, K. and Kutser, T. (2017). Contrasting seasonality in optical-biogeochemical properties of the baltic sea, *PLOS ONE* **12**: 1–31.

Simis, S. G., Ruiz-Verdu, A., Dominguez-Gomez, J. A., Pena-Martinez, R., Peters, S. W. and Gons, H. J. (2007). Influence of phytoplankton pigment composition on remote sensing of cyanobacterial biomass, *Remote Sensing of Environment* **106**(4): 414 – 427.

Spyrakos, E., O Donnell, R., Hunter, P. D., Miller, C., Scott, M., Simis, S. G. H., Neil, C., Barbosa, C. C. F., Binding, C. E., Bradt, S., Bresciani, M., Dall Olmo, G., Giardino, C., Gitelson, A. A., Kutser, T., Li, L., Matsushita, B., Martinez-Vicente, V., Matthews, M. W., Ogashawara, I., Ruiz-Verdú, A., Schalles, J. F., Tebbs, E., Zhang, Y. and Tyler, A. N. (2017). Optical types of inland and coastal waters, *Limnology and Oceanography* pp. 1–25.

Stumpf, R. P. and Tyler, M. A. (1988). Satellite detection of bloom and pigment distributions in estuaries, *Remote Sensing of Environment* **24**(3): 385 – 404.

Sun, D., Li, Y. and Wang, Q. (2009). A unified model for remotely estimating chlorophyll a in lake taihu, china, based on svm and in situ hyperspectral data, *IEEE Transactions on Geoscience and Remote Sensing* **47**(8): 2957–2965.

Tewari, A. and Bartlett, P. L. (2014). Learning theory, *in* P. S. Diniz, J. A. Suykens, R. Chellappa and S. Theodoridis (eds), *Academic Press Library in Signal Processing: Volume 1*, Vol. 1 of *Academic Press Library in Signal Processing*, Elsevier, pp. 775 – 816.

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

Toole, D. A., Siegel, D. A., Menzies, D. W., Neumann, M. J. and Smith, R. C. (2000). Remote-sensing reflectance determinations in the coastal ocean environment: impact of instrumental characteristics and environmental variability, *Appl. Opt.* **39**(3): 456–469.

Tyler, A. N., Hunter, P. D., Spyrakos, E., Groom, S., Constantinescu, A. M. and Kitchen, J. (2016). Developments in earth observation for the assessment and monitoring of inland, transitional, coastal and shelf-sea waters, *Science of The Total Environment* **572**: 1307 – 1321.

USEPA (2002). *Federal Water Pollution Control Act (CWA)*, U.S. Environmental Protection Agency. Washington, D.C.

USEPA (2009). *National Lakes Assessment: A Collaborative Survey of the Nation's Lakes*, U.S. Environmental Protection Agency. EPA 841-R-09-001. Washington, D.C.

USEPA (2012). *2012 National Lakes Assessment, Field Operations Manual*, U.S. Environmental Protection Agency. EPA 841-B-11-003. Washington, D.C.

Vapnik, V. N. (1999). An overview of statistical learning theory, *IEEE Transactions on Neural Networks* **10**(5): 988–999.

Vilas, L. G., Spyrakos, E. and Palenzuela, J. M. T. (2011). Neural network estimation of chlorophyll a from meris full resolution data for the coastal waters of galician rias (nw spain), *Remote Sensing of Environment* **115**(2): 524 – 535.

von Luxburg, U. and Schölkopf, B. (2011). *Statistical Learning Theory: Models, Concepts, and Results*, Vol. 10, Elsevier North Holland, Amsterdam, Netherlands, pp. 651–706.

Wang, L., Zhou, Y., Zhou, W. and Wang, S. (2005). Estimation of trophic state of inland lake using remote sensing data, *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS '05.*, Vol. 8, pp. 5691–5694.

Wang, Q., Sun, D., Li, Y., Le, C. and Huang, C. (2010). Mechanisms of remote-sensing reflectance variability and its relation to bio-optical processes in a highly turbid eutrophic lake: Lake taihu (china), *IEEE Transactions on Geoscience and Remote Sensing* **48**(1): 575–584.

Wei, J., Lee, Z. and Shang, S. (2016). A system to measure the data quality of spectral remote-sensing reflectance of aquatic environments, *Journal of Geophysical Research: Oceans* **121**(11): 8189–8207.

Wilkinson, G. M. (2017). Eutrophication of freshwater and coastal ecosystems, *in* M. A. Abraham (ed.), *Encyclopedia of Sustainable Technologies*, Elsevier, Oxford, pp. 145 – 152.

Witt, H. (1998). *Die spektralen und räumlichen Eigenschaften von Fernerkundungssensoren bei der Ableitung von Landoberflächenparametern*, PhD thesis, Freie Universität Berlin - Institut für Weltraumsensorik.

Xi, H., Hieronymi, M., Krasemann, H. and Röttgers, R. (2017). Phytoplankton group identification using simulated and in situ hyperspectral remote sensing reflectance, *Frontiers in Marine Science* **4**: 272.

Xi, H., Hieronymi, M., Röttgers, R., Krasemann, H. and 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.

Xu, J.-P., Li, F., Zhang, B., Gu, X.-F. and Yu, T. (2010). Remote chlorophyll-a retrieval in case-ii waters using an improved model and irs-p6 satellite data, *International Journal of Remote Sensing* **31**(17-18): 4609–4623.

Yacobi, Y. Z., Moses, W. J., Kaganovsky, S., Sulimani, B., Leavitt, B. C. and 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.

Yang, W., Matsushita, B., Chen, J. and Fukushima, T. (2011). Estimating constituent concentrations in case ii waters from meris satellite data by semi-analytical model optimizing and look-up tables, *Remote Sensing of Environment* **115**(5): 1247 – 1259.

Yang, X., Wu, X., Hao, H.-L. and He, Z.-L. (2008). Mechanisms and assessment of water eutrophication, *Journal of Zhejiang University SCIENCE B* **9**(3): 197–209.

Yousef, F., Kerfoot, W. C., Shuchman, R. and Fahnenstiel, G. (2014). Bio-optical properties and primary production of lake michigan: Insights from 13-years of seawifs imagery, *Journal of Great Lakes Research* **40**(2): 317 – 324.

Zhai, P.-W., Hu, Y., Trepte, C. R. and Lucker, P. L. (2009). A vector radiative transfer model for coupled atmosphere and ocean systems based on successive order of scattering method, *Opt. Express* **17**(4): 2057–2079.

Zheng, G. and DiGiacomo, P. M. (2017). Remote sensing of chlorophyll-a in coastal waters based on the light absorption coefficient of phytoplankton, *Remote Sensing of Environment* **201**: 331 – 341.

Zibordi, G. and Bulgarelli, B. (2007). Effects of cosine error in irradiance measurements from field ocean color radiometers, *Appl. Opt.* **46**(22): 5529–5538.