UiT
THE ARCTIC UNIVERSITY OF NORWAY

Faculty of Science and Technology

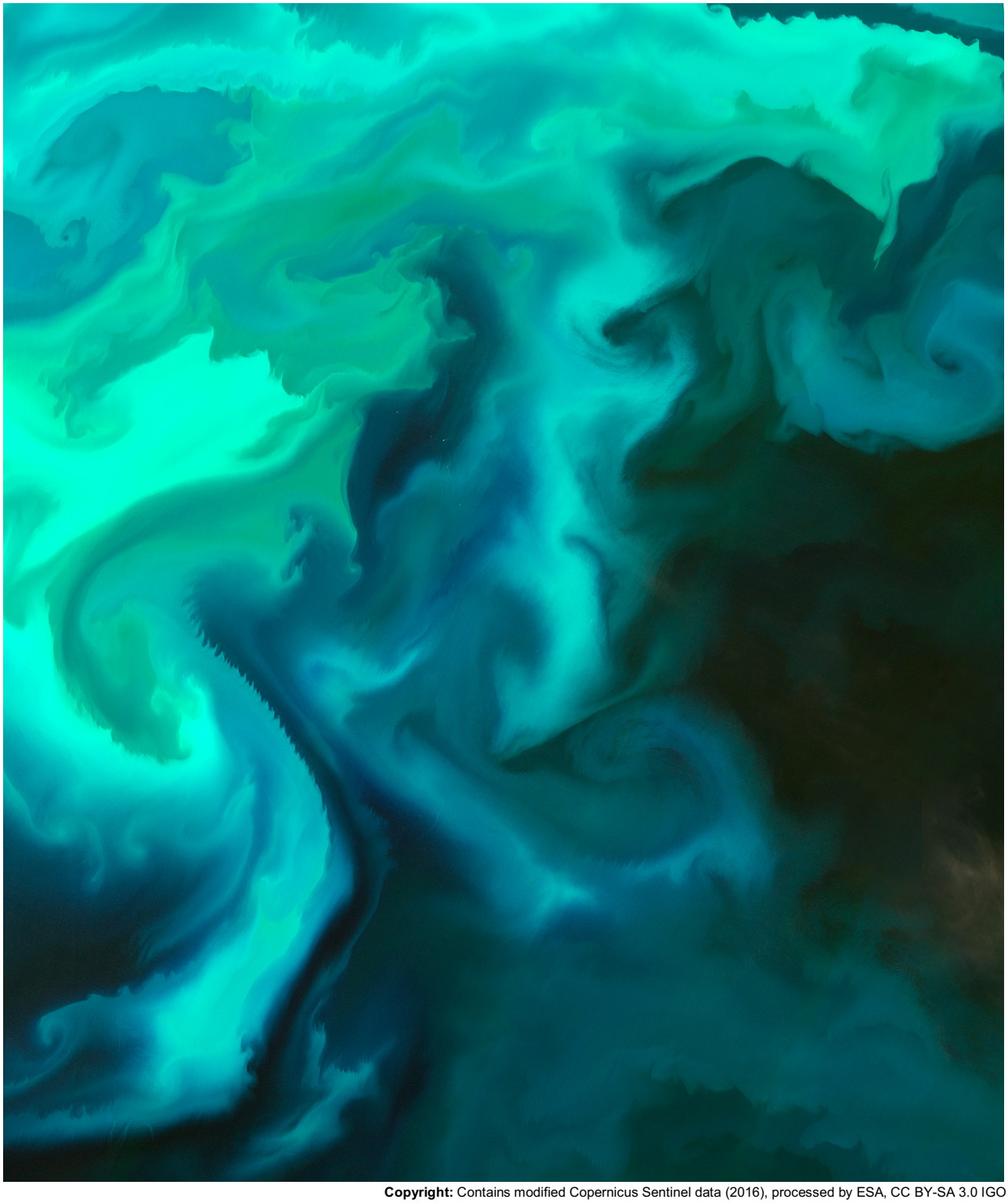
Department of Physics and Technology

Machine Learning Water Quality Monitoring

—

Katalin Blix

A dissertation for the degree of Philosophiae Doctor – April 2019



UNIVERSITY OF TROMSØ

To my grandmother, mother, sister and Kincső Katalin

"Ha én zászló volnék, sohasem lobognék,
Mindenféle szélnek haragosa volnék,
Akkor lennék boldog, ha kifeszítenének,
S nem lennék játéka mindenféle szélnek."
Koncz Zsuzsa

"Mondottam ember: küzdj es bízva bízzál."
Madách Imre

"All we have to decide is what to do with the time given to us."
J. R. R. Tolkien

Abstract

This work utilizes Machine Learning (ML) regression and feature ranking techniques for water quality monitoring from remotely sensed data. The investigated regression methods include the Gaussian Process Regression (GPR), Suport Vector Regression (SVR) and Partial Least Squares Regression (PLSR). Feature relevance in the GPR model is assessed by the probabilistic Sensitivity Analysis (SA) approach.This thesis introduces the SA of the predictive mean and variance functions of the GPR, which reveals the relevance of the input features and the spectral spacing of the input space, respectively. The approach was applied to both controlled and Chlorophyll-a (Chl-a)/ Remote sensing reflectance (Rrs) matchup datasets with promising results.

The SA of the predictive mean function of the GPR was compared and evaluated with the Automatic Relevance Determination (ARD) and Variable Importance in Projection (VIP) feature ranking methods. The ARD is associated with GPR model, and the VIP is used to assign relevance to the input features in the PLSR model. The comparison results showed that feature ranking methods can not only be used to reduce dimension, while still obtaining satisfactory regression, but also to reveal the underlying biophysical properties of aquatic environments.

Feature ranking methods and ML regression models were combined to design an Automatic Model Selection Approach (AMSA). AMSA automatically compares and validates regression models by evaluating the number and combination of ranked input features. The output of AMSA is a regression model and the number and position of features used for obtaining the strongest model based on user defined statistical measures. AMSA was tested on several Chl-a/ Rrs matchups representing various water conditions.

Finally, AMSA was applied to an aquatic environment showing a large variety of water conditions. The chosen test site was Lake Balaton, due to its unique optical properties. Lake Balaton represents eutrophic, oligotrophic, turbid and clear, open ocean like conditions. Thus, being able to retrieve water quality by using a unified model established by AMSA, for all these different water conditions, might allow a more extensive use of the model.

Acknowledgments

I would like to express my deepest gratitude to the UiT the Arctic University of Norway for employing me, whenever I was seeking a job. Thank you for educating me and giving me the possibility to travel and obtain knowledge. In particular, I would like acknowledge the Department of Physics and Technology for allowing me to submit my work for evaluation before due time. I truly appreciate that you finally allow me to move on.

Foremost, I would like to express my sincerest gratitude to my supervisor Torbjørn Eltoft. Thank you for giving me the unconditional freedom in research. Thank you for giving me back the faith in research and science. Thank you for all the hard work and countless hours. Thank you for being available 24/7. Thank you for standing always right behind me along this rocky way. Thank you for not trying to hold me back, and letting me be who I am.

I would like to acknowledge my co-supervisor Nima Pahlevan for all the useful discussions and conversations. Thank you for sharing your knowledge with me, and providing me the simulated data.

Viktor R. Tóth and Károly Pálffy, thank you for receiving me in your laboratory. I will always treasure the time I spent in Tihany.

I would like to express my appreciations to Gustau Camps-Valls and Robert Jenssen for the fruitful discussions in the first part of my work. It was a pleasure working with you. Thank you Gustau for providing me the matchups.

Thank you Marit Krogstad Hall and Geir Antonsen for your support and guidance during the years I spent by the UiT the Arctic University of Norway.

I would like to thank my family for their endless love and support. Thank you Bogi for having faith in me, being my best friend, consultant and sister. Thank you my mother for your inexhaustible energy, and never hearing about doubts and challenges. I will hold on to this attitude. Kincső Katalin, you are my inspiration and motivation. You have been listening to my talks by paying attention, and showing curiosity and interest since you were 2 years old. That is my greatest achievement. Finally, I would like to thank my genius grandmother, who not only introduced me to science and taught me to speak multiple languages in a very early age, but also believed in me. You have always known that research is my destiny. I will try to remember this.

Preface

Why did I choose to work with Gaussian Processes (GPs), when the trend in Machine Learning is artificial Neural Networks (aNNs) in a deep context?

Deep aNNs started to become very popular and used by several big companies only in the 2010's. Although aNNs have been around for decades, they had not shown a significant impact. Recent advances in the development of aNNs has now led to the desired breakthrough. However, these aNNs are often referred to as *black box*, since the internal architecture of the network usually stays hidden. This often causes concerns about the future of the development in artificial intelligence and machine learning.

I am certain that kernel machines, for instance GPs, will have their comeback, just as it was in the case of aNNs. Then, having an approach already available that reveals the driving mechanism of these kernel methods, is highly advantageous. The *black box* syndrome will be avoided. This would mean, that we not only can benefit from the use of machine learning methods, but also have full control over the internal information extraction mechanisms involved.

Contents

List of Tables

List of Figures

List of Abbreviations

AMSA Automatic Model Selection Approach

ARD Automatic Relevance Determination

CDOM Colored Dissolved Organic Matter

Chl-a Chlorophyll-a

CO_2 Carbon Dioxide

GPR Gaussian Process Regression

GPs Gaussian Processes

MERIS MEdium Resolution Imaging Spectroradiometer

MIZ Marginal Ice Zone

ML Machine Learning

MODIS Moderate Resolution Imaging Spectroradiometer

NIR Near Infra Red

NNs Neural Networks

NRMSE Normalized Root Mean Squared Errors

OC Ocean Color

OLCI Ocean and Land Color Instrument

PLSR Partial Least Square Regression

Rrs Remote sensing reflectance

R^2 Pearson correlation coefficient

S3 Sentinel 3

SA Sensitivity Analysis

SE Squared Exponential

SeaBASS SeaWiFS Bio-optical Archive and Storage System

SeaWiFS Sea-Viewing Wide Field-of-View Sensor

SVM Support Vector Machine

SVR Support Vector Regression

TSM Total Suspended Matter

VIP Variable Importance in Projection

VIS VISible

Chapter 1

Introduction

1.1 Motivation

The general advances in data technology and the society's ever-increasing demand for information have led to an enormous increase in the amount of data that is continuously being collected. This Big Data revolution, together with the rapid advancements in computer science technologies, have challenged the traditional way data has previously been processed for retrieving information, and resulted in the development of a manifold of Machine Learning (ML) algorithms. By today, there exists a vast number of these ML methods, many of which are targeted towards applications in regression and classification problems.

This thesis is focusing on the ML Gaussian Process Regression method. The ML Gaussian Process Regression (GPR) has been experiencing tremendous success in the past decade [1–3]. ML GPR has shown to have outstanding regression power, it is stable, robust, fast and has the property of also providing the variance of the estimated output. Most importantly in the context of this thesis, ML GPR has been successfully applied to biophysical parameter estimation from remotely sensed data [4,5].

ML algorithms, including ML GPR, which is a non-linear kernel method, are often referred to as *black boxes*. The *black box* here means that despite the successful learning, the driving mechanism of the method is not well, or not at all understood.

The two main goal of this work was:

1: To reveal the driving mechanism of the ML GPR, and

2: To use the developed method for water quality monitoring from remotely sensed data.

The reason that this particular application was chosen is that there is general consensus in the society that water quality monitoring needs to have prioritized attention. The Earth's water reservoirs have been going through rapid and significant deterioration in the last decades due to the continuously increasing anthropogenic impact and climate change. Being able to monitor these ongoing changes on a large scale would

help us to locate vulnerable waters, which would be an important aid in environmental research, to monitor industrial activities, and for policy makers.

The most important water quality parameter is Chlorophyll-a (Chl-a). Chl-a can be found in phytoplankton, which is an aquatic photosynthetic organism. Phytoplankton forms the basis of the aquatic food-web. Without its presence neither marine nor fresh water ecosystems would occur or sustain.

Continuous monitoring of phytoplankton through Chl-a allows us to understand the occurrence and spatial distribution of aquatic ecosystems. This is highly important from an environmental perspective, but has also relevance for industries, for instance aquaculture and fisheries. At the same time other industries, such as the offshore oil and gas industry, shipping and tourism can take the location of highly vulnerable areas into consideration, when planning their operations.

Phytoplankton takes up Carbon-dioxide (CO_2) during photosynthesis in order to live and grow [6]. Part of this CO_2 sinks to the bottom of the oceans and will be buried in the sediments. Hence, phytoplankton is also referred to as a CO_2 pump, since it removes CO_2 from the atmosphere. The continuous monitoring of its presence and amount is an important contribution in climate studies [7–9].

The amount of in water Chl-a is also used for determining eutrophication. This is frequently observed in inland and coastal waters [10, 11]. Remote sensing to monitor Chl-a is an efficient tool to detect the worsening of water quality.

Chl-a monitoring from space is done by optical imaging sensors onboard satellites. These sensors measure the spectral radiance on several wavelength in the visible (VIS) and near infrared (NIR) part of the electromagnetic spectrum, and by incorporating atmospheric correction procedures, the water leaving radiance is retrieved.

This signal carries the signature of the water bodies. Although the number, position and width of the spectral bands differ by sensors, there are certain wavelengths measured by all instruments, namely the bands that characterize the absorption spectral curve of the Chl-a [12]. This is situated in the blue (first absorption peak) and green (little or no absorption) part of the VIS.

It is in common practice to relate the measured *so called* Remotesensing Relfectance (Rrs) on these spectral bands to the amount of in-water Chl-a, so that a statistical functional relationship can be established [13,14]. Then, this relationship is used to estimate Chl-a from the remotely sensed data. This widely used and state-of-the-art approach is often referred to as the Ocean Color (OCx) algorithm [15], where x = 2, 3 or 4, and refers to the number of bands used in the OCx model.

Although these parametric bio-optical OCx models are simple, and have been shown to be reliable approaches in phytoplankton dominated open oceans, they have certain disadvantageous properties. They are based on the assumption that there is an explicit relationship between the predefined spectral bands of the sensor and Chl-a content, and model coefficients need to be adjusted by extending the training data. Good performance of these models is limited to waters, where there are no or little influence of other water constituents [12]. Hence, they are not recommended to be used for complex water monitoring, such as coastal and inland waters [16]. Furthermore, since aquatic envir-

onments are experiencing changes, OCx algorithms often result in erroneous Chl-a retrieval, when the waters to be monitored are in transition to conditions with increasing complexity.

To overcome these difficulties, ML approaches have been introduced for water quality monitoring. Many ML methods have been investigated with promising results. Some prominent examples are, support vector regression (SVR) [17–19], relevance vector machine [20] and Neural Networks (NNs) [21], the latter have even become the state-of-the-art approach for estimating water quality parameters, including Chl-a, in complex waters from data acquired by the Ocean and Land Color Instrument (OLCI) onboard Sentinel-3A and B (S3) satellites launched in 2016 and 2018, respectively [22,23]. This clearly shows that ML algorithms have become of great importance in the monitoring of water quality, especially in areas, where the traditional approaches fail.

Although NNs have been successfully utilized to monitor complex waters, the validation of these complex water products has revealed erroneous retrievals [24,25]. In [26], it was found that NNs could not estimate Chl-a content correctly in an aquatic environment with large variation of water complexity. In this case, the analysis indicated that this was due to the fact that the estimated Chl-a amount was sensitive to suspended sediments in the water body.

Furthermore, it is often challenging to classify the type of the water in advance, due to changes and/ or lack of information about the given aquatic environment. Thus, having one unified algorithm, which could retrieve water quality from remotely sensed data under a large variety of water conditions, would be highly desired.

In this work, these aforementioned issues were addressed by using the ML GPR model to retrieve information about water quality. The objectives of this thesis are as follows.

Objectives

- To introduce an approach which reveals the driving mechanism of the GPR model.
- To create a model selection tool that combines information retrieval with machine leaning regression methods, including the GPR and the associated feature ranking methods, to output the most suitable model for the given data
- To use the tool to establish a unified model to retrieve information about water quality from remotely sensed data in both complex and clear waters

To achieve these objectives, firstly the Sensitivity Analysis (SA) of the GPR for both the predictive mean and variance functions were introduced. The approach is based on approximating the expected value of the squared partial derivatives of the GPR mean and variance functions with respect to the given dimension. The SA of the GPR mean function outputs the relative relevance of the input features, and the SA of the GPR variance function reveals the spectral spacing of the input space. Note, that the SA of the GPR variance is independent of the observed output, hence it can be used without

having the ground truth available. The SA was evaluated and tested on both simulated controlled data and Chl-a/Rrs matchups.

To visualize the practical application of the approach, sensitivity maps were presented for Chesapeake Bay, which is known to have highly complex water. The sensitivity maps could reveal how the most important spectral bands change with varying water conditions. The SA of the GPR mean function assigned highest relevance to the red bands in complex waters. By using the sensitivity maps and revealing areas, where red bands were given highest importance, we were able to detect areas of complex waters. This is considered to be a helpful tool in the understanding of the type of the water body and if the water is in transition.

The SA of the GPR variance function provides information about the spectral spacing of the given band. This means, if the measured reflectance in the given band show similarities, the sensitivity will be low, and vise verse. This can be an important additional information.

In the next analysis step of the thesis, the goal was to compare and evaluate some selected feature ranking and regression methods, including the SA and GPR. The outcome of this study was that feature ranking could not only improve Chl-a retrieval, and at the same time reduce the number of input features, but it also reflected that the method could provide insight into the underlying biophysical properties.

This motivated the author to automatize the methodology, and to create an Automatic Model Selection Approach (AMSA), which was designed to output the most suitable regression model to predict water quality from a given library of regression models, with associated feature ranking methods. AMSA uses a training data set for the area of interest, to automatically return the most suitable regression model, together with the associated most relevant features, and the numerical value of the performance measures. AMSA was tested on synthetic and real data, representative for global and complex waters. The experiments demonstrated that the approach worked well for the test cases, which suggests that AMSA should be implemented and applied in practice.

Having the AMSA tool available, the final objective of the thesis was to create a unified regression model for highly varying water quality conditions. The chosen test site was Lake Balaton in Hungary, which has a great variety of water conditions. The optical properties of Lake Balaton represents several trophic states, such as eutrophic, mesotrophic and oligotrophic, and turbid and clear waters. The collected in situ water quality data from the lake provided a unique possibility for using AMSA to develop and evaluate a unified regression model. The model was developed for Sentinel 3 OLCI sensor, which has quite advantageous spatial and spectral properties. AMSA resulted in a successful model that seemed to be able to differentiate between Chl-a and Total Suspended Matter (TSM), in contrast to the state-of-the-art NNs. We refer to this unified model as Balaton model. It was tested on a S3 OLCI image, acquired when the lake was in its most complex state with high turbidity, and the Chl-a map produced by the Balaton model showed good correspondence with dynamic processes and limnological properties of the lake. This model is described in [26].

Currently the Balaton model is under testing in Arctic coastal and open waters, and

for the Marginal Ice Zone. Preliminary results suggest, that the unified model can estimate Chl-a content in both complex and open Arctic waters. Hence, the Balaton model may be a very useful tool in future studies of Arctic marine ecosystems.

1.2 Thesis outline

The rest of this work is organized as it follows. Chapter 2 gives on overview about the principles of water quality monitoring. Chapter 3 presents the datasets used in this thesis and explains how the Balaton data was obtained and processed during the Balaton project. Chapter 4 discusses the ML methods used in this work, with focus to the SA, GPR and AMSA. Chapter 5 gives on overview of the publications included in this thesis, and lists other contributions, which are not discussed in this work. Chapters 6, 7, 8 and 9 present the four peer-reviewed published papers, and Chapter 10 concludes this thesis and outlines future research directions.

Chapter 2

Ocean color monitoring

2.1 Principles

Ocean color monitoring uses passive remote sensing techniques to retrieve information about water bodies. Optical imaging sensors onboard satellites measure the radiometric flux at the sensor on predefined wavelengths in the VIS and NIR part of the electromagnetic spectrum. The source of illumination is the Sun itself. However, the Sun-rays follow various paths before they reach the sensor. Figure 2.1 shows the simplified composition of the total measured radiance at sensor L_T, which can be written by

$$L_T = L_p + L_s + L_b + L_w, \tag{2.1}$$

where L_p is the path radiance, which is the contribution of the atmosphere to the propagating electromagnetic radiation. L_s and L_b are the reflected radiance by the water surface and bottom, respectively [27]. L_w is the water-leaving radiance, which interacts with the water-constituents, and this is the signal that ocean color monitoring aims to measure. L_w can be mathematically expressed by rearranging Eq. (2.1), which yields $L_w = L_T - L_p - L_s - L_b$. L_w is retrieved by using radiometric processing [27]. [1]

The light (L_w) that penetrates into the water bodies, interacts with the water-constituents and reaches the sensor can be seen in Fig. 2.2. The most important and common water-constituents are Chl-a, which occurs in phytoplankton, Colored Dissolved Organic Matter (CDOM) and Total Suspended Matter (TSM). Chl-a and CDOM absorbs photons from the incoming solar radiation with certain frequency, whereas TSM scatters the penetrating light. Figure 2.2 illustrates the different processes. Hence L_w contains the biophysical signature of the water bodies.

[1] Note, this research was not focusing on radiometric correction algorithms. The data was already processed and has gone through atmospheric correction.

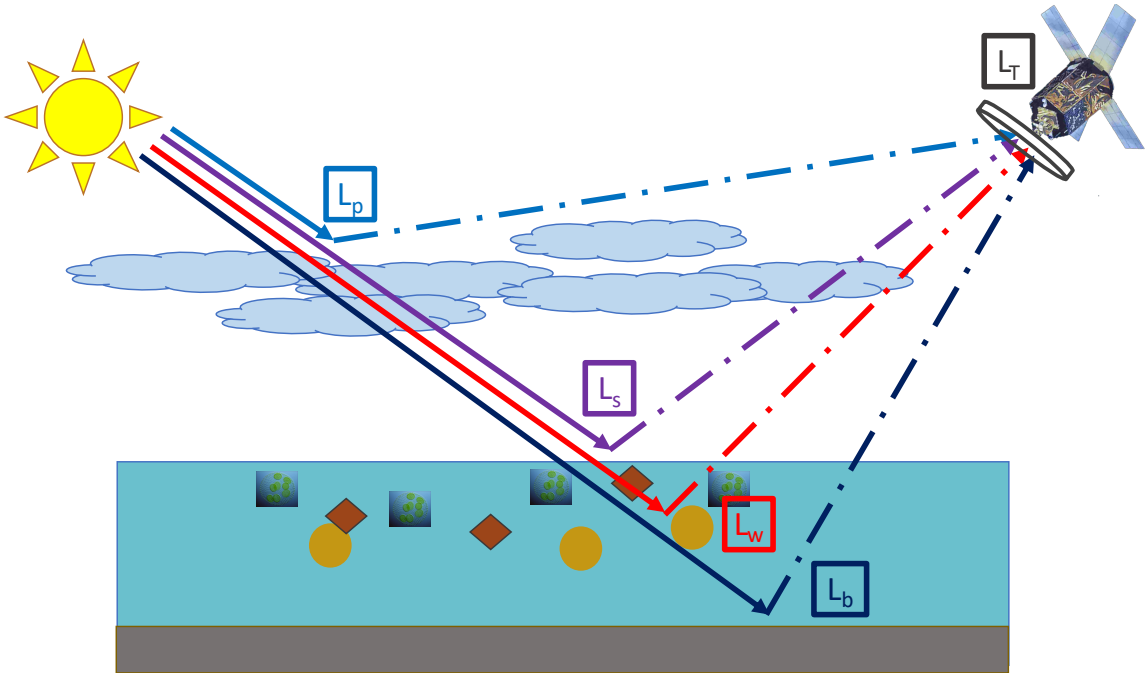


Figure 2.1: The components of the measured signal at sensor.

2.2 Water-constituents

There is a great variety of water constituents. In this work, the focus was on Chl-a, CDOM and TSM, which are commonly used to describe water quality.

Chl-a has a characteristic absorption spectrum, with its peaks positioned at wavelengths around 443 nm and 675 nm. However, these peaks can be shifted and broadened due to the various processes, which might occur in the phytoplankton communities [12].

CDOM is the composition of humic and fluvic acids, originating from decaying marine and terrestrial matter [12]. CDOM absorbs in the blue part of the visible spectrum, and tends to mask the first absorption peak of the Chl-a.

Figure 2.3 shows an example of the absorption spectrum of different amounts of Chl-a concentration in the presence of CDOM [28]. (Figure 2.3 is from [28].) It can be seen how the shapes and positions of the peaks are displaced.

TSM includes re-suspended bottom sediments, river-borne particles and even atmospheric particulates. The type, size and amount of TSM shows great variations resulting in difficulties to establish a characteristic absorption/ scattering spectrum.

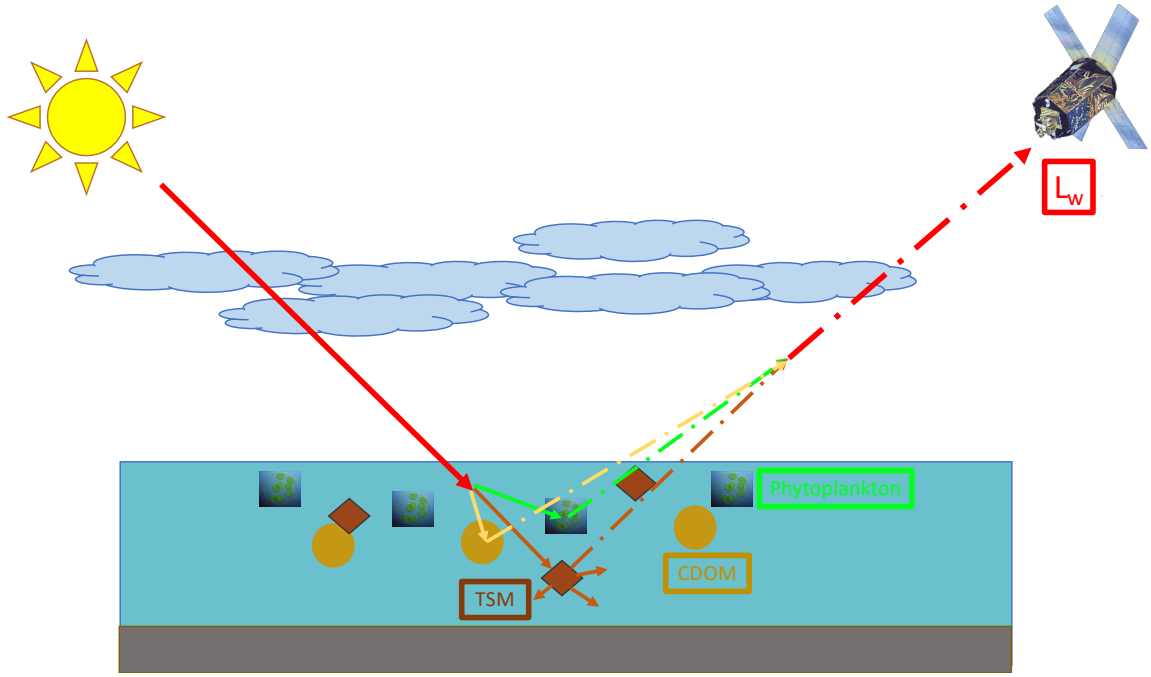


Figure 2.2: The components of L_w.

2.3 Water types

There are many different kinds of water bodies. However, it is common to classify water types based on the occurrence, amount, type and distribution of water-constituents, which again determine the composition of the received signal, hence the color of the water. (Note, there are other factors, which can also influence the water color, for instance bottom reflectance, which is common is shallow transparent waters.)

Water color shows great variations. It has been common practice to classify water bodies into two types: Case 1 and Case 2 waters [29]. Case 1 waters are dominated by phytoplankton and products associated with these primary producers. Case 2 waters are optically complex waters, consisting of additional water-constituents.

Case 1 conditions are usually representative for open oceans, whereas Case 2 conditions often are assumed to be coastal waters. In this work, Case 1 and Case 2 waters refer to open and complex waters, respectively. Under complex waters, coastal and Arctic waters, and shallow inland lakes are assumed.

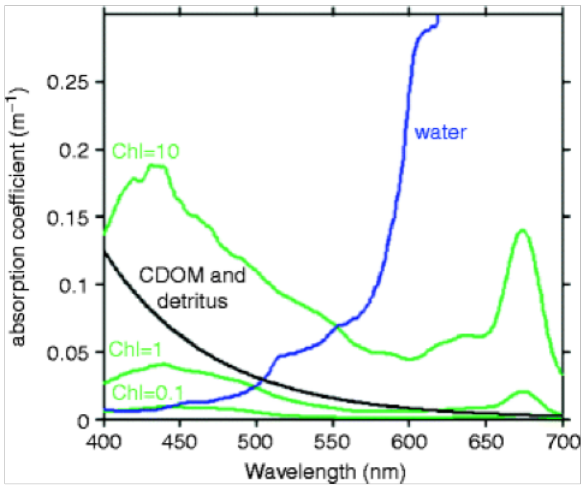


Figure 2.3: An example of absorption spectra for various amounts of Chl-a and CDOM. Figure is from H. M. Dierssen and K. Randolph, 2013.

Chapter 3

Description of the data

The datasets consist of in situ Chl-a, CDOM and TSM observations, and corresponding satellite measurements, Rrs, which are referred to as outputs $\{y_n\}_{n=1}^N$ and inputs $\{\mathbf{x}_n \in \mathbb{R}^D\}_{n=1}^N$, respectively, where N is the total number of samples.

The in situ Chl-a samples listed in Table 3.1 are surface oceanic water measurements taken from the upper water layer, corresponding to the photic zone. The Chl-a, CDOM and TSM measurements in Table 3.2 are integrated water column values from Lake Balaton.

The Rrs originates from various operational and non-operational sensors, with different spectral and spatial resolutions. It is Level-2 data, hence it has already gone through sensor calibration and atmospheric correction.

Both real and synthetic datasets were used. The term *"synthetic resampled"* in Table 3.1 refers to the synthetized hyper-spectral IOCCG dataset [30], which were resampled to match the spectral resolutions of the sensors of interest.

The following sensors were used in this work: Sea-Viewing Wide Field-of-View Sensor (SeaWiFS) on GeoEye's OrbView-2 satellite, Moderate Resolution Imaging Spectroradiometer (MODIS) onboard Aqua, MEdium Resolution Imaging Spectroradiometer (MERIS) on Envisat, and the Ocean and Land Color Instrument (OLCI) on Sentinel-3A.

The summary of the sensors and datasets are listed in Table 3.1 and Table 3.2. Two additional HidroLight simulated datasets for MERIS and OLCI were also used, and these are referred to as MERIS synthetic and OLCI synthetic.

The datasets include a large variety of aquatic environments representing both open and complex waters.

The SeaWiFS, MODIS-Aqua and MERIS datasets can be freely downloaded and obtained from NASA's SeaWiFS Bio-optical Archive and Storage System (SeaBASS).

Data collection at Lake Balaton

Lake Balaton provides a unique environment to train and evaluate water quality parameter retrieval models for waters including a wide range of optical properties. Figure 3.1 ([26]) shows the color transitions along the South West (SW) - North East (NE) axis.

Table 3.1: Summary of the datasets.

SeaBAM	
Bands (λ_c (nm))	412 443 490 510 555
Band width	20
Spatial resolution	1100 m
Chl-a range (mgm^{-3})	0.019 - 32.787
Nr. of samples	919
SeaWiFS	
Bands (λ_c (nm))	412 443 490 510 555 670
Band width	20
Spatial resolution	1100 m
Chl-a range (mgm^{-3})	0.024 - 129.332
Nr. of samples	1465
MODIS-Aqua	
Bands (λ_c (nm))	412 443 488 531 551 667 678
Band width	10 nm, 15 nm
Spatial resolution	1000 m
Chl-a range (mgm^{-3})	0.0153-25.4985
Nr. of samples	579
Synthetic resampled MODIS-Aqua	
Chl-a range (mgm^{-3})	0.03 - 30
a_{CDOM} (m^{-1})	0.0025 - 2.3677
Nr. of samples	478
MERIS	
Bands (λ_c (nm))	413 443 490 510 560 620 665 681
Band width	10 nm and 7.5 nm
Spatial resolution	300 m
Chl-a range (mgm^{-3})	0.017 - 40.23
Nr. of samples	557
MERIS synthetic	
Chl-a range (mgm^{-3})	0.021 - 53.4429
Nr. of samples	5000
Synthetic resampled MERIS	
Chl-a range (mgm^{-3})	0.03 - 30
Nr. of samples	478

The main tributary is the Zala river, entering the lake at the SW part of the lake (station 1 in Fig. 3.1). This is an eutrophic area, which has usually high CDOM and Chl-a con-

Table 3.2: Summary of the Balaton data.

OLCI	
Bands (λ_c (nm))	412.5 442.5 490 510 560 620 665 673.25 681.25
Band width	15 nm, 10 nm and 7.5 nm
Spatial resolution	300 m
Chl-a range (mgm^{-3})	2 - 55
CDOM range (g Ptm^{-3})	2 - 124
TSM range (gm^{-3})	2 - 60
Nr. of samples	36
OLCI synthetic	
Chl-a range (mgm^{-3})	2 - 55
Nr. of samples	624

centrations. The trophic gradient decreases along the SW - NE axis, and at the NE part the lake shows oligotrophic conditions (station 5 in Fig. 3.1).

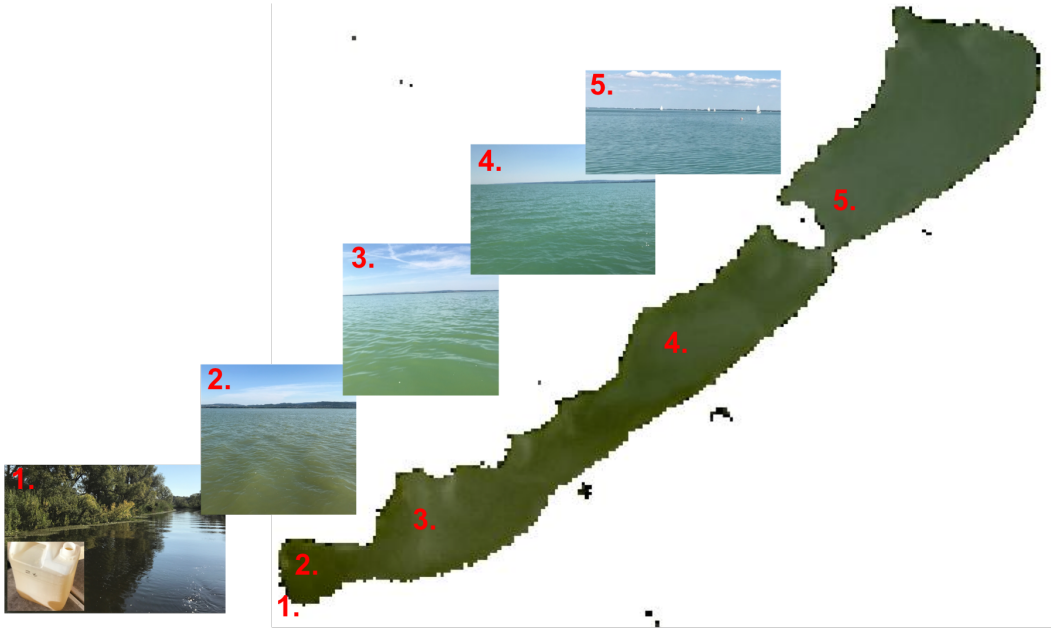


Figure 3.1: Illustrating the unique optical properties of Lake Balaton.

The Hungarian Academy of Sciences (HAS), Center for Ecological Research (CER), Balaton Limnological Institute (BLI) conducts regular data collections. To illustrate the in-situ data collection, Figure 3.2 shows the study site, boat, field work, water samples and the team. The author was a visiting fellow at the institute for one year, and parti-

cipated in the water sample collection and processing.

There are a series of measurements taken at each station (Fig. 3.1), and these are used to retrieve water quality parameters. The three parameters of interest were Chl-a, CDOM and TSM. Chl-a is retrieved by filtering a known volume of three replicates of water samples through a Whatman filter, then spectrophotometrically measuring it after hot methanol extraction [31]. The unit of Chl-a is mg m^{-3}. CDOM concentration is retrieved from water samples of known volume, which are filtered through a 0.45 μm pore size cellulose acetate filter, buffered with borate buffer and measured against a blank of buffered Milli-Q water at 440 nm and 750 nm using a Shimadzu UV 160A spectrophotometer. Then CDOM concentration is measured in platina (Pt) units, which are calculated from the absorbance values [32]. The Pt units of CDOM is mg Pt L^{-1}. Finally, TSM is determined gravimetrically after sample filtration through a 0.4 μm pore size cellulose acetate filter.

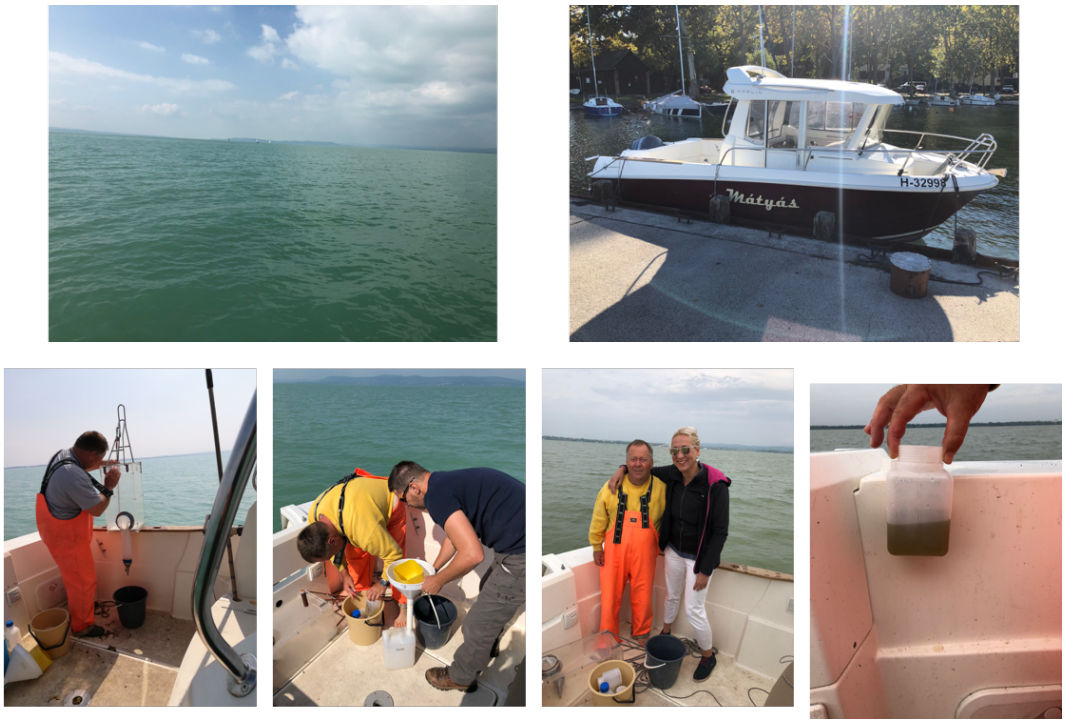


Figure 3.2: Data collection at Lake Balaton.

These measurements of water quality parameters were used to produce matchups for S3 OLCI, where the standard practice of extracting level 2 Rrs measured at bands in the VIS spectral range were followed. The matchups were used for validation of the level 2 water quality products. The matchup data was subsequently merged with the synthetic OLCI data (Table 3.2), and used for establishing the Balaton model by AMSA.

Chapter 4

Machine Learning algorithms for water quality parameter retrieval from remotely sensed data

4.1 Machine Learning for regression

ML regression methods are based on learning the relationship between the input and output training data, and then using this for predicting unseen outputs from new observed inputs. Figure 4.1 illustrates the learning. The example shows an input data matrix $\mathbf{X}$ (stars), consisting of three observations of two dimensions, and the corresponding output vector $\mathbf{y}$ (solid circles) holding three elements. The input training data matrix is denoted $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3]$, consisting of three two dimensional input feature vectors, and the corresponding output vector is $\mathbf{y} = [y_1\ y_2\ y_3]$.

ML regression learns the relationship between $\mathbf{X}$ and $\mathbf{y}$. This is used for prediction of outputs for new input data.

In this work, the training data will consist of input Rrs measured on the spectral bands of the given sensor in VIS, and in some cases, additional features. These additional features are band ratios, used in the parametric band ratio models. The corresponding output is the water quality parameters, which can be either Chl-a or CDOM or TSM. The training input and output data pairs are denoted by X and y, respectively, and they may be written as a matrix (upper case bold), a vector (lower case bold) or a scalar (plain text). The test input and output are symbolized with a star symbol.

Figure 4.2 illustrates the approach for water quality remote sensing. The training data is illustrated with the crosses, and the input is observed on three dimensions (bands), while the output here is Chl-a. The predicted values are the pixels outside the crosses in the output image.

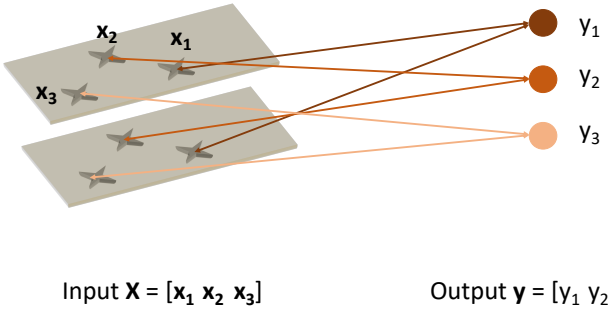


Figure 4.1: Illustrating the learning of ML regression.

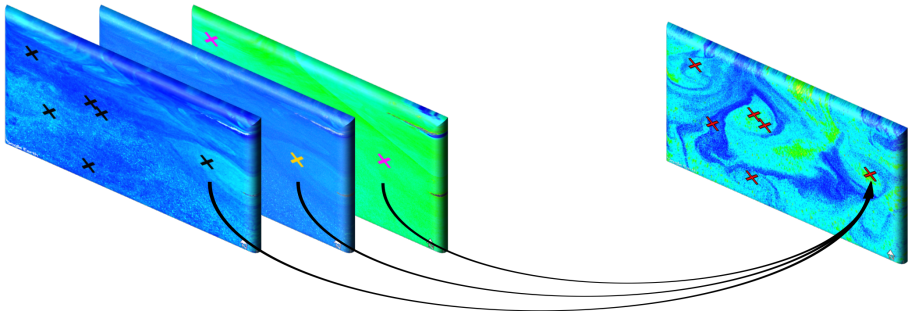


Figure 4.2: Illustrating the ML regression approach for water quality remote sensing.

4.2 Gaussian Process Regression

Let us define the observed training data by $\mathcal{D} \equiv \{\mathbf{x}_n, y_n | n = 1, \ldots N\}$, where $\mathbf{x}_n$ is the input d-dimensional feature vector, y_n is the corresponding output point, and $n = 1, \ldots N$ is the number of observations. We assume that the output is a function of the inputs and a Gaussian noise ε, which can be written by $y_n = f(\mathbf{x}_n) + \varepsilon_n$, where $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$. The Gaussian Process ($\mathcal{GP}$) uses Bayesian inversion [33, 34] to estimate the output. This is done by placing a zero mean $\mathcal{GP}$ prior on the latent function $f(\mathbf{x})$ and a Gaussian prior over the noise ε, i.e. $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}'))$, where $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ is a kernel function used for computing the elements of the covariance matrix. The symbols $\boldsymbol{\theta}$ and σ^2 are the hyper-parameters of the kernel function $k_{\boldsymbol{\theta}}$ and the distribution of the noise ε, respectively. Observations drawn from the $\mathcal{GP}$ function at $\{\mathbf{x}_n\}_{n=1}^N$ locations will be jointly multivariate Gaussian distributed with zero mean and covariance matrix $\mathbf{K}_{\mathbf{ff}}$, where the elements of the covariance matrix are computed by the kernel function $k_{\boldsymbol{\theta}}$, and are expressed by $[\mathbf{K}_{\mathbf{ff}}]_{pq} = k_{\boldsymbol{\theta}}(\mathbf{x}_p, \mathbf{x}_q)$. Then for a new input $\mathbf{x}_*$, the posterior

distribution of the corresponding output value y_* is computed analytically by

$$\begin{aligned} p(y_*|\mathbf{x}_*, \mathcal{D}) &= \mathcal{N}(y_*|\mu_{\text{GP}*}, \sigma^2_{\text{GP}*}) \\ \mu_{\text{GP}*} &= \mathbf{k}_{\mathbf{f}*}^\top (\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} = \mathbf{k}_{\mathbf{f}*}^\top \boldsymbol{\alpha} \\ \sigma^2_{\text{GP}*} &= \sigma^2 + k_{**} - \mathbf{k}_{\mathbf{f}*}^\top (\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{k}_{\mathbf{f}*} \\ &= \sigma^2 + k_{**} - \mathbf{k}_{\mathbf{f}*}^\top \mathbf{A} \mathbf{k}_{\mathbf{f}*}, \end{aligned}$$

where $\mu_{\text{GP}*}$ and $\sigma^2_{\text{GP}*}$ are the predictive mean and variance functions, respectively. $\mathbf{k}_{\mathbf{f}*}$ is the covariance between the training vector and the test point, $\boldsymbol{\alpha} = (\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y}$ is the weight vector of the $\mathcal{GP}$ mean, k_{**} is the covariance between the test point with itself, and $\mathbf{A} = (\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_n)^{-1}$ is the weight matrix of the $\mathcal{GP}$ variance.

This means that the approach has an analytic closed form solution, which makes it trackable, and it automatically outputs the variance, allowing to assess the certainty level of the estimates. These are advantageous properties, and usually not easily accessible in other machine learning algorithms.

There is a great selection for kernel functions. In this work, the Squared Exponential (SE) kernel function was used, which can be expressed by

$$k(\mathbf{x}_p, \mathbf{x}_q) = \nu^2 \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \left(\frac{x_p^d - x_q^d}{\lambda_d} \right)^2 \right), \tag{4.1}$$

where λ_d is the length-scale for feature d and ν is a positive scaling factor.

The SE kernel function has several advantageous properties. It is exponential, hence infinitely differentiable, which is an important property in the sensitivity analysis of the $\mathcal{GP}$. Furthermore, the inverse of the optimized length-scale hyperparameter(s) in Eq. (4.1) can provide feature relevance.

The optimization of the hyper-parameters, ν, λ_d and σ^2 is achieved by maximizing the negative log-marginal likelihood function with respect to the hyper-parameters. Note, the optimization may be trapped in local maxima, which might lead to in-correct ranking of the spectral bands [5].

4.2.1 Other Machine Learning regression methods

Although this thesis focuses on the GPR model, two other regression methods are briefly described here. These are the Support Vector Regression (SVR) and Partial Least Square Regression (PLSR) models. The reason that these methods are included is that, beside their different kind of advantageous properties, feature relevance can be assessed in both of them.

The SVR has been successfully applied for ocean color applications [17–19]. Since the kernel SVR is also a non-linear kernel method, the sensitivity analysis could be extended to the SVR.

The PLSR has also been applied for water quality parameter retrieval from remotely sensed data [35]. Feature relevance in the PLSR can be assessed through the Variable

Importance in Projection (VIP). PLSR can handle multiple outputs, reduce noise and co-linearity in the data. It can handle high dimensional data, where the number of dimensions exceeds the number of observations. This can occur in hyper spectral water quality matchups, due to the challenges of obtaining the data. For future work, it has been planned to work with hyper-spectral data, where the number of observations might be low in comparison to the number of input features. Therefore, the PLSR would be a potential candidate to be used.

Support Vector Regression

The SVR model assumes that the output can be computed by $y_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b$, where $\mathbf{w}^{\mathrm{T}}$ is the transposed weight vector and b is the bias term [36–39].

The SVR model uses the so-called ϵ-intensitive loss function to obtain estimates by penalizing errors exceeding an ϵ limit and at the same time obtaining a regression function as flat as possible. The weights are estimated by minimizing $J = \frac{1}{\beta}\sum_{n=1}^{N}(\zeta_n^+ + \zeta_n^-) + \frac{1}{2}||\mathbf{w}||^2$, also called the objective function, with respect to $\mathbf{w}$, ζ_n^+, ζ_n^-, and constrained to

$$y_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n - b \leq \epsilon + \zeta_n^+ \qquad \text{for} \qquad n = 1, ..., N \tag{4.2}$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b - y_n \leq \epsilon + \zeta_n^- \qquad \text{for} \qquad n = 1, ..., N \tag{4.3}$$

$$\zeta_n^+, \zeta_n^- \geq 0 \qquad \text{for} \qquad n = 1, ..., N. \tag{4.4}$$

ζ_n^+ and ζ_n^- are called slack variables, and allow measurements to be larger than ϵ, and $\beta > 0$ is a constant controlling the trade-off between the flatness of the regression function and the magnitude of the deviations from ϵ.

The optimal solution for the weights are obtained by constructing a Lagrange function from the objective function. This can be written by $\hat{\mathbf{w}} = \sum_{n=1}^{N}(\alpha_n^+ - \alpha_n^-)\,\mathbf{x}_n$, where α_n^+ and α_n^- are the Lagrange multipliers, also called support vectors. Defining $a_n = \alpha_n^+ - \alpha_n^-$, and collecting the estimated output values $\hat{y}_n$ into a vector $\hat{\mathbf{y}}$, the estimated output can be written by

$$\hat{\mathbf{y}} = \hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x} + \hat{\mathbf{b}} = \sum_{n=1}^{N} a_n \mathbf{x}_n^{\mathrm{T}}\mathbf{x} + \hat{\mathbf{b}}. \tag{4.5}$$

Applying the SE kernel function (Eq. (4.1)) to $\mathbf{x}_n^{\mathrm{T}}\mathbf{x}$ results in the expression for the estimated output:

$$\hat{\mathbf{y}} = \sum_{n=1}^{N} a_n k(\mathbf{x}_n, \mathbf{x}) + \hat{\mathbf{b}}. \tag{4.6}$$

Partial Least Square Regression

The training data holding the input and output observations is $D \equiv \{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{X}$ is an $N \times D$ input data-matrix consisting of $d = 1, ..., D$ features and $n = 1, ..., N$

observations, and the output $\mathbf{y}$ is the corresponding $N \times 1$ output-vector consisting of $n = 1, ..., N$ observations.

The PLSR model relates the input $\mathbf{X}$ and the output $\mathbf{y}$ through a latent-space [40,41] by introducing latent variables $\mathbf{T}$ ($N \times H$), which are representing both $\mathbf{X}$ and $\mathbf{y}$ in the latent-space, so that the covariance between the projection of $\mathbf{X}$ and $\mathbf{y}$ in this latent-space is maximized. The PLSR model can be written by

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \\ \mathbf{y} &= \mathbf{T}\mathbf{c} + \mathbf{f} \\ \mathbf{T} &= \mathbf{X}\mathbf{W}^{\star} \\ \mathbf{W}^{\star} &= \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}, \end{aligned} \tag{4.7}$$

where $\mathbf{P}$ ($D \times H$) is a matrix of the X-loadings and $\mathbf{c}$ ($H \times 1$) is the y-loadings. They are good representations of $\mathbf{X}$ and $\mathbf{y}$ in the latent space, respectively. The term $\mathbf{W}^{\star}$ ($D \times H$) holds the weights of $\mathbf{X}$, and defines the common latent-space. The error terms, $\mathbf{E}$ ($N \times D$) and $\mathbf{f}$ ($N \times 1$), are assumed to be iid. $\sim N(0, \sigma^2)$. The estimated output $\mathbf{y}$ can be written by

$$\mathbf{y} = \mathbf{X}\mathbf{W}^{\star}\mathbf{c} + \mathbf{f} = \mathbf{X}\mathbf{b} + \mathbf{f}, \tag{4.8}$$

where $\mathbf{b} = \mathbf{W}^{\star}\mathbf{c}$ and $\mathbf{W}$ ($D \times H$) is the weight matrix consisting of the eigenvectors of the variance-covariance matrix $\mathbf{X^T Y Y^T X}$. Minimizing the error term $\mathbf{f}$ in the PLSR model results the most optimal regression. Details on the PLSR model and algorithms can be found in [42–47].

4.3 Feature ranking for information retrieval

Feature ranking methods can be used for information retrieval, namely to understand the contribution of the input features to the output. In this work, a feature ranking method for the GPR model was introduced. This was the Sensitivity Analysis (SA), which was further extended to the SVR model. The method can be generalized to kernel methods satisfying certain criteria. The generalization of the SA is out of the scope of this thesis. Here, the application of the methodology in water quality remote sensing was the focus. Two additional feature ranking methods are included, the ARD and the VIP, which are associated with the GPR and PLSR, respectively.

4.3.1 SA of Kernel Machines: SA GPR and SA SVR

The SA feature ranking method for the SVR and GPR models are based on the same concept. Although both the SVR and GPR are non-linear kernel machines, their underlying principles differ. The SA of the GPR model was introduced in [48] and [49], while the SA of the Support Vector Machine (SVM) for classification purposes was described

in [50], and extended to the SVR in [51]. The sensitivity of feature j is defined as

$$s_j = \int \left(\frac{\partial \boldsymbol{\phi}(\mathbf{x})}{\partial x_j} \right)^2 p(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{4.9}$$

where $p(\mathbf{x})$ is the probability density function of the D-dimensional input vector $\mathbf{x} = [x_1, \ldots, x_D]^\top$, and $\boldsymbol{\phi}(\mathbf{x})$ represents either the predictive mean $\mu_{\mathrm{GP}*}$ or variance $\sigma_{\mathrm{GP}\star}$ function of the GPR, or the function used to estimate the output $\hat{y}$ in the SVR. The sensitivity of the feature j can be interpreted as a measure of the average gradient in the given dimension. In practice, the gradient measures changes of the function in direction j. This can take both positive and negative values, which by the integration may cancel out each other. Therefore, the derivatives are squared, which means that the sensitivity can only take positive values. The empirical estimate of the sensitivity for the j^{th} feature is written by

$$s_j = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\partial \boldsymbol{\phi}(\mathbf{x}_n)}{\partial x_n^j} \right)^2, \tag{4.10}$$

where N denotes the number of training samples.

Applying the SA (Eq. (4.10)) to the GPR mean yields:

$$\begin{aligned} s^j_{\mu_{\mathrm{GP}\star}} &= \frac{1}{N} \sum_{q=1}^{N} \left(\frac{\partial \boldsymbol{\phi}(\mathbf{x}_q)}{\partial x_q^j} \right)^2 \\ &= \frac{1}{N} \sum_{q=1}^{N} \left(\frac{\partial \sum_{p=1}^{N} \alpha_p k(\mathbf{x}_p, \mathbf{x}_q)}{\partial x_q^j} \right)^2 \\ &= \frac{1}{N} \sum_{q=1}^{N} \left(\sum_{p=1}^{N} \frac{\alpha_p (x_p^j - x_q^j)}{\lambda_j^2} k(\mathbf{x}_p, \mathbf{x}_q) \right)^2, \end{aligned} \tag{4.11}$$

for the GPR variance is:

$$s^j_{\sigma_{\mathrm{GP}\star}} = -2N\nu^2 \sum_{q=1}^{N} \left(\sum_{p,q=1}^{N} A_{pq} (x_p^j - x_q^j) k(\mathbf{x}_p, \mathbf{x}_q)^2 / \lambda_j^2 \right)^2.$$

and for the SVR model is:

$$s^j_{SVR} = \frac{1}{N} \sum_{q=1}^{N} \left(\sum_{p=1}^{N} \frac{a_p (x_p^j - x_q^j)}{\lambda_j^2} k(\mathbf{x}_p, \mathbf{x}_q) \right)^2. \tag{4.12}$$

Here, the kernel function is the SE kernel (Eq. (4.1)), which is an exponential function, hence it can be infinitely differentiated.

4.3.2 ARD

The SE kernel function (Eq. (4.1)) provides the possibility to assess feature relevance. This can be done through the optimization of the length-scale hyper-parameter λ_d. Then, the inverse of the optimized length-scale hyper-parameter provides the relative relevance of the given input feature. The ARD method is limited to the use of the SE kernel function.

4.3.3 VIP

The VIP feature ranking method is specifically derived for the PLSR model, and it measures the contribution to the total variance of the j^{th} input feature ($j = 1, ..., D$) [52], [53].

The VIP can be expressed in term of Sum-of-Squares [54] by

$$\text{VIP}_j = \sqrt{D \sum_{h=1}^{H} SS_h (w_{hj} / \parallel w_j \parallel^2) / \sum_{h=1}^{H} SS_h}, \tag{4.13}$$

where SS_h is the percentage of the output explained by the hth latent variable and w_j the jth weight of the PLSR model (see Eq. (4.7)).

4.3.4 Illustrating feature ranking methods for water quality remote sensing

This example illustrates how the feature ranking methods assign relevance to spectral bands for various amount of water constituents. The IOCCG dataset [30] was used and resampled to correspond to the spectral bands of OLCI. This dataset was designed to imitate low and increasing water complexity. The chosen threshold for the absorption of CDOM was 0.06 m^{-1} and for the amount of Chl-a 0.7 mg m^{-3}. Observations below these thresholds are assumed to represent open water conditions, and above water conditions with increasing complexity.

Figure 4.3 shows the Rrs spectra for certain Chl-a values for open water conditions, and Fig. 4.4 represents the more complex waters. It can be seen how the Rrs spectra changes for a certain Chl-a value due to the contribution of other water constituents. The number and position of bands along the x-axis correspond to the ten OLCI bands in the VIS.

Then the SA of the GPR, SVR and the VIP feature ranking methods were applied to these datasets. First, the feature ranking methods were used only for the Chl-a values indicated on the y-axis. This can be seen in Fig. 4.5. The color of the images shows the assigned relative importance of the OLCI bands, yellow indicates high importance and blue represents low relevance. For the open water like conditions, all the three feature ranking methods assigned high relevance to the lower bands (Fig. 4.5 left column). They are capturing the Rrs spectra for low Chl-a and CDOM concentrations. This is in

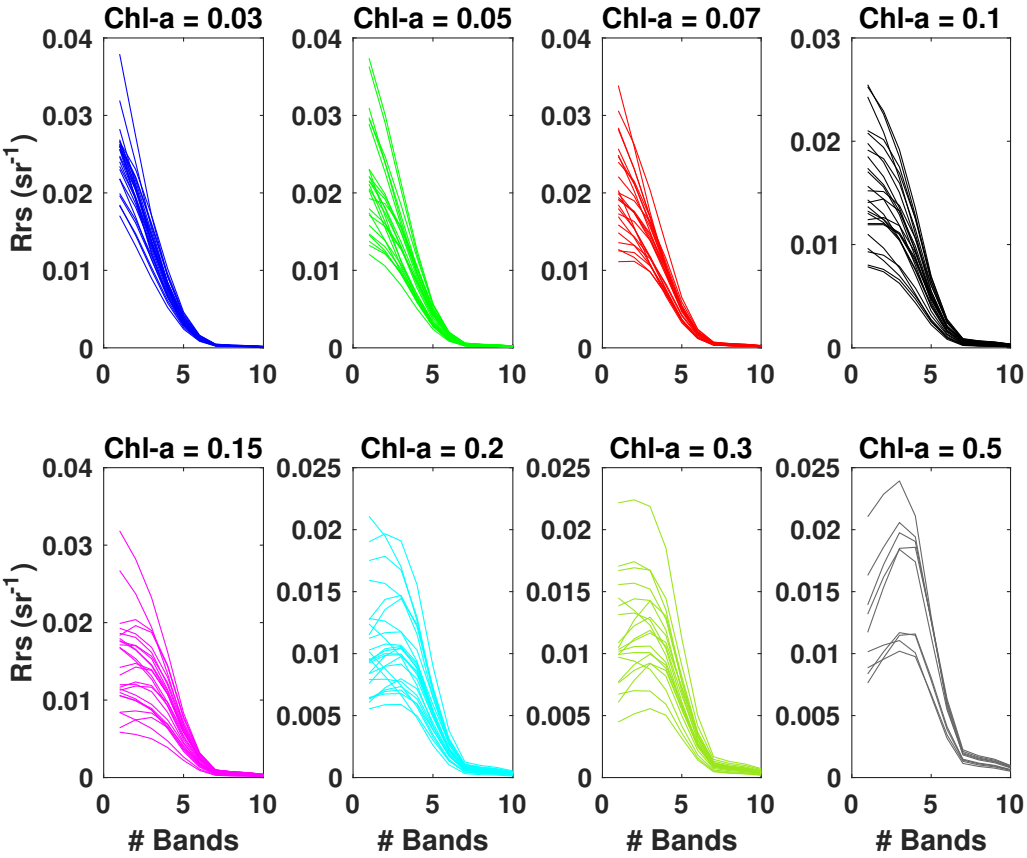


Figure 4.3: Rrs values for low Chl-a content for open water like conditions.

contrast to water conditions of increasing complexity (Fig. 4.5 right column). In this case the importance of the bands is shifted towards longer wavelength, once again mirroring the Rrs spectra. Note, how both the SA GPR and SVR favor the red bands, when Chl-a concentration is the highest, 30 mg m^{-3}.

Figure 4.6 shows the behavior of the feature ranking methods, when continuously adding Chl-a contents. This was done by starting with the lowest Chl-a value, computing the relevance of the band, then adding the next range, applying the feature ranking methods and so forth. For open water conditions (Fig. 4.6 left column), although still the bands corresponding to lower wavelengths were favored, the SA GPR and SVR assigned highest relevance to bands centered 510 and 560 nm, above a certain Chl-a content. It can be seen in Fig. 4.3 that this corresponds to the changes in the Rrs spectra due to the increasing Chl-a content. This shows the underlying principles of the SA, namely that it responds to changes of the function in the input space (the derivatives on the given spectral band). This is also the case for the water conditions with increasing complexity (Fig. 4.6 right column). Both the SA GPR and SVR assign highest relevance to red bands, after a certain range of Chl-a is added. This illustration shows how the

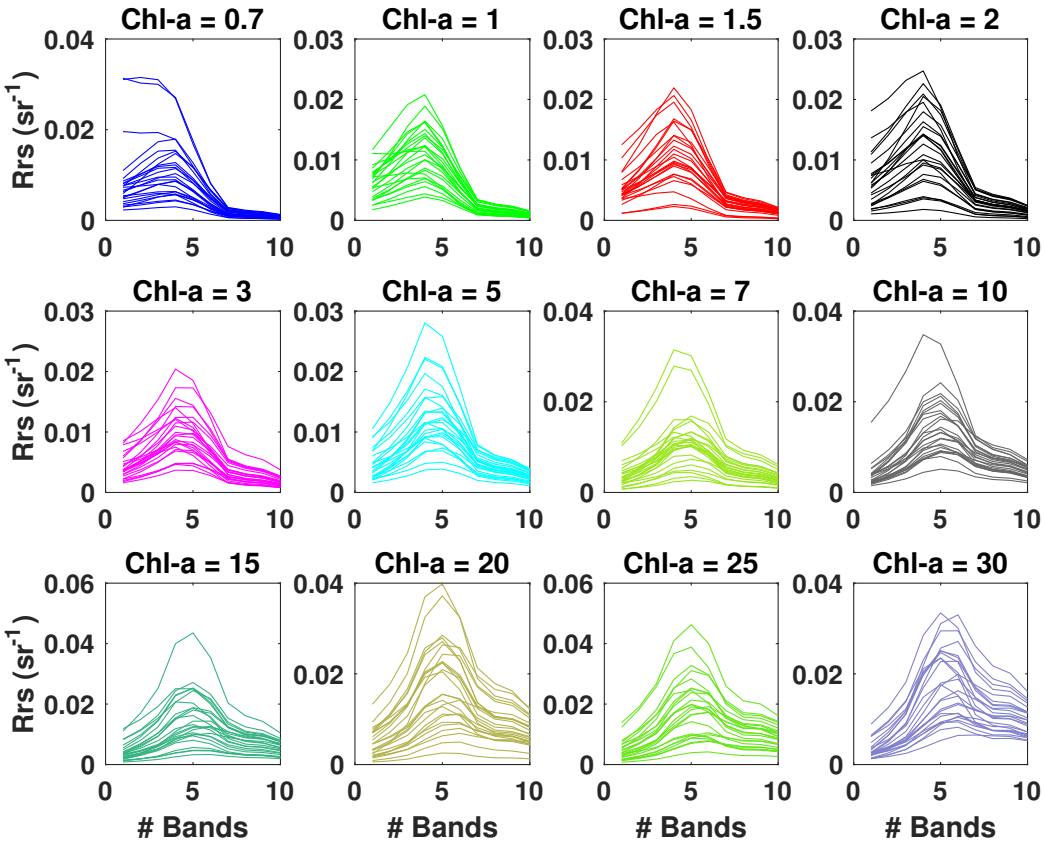


Figure 4.4: Rrs values for higher Chl-a content for water conditions with increasing complexity.

SA can return the variations in the input space by quantifying functional changes in the given dimension.

4.4 Automatic Model Selection Algorithm

The Automatic Model Selection Algorithm (AMSA) combines feature ranking and regression methods to select the most suitable model for a given data. AMSA uses two stages: the first stage is feature ranking and the second is regression. In this work, AMSA was built by using the ML regression models and the associated feature ranking methods discussed in this thesis. AMSA was applied to Rrs/ Chl-a matchups.

Figure 4.7 shows the concept of AMSA. (Figure 4.7 is from [51].) AMSA uses in Stage 1 the Chl-a/Rrs matchup dataset to rank the features by using the SA GPR, SA SVR, ARD and VIP feature ranking methods. Stage 1 results in four sets of ranked features in a decreasing order. In Stage 2, the dataset is split into a training and a test set

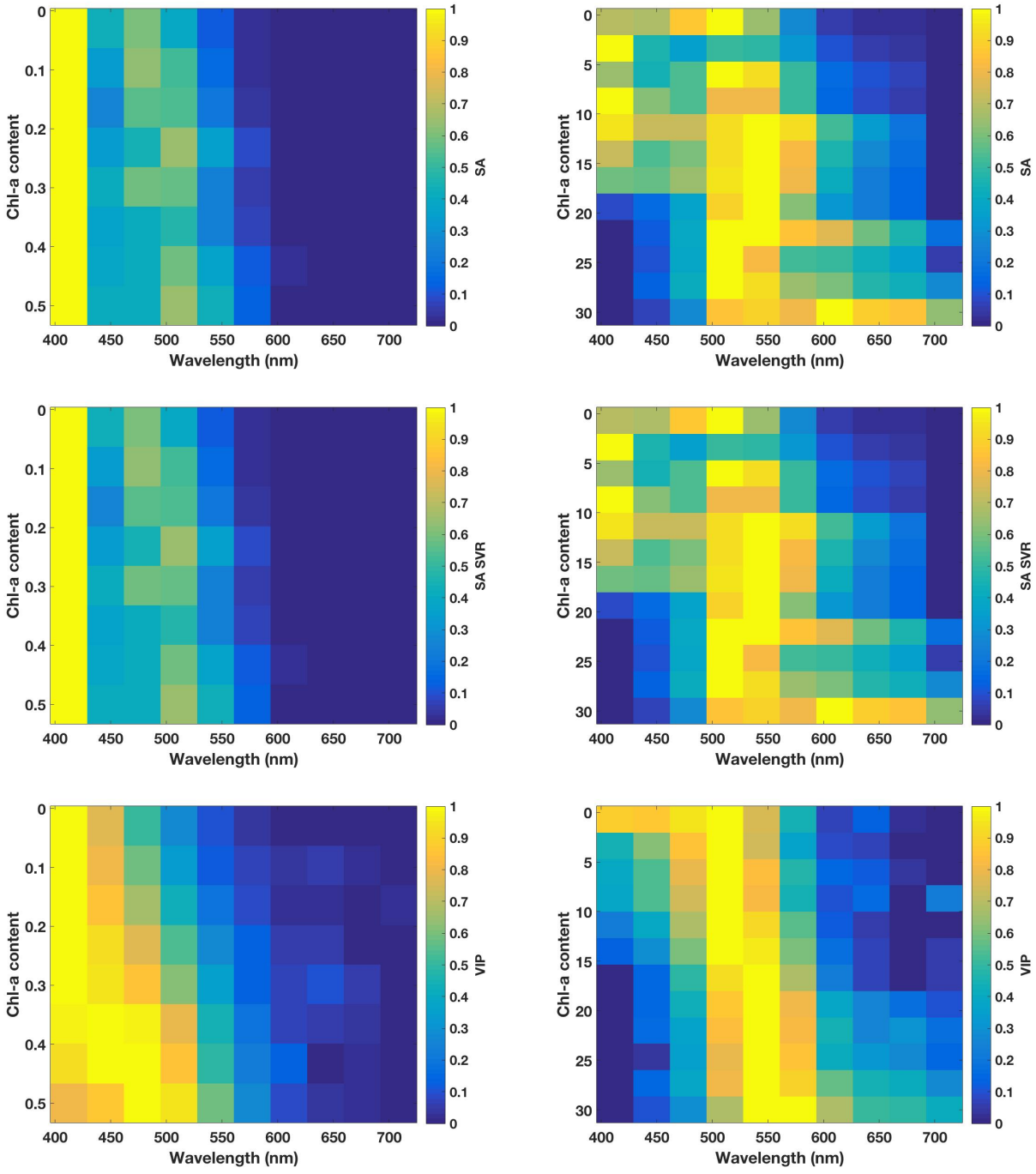


Figure 4.5: SA of the GPR (top row) and SVR (middle row), and the VIP (bottom-row) for open (left column) and complex water (right column) conditions. Feature ranking was computed for a certain Chl-a content value (corresponding to Fig. 4.3 and 4.4).

to perform regression by the GPR, SVR and PLSR models. For evaluating model per-

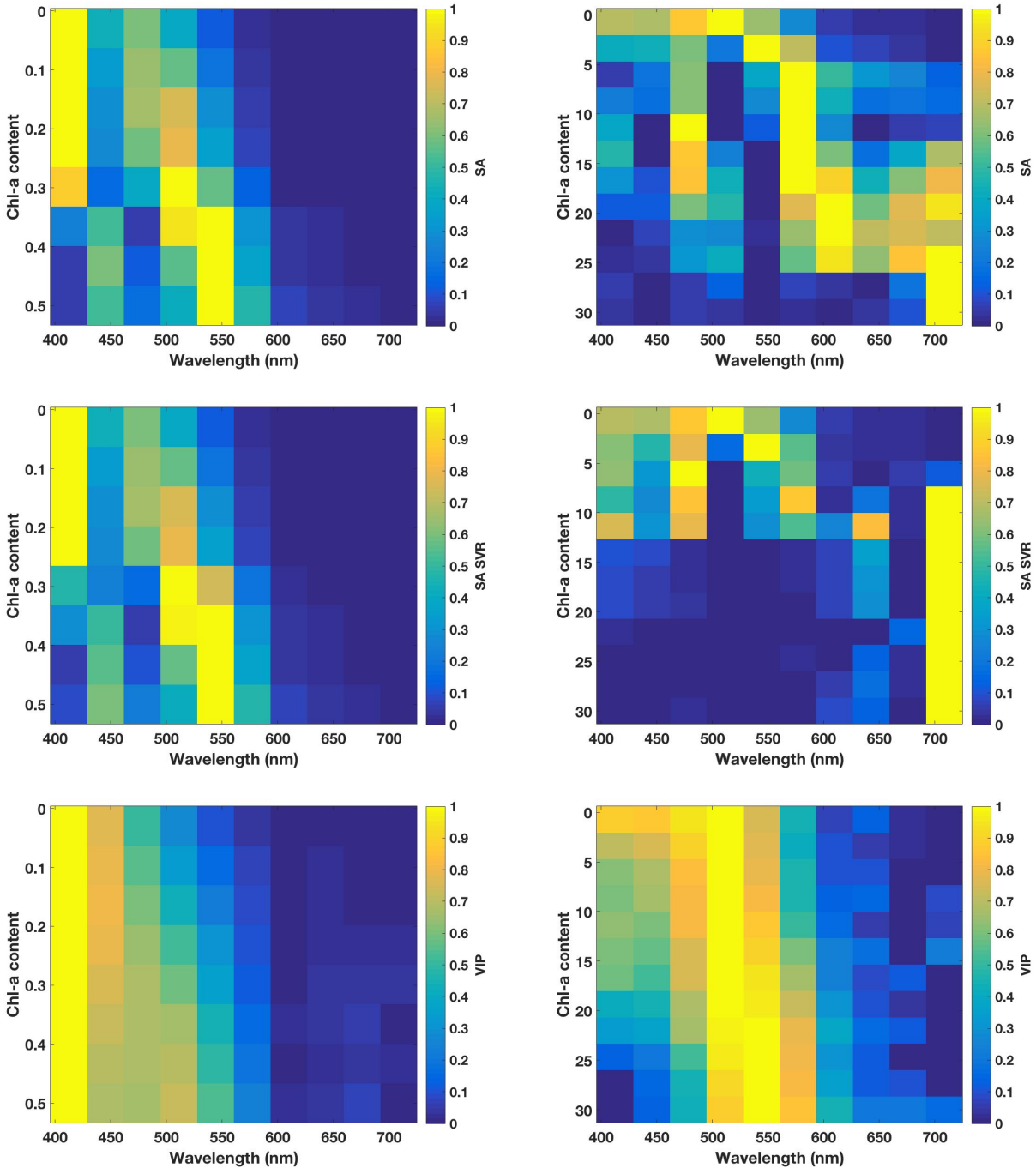


Figure 4.6: SA of the GPR (top row) and SVR (middle row), and the VIP (bottom-row) for open (left column) and complex (right column) water conditions. Feature ranking was computed by continuously adding Chl-a content ranges.

formance, statistical measures are predefined. In this case, the chosen measures are the

Normalized Root Mean Squared Errors (NRMSE) and the Pearson correlation coefficient (R^2). Stage 2 starts with training Regression model 1 by taking the most important feature from ranked feature set 1. Then statistical measures are computed on the test set, and stored. Then it continues by taking the next ranked feature and doing the same procedure. Regression model 1 stops, when no improvements can be detected when adding more features from feature set 1. Then Regression model 1 repeats the same with the all the feature sets. This is done for all the three regression model.

Finally, the model with lowest NRMSE and highest R^2 is returned. This is the most suitable model for the data. AMSA not only provides a model, but also a set of features needed to obtain that particular model. Figure 4.8 shows an illustrative example, how AMSA is used on a real data set. (Figure 4.8 is from [51].)

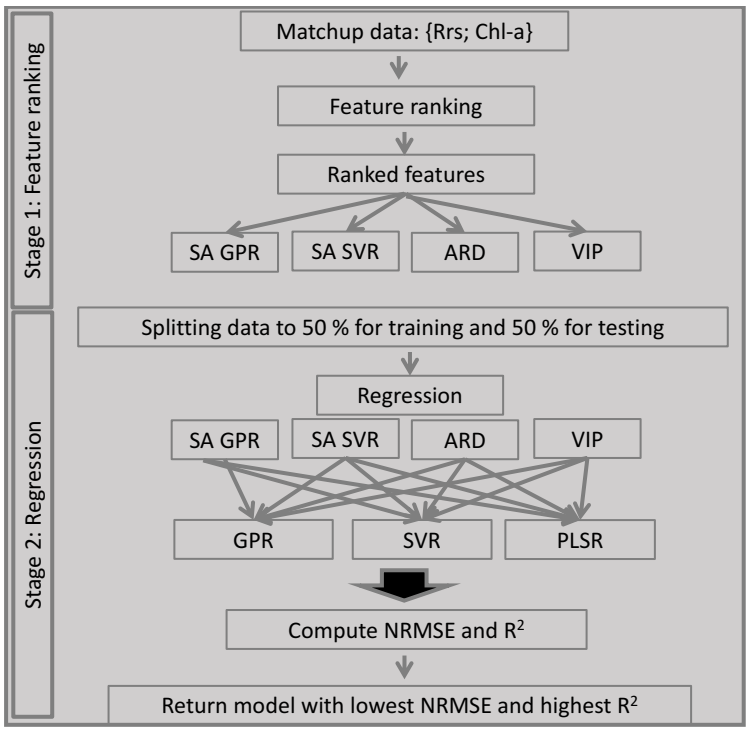


Figure 4.7: The Machine Learning AMSA for oceanic Chl-a content estiamtion.

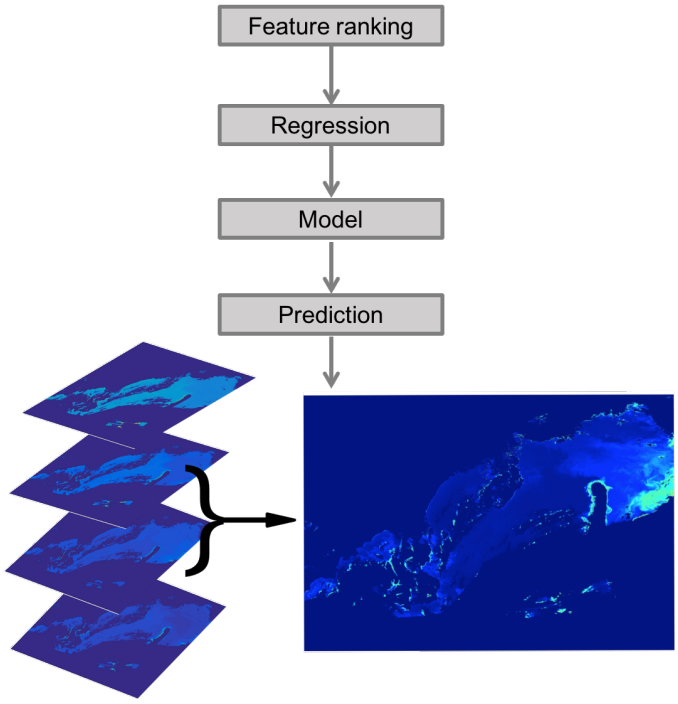


Figure 4.8: Illustration of the AMSA for application.

Chapter 5

Overview of publications

5.1 Short summary of the published papers

5.1.1 Paper 1: Gaussian Process Sensitivity Analysis for Oceanic Chlorophyll Estimation

The GPR is a non-linear kernel regression method, which does not make the relevance of the input features directly accessible.

The objective of Paper 1 was to reveal the driving mechanism of the GPR. This was done by deriving and evaluating the SA of the GPR for the predictive mean and variance functions. The SA is a gradient based method, including a partial derivative of the model's output with respect to the given dimension. The SA of the GPR's mean function outputs the relative relevance of the input features, and the SA of the GPR's variance shows the spacing of the input space.

This work evaluates the approach on controlled toy data and on five Chl-a relevant matchups. A controlled data was generated by creating an output, which is a function of a relevant and an irrelevant input feature. This allows us to evaluate how the SA of the GPR's mean function can capture the relevant input feature. In addition, while generating the data, the spacing of the inputs were controlled. A part of the data was evenly spaced, while the other part was unevenly. In this way, the behavior of the SA of the GPR's variance function was studied. The results of the experiment were very convincing, both the SA of the GPR mean and variance functions performed as expected.

Therefore, the methodology was further evaluated on Chl-a datasets for various sensors, and the GPR model was compared to commonly known parametric models. Finally, sensitivity maps were generated for the Chesapeake Bay to present potential possibilities of the method. These maps reveal how the most important feature changes in different regions of the Chesapeake Bay. In practice, this analysis showed that the SA is a useful tool in the monitoring of changes in the given aquatic environment.

The conclusion in Paper 1 was, that the SA approach is a powerful method, which

can be extended to any differentiable kernel function.

It was shown that the approach can contribute to the understanding of an aquatic environment by creating sensitivity maps. Using feature ranking can reveal which spectral band is the most relevant to estimate Chl-a for the given water type. Then this information can be used in the sensitivity maps to spatially visualize how the most relevant feature changes, when water conditions are changing.

The SA of the GPR variance function can reveal the spectral spacing of the given dimension, which is highly advantageous, especially that the the computation of it does not involve the output. Hence, it does not require available ground truth.

Author's contribution

The idea was developed in collaboration with Robert Jenssen and Gustau Camps-Valls. I performed the analysis and implementations, and wrote the paper.

5.1.2 Paper 2: Evaluation of Feature Ranking and Regression Methods for Oceanic Chlorophyll-a Estimation

The objectives of Paper 2 was to further evaluate feature ranking and selection for several regression methods for the application to oceanic Chl-a content estimation from remotely sensed data. The goal here was not only to compare feature ranking and regression methods, but also to understand the benefits of these analysis in Chl-a estimation from optical imaging data.

The state-of-the art method, when it comes to the determination of feature relevance in the GPR model, is to optimize the length-scales hyper-parameters in the squared exponential kernel function. This method is called ARD. It was included in the comparison.

An additional regression method, PLSR, and its associated feature ranking, namely VIP, was included in the analysis. The PLSR method has several advantageous properties, which are important, especially for high dimensional correlated data. Most importantly, PLSR provides the possibility to rank input features through the VIP method.

The two regression models, GPR and PLSR, and the three feature ranking methods, SA, ARD and VIP, were tested on a toy data and a real Chl-a matchup for the MERIS sensor. The results on the simulated data showed once again that the feature ranking methods can successfully assign relevance to the important features, and the evaluation confirmed excellent regression strength of the GPR.

A sequential evaluation of the ranking algorithms of the regression methods was conducted. Starting with the highest ranked feature, and adding one more at the time in decreasing order of relevance, the regression performance of the regression models were compared by using quantitative performance measures.

This showed that using only two features (spectral bands) as input to the GPR, can already compete with the state-of-the-art model used for Chl-a estimation. More in-

terestingly, these two spectral bands mirror the biophysical properties of ocean. The conclusion in Paper 2 is that feature ranking and selection can not only reduce the number of input features and improve regression, but can also be used to understand the underlying biophysical properties of water bodies.

Author's contribution

The idea of including PLSR and VIP was developed in collaboration of the authors. I developed the approach, performed the analysis and implementations and I wrote the paper.

5.1.3 Paper 3: Machine Learning Automatic Model Selection Algorithm for Oceanic Chlorophyll-a Content Retrieval

Paper 3 introduces the Automatic Model Selection Approach (AMSA) for Chl-a content estimation. This work builds on the first two papers. Here the goal was to build an approach, which combines feature ranking and selection algorithms with regression methods to output the most suitable model for a given data.

AMSA needs an input and an output dataset to determine the most suitable model. Firstly, the whole dataset is used to rank features by various methods. These ranked feature sets are sequentially evaluated for different regression models. At this step only part of the dataset is used for training models, while the other part is used for testing. This means, that AMSA validates itself, while it determines the most suitable model. The returned model includes the type of regression model, the features to be used in the model to obtain the strongest regression, and also the computed statistical measures computed under the validation process.

AMSA was tested on several Chl-a relevant matchups for various sensors on both real and synthetic datasets. Aquatic environments show large variations in their optical properties, which makes monitoring quite challenging. It is often difficult to determine, which model to use for a certain area. The AMSA approach allows for fair model comparison, which can be very useful, when we want to evaluate a candidate model and compare its performance with the state-of-the-art methods.

The conclusion of Paper 3 was that the AMSA approach appears to be a suitable tool for water quality monitoring from remotely sensed data. It is helpful for algorithm development since models can objectively be compared. Finally, AMSA gives an insight about the optical composition of the aquatic environment due to the feature ranking and selection stage of the approach.

Author's contribution

I conceived and developed the idea and the approach. I performed the analysis and implementations and I wrote the paper.

5.1.4 Paper 4: Remote Sensing of Water Quality Parameters over Lake Balaton by Using Sentinel-3 OLCI

Paper 4 exploits the possibilities to use the recently available data acquired by the Ocean and Land Color Instrument (OLCI) onboard the Sentinel - 3 (S3) satellite for monitoring waters with a wide range of optical complexities.

For this purpose, the chosen test site was Lake Balaton, which represents water bodies in different trophic states, turbid and clear waters and also shallow and relative deep waters. Lake Balaton is an excellent environment for product validation and model training.

This work had two objectives: the first was to validate water quality products retrieved by OLCI, and the second was to use AMSA to determine a unified model for Lake Balaton.

The water quality parameters studied here were CDOM, TSM and Chl-a, collected during the year 2017 at regions, which represent characteristic optical properties of the lake. These parameters were compared with the OLCI complex water products, which are estimated by using NNs. The validation results revealed erroneous OLCI estimates. In case of Chl-a, this could be explained by the the sensitivity of the NNs to the TSM.

AMSA was applied to investigate, whether its model selection approach could lead to improvements, and help to understand the optical properties of the lake.

The results showed both significant improvements in the estimation of the Chl-a water quality parameter, and the resulting maps were in good correspondence with the limnological properties of the lake.

The conclusion of the paper was that the model determined for Lake Balaton by using AMSA for S3 OLCI data opens the possibility to design **one** unified algorithm for Chl-a estimation for various complex and open waters. This model can potentially be used globally, and hence represent the fulfillment of a main objective of the thesis.

Author's contribution

I developed the idea and the approach. I performed the analysis and implementations and I wrote the paper.

5.2 List of other publications and contributions

5. K. Blix, G. Camps - Valls and R. Jenssen, "Sensitivity Analysis of Gaussian Processes for Oceanic Chlorophyll Prediction", 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, 2015, pp. 996-999, doi: 10.11009/IGARSS.215.7325936 (oral presentation)

6. K. Blix and T. Eltoft, "Ocean Color Remote Sensing in the Marginal Ice Zone in the Arctic", 2016 Colour and Light in the Ocean from Earth Observation (CLEO), Frascati, 2016 (poster presentation)

7. K. Blix, "Ocean Color Monitoring in the Arctic", 2016 Center for Integrated Remote Sensing and Forecasting for Arctic Operations (CIRFA) annual conference, Sommarøy, 2016 (poster presentation)

8. K. Blix and T. Eltoft, "Monitoring primary productivity through Chlorophyll-a content estimation in the Arctic", 2017 Arctic Frontiers, Tromsø, 2017 (poster presentation, winner of the *Outstanding Poster Award Overall Winner*)

9. K. Blix and T. Eltoft, "An alternative Chl-a content retrieval algorithm for MERIS/ OLCI", 2017 S3 Validation Team Meeting, Frascati, 2017 (poster presentation)

10. K. Blix, M. M. Espeseth and T. Eltoft, "ML simulation of quad-pol features from dual-pol data", 2017 Center for Integrated Remote Sensing and Forecasting for Arctic Operations (CIRFA) annual conference, Sommarøy, 2017 (poster presentation, winner of the *Best Poster Award*)

11. K. Blix and T. Eltoft, "Model selection algorithm for Chlorophyll-a content retrieval", 2017 High Spatial and temporal Resolution Ocean Color products and services conference (HIGHROC), Brussels, 2017 (oral and poster presentation)

12. K. Blix, "Validation of Sentinel-3A Chlorophyll-a and Total Suspended Matter retrieval over Lake Balaton", 2018 Balaton Limnological Institute annual conference, Tihany, 2018 (oral presentation)

13. K. Blix, T. Eltoft and V. R. Tóth, "Validation of Sentinel-3A OLCI Level-2 water-quality products over Lake Balaton", 2018 S3 Validation Team Meeting, Darmstadt, 2018 (oral presentation)

14. K. Blix, M. M. Espeseth and T. Eltoft, "Machine Learning simulations of quad-polarimetric features from dual-polarimetric measurements over sea ice", 12th European Conference on Synthetic Aperture Radar (EUSAR), Aachen, 2018 (oral presentation)

15. K. Blix, M. M. Espeseth and T. Eltoft, "Up-scaling from quad-polarimetric to dual-polarimetric SAR data using Machine Learning Gaussian Process Regression", 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, 2018 (poster presentation)

16. K. Blix, "Machine Learning Gaussian Process Regression for Remote Sensing Applications", 2018 Center for Integrated Remote Sensing and Forecasting for Arctic Operations (CIRFA) annual conference, Sommarøy, 2018 (oral and poster presentation)

17. K. Blix, K. Pálffy and V. R. Tóth, "Machine Learning to monitor water quality in Lake Balaton", LX. Hydrobiology Conference, Tihany, 2018 (oral presentation)

18. K. Blix, "Remote Sensing of Arctic Environments: Monitoring Ecosystems by Using Satellites", Oslo, 2018, invited speaker at the Hungarian Embassy for Norway

19. K. Blix and T. Eltoft, "Machine Learning Remote Sensing for Ecosystem Monitoring in the Arctic", 2019 Arctic Frontiers, Tromsø, 2019 (oral presentation)

20. K. Blix, "Machine Learning for Remote Sensing of Arctic Waters", Rimouski, 2019, invited speaker at Quebec Ocean, UQAR

21. K. Blix, K. Pálffy, V. R. Tóth and t. Eltoft, "Machine Learning for Remote Sensing Complex Waters", 2019 International Ocean Colour Science Meeting, Busan, 2019 (poster presentation, recipient of travel award)

22. K. Blix, M. Babin, P. Massicotte and T. Eltoft, "Machine Learning for Monitoring Arctic Waters by Using Sentinel 3 OLCI", S3 Validation Team Meeting, Frascati, 2019 (submitted)

23. K. Blix and T. Eltoft, "A generalized chlorophyll-a estimation model for complexity-diverse Arctic waters", 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, 2019 (submitted)

Chapter 6

Paper 1:
Gaussian Process Sensitivity Analysis for Oceanic Chlorophyll Estimation

Gaussian Process Sensitivity Analysis for Oceanic Chlorophyll Estimation

Katalin Blix, Gustau Camps-Valls, *Senior Member, IEEE*, and Robert Jenssen, *Member, IEEE*

***Abstract*—Gaussian process regression (GPR) has experienced tremendous success in biophysical parameter retrieval in the past years. The GPR provides a full posterior predictive distribution so one can derive mean and variance predictive estimates, i.e., point-wise predictions and associated confidence intervals. GPR typically uses translation invariant covariances that make the prediction function very flexible and nonlinear. This, however, makes the relative relevance of the input features hardly accessible, unlike in linear prediction models. In this paper, we introduce the sensitivity analysis of the GPR predictive mean and variance functions to derive feature rankings and spectral spacings, respectively. The methodology can be used to uncover knowledge in any kernel-based regression method, it is fast to compute, and it is expressed in closed-form. The methodology is evaluated on GPR for global ocean chlorophyll prediction, revealing the most important spectral bands and their spectral spacings. We illustrate the (successful) methodology in several datasets and sensors.**

***Index Terms*—Gaussian process regression (GPR), kernel methods, oceanic chlorophyll prediction, sensitivity analysis (SA).**

I. Introduction

BEING able to monitor ocean chlorophyll content from remotely sensed data provides the possibility of monitoring the health status of oceans through the photosynthetic activity [1]. Changes in the photosynthetic activity result in changes in the chlorophyll fluorescence [2], [3]. Therefore detecting chlorophyll fluorescence from space can reveal the distribution of the marine primary producers, the phytoplankton [4]–[7]. This has deep ecological [8] and economic implications.[1] In addition, monitoring ocean chlorophyll content also provides a tool to achieve deeper understanding of the contribution of CO_2 to the climate [9]–[11].

In this scenario, ocean chlorophyll estimation from space requires accurate and fast mapping algorithms. It is a standard practice to use parametric bio-optical models, such as the OC2 and OC4 models [12], [13]. They, however, assume explicit relationships between the input reflectance bands and the chlorophyll content. They are relatively simple (empirical) nonlinear mapping functions (most of the times simple band ratios and polynomial functions), and very importantly, the model complexity must be controlled based on experience. In recent years, many alternative algorithms have appeared using statistical machine learning algorithms for chlorophyll content estimation from multi- and hyperspectral data. Many methods have been studied: neural networks [14], support vector regression [15]–[17], and the relevance vector machine [18]. The recently introduced Gaussian process regression (GPR) model has been shown to outperform other oceanic chlorophyll content estimation methods [19]. For oceanic chlorophyll content estimation in remotely sensed data, GPR framework has been successfully applied in [20] and [21]. GPR differs from other machine learning methods not only in its predictive power, but also in its underlying fundamental principles [22], [23]. The other advantageous property of GPR is that it provides additional information about the prediction: the predictive variance. Thus, the output of the regression is not only the estimated chlorophyll content, but also the estimated variance, which reveals the confidence of the prediction.

Although GPR has shown an excellent predictive performance, the information about the relative relevance of the features being used for regression is lost, since the model is a nonlinear kernel method that defines an implicit (not accessible) feature mapping. In [24], we presented the *sensitivity analysis* (SA) of the mean function in the GPR model and applied it to three oceanic chlorophyll matchup datasets.

Our contribution in this paper is the extension of our method to the predictive variance function of the GPR model. In addition, we further exploit the SA for both the predictive mean and variance function by evaluating their performances on both controlled examples and new updated global oceanic chlorophyll relevant datasets. We compare the methodology with state-of-the-art oceanic chlorophyll content estimation models for the Sea-Viewing Wide Field-of-View Sensor (SeaWIFS), the MEdium Resolution Imaging Spectrometer (MERIS), and the NASA operational Moderate Resolution Imaging Spectroradiometer onboard AQUA (MODIS-Aqua). Furthermore, we present our results on sensitivity maps, which are aimed to show the possibility of applying our method for practical purposes.

SA reduces to study the variance (uncertainty) of the predictive function in terms of the uncertainties of the input features.

Manuscript received July 6, 2016; revised September 27, 2016; accepted December 7, 2016. Date of publication January 3, 2017; date of current version March 22, 2017. This work was supported in part by the Research Council of Norway under FRIPRO Grant 239844 (RJ) and in part by the European Research Council under the ERC Consolidator Grant SEDAL-647423. *(Corresponding author: Katalin Blix.)*

K. Blix is with the Department of Physics and Technology, University of Tromsø—The Arctic University of Norway, Tromso 9037, Norway (e-mail: katalin.blix@uit.no).

G. Camps-Valls is with the Image Processing Laboratory, Universitat de València, València 46980, Spain (e-mail: gustau.camps@uv.es).

R. Jenssen is with the Machine Learning @ UiT Lab, Department of Physics and Technology, University of Tromsø—The Arctic University of Norway, Tromso 9037, Norway (e-mail: robert.jenssen@uit.no).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2016.2641583

[1] http://oceancolor.gsfc.nasa.gov/

The family of SA methods is vast and depends on the number of problem *constraints*, also known as *settings*. In this paper, we focus on the particular field of *local methods*, which involve taking the partial derivative of model's output with respect to input features to assess its impact. Interestingly, in the case of the GPR model, such gradients can be computed in closed form for most of the covariance functions.

Introducing the SA for the GPR mean function for determining feature relevance has the advantage that it is not limited to a specific kernel function, such as in the case of the automatic relevance determination, where the length-scale parameters of the squared exponential kernel are optimized in order to assign feature relevance [25], [26]. Furthermore, the SA of the GPR variance function, to the best of our knowledge, is the only existing method that can reveal the spectral spacing of the input space in the GPR model.

In order to gain an intuitive interpretation, we first present the SA on a controlled example, and then apply it to five global chlorophyll related datasets from different sensors: SeaBAM, SeaWIFS, MODIS-Aqua, and two MERIS complementary datasets. SA can efficiently reveal the most important spectral bands for chlorophyll content prediction, and the spectral spacing of the input space globally. In addition, we compare GPR using only the most relevant spectral bands for regression with spectral-band-ratio models. We also present sensitivity maps for both the GPR predictive mean and variance, which open the possibility of presenting the distribution of the most relevant wavelengths on a global scale, and also to access information about the (spatially resolved) distribution of the spectral sampling of the inputs. The sensitivity maps might indicate the detection of the distribution of chlorophyll fluorescence, thus opening the possibility of monitoring ocean status through a fundamentally different, mathematically solid, approach.

Finally, we validate the results of the SA of the GPR mean function on a global scale by producing global chlorophyll content maps, allowing visual comparison with the actual measured chlorophyll content maps and predicted chlorophyll content maps computed with parametric models.

The objective of performing the SA of the GPR model on oceanic chlorophyll datasets was that GPR has been shown to have a strong regression capacity in the estimation of biophysical parameters, therefore the methodology could be efficiently used in practice for oceanic chlorophyll content estimation from remotely sensed data. However, the driving mechanisms of the GPR model has not been fully understood yet. This is in contrast to parametric models, where the estimated coefficients (weights) allow the direct interpretation of the relevance of the spectral bands. Applying the SA to the GPR mean function for oceanic matchup datasets revealed the most important spectral bands in the regression model. This not only shows that the algorithm performs well and can be extended to a variety of kernel methods, but also provides a tool for having a deeper understanding in the optical properties of the oceans. Furthermore, using the SA of the GPR variance function for these datasets results in a unique interpretation of the spectral spacing of the input space. In addition, our aim by presenting sensitivity maps for the oceanic chlorophyll content datasets was to show that the SA could be used for mapping the most relevant bands and their spectral spacing that might be important when the biophysical and optical properties of the oceans are in focus. These sensitivity maps can be used for information retrieval purposes from remotely sensed data in the very important task of oceanic chlorophyll content estimation.

The reminder of the paper is organized as follows. Section II reviews the GPR model, presents the SA, and an illustrative toy example. Section III details the data collection and experimental setup used in this paper. Section IV gives the experimental results for the estimation of ocean chlorophyll content, comparison of the GPR (using only those bands which were ranked as most relevant of the SA) with parametric state-of-the-art models for ocean chlorophyll content estimation, sensitivity maps, and validation maps. Finally, Section V concludes the paper and outlines the further work.

II. SA in Gaussian Processes

We first review the standard formulation of the GPR model briefly, then present the SA of the GPR predictive mean and variance, and illustrate its performance in a toy example.

A. Regression With Gaussian Processes

Standard regression approximates observations (often referred to as *outputs*) $\{y_n\}_{n=1}^N$ as the sum of some unknown latent function $f(\mathbf{x})$ of the inputs $\{\mathbf{x}_n \in \mathbb{R}^D\}_{n=1}^N$ plus constant power Gaussian noise, i.e., $y_n = f(\mathbf{x}_n) + \varepsilon_n$, $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$. Instead of proposing a parametric form for $f(\mathbf{x})$ and learning its parameters in order to fit observed data well, GPR proceeds in a Bayesian, nonparametric way [22], [23]. A zero mean[2] Gaussian Process ($\mathcal{GP}$) prior is placed on the latent function $f(\mathbf{x})$ and a Gaussian prior is used for each latent noise term ε_n, $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}'))$, where $k_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$ is a covariance function parameterized by $\boldsymbol{\theta}$, and σ^2 is a hyperparameter that specifies the noise power. Essentially, a $\mathcal{GP}$ is a stochastic process whose marginals are distributed as a multivariate Gaussian. In particular, given the priors $\mathcal{GP}$, samples drawn from $f(\mathbf{x})$ at the set of locations $\{\mathbf{x}_n\}_{n=1}^N$ follow a joint multivariate Gaussian with zero mean and covariance matrix $\mathbf{K}_{\mathbf{ff}}$ with $[\mathbf{K}_{\mathbf{ff}}]_{ij} = k_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)$.

If we consider a test location $\mathbf{x}_*$ with corresponding output y_*, the $\mathcal{GP}$ defines a joint prior distribution between the observations $\mathbf{y} \equiv \{y_n\}_{n=1}^N$ and y_*.

Collecting available data in $\mathcal{D} \equiv \{\mathbf{x}_n, y_n | n = 1, \ldots N\}$, it is possible to analytically compute the posterior distribution over the output y_* as

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \mathcal{N}(y_*|\mu_{\text{GP}*}, \sigma^2_{\text{GP}*}) \tag{1}$$

$$\mu_{\text{GP}*} = \mathbf{k}_{\mathbf{f}*}^\top (\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} = \mathbf{k}_{\mathbf{f}*}^\top \boldsymbol{\alpha} \tag{2}$$

$$\begin{aligned} \sigma^2_{\text{GP}*} &= \sigma^2 + k_{**} - \mathbf{k}_{\mathbf{f}*}^\top (\mathbf{K}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{k}_{\mathbf{f}*} \\ &= \sigma^2 + k_{**} - \mathbf{k}_{\mathbf{f}*}^\top \mathbf{A} \mathbf{k}_{\mathbf{f}*}, \end{aligned} \tag{3}$$

[2]It is customary to subtract the sample mean to data $\{y_n\}_{n=1}^N$, and then to assume a zero mean model.

where $\mathbf{k}_{\mathbf{f}*}$ is the covariance between the training vector and the test point, $\boldsymbol{\alpha} = (\mathbf{K}_{\mathbf{ff}} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}$ is the weight vector of the GPR mean, k_{**} is the covariance between the test point with itself, and $\mathbf{A} = (\mathbf{K}_{\mathbf{ff}} + \sigma^2\mathbf{I}_n)^{-1}$ is the weight matrix of the GPR variance.

Note that the predictive mean in (2) $\mu_{\text{GP}*}$ depends on the observations through the weight vector $\boldsymbol{\alpha}$, while the confidence intervals $\sigma^2_{\text{GP}*}$ [see (3)] only depend on the inverse of the regularized covariance function $\mathbf{A}$.

B. SA of Features From Gaussian Processes

GPR offers some advantages over other regression methods. Since they yield a full posterior predictive distribution over y_* [see (1)], it is possible to obtain not only mean predictions for test data $\mu_{\text{GP}*}$ (2), but also the so-called "error-bars," assessing the uncertainty of the mean prediction $\sigma^2_{\text{GP}*}$ [see (3)]. In this paper, we focus on extracting knowledge from trained GPR model. To do so, let us define the sensitivity of feature j as

$$s_j = \int \left(\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}_j} \right)^2 p(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{4}$$

where $p(\mathbf{x})$ is the probability density function over the D-dimensional input vector $\mathbf{x}_n = [x_n^1, \ldots, x_n^D]^\top$, and $\phi(\mathbf{x})$ represents either the predictive mean $\mu_{\text{GP}*}$ or variance $\sigma^2_{\text{GP}*}$. Intuitively, the objective of the SA of features is to measure the changes of the derivative of the function $\phi(\mathbf{x})$ in the jth direction. In order to avoid the possibility of cancelation of the terms due to its signs, the derivatives are squared. Therefore, the resulting sensitivities will be positive $s_j \geq 0$ for all bands. The empirical estimate of the sensitivity for the jth feature can be written as

$$s_j = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\partial \phi(\mathbf{x}_n)}{\partial x_n^j} \right)^2, \tag{5}$$

where N denotes the number of training samples. Before calculating the sensitivity, let us define the covariance prior that we used in this paper, the standard isotropic-scaled Gaussian kernel function

$$k(\mathbf{x}_m, \mathbf{x}_n) = \nu^2 \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \left(\frac{x_m^d - x_n^d}{\lambda_d} \right)^2 \right), \tag{6}$$

where λ_d is the length scale for the dimension d, and ν is a positive scale factor. The hyperparameters of this GP prior are collectively grouped in $\boldsymbol{\theta} = [\nu, \sigma, \lambda_1, \ldots, \lambda_D]$.

The resulting empirical estimate of the GPR mean sensitivity is given as

$$\begin{aligned} s^j_{\mu_{\text{GP}\star}} &= \frac{1}{N} \sum_{q=1}^{N} \left(\frac{\partial \phi(\mathbf{x}_q)}{\partial x_q^j} \right)^2 \\ &= \frac{1}{N} \sum_{q=1}^{N} \left(\frac{\partial \sum_{p=1}^{N} \alpha_p k(\mathbf{x}_p, \mathbf{x}_q)}{\partial x_q^j} \right)^2 \\ &= \frac{1}{N} \sum_{q=1}^{N} \left(\sum_{p=1}^{N} \frac{\alpha_p (x_p^j - x_q^j)}{\lambda_j^2} k(\mathbf{x}_p, \mathbf{x}_q) \right)^2, \end{aligned} \tag{7}$$

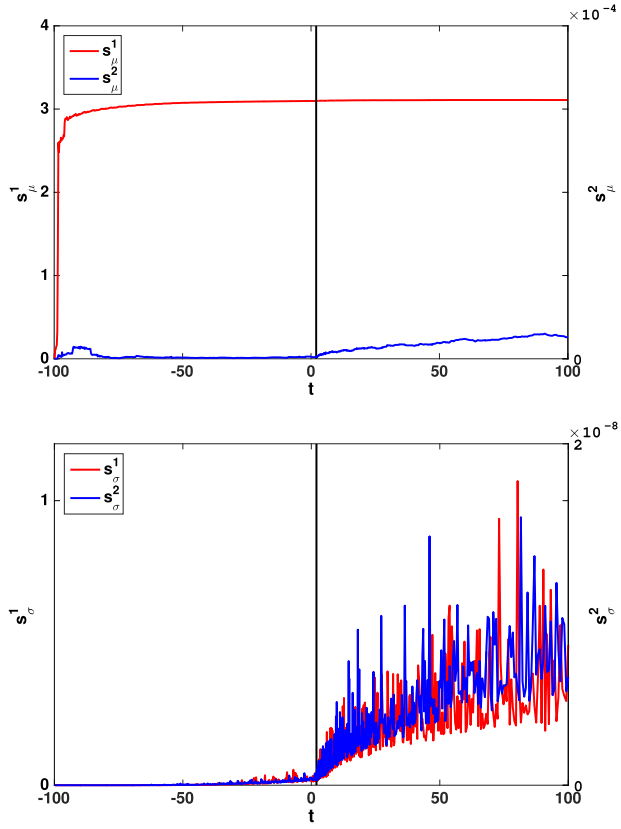


Fig. 1. Evolution of the sensitivities through time of the GPR mean (top) and variance (bottom) for the relevant (red) and irrelevant (blue) feature.

and for the GPR variance sensitivity is

$$\begin{aligned} s^j_{\sigma_{\text{GP}\star}} &= -2N\nu^2 \\ &\quad \times \sum_{q=1}^{N} \left(\sum_{p,q=1}^{N} A_{pq} (x_p^j - x_q^j) k(\mathbf{x}_p, \mathbf{x}_q)^2 / \lambda_j^2 \right)^2. \end{aligned} \tag{8}$$

Note that the calculation of the empirical sensitivity is computed in closed form using only training data points and the inferred $\boldsymbol{\alpha}$ and $\mathbf{A}$. The SA derived here is inspired by Rasmussen *et al.* [27] who, however, only regarded a support vector machine and brain research context, and who did not extend the analysis to variance.

C. Proof of Concept

We show the concept of the SA on a synthetic example. The goal of this experiment is to examine whether the SA of the GPR mean computed by (7) function can identify the relevant feature in the regression process. At the same time, we compute the SA of the GPR variance function by using (8), so that the spacing t of the input features can be revealed.

Assume that the input consists of two features $\mathbf{x}_n = [x_n^1, x_n^2]$, where $x_n^1 = A \sin(2\pi t)$ is the relevant feature, $x_n^2 \sim \mathcal{N}(0, \sigma^2)$ is irrelevant, and $A \gg \sigma$. The output is the sum of the two input features, $y_n = x_n^1 + x_n^2$. Time sampling is uniform for $t \leq 0$ and logarithmically for $t > 0$. In order to trace the evolution of the sensitivities as t grows, we compute s^i_μ and s^i_σ by using (7) and

TABLE I
DESCRIPTIVE STATISTICS OF THE FIVE DATASETS

	SeaBAM		**SeaWIFS**	
Chlorophyll range (mgm^{-3})	0.019–32.787		0.024–129.332	
Nr. of samples	919		1465	
Band (λ_c (nm))	μ	σ^2	μ	σ^2
412	0.0066	$0.1374{\cdot}10^{-4}$	0.0036	$0.1001{\cdot}10^{-4}$
443	0.0059	$0.0969{\cdot}10^{-4}$	0.0038	$0.0625{\cdot}10^{-4}$
490	0.0049	$0.039{\cdot}10^{-4}$	0.0041	$0.0468{\cdot}10^{-4}$
510	0.0032	$0.0163{\cdot}10^{-4}$	0.0038	$0.047{\cdot}10^{-4}$
555	0.002	$0.0163{\cdot}10^{-4}$	0.0038	$0.0910{\cdot}10^{-4}$
670			0.001	$0.0178{\cdot}10^{-4}$
	MERIS (synthetic)		**MERIS (real)**	
Chlorophyll range (mgm^{-3})	0.021–53.4429		0.017–40.23	
Nr. of samples	5000		567	
Band (λ_c (nm))	μ	σ^2	μ	σ^2
413	0.0258	0.0006	−1.7594	1760.2
443	0.0323	0.008	0.0031	$0.0597{\cdot}10^{-4}$
490	0.0476	0.0018	0.0042	$0.0624{\cdot}10^{-4}$
510	0.0524	0.0022	0.0045	$0.0793{\cdot}10^{-4}$
560	0.0606	0.0033	0.0057	$0.1784{\cdot}10^{-4}$
620	0.0285	0.0012	0.003	$0.134{\cdot}10^{-4}$
665	0.0222	0.0008	0.0022	$0.095{\cdot}10^{-4}$
681	0.0234	0.0007	0.0022	$0.0873{\cdot}10^{-4}$
	MODIS-Aqua			
Chlorophyll range (mgm^{-3})	0.0153–25.4985			
Nr. of samples	579			
Band (λ_c (nm))	μ	σ^2		
412	0.0028	$0.8138{\cdot}10^{-5}$		
443	0.0032	$0.4778{\cdot}10^{-5}$		
488	0.0036	$0.302{\cdot}10^{-5}$		
531	0.0037	$0.4422{\cdot}10^{-5}$		
547	0.0037	$0.5556{\cdot}10^{-5}$		
667	0.0009	$0.1302{\cdot}10^{-5}$		
678	0.001	$0.1186{\cdot}10^{-5}$		

(8), respectively, for $i = 1, 2$ through time t. Fig. 1 [top] shows the sensitivities for the GPR mean [see (7)] for the relevant feature s_μ^1 and for the noise s_μ^2, respectively. It can be observed that the SA could consistently identify the relevant feature. The sensitivities of the GPR variance [computed by (8)] are shown in Fig. 1 [bottom]. It can be seen how it correctly captures the change at $t = 0$ related to the sampling rate. The SA of the GPR variance, as expected, assigned greater values to the relevant feature. This example shows how the SA of the GPR can be used for determining the most relevant features, and to uncover the sampling rates of the input variables by using (7) and (8), respectively.

III. DATA COLLECTION

In this paper, we show results of the SA in five chlorophyll relevant datasets, acquired by different sensors and thus different spectral resolutions and complexity [28]: SeaBAM, SeaWIFS, MODIS-Aqua, and two complementary MERIS datasets. (For further details on the SeaBAM and MERIS (synthetic) datasets, we refer to [13], [14], and [17]–[19]. The SeaWIFS, MODIS-Aqua, and MERIS (real) datasets can be obtained from the SeaBASS database.[3]) Table I summarizes the main parameters of the descriptive statistics of these datasets, such as the center wavelengths λ_c, the mean μ and variance σ of each channel, the range of the chlorophyll-a concentrations, and the total number of samples. Note that we used the reflectances measured in Remote sensing reflectance (Rrs) for chlorophyll content prediction purposes. The SeaBAM dataset gathers 919 ocean chlorophyll measurements around the United States and Europe. The matchup dataset consists of coincident *in situ* remote sensing reflectance on five channels, which correspond to some of the SeaWIFS channels and chlorophyll-a concentration measurements. The bandwidths of the channels are 20 nm, and they are situated in the range between 402 and 565 nm. The chlorophyll-a concentrations range between 0.019 and 32.787 mgm^{-3}. In addition, we applied the SA of features to three global remote sensing ocean chlorophyll data, the SeaWIFS, the MODIS-Aqua, and the MERIS dataset [29]. The SeaWIFS dataset covers the spectral region between 402 and 680 nm on six channels. We used 1465 chlorophyll-a measurements with coincident Rrs between September 1997 and November 2010. Chlorophyll-a concentrations span a quite wide range, between 0.024 and 129.332 mgm^{-3}. The MODIS-Aqua dataset has seven channels ranging from 405 to 683 nm. The data we used here have 579 measurements between July 2002 and November 2012, where the chlorophyll-a molecule concentrations are between 0.0153 and 25.4985 mgm^{-3}. Finally, the MERIS dataset has the same channels as the synthetic MERIS data, consisting of 567 measurements between April 2002 and March 2012, where the range of the chlorophyll-a concentration is between 0.017 and 40.23 mgm^{-3}. We applied the SA to these global data and computed sensitivity maps for an extracted area, East-USA. An additional MERIS dataset is formed by synthetic data, where 5000 coincident chlorophyll-a concentrations and Rrs were simulated [17]. The chlorophyll-a concentrations range between 0.021 and 53.4429 mgm^{-3}. The Rrs were simulated on eight channels. The channels are placed between 407.5 and 685 nm, with a bandwidth of 10 and 7.5 nm. The means and the variances of the channels show similar values for all the five datasets. Generally, the means are situated close to zero and the variances are small. Note that the MERIS (real) dataset's mean value of band 1 differs from the rest of the means. The corresponding variance is large. This might indicate fault measurement(s) in the dataset at this band.

TABLE II
SUMMARY OF THE TEST RESULTS IN THE FIVE DATASETS

Database	ME	RMSE	MAE	ρ
SeaBAM (2, 4, and 5)	+0.0037	0.1493	0.1104	0.9679
SeaWIFS (4, 5, and 6)	–0.0887	0.3149	0.2361	0.9236
MODIS-Aqua (4, 5, and 6)	+0.0229	0.2461	0.1866	0.9188
MERIS (synthetic) (5, 6, 7, and 8)	0.004	0.084	0.0232	0.9996
MERIS (real) (5, 6, 7, and 8)	$< 10^{-7}$	0.21	0.1464	0.9261

(*The numbers in the parentheses refer to the most relevant channels which were used as inputs in the GPR.)*

[3]http://seabass.gsfc.nasa.gov/seabasscgi/search.cgi.

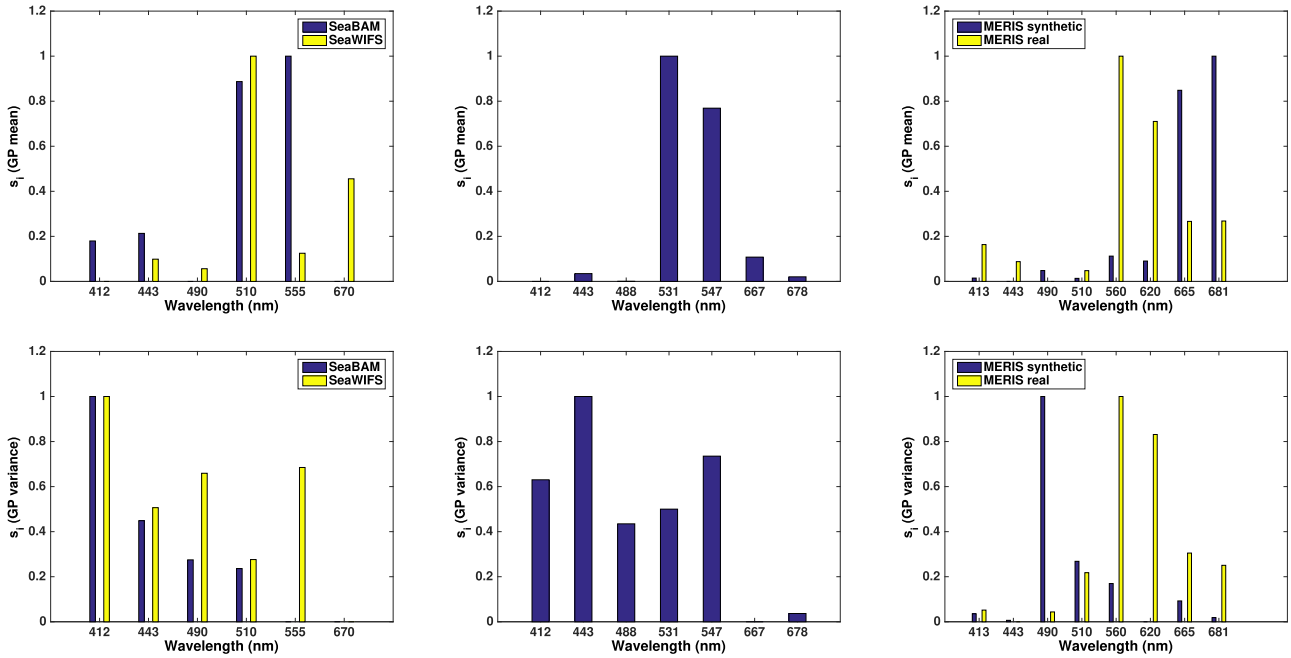


Fig. 2. SA of the GPR mean (top row) and variance (bottom row) for the SeaWIFS and SeaBAM dataset (left column), MODIS-Aqua dataset (middle column), and MERIS dataset (right column).

IV. Experimental Results

Here, we present the experimental results on the previous five different datasets. We first describe the experimental setup, and then study the provided feature ranking from the GPR model. Furthermore, we compare GPR, using only those bands for regression which were assigned to greatest relevance by the SA, with commonly applied parametric models. Finally, we provide spatially explicit SA maps for both the predictive mean and variance.

A. Experimental Setup

We trained five different GPR models for the corresponding datasets. In all cases, we standardized the input features and removed the mean of the observed chlorophyll content. We split the available data randomly into a training set (50%) and a test (hold-out) set. The hyperparameters θ were optimized by maximizing the marginal log-likelihood [22] using the training set. Results of the best models are shown for the test set in Table II. We show different quality measures for the models: bias (mean error, ME), accuracy (root-mean-square error, RMSE, and mean absolute error, MAE), and goodness of fit (Pearson's correlation coefficient ρ). It can be noted that in all cases, the GPR models are accurate and generally unbiased, so an SA is feasible.

B. SA of the Five Datasets

We perform the SA of the GPR mean and variance functions for all five datasets (results are given in Fig. 2). For the SeaWIFS dataset, the SA of the GPR mean revealed that band 4 (510 nm) is the most sensitive (see Fig. 2 [top-left]), which matches

TABLE III
Model Comparison of the Test Results for Bio-Optical Models and GPR for All Datasets

Model	ME	RMSE	MAE	ρ
		SeaBAM		
Morel-1	–0.0289	0.18	0.1404	0.9558
Morel-3	–0.0309	0.1844	0.1432	0.954
CalCOFI 2-band cubic	–0.056	0.1791	0.1424	0.9598
CalCOFI 2-band linear	+0.0729	0.3209	0.2539	0.9558
Ocean chlorophyll 2, OC2	–0.075	0.1856	0.1456	0.9593
Ocean chlorophyll 4, OC4	–0.0835	0.1811	0.1451	0.9652
GPR (2, 4, and 5)	$< 10^{-16}$	0.0117	0.0047	1.0000
		SeaWIFS		
Ocean chlorophyll 2, OC2	–0.376	0.308	0.2312	0.9025
Ocean chlorophyll 3, OC3	–0.0297	0.3046	0.2269	0.9048
Ocean chlorophyll 4, OC4	–0.0194	0.2839	0.2129	0.9165
GPR (4, 5, and 6)	$< 10^{-14}$	0.149	0.035	0.9994
		MODIS-Aqua		
Ocean chlorophyll 2, OC2	–0.0788	0.3283	0.2319	0.8802
Ocean chlorophyll 3, OC3	–0.0742	0.3236	0.2328	0.885
GPR (4, 5, and 6)	$< 10^{-15}$	0.0345	0.0078	0.9999
		MERIS (synthetic)		
Ocean chlorophyll 2, OC2	–0.5397	0.6489	0.5634	0.6799
Ocean chlorophyll 3, OC3	–0.5606	0.667	0.5813	0.6795
Ocean chlorophyll 4, OC4	–0.5439	0.6506	0.5653	0.6862
GPR (5, 6, 7, and 8)	$< 10^{-11}$	0.0144	0.0073	1.0000
		MERIS (real)		
Ocean chlorophyll 2, OC2	–0.0719	0.3699	0.2715	0.8549
Ocean chlorophyll 3, OC3	–0.0668	0.3571	0.2654	0.8641
Ocean chlorophyll 4, OC4	–0.0315	0.3100	0.2311	0.8853
GPR (5, 6, 7, and 8)	$< 10^{-15}$	0.0081	0.0022	1.0000

The computed model measures are the mean values of 100 bootstrap samples.

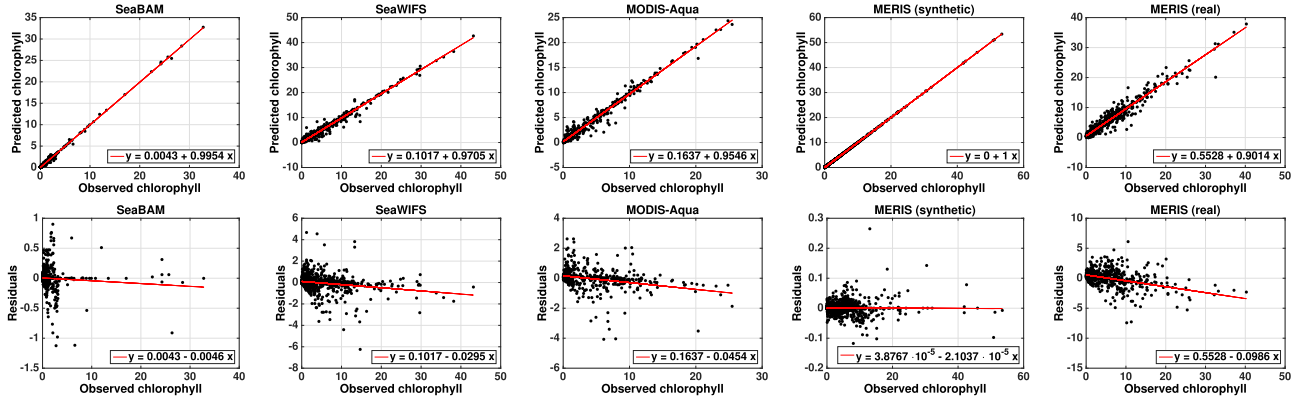


Fig. 3. Using only the most relevant bands in the GPR model for the five datasets. Observed versus predicted chlorophyll (top row) and observed chlorophyll versus residuals (bottom row).

previous results [18], and the accurate bio-optical model OC4. The second most sensitive band corresponds to 670 nm center wavelength (band 6). This is in good correspondence with the three-band reflectance difference model [30], where it was shown that adding band 6 is advantageous and results improved chlorophyll content prediction, especially when chlorophyll content increases. The inclusion of band 6 is based on similar principles as using band 5 (555 nm). The SA assigned the third greatest relevance to band 5, which is commonly used in band-ratio models, such as the OC2/OC3/OC4 and the three-band reflectance difference model [31] and [30]. Fig. 2 [bottom-left] shows the result of the SA of the variance, where the most stable spectral band was band 6. Similar results are obtained in the SeaBAM dataset, where the SA assigned the greatest importance to bands 5 (555 nm), 4 (510 nm), and 2 (443 nm), see Fig. 2 [top-left]). Again, these results matches Morel, CalCOFI-2, and Ocean Color (OC) parametric models. The SA of the variance resulted that band 5 (555 nm) has the most stable spectral variance (Fig. 2 [bottom-left]).

For the MODIS-Aqua dataset, bands 4, 5, and 6 were found to have the highest sensitivities of the GPR mean (Fig. 2 [top-middle]). These channels correspond to 531, 547, and 667 nm, respectively. Band 5 also used in the OC2 and OC3 parametric models. The position of band 4 on MODIS-Aqua was selected to improve the detection of the accessory pigments [32], while band 6 is one of the MODIS-Aqua channels to detect chlorophyll fluorescence [33], [34]. Nevertheless, the SA of the variance, Fig. 2 [bottom-middle], assigned the lowest sensitivity to channel 6.

For the MERIS dataset, we performed the SA on both synthetic and real datasets. For the synthetic dataset, the SA resulted that band 8 (681 nm) has the greatest importance (see Fig. 2 [top-right]) with relative low spectral variance (see Fig. 2 [bottom-right]). This might be the indication of chlorophyll fluorescence [35]–[37]. Channels 7 and 8 were included on MERIS for the detection of the chlorophyll fluorescence signal. Being able to detect chlorophyll fluorescence has special importance when chlorophyll content mapping in coastal waters is in focus, since the presence of gelbstoff and suspended matter might mask the water-leaving radiance from chlorophyll-a, when spectral-band ratios are applied [35].

For the real MERIS dataset, the SA of the GPR mean resulted in that band 5 (560 nm) is the most sensitive (see Fig. 2 [top-right]), also with the highest sensitivity of the GPR variance (see Fig. 2 [bottom-right]). This result is in good correspondence with the OC2/OC3/OC4 parametric models. Note that band 6 (620 nm), 7 (665 nm), and 8 (681 nm) were also found to have high sensitivities in comparison to the rest of the channels. Looking at the sensitivities of the variance for these four channels reveals that band 8 has the lowest sensitivity of the predictive variance.

Applying the SA of features for the GPR mean for these global datasets might reveal the most relevant spectral band for global oceanic chlorophyll prediction. Apart from the synthetic MERIS dataset, in all cases, the most sensitive band fell into the spectral region between 510 and 560 nm. The SA of the GPR predictive variance opens the possibility of accessing the spectral sampling of the channels. This additional information might help selecting channels for analysis in an automated way, since channels with high sensitivity of the GPR mean and low sensitivity of the GPR variance should be preferred.

C. Comparison of Methods

We compared the performance of the GPR (using only the most sensitive bands) with parametric bio-optical models ([13], [31], and [38]) in all the five datasets. These models can be written as follows [18]: Morel-1 and CalCOFI 2-band linear are expressed by $C = 10^{a_0 + a_1 R}$, Morel-3 and CalCOFI 2-band cubic interpolators are $C = 10^{a_0 + a_1 R + a_2 R^2 + a_3 R^3}$, and models OC2/OC3/OC4 are described by $a_0 + \sum_{i=1}^{4} a_i \log_{10} R^i$, where R indicates the logarithmic ratio between the blue and green wavelengths, and a_i are the coefficients. Note that the coefficients and the wavelengths used for determining R are sensor specific (and they can be found at NASA's ocean color web site http://oceancolor.gsfc.nasa.gov/). Model performances were evaluated by computing the same measures as in Section IV-A. The goal of this comparison study was to evaluate the regression strength of the GPR by using only the most important spectral bands, and compare them with the commonly used state-of-the-art algorithms. Therefore, the measures were computed by using the available datasets for both training and

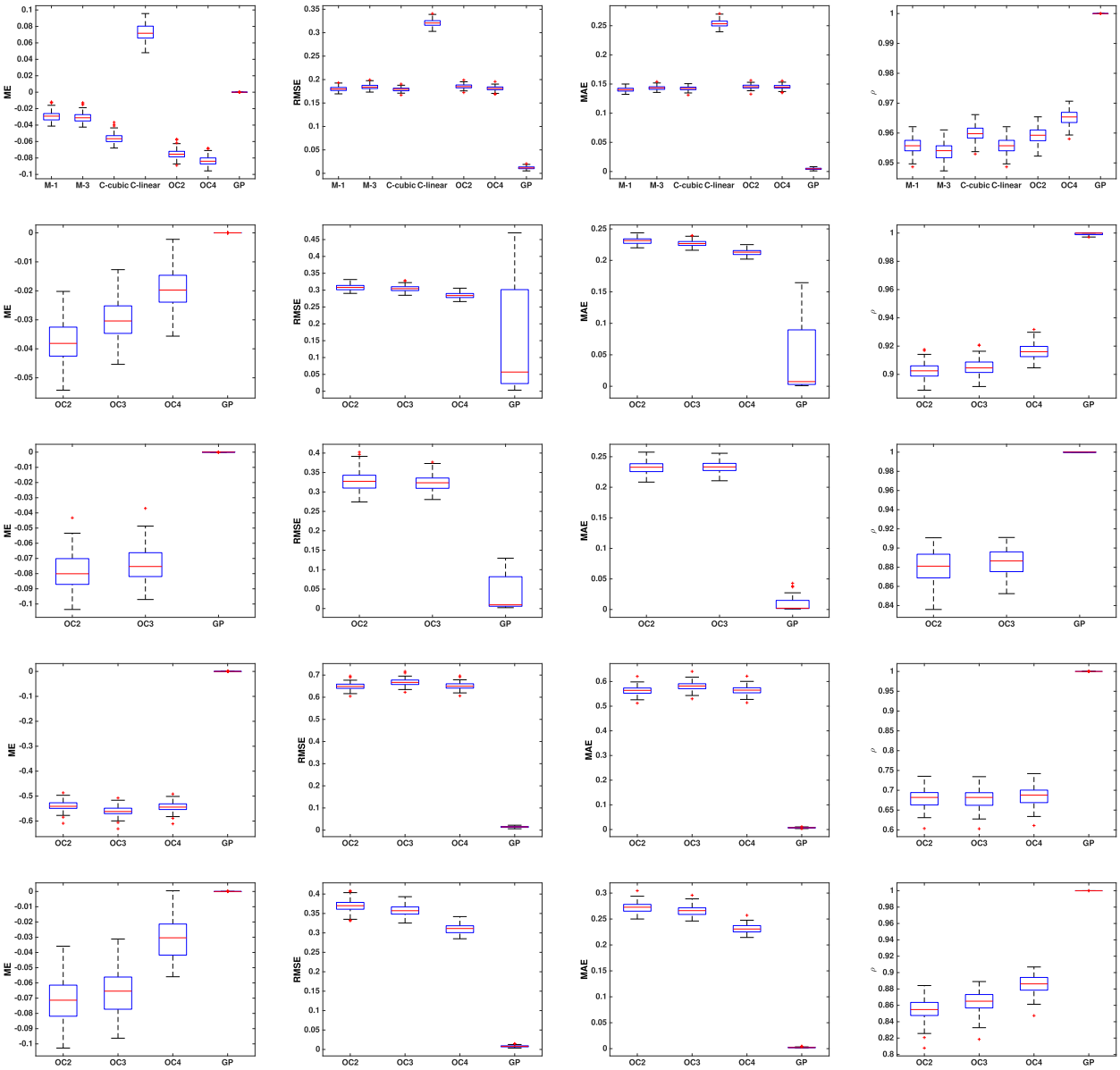


Fig. 4. Box plots of the bootstrap model criteria for the SeaBAM (top row), SeaWIFS (second row), MODIS-Aqua (third row), MERIS synthetic (fourth row), and MERIS real (bottom row) datasets. The models for the SeaBAM dataset are indicated by M-1 and M-3 for Morel-1 and Morel-3, and C-linear and C-cubic for CalCOFI 2-band linear and CalCOFI 2-band cubic, respectively. The OC models are the OC2, OC3, and OC4 algorithms, and the GPR model using the most relevant bands is denoted by GP.

testing. We used bootstrapping for accessing model performance. (Note that in Table II, the prediction strength of the method was in focus, therefore the available datasets were randomly divided into training-testing data, as described in Section IV-A.) The results of the models for the five datasets can be seen in Table III. The measures are the mean values of 100 bootstrap samples. The distribution of the bootstrap measures is presented in Fig. 4. The box plots reveal that the computed model measures from the bootstrap samples for the GPR model (indicated by GP in Fig. 4), has a narrow range (except for the RMSE and MAE in the case of the SeaWIFS and MODIS-Aqua datasets), low bias, and high accuracy.

Applying only the most sensitive bands to the GPR can outperform other commonly used parametric models, which indicates the strength of the SA.

TABLE IV
SUMMARY OF THE STATISTICAL ANALYSIS (ONE-WAY-ANOVA) IN THE FIVE DATASETS

	Bias		Accuracy	
Database	F-value	p-value	F-value	p-value
SeaBAM	67.76	<0.001	156.45	<0.001
SeaWIFS	2.38	<0.1	18.52	<0.001
MODIS-Aqua	5.12	<0.01	31.94	<0.001
MERIS (synthetic)	3243.61	0	4441.51	0
MERIS (real)	0.99	<0.4	202.93	<0.001

Note that parametric models have been previously compared to machine learning methods, for example, in [18], where no statistically significant difference was found. Furthermore, GPR

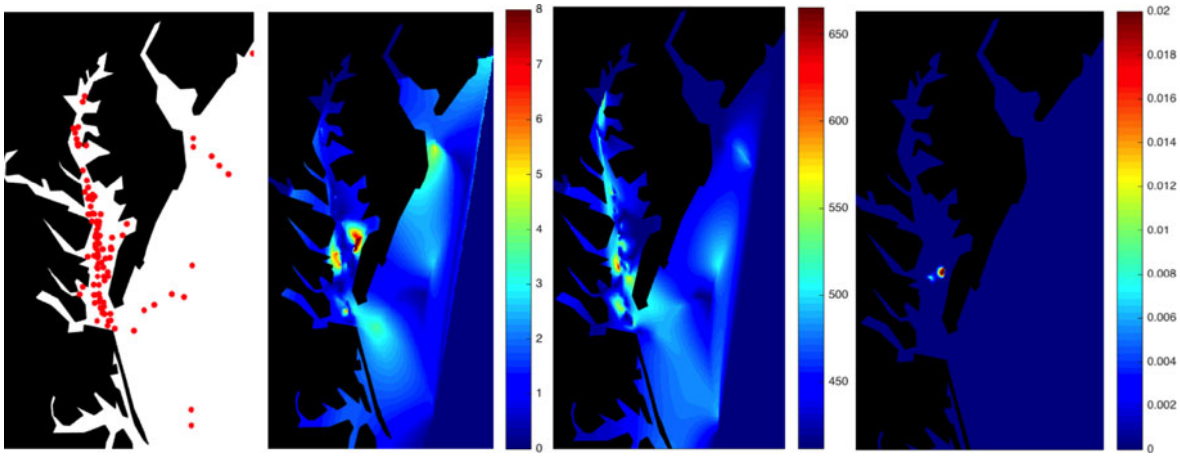


Fig. 5. Position of the *in situ* chlorophyll measurements marked by red points (left), chlorophyll content map in mgm^{-3} (s), sensitivity map of the GPR mean (third), and variance (right) for the SeaWIFS dataset.

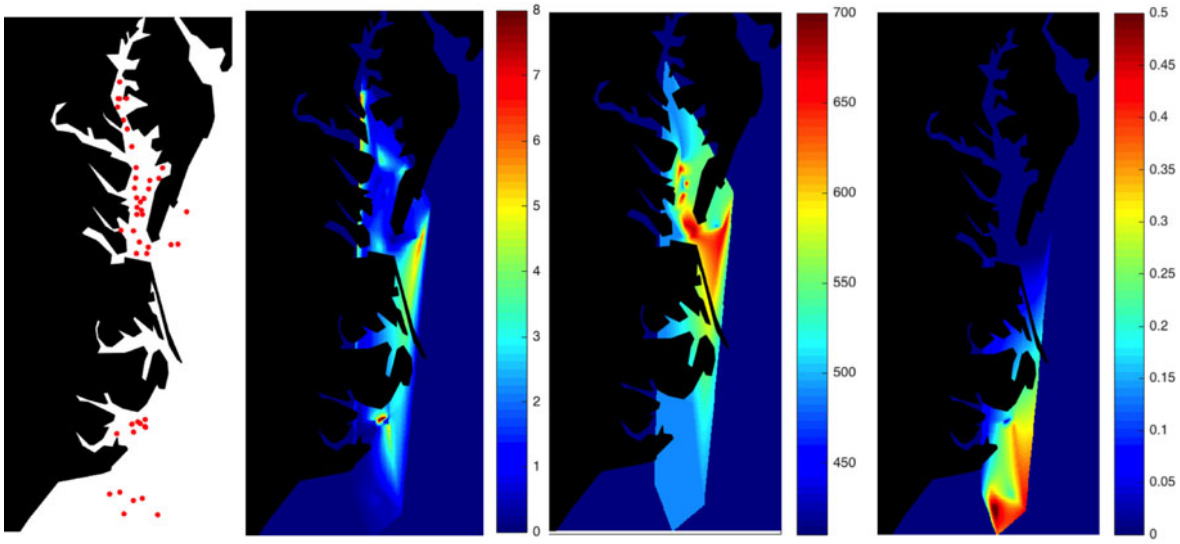


Fig. 6. Position of the *in situ* chlorophyll measurements marked by red points (left), chlorophyll content map in mgm^{-3} (s), sensitivity map of the GPR mean (third), and variance (right) for the MODIS-Aqua dataset.

model using all available features has been shown to outperform other machine learning methods [25]. The goal of our comparison study is to show the strength of the SA in the important task of chlorophyll content estimation from remotely sensed data. In addition, to find the most relevant features for chlorophyll content estimation, we also examine how the GPR model using only the most important spectral bands performs in comparison to the state-of-art algorithms.

We tested the statistical significance of model's difference by performing a one-way Analyzis of Variance (ANOVA) on the estimates. We performed the statistical analysis of the bias and accuracy of the residuals by computing the F-value and p-value for each cases [39]. Table IV shows the results of the ANOVA analysis for the five datasets. Significant statistical differences can be observed for both the bias and accuracy for the SeaBAM, MODIS-Aqua, and MERIS (synthetic) dataset. In the case of the SeaWIFS and MERIS (real) datasets, the statistical analysis could not reveal any difference in the bias between the GPR with the most relevant bands and the rest of the models. However, the accuracy shows a great deviation between the models for these datasets. Fig. 3 presents the scatter plots of the observed versus predicted chlorophyll values (top row) and the observed chlorophyll versus residuals (bottom row) of the five datasets. Good linear agreement can be observed on the observed versus predicted chlorophyll scatter plots. The observed chlorophyll versus residuals scatter plots show a random scattering around zero with a relative small variance.

D. Sensitivity Maps for the Predictive Mean and Variance

We illustrate the performance of the SA by computing sensitivity maps for the SeaWIFS, MODIS-Aqua, and MERIS dataset. The sensitivity maps were computed by extracting a coastal area of the Eastern USA. Then, the SA of the GPR mean and variance for the measurements in this area were computed. We chose the k-nearest neighbors for each data points,

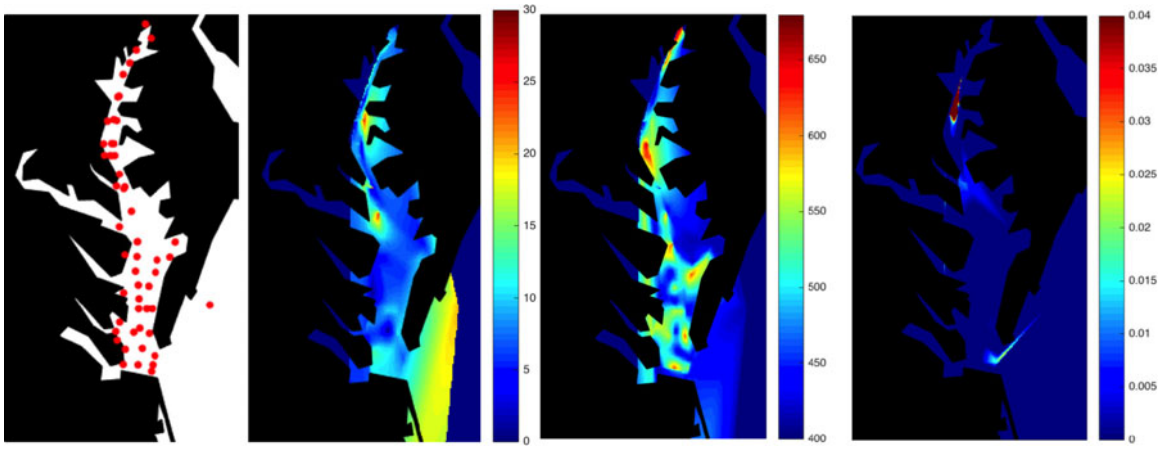


Fig. 7. Position of the *in situ* chlorophyll measurements marked by red points (left), chlorophyll content map in mgm^{-3} (s), sensitivity map of the GPR mean (third), and variance (right) for the MERIS dataset.

performed the SA on the group of these data points through an iteration process, and picked the most sensitive band for the GPR mean and the least sensitive for the GPR variance. Finally, we used these bands to produce sensitive maps by spatial interpolation. We used natural neighbor interpolation [40] for spatial illustration of the SA. Although other interpolation methods, such as kriging, are also commonly used, natural neighbor interpolation has been showed to have a good performance for this type of data as well [41], [42].

Results are shown in Figs. 5–7. The left maps show the positions of the *in situ* chlorophyll measurements (red dots), the second figures illustrate the interpolated measured chlorophyll values, while the third and right figures show the sensitivity maps for the GPR predictive mean and variance, respectively.

The sensitivity maps show that the SA of the GPR mean assigns higher wavelengths to areas where the chlorophyll is present. Interestingly, it can be observed that there are areas with low chlorophyll content (second column) and corresponding higher wavelengths (third column) (in Fig. 6, middle part). This might indicate the presence of suspended particulate materials, which tend to result higher values in the reflectance spectra with increasing concentration [28]. Therefore, computing sensitivity maps in addition to estimated chlorophyll content maps might open the possibility of retrieving further information about the constituents of the oceans through their optical properties. Looking at the chlorophyll content maps together with the SA of the GPR mean maps might give an intuition about the connection between the amount of chlorophyll and the most important wavelengths. The SA of the GPR mean maps represent the geographical distribution of the most important wavelengths, while the SA of the GPR variance maps show how the distribution of the spectral spacing varies in the same area.

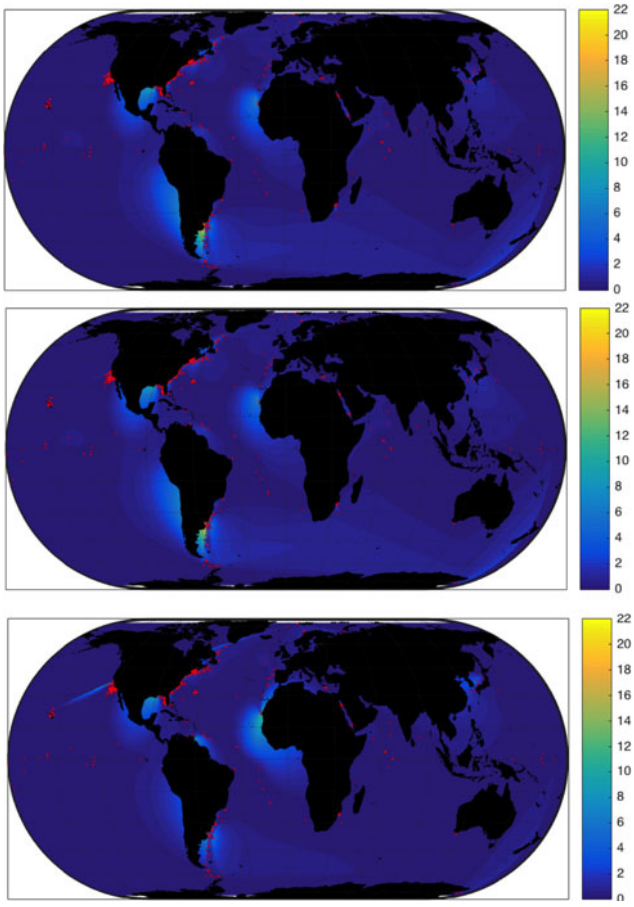


Fig. 8. Global *in situ* chlorophyll content map (top), predicted chlorophyll content map for the GPR with the most relevant bands (middle), and predicted chlorophyll content maps for the OC4 parametric model (bottom) for the SeaWIFS dataset. The total number of samples is 1465 and the unit of the chlorophyll content is mgm^{-3}. The red dots indicate the position of the measurements (interpolation points).

E. Verifying the Results of the SA of the GPR Mean on the SeaWIFS Dataset

In order to validate the results on a global scale, we present global chlorophyll content maps (see Fig. 8) for the SeaWIFS dataset. The global validation maps for the MODIS-Aqua and MERIS (real) datasets and the local validation maps for the SeaWIFS, MODIS-Aqua and MERIS (real) datasets can be found under the appendix. We use the same procedure for spatial interpolation as in Section IV-D. Fig. 8 shows the results of the *in situ* chlorophyll

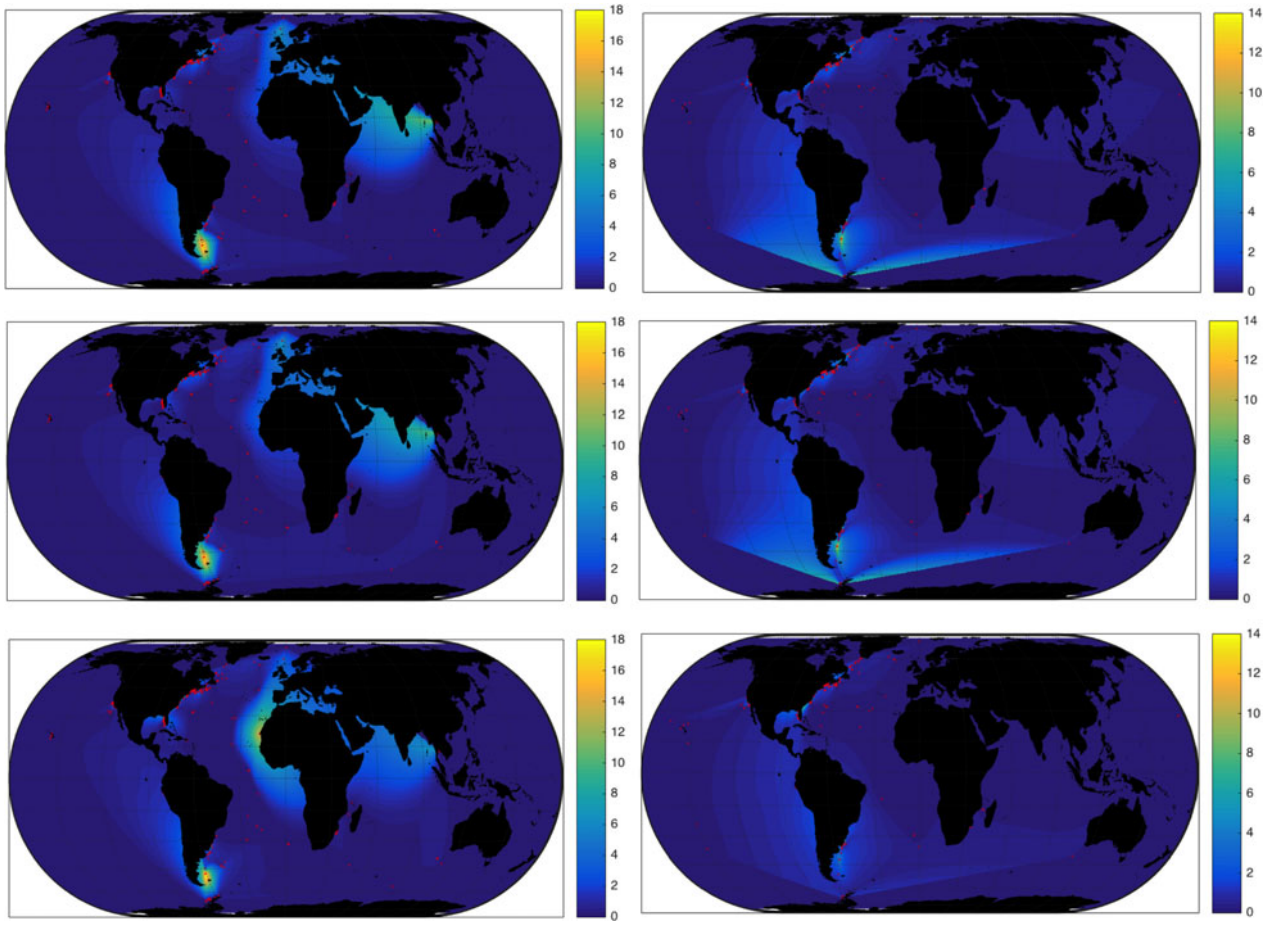


Fig. 9. Global *in situ* chlorophyll content maps (top row), predicted chlorophyll content map for the GPR with the most relevant bands (middle row), and predicted chlorophyll content maps for the OC3 (MODIS-Aqua) and OC4 (MERIS real) parametric model (bottom row) for the MODIS-Aqua (left column) and MERIS real (right column) datasets. The total number of samples is 579 for the MODIS-Aqua dataset and 567 for MERIS real dataset. The unit of the chlorophyll content is mgm^{-3}. The red dots indicate the position of the measurements (interpolation points).

content map (top), the predicted chlorophyll content map using GPR with the most relevant bands (middle), and the predicted chlorophyll content map applying a parametric model (bottom). We chose the parametric model with the lowest RMSE value (see Table III). It can be observed that the chlorophyll content map of the GPR with the most relevant bands looks almost identical as the true chlorophyll content map, while the parametric model seems to overestimate the predicted values. Thus, the SA of GPR can be used to determine feature relevance and selection.

Note that the aim of presenting validation maps is to visualize the strength of the SA rather than to produce accurate global chlorophyll content maps, which would have been challenging for these datasets due to the number of samples and the wide time frame the chlorophyll samples were taken at. Our focus was to illustrate that using the SA of the GPR mean function for identifying the most important spectral bands in the regression process and using only these bands as inputs for the GPR for chlorophyll content estimation can compete with the frequently applied parametric models. Therefore, the methodology opens the possibility for practical application purposes.

V. Conclusion and Further Work

We derived empirical estimates for the sensitivity of the GPR predictive mean and variance functions. After applying the SA to a controlled example, we illustrated the performance of the method on five global datasets. We found that the SA of the GPR mean assigned the highest sensitivity to bands in the range between 510 and 560 nm. This is in good correspondence with the reflectance spectra of the chlorophyll. Bands positioned on higher wavelengths also got ranked as relevant bands for chlorophyll content prediction. This might indicate the preference for bands associated with chlorophyll fluorescence. Being able to monitor chlorophyll fluorescence allows the possibility of detecting changes in photosynthesis, thus to monitor the health status of oceans. In addition, the detection of chlorophyll fluorescence might be a useful tool, when other substances beside chlorophyll are also present. This might be the special case for coastal waters. Besides the sensitivity of the GPR mean, we also derived the SA of the GPR variance for the five global datasets, and uncovered the relevance of the (spectral) sampling of the bands. Knowing the spectral distribution of the inputs might allow the deeper understanding of the underlying biophysics, as

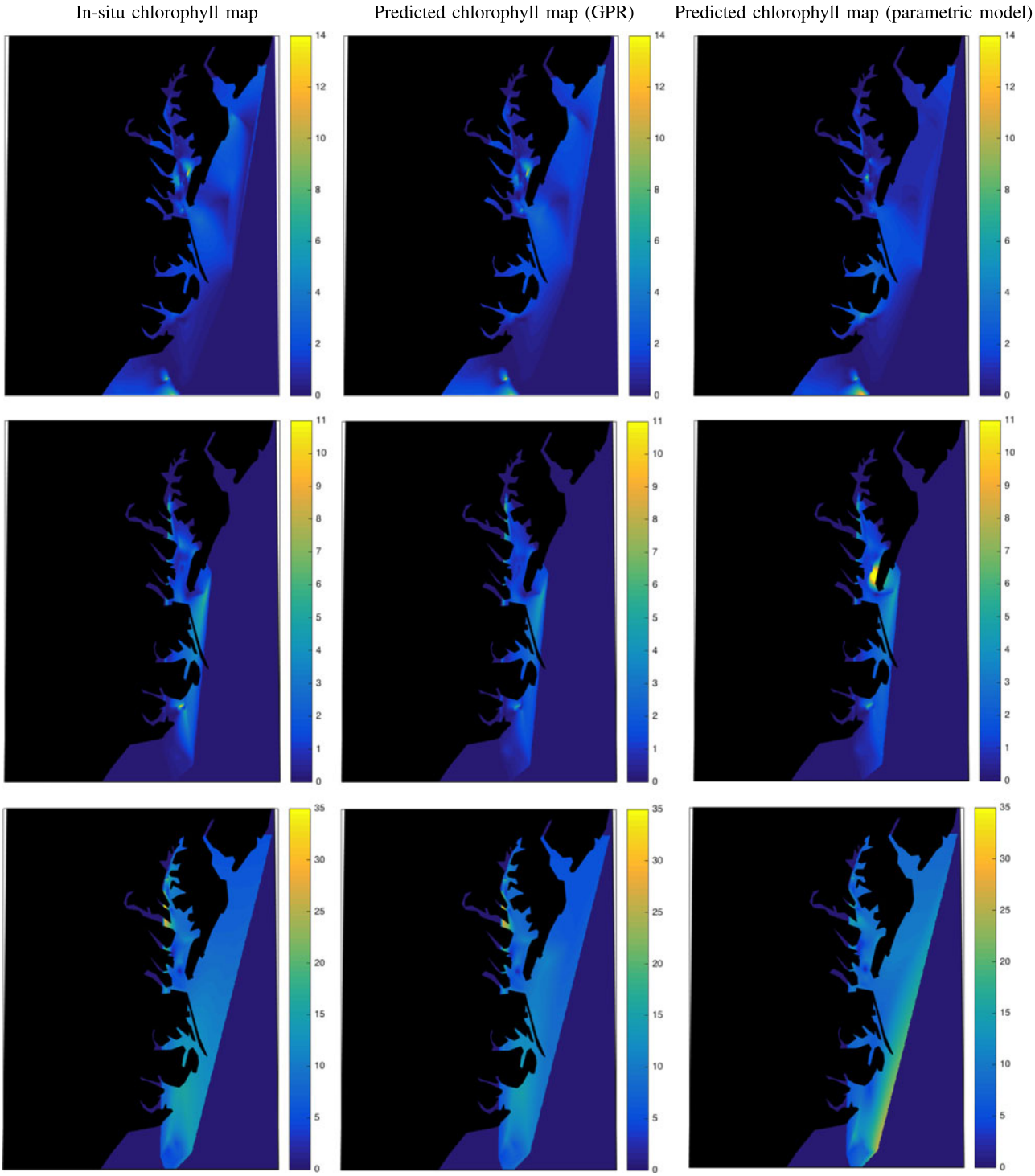


Fig. 10. Local *in situ* chlorophyll content maps (left column), predicted chlorophyll content map for the GPR with the most relevant bands (middle column), and predicted chlorophyll content maps for the best parametric model (right column) for the SeaWIFS (top row), MODIS-Aqua (middle row), and MERIS real (bottom row) datasets. The unit of the chlorophyll content is mgm^{-3}.

well as the design of further sensors. Furthermore, we compared the performance of the GPR using only the most sensitive bands for regression with parametric models. The computed measures revealed that the SA could identify the most important features, and thus using only these features as inputs to the GPR could outperform other models.

Finally, we presented the SA of the GPR on sensitivity maps for a given region. These spatially-explicit maps highlight the usefulness of the GPR SA to study the distribution of the most relevant wavelengths and to reveal the optimality of the spectral sampling density. In addition, we compared the SA of the GPR mean function for chlorophyll content prediction on a global

scale by computing global chlorophyll content maps for the actual chlorophyll content, the GPR model with the most sensitive bands, and with the parametric model of the lowest computed RMSE value. These global maps confirmed that using the bands that were assigned to have the greatest relevance to perform GPR shows good correspondence and spatial comparability. For future work, we plan to produce sensitivity maps on a time scale as well, with the aim of detecting changes in oceanic chlorophyll fluorescence. It does not escape our notice that the methodology can be used for global SA of radiative transfer models, as well as to further evaluate current GPR emulators.

APPENDIX
GLOBAL AND LOCAL VALIDATION MAPS

Fig. 9 shows the global validation maps for the MODIS-Aqua and MERIS (real) datasets. In the case of the MODIS-Aqua dataset (left column), it seems that both the GPR and the parametric model results overestimates along the Western coast of Europe and Africa and underestimates around the Northern part the Indian ocean. The predicted chlorophyll contents show good correspondence with the true values along the coasts of America for both models. Comparing the GPR and the parametric model with the *in situ* chlorophyll content map for the MERIS real dataset (right column) reveals an overall overestimation and underestimation in the predicted chlorophyll contents, respectively. However, the distribution of the chlorophyll seems to follow the same pattern as the true chlorophyll content map for both cases. In general, it can be concluded that the predicted chlorophyll contents are in good correspondence with the true values. Even though there might occur over- and underestimates in the predicted values, using the most sensitive bands to perform GPR for chlorophyll content prediction on a global scale shows just as good performance as the parametric model (with the lowest RMSE value). Therefore, the SA of GPR can be used to determine feature relevance and selection.

The validation maps were also implemented for the same area as in Section IV-D. Fig. 10 shows the results.

ACKNOWLEDGMENT

The authors would like to thank Dr. G. Elvebakk in the Department of Mathematics and Statistics (UiT the Arctic University of Norway) and Dr. A. Ruescas-Orient in the Image Processing Laboratory (Universitat de València) for their useful comments and discussion on the results.

REFERENCES

[1] D. Blondeau-Patissier, J. F. Gower, A. G. Dekker, S. R. Phinn, and V. E. Brando, "A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans," *Progress Oceanography*, vol. 123, pp. 123–144, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0079661114000020

[2] Govindjee,*Bioenergetics of Photosynthesis*. New York, NY, USA: Academic, 1975.

[3] P. Falkowski and D. A. Kiefer, "Chlorophyll a fluorescence in phytoplankton: Relationship to photosynthesis and biomass," *J. Plankton Res.*, vol. 7, pp. 715–731, 1985.

[4] S. M. McKibben, P. G. Strutton, D. G. Foley, T. D. Peterson, and A. E. White, "Satellite-based detection and monitoring of phytoplankton blooms along the Oregon coast," *J. Geophysical Res.*, vol. 117, pp. 1048–1049, 2012.

[5] A. Gitelson, M. Mayo, Y. Z. Yacobi, A. Parparov, and T. Berman, "The use of high-spectral-resolution radiometer data for detection of low chlorophyll concentrations in Lake Kinneret," *J. Plankton Res.*, vol. 16, pp. 993–1002, 1994.

[6] J. Fischer and U. Kronfeld, "Sun-stimulated chlorophyll fluorescence 1: Influence of oceanic properties," *Int. J. Remote Sens.*, vol. 11, no. 12, pp. 2125–2147, 1990. [Online]. Available: http://dx.doi.org/10.1080/01431169008955166

[7] M. J. Behrenfeld *et al.*, "Satellite-detected fluorescence reveals global physiology of ocean phytoplankton," *Biogeosciences*, vol. 6, no. 5, pp. 779–794, 2009.

[8] C. S. Reynolds, *The Ecology of Phytoplankton*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[9] K. R. Arrigo *et al.*, "Phytoplankton community structure and the drawdown of nutrients and CO2 in the Southern Ocean," *Science*, vol. 283, no. 5400, pp. 365–367, 1999. [Online]. Available: http://science.sciencemag.org/content/283/5400/365

[10] M. Hein and K. Sand-Jensen, "CO2 increases oceanic primary production," *Nature*, vol. 388, pp. 526–527, 1997.

[11] M. Hofmann, B. Worm, S. Rahmstorf, and H. J. Schellnhuber, "Declining ocean chlorophyll under unabated anthropogenic CO2 emissions," *Environ. Res. Lett.*, vol. 6, no. 3, pp. 034–035, 2011. [Online]. Available: http://stacks.iop.org/1748-9326/6/i=3/a=034035

[12] H. R. Gordon, O. B. Brown, R. H. Evans, J. W. Brown, and R. C. Smith, "A semianalytic radiance model of ocean color," *J. Geophysical Res.*, vol. 93, pp. 10 909–10 924, 2011.

[13] O. Reilly *et al.*, "SeaWiFS postlaunch calibration and validation analyses, Part 3," NASA Tech. Memo. 2000-206892, vol. 11, 2000.

[14] P. Cipollini, G. Corsini, M. Diani, and R. Grass, "Retrieval of sea water optically active parameters from hyperspectral data by means of generalized radial basis function neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1508–1524, Jul. 2001.

[15] H. Zhan, P. Shi, and C. Chen, "Retrieval of oceanic chlorophyll concentration using support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2947–2951, Dec. 2003.

[16] E. J. Kwiatkowska and G. S. Fargion, "Application of machine-learning techniques toward the creation of a consistent and calibrated global chlorophyll concentration baseline dataset using remotely sensed ocean color data," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2844–2860, Dec. 2003.

[17] G. Camps-Valls, J. Muñoz-Marí, K. R. L. Gómez-Chova, and J. Calpe-Maravilla, "Biophysical parameter estimation with a semisupervised support vector machine," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 248–252, Apr. 2009.

[18] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, J. Amorós-López, and J. Calpe-Maravilla, "Retrieval of oceanic chlorophyll concentration with relevance vector machines," *Remote Sens. Environ.*, vol. 105, no. 1, pp. 23–33, 2006.

[19] L. Pasolli, F. Melgani, and E. Blanzieri, "Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing data," *IEEE Geoscience Remote Sens. Lett.*, vol. 7, no. 3, pp. 464–468, Jul. 2010.

[20] Y. Bazi, N. Alajlan, and F. Melgani, "Improved estimation of water chlorophyll concentration with semisupervised Gaussian process regression," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 7, pp. 2733–2743, Jul. 2012.

[21] Y. Bazi, N. Alajlan, F. Melgani, H. AlHichri, and R. R. Yager, "Robust estimation of water chlorophyll concentrations with Gaussian process regression and IOWA aggregation operators," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 7, pp. 3019–3028, Jul. 2014.

[22] C. E. Rasmussen and C. K. I. Williams, *Gaussian Process for Machine Learning*. Cambridge MA, USA: MIT Press, 2006.

[23] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," *Adv. Neural Inf. Process. Syst.*, vol. 8, pp. 514–520, 1996.

[24] K. Blix, G. Camps-Valls, and R. Jenssen, "Sensitivity analysis of Gaussian processes for oceanic chlorophyll prediction," in *Proc. 2015 IEEE Int. Geosci. Remote Sens. Symp.*, Milan, Italy, Jul. 26–31, 2015, pp. 996–999. [Online]. Available: http://dx.doi.org/10.1109/IGARSS.2015.7325936

[25] J. Verrelst, J. Muñoz, L. Alonso, J. Delegido, J. P. Rivera, G. Camps-Valls, and J. Moreno, "Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for sentinel-2 and -3," *Remote*

Sens. Environ., vol. 118, pp. 127–139, 15 Mar. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003442571100397X
[26] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of vegetation biophysical parameters using Gaussian process techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5 PART 2, pp. 1832–1843, May 2012. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84860332507&partnerID=40&md5=8f89c24c1927827bf249a795b35098fc
[27] P. M. Rasmussen, K. H. Madsen, T. E. Lund, and L. K. Hansen, "Visualization of nonlinear kernel models in neuroimaging by sensitivity maps," *NeuroImage*, vol. 55, no. 3, pp. 1120–1131, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811910016198
[28] I. S. Robinson, *Measuring the Oceans From Space: The Principles and Methods of Satellite Oceanography*. Athens, Greece: Praxis, 2004.
[29] P. J. Werdell and S. W. Bailey, "The seaWiFS bio-optical archive and storage system (seabass): Current architecture and implementation," NASA Goddard Space Flight Center, Greenbelt, MD, USA, NASA Tech. Memo. 2002-211617, p. 45, 2002.
[30] C. Hu, Z. Lee, and B. Franz, "Chlorophyll a algoritms for oligotrophic oceans: A novel approach based on three-band reflectance difference," *J. Geophysical Res.*, vol. 117, 2012, Art. no. C01011.
[31] J. E. O'Reilly *et al.*, "Ocean color chlorophyll algorithms for SeaWiFS," *J. Geophysical Res.*, vol. 103, pp. 24 937–24 953, 1998.
[32] W. E. Esaias *et al.*, "An overview of MODIS capabilities for ocean science observations," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 4, pp. 1250–1265, Jul. 1998.
[33] J. F. R. Gower, L. Brown, and G. A. Borstad, "Observation of chlorophyll fluorescence in west coast waters of Canada using MODIS satellite sensor," *Can. J. Remote Sens.*, vol. 30, no. 1, pp. 17–25, 2004.
[34] C. Hu *et al.*, "Red tide detection and tracing using MODIS fluorescence data: A regional example in SW Florida coastal waters," *Remote Sens. Environ.*, vol. 97, pp. 311–321, 2005.
[35] J. F. R. Gower, R. Doerffer, and G. A. Borstad, "Interpretation of the 685 nm peak in water-leaving radiance spectra in terms of fluorescence, absorption and scattering, and its observation by MERIS," *Int. J. Remote Sens.*, vol. 20, no. 9, pp. 1771–1786, 1999. [Online]. Available: http://dx.doi.org/10.1080/014311699212470
[36] J. Gower, S. King, W. Y. G. Borstad, and L. Brown, "Use of 709 nm band of MERIS to detect intense plankton blooms and other conditions in coastal waters," in *Proc. ESA, MERIS User Workshop*, 2003, pp. 57–61.
[37] X.-G. Xing, D.-Z. Zhao, Y.-G. Liu, J.-H. Yang, P. Xiu, and L. Wang, "An overview of remote sensing of chlorophyll fluorescence," *Ocean Sci. J.*, vol. 42, no. 1, pp. 19–59, 2007.
[38] P. J. Werdell and S. W. Bailey, "An improved bio-optical data set for ocean color algorithm development and satellite data product validation." *Remote Sens. Environ.*, vol. 98, pp. 122–140, 2005.
[39] S. Salcedo-Sanz, C. Casanova-Mateo, J. Muñoz-Marí, and G. Camps-Valls, "Prediction of daily global solar irradiation using temporal gaussian processes," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 11, pp. 1936–1940, Nov. 2014.
[40] J.-D. Boissonnat and F. Cazals, "Natural neighbor coordinates of points on a surface," *Comput. Geom. Theory Appl.*, vol. 19, pp. 155–173, 2001.
[41] A. K. Mishra and N. Garg, "Analysis of trophic state index of Nainital Lake from Landsat-7 ETM data," *J. Indian Soc. Remote Sens.*, vol. 39, no. 4, pp. 463–471, 2011. [Online]. Available: http://dx.doi.org/10.1007/s12524-011-0105-3
[42] M. Rowe, E. Anderson, J. Wang, and H. Vanderploeg, "Modeling the effect of invasive quagga mussels on the spring phytoplankton bloom in Lake Michigan," *J. Great Lakes Res.*, vol. 41, Suppl. 3, pp. 49–65, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0380133015000040

Katalin Blix received the B.S. degree in geosciences in 2010 from the Sogn og Fjordane University Collage, Førde, Norway, and a Civil Engineer/ M.S. degree in Technology with specialization to applied physics and mathematics in 2014 from UiT the Arctic University of Norway, Tromsø, Norway, where she is currently working toward the Ph.D. degree in remote sensing.

Gustau Camps-Valls (SM'07) received the B.Sc. degrees in physics in 1996 and in electronics engineering in 1998, and the Ph.D. degree in physics in 2002, all from the Universitat de València, Valencia, Spain.

He is currently an Associate Professor (hab. Full professor) in the Department of Electronics Engineering. His is a research coordinator in the Image and Signal Processing Group. He has been a Visiting Researcher in the Remote Sensing Laboratory (University of Trento, Trento, Italy) in 2002, the Max Planck Institute for Biological Cybernetics (Tübingen, Germany) in 2009, and as an Invited Professor in the Laboratory of Geographic Information Systems of the École Polytechnique Fédérale de Lausanne (Lausanne, Switzerland) in 2013. He is interested in the development of machine learning algorithms for geoscience and remote sensing data analysis. He is an author of 120 journal papers, more than 150 conference papers, 20 international book chapters, and editor of the books "Kernel Methods in Bioengineering, Signal and Image Processing" (IGI, 2007), "Kernel Methods for Remote Sensing Data Analysis" (Wiley, 2009), and "Remote Sensing Image Processing" (MC, 2011). He is a Coeditor of the forthcoming book "Digital Signal Processing With Kernel Methods" (Wiley, 2015). He holds a Hirsch's index h=42 (see Google Scholar page), entered the ISI list of Highly Cited Researchers in 2011, and Thomson Reuters ScienceWatch identified one of his papers on kernel-based analysis of hyperspectral images as a Fast Moving Front research.

Dr. Camps-Valls received the prestigious European Research Council (ERC) Consolidator Grant on Statistical Learning for Earth Observation Data Analysis in 2015. He is a Referee and Program Committee member of many international journals and conferences. Since 2007, he has been a member of the Data Fusion Technical Committee of the IEEE GRSS, and since 2009 of the Machine Learning for Signal Processing Technical Committee of the IEEE SPS. He is member of the MTG-IRS Science Team of EUMETSAT. He is an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE SIGNAL PROCESSING LETTERS, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and an invited Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING in 2012 and the IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE in 2015.

Robert Jenssen (M'02) received the PhD in Electrical Engineering from the University of Tromsø, in 2005.

Currently, he is an Associate Professor with the Department of Physics and Technology, UiT the Arctic University of Norway, Tromsø, Norway. He directs the Machine Learning @ UiT research group: http://site.uit.no/ml. The group's focus is on innovating information theoretic learning, kernel machines, and deep learning research, an activity that is funded, e.g., over the Norwegian Research Council's prestigious FRIPRO program, grant 239844 (Next-Generation Learning Machines). He is a Research Professor with the Norwegian Computing Center, Oslo, Norway, and a Senior Researcher in the Norwegian Center on E-Health Research. He was a Guest Researcher with the Technical University of Denmark, Kongens Lyngby, Denmark, from 2012 to 2013, Technical University of Berlin, Berlin, Germany, from 2008 to 2009, and the University of Florida, Gainesville, FL, USA, from 2002 to 2003, Ph.d. in electrical engineering, university of tromsø, 2005.

Mr. Jenssen received the Honorable Mention for the 2003 Pattern Recognition Journal Best Paper Award, the 2005 IEEE ICASSP Outstanding Student Paper Award, and the 2007 UiT Young Investigator Award. His paper "Kernel Entropy Component Analysis" was the featured paper of the May 2010 issue of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. In addition, his paper "Kernel Entropy Component Analysis for Remote Sensing Image Clustering" with L. Gomez-Chova and G. Camps-Valls received the IEEE Geoscience and Remote Sensing Society Letters Prize Paper Award in 2013.

Chapter 7

Paper 2:
Evaluation of Feature Ranking and Regression Methods for Oceanic Chlorophyll-a Estimation

Evaluation of Feature Ranking and Regression Methods for Oceanic Chlorophyll-a Estimation

Katalin Blix and Torbjørn Eltoft, *Member, IEEE*

***Abstract*—This paper evaluates two alternative regression techniques for oceanic chlorophyll-a (Chl-a) content estimation. One of the investigated methodologies is the recently introduced Gaussian process regression (GPR) model. We explore two feature ranking methods derived for the GPR model, namely sensitivity analysis (SA) and automatic relevance determination (ARD). We also investigate a second regression method, the partial least squares regression (PLSR) for oceanic Chl-a content estimation. Feature relevance in the PLSR model can be accessed through the variable importance in projection (VIP) feature ranking algorithm. This paper thus analyzes three feature ranking models, SA, ARD, and VIP, which are all derived from different fundamental principles, and uses the ranked features as inputs to the GPR and PLSR to assess regression strengths. We compare the regression performances using some common performance measures, and show how the feature ranking methods can be used to find the lowest number of features to estimate oceanic Chl-a content by using the GPR and PLSR models, while still producing comparable performance to the state-of-the-art algorithms. We evaluate the models on a global MEdium Resolution Imaging Spectrometer matchup dataset. Our results show that the GPR model has the best regression performance for most of the input feature sets we used, and our conclusion is this model can favorably be used for Chl-a content retrieval, already with two features, ranked by either the SA or ARD methods.**

***Index Terms*—Arctic, environmental monitoring, gaussian processes, optical imaging, ranking, regression analysis.**

I. Introduction

CONTINUOUS monitoring of the occurrence and distribution of phytoplankton has high ecological [1] and economical importance (http://oceancolor.gsfc.nasa.gov/). Phytoplankton content can be indirectly estimated from the chlorophyll-a (Chl-a) concentration. Similar to terrestrial plants, phytoplankton also use photosynthesis in order to live and grow. Chl-a is the key molecule for capturing light, which is the driving of photosynthesis [2]. Hence, Chl-a content is used as an indicator for several biophysical processes, which can be used for various applications.

Phytoplankton removes CO_2 from the atmosphere, through the photosynthetic process [3], and therefore the monitoring of phytoplankton via Chl-a has important relevance in climate studies [4]–[6].

Chl-a is also used to determine water quality. Eutrophication of coastal waters and lakes has been increasing in the past decades, leading to degraded water quality [7], [8]. A symptom of degraded water quality is an increase of algae biomass, which may be measured by the concentration of Chl-a. Hence, estimates of aquatic Chl-a concentration may also be used to derive information about water quality in coastal waters.

Monitoring can be achieved by optical sensors onboard satellites. It is often required to have high-spatial resolution in order to monitor water quality on a finer scale. However, optical remote sensing has its limitations with regard to spectral–spatial resolution [9], [10]. In order to achieve high-spatial resolution, the number of spectral bands is limited. Therefore, it is critical to know the number, position, and width of the bands required to retrieve Chl-a for the given aquatic condition, without loosing accuracy in the estimation.

Satellite derived Chl-a concentration is usually based on globally tuned parametric bio-optical models, such as NASA's Ocean Color (OC) models [11]–[15]. In the remainder of this paper, we refer to these models as the OC algorithms. The OC algorithms are polynomial regression models, which are trained by relating *in situ* Chl-a content to remote sensing reflectance $\mathrm{Rrs}(\lambda)$ (sr^{-1}), measured at predefined wavelengths through a so-called band ratio R. There is a variety of Chl-a content retrieval models based on band ratios [16]. In this paper, we will confine ourselves to band ratios used in NASA's OC algorithms. This band ratio is calculated at the spectral position of the Chl-a absorption peak [17], and given by $\mathrm{Rrs}(\lambda_{\mathrm{blue}})/\ \mathrm{Rrs}(\lambda_{\mathrm{green}})$ [13]. Even though these algorithms are fast, simple, and reflect the biophysical properties of aquatic Chl-a, they have certain weaknesses. This is due to the fact that the absorption spectrum varies with the amount of Chl-a concentration in the water, and it is also affected by the amount of other surfactant materials in the ocean waters near to the surface [17]. Furthermore, the coefficients of the polynomial in the OC regression models are determined by using a global training dataset. In order to allow a model to adapt to local variations, the coefficients need to be adjusted by extending the training data with measurements from the region of interest. Several studies have shown that the algorithms based on band ratios result in erroneous retrieval of Chl-a content [16]

Manuscript received September 21, 2017; revised November 28, 2017 and January 19, 2018; accepted February 22, 2018. Date of publication March 21, 2018; date of current version May 1, 2018. This work was supported by the CIRFA partners and the Research Council of Norway under Grant 237906. *(Corresponding author: Katalin Blix.)*

K. Blix is with the Department of Physics and Technology, University of Tromsø—The Arctic University of Norway, Tromsø 9037, Norway (e-mail: katalin.blix@uit.no).

T. Eltoft is with the Centre for Integrated Remote Sensing and Forecasting for Arctic Operations and the Department of Physics and Technology, University of Tromsø—The Arctic University of Norway, Tromsø 9037, Norway (e-mail: torbjorn.eltoft@uit.no).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2018.2810704

due to the regional variations of the optical properties of ocean waters. In order to overcome these difficulties in the retrieval of Chl-a content from remotely sensed data with OC models, it is important to use the correct combination of spectral bands in the computation of the band ratios.

Several studies have proposed alternative regression models for increasing the accuracy, reliability, and effectiveness in the monitoring of oceanic Chl-a content from optical remote sensing data. (For further details, we refer to [16] for a review of these algorithms.) Machine learning regression methods are known to have strong regression capabilities, and several such algorithms have been studied for Chl-a content estimation. The investigated models include neural networks [18]–[20], support vector regression [21]–[23], relevance vector regression [24], and the lately introduced Gaussian process regression (GPR) algorithm [25].

The objectives of this paper are as follows. First, we study the relevance of features (i.e., spectral bands and/or band ratios), and regression performances of two regression methods, namely the GPR [26], [27] and the partial least squares regression (PLSR) [28] models, when applied for Chl-a content estimation from satellite-based optical measurements. Both of these two models are known to have good regression performance, and both have model-tailored methods for assessing the relevance of input feature.

The GPR model uses a Bayesian approach to learn the nonlinear functional relationship between the input feature vectors and the output Chl-a measurements, and feature ranking[1] can be conducted using the automatic relevance determination (ARD) and sensitivity analysis (SA). PLSR is a well-known linear regression model, which uses a so-called latent variable space to relate the input features to the Chl-a measurement. In PLSR, feature relevance is analyzed using a ranking method called variable importance in projection (VIP). Second, using a set of regression performance measures, we evaluated the regression strength of individual spectral features and sets of spectral features, and used the performance tests to propose a lowest number of spectral bands and/or features needed to estimate Chl-a content without any significant loss of accuracy compared to the state-of-the-art OC algorithms. Finally, we include an assessment of the uncertainty level of some Chl-a estimates.

The GPR model differs from other machine learning and parametric methods in its underlying fundamental principles. Instead of proposing a function to relate Rrs to Chl-a, the GPR model learns the function by using a Bayesian approach, which has an analytic closed-form solution.

The GPR model has been shown to perform better than other machine learning methods [29] and parametric models [30] in terms of accuracy and speed for the retrieval of biophysical parameters. In addition to the estimated Chl-a content, the GPR model is able to output the certainty level of the estimates.

The relative relevance of the features being used in the regression process is not directly accessible in GPR, since it is a nonlinear kernel method. Feature relevance of Gaussian processes (GPs) in land Chl-a content estimation was proposed, computed by the so-called ARD method in [30] and [31]. Another method, the SA of GPs was introduced in [32] for oceanic Chl-a content estimation.

PLSR is an iterative statistical model, which has several advantageous properties. It can reduce colinearity and noise in the dataset, and it can provide multidimensional outputs. Feature relevance can be accessed through a measure denoted by the VIP. Lately, another method for band selection in PLSR (and random forest and support vector machine regression) was proposed in [33], the so-called ensemble approach. This study was conducted for leaf Chl-a content estimation. PLSR has been widely used in chemometrics [28], [34], and in several fields where there are a large amount of control variables with corresponding multidimensional outputs, for example, in controlling and monitoring industrial processes [35]. The PLSR model has also been successfully applied for Chl-a content estimation in optically challenging oceanic waters [36].

In this paper, we first demonstrate feature ranking by the ARD, SA, and VIP methods on two simulated datasets: a simple low-dimensional dataset, and a more complicated test example, with a very high-dimensional feature space. The purpose of these controlled experiments is to give the readers some confidence in the applied methods.

Then, we use a global Chl-a validated SeaBASS dataset [37], [38] to train the regression models and to evaluate the feature ranking methods. We conduct a performance study of the regression models discussed above with respect to estimation of Chl-a based on a MEdium Resolution Imaging Spectrometer (MERIS) dataset, and we compare feature ranking by SA, ARD, and VIP for GPR, and PLSR. Finally, we demonstrate how uncertainty can be accessed for the proposed models. Note, we have performed the same study for two additional global datasets for the SeaWiFS (Sea-Viewing Wide Field-of-View Sensor) and MODIS-Aqua (MODerate-resolution Imaging Spectroradiometer) sensors. These results are in correspondence with the results for the MERIS dataset, and presented in the Appendix.

The remainder of this paper is organized as follows. Section II reviews the GPR and PLSR models and the associated feature ranking methods. Section III illustrates the concept of the feature ranking methods on two simulated examples. Section IV details the experimental setup of this study. Section V evaluates and compares the performance of the feature ranking methods and regression models. Section VI gives the illustrative example. Finally, Section VII concludes this paper and outlines future work.

II. Feature Ranking Methods for Regression

A. Gaussian Process Regression

Here, we apply regression in the context of estimating oceanic Chl-a contents (outputs) from Rrs values (inputs) by fitting a flexible GPR model to the training data. This training dataset consists of *in situ* Chl-a contents and corresponding Rrs values measured in mgm^{-3} and sr^{-1}, respectively. Furthermore, denote Chl-a by $\{y_n\}_{n=1}^N$ and Rrs by $\{\mathbf{x}_n \in \mathbb{R}^D\}_{n=1}^N$, where $n = 1, \ldots, N$ is the number of measurements, and $d = 1, \ldots, D$ is the number of spectral bands. The GPR model assumes that the observed Chl-a content is a function (also called a latent function) of the Rrs values, and the latent function values or

[1] Feature ranking methods have refer to methods that assign relative relevance to the input features.

outputs follow a multivariate joint Gaussian distribution, if $(f(\mathbf{x}_1),\ldots,f(\mathbf{x}_N)\sim\mathcal{N}(\mathbf{0},\mathbf{K}))$, with zero mean and covariance matrix $\mathbf{K}$. The observed outputs are usually contaminated by noise ε_n, thus $y_n=f(\mathbf{x}_n)+\varepsilon_n$ for $n=1,\ldots,N$. The noise terms are assumed to be additive, independently, identically Gaussian distributed with zero mean and constant variance, i.e., $\varepsilon_n\sim\mathcal{N}(0,\sigma^2)$.

Consider now a new input Rrs data $\mathbf{x}_*$, where the goal is to estimate the corresponding output Chl-a content y_*. Then, the GP defines a joint prior distribution of the available Chl-a observations $\mathbf{y}\equiv\{y_n\}_{n=1}^N$ and the unseen y_*. This can be written by

$$\begin{bmatrix}\mathbf{y}\\ y_*\end{bmatrix}\sim\mathcal{N}\left(\mathbf{0},\begin{bmatrix}\mathbf{K}+\sigma^2\mathbf{I}_n & \mathbf{k}*\\ \mathbf{k}_*^\top & k_{**}+\sigma^2\end{bmatrix}\right) \tag{1}$$

where $\mathbf{k}_*$ is the covariance between the training vector and the test point, k_{**} is the covariance between the test point with itself, and $\mathbf{K}+\sigma^2\mathbf{I}_n$ is the noisy covariance matrix of the training inputs. Applying Bayesian inversion, it is possible to analytically compute the posterior distribution over the output y_*, given the new input, and the training dataset $\mathcal{D}$

$$p(y_*|\mathbf{x}_*,\mathcal{D})=\mathcal{N}(y_*|\mu_{\text{GP}*},\sigma^2_{\text{GP}*}) \tag{2}$$

$$\mu_{\text{GP}*}=\mathbf{k}_*^\top(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{y}=\mathbf{k}_*^\top\boldsymbol{\alpha} \tag{3}$$

$$\begin{aligned}\sigma^2_{\text{GP}*}&=\sigma^2+k_{**}-\mathbf{k}_*^\top(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k}_*\\&=\sigma^2+k_{**}-\mathbf{k}_*^\top\mathbf{A}\mathbf{k}_*\end{aligned} \tag{4}$$

where $\mathcal{D}=\left\{\mathbf{x}_n\in\mathbb{R}^D;y_n\right\}_{n=1}^N$ is the training data, $\boldsymbol{\alpha}=(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}\mathbf{y}$ is the weight vector of the GP mean, and $\mathbf{A}=(\mathbf{K}+\sigma^2\mathbf{I}_n)^{-1}$ is the weight matrix of the GP variance.

Note that the predictive mean $\mu_{\text{GP}*}$ depends on the observations through the weight vector $\boldsymbol{\alpha}$, whereas the predictive variance $\sigma^2_{\text{GP}*}$ only depends on the inverse of the covariance function $\mathbf{A}$, and σ^2 is a regularization factor. Intuitively, the predicted Chl-a content in (3) is a linear combination of the observed Chl-a content values, whereas the certainty level, (4), only depends on the Rrs values, as seen from (5). In this paper, we use the squared exponential kernel function to access similarity in the data by computing the elements of the covariance matrices. This can be written by

$$k(\mathbf{x}_m,\mathbf{x}_n)=\nu^2\exp\left(-\frac{1}{2}\sum_{d=1}^D\left(\frac{x_m^d-x_n^d}{\lambda_d}\right)^2\right) \tag{5}$$

where the length scale for band d, λ_d, and the positive scale factor, ν, are two hyperparameters of the kernel function. These hyperparameters, together with the noise variance σ^2, are optimized by maximizing the marginal likelihood of the training data. For further details on the GPR model, we refer to [26].

B. Feature Ranking for GPR

1) Automatic Relevance Determination: Relative relevance of the features can be accessed though optimizing the length-scale hyperparameters of the kernel function in (5) [30]. Since these hyperparameters control the spread of the inputs on each spectral band, small values of λ_d indicate greater relevance. Therefore, the inverses of the optimized parameters allow the ranking of the spectral bands used in the GPR model. The length-scale hyperparameter is optimized through the maximization of the marginal likelihood function with respect to the given parameter. The optimization is achieved by computing the partial derivatives with respect to λ_d of the negative log-marginal likelihood function. However, this method can result local maxima, which might lead to incorrect ranking of the spectral bands [30].

2) Sensitivity Analysis: We want to analyze the importance of spectral bands and features for a given function $\phi(\mathbf{x})$ by using a trained GPR model. To do so, let us define the sensitivity of spectral band (also called feature) j as

$$s_j=\int\left(\frac{\partial\boldsymbol{\phi}(\mathbf{x})}{\partial x_j}\right)^2p(\mathbf{x})d\mathbf{x} \tag{6}$$

where $p(\mathbf{x})$ is the probability density function over the D-dimensional input vector $\mathbf{x}_n=[x_n^1,\ldots,x_n^D]^\top$. Intuitively, the objective of the SA is to evaluate changes of the function $\boldsymbol{\phi}(\mathbf{x})$ in the jth direction. In order to avoid the possibility of cancelation of the terms due to its signs, the derivatives are squared. Therefore, the resulting sensitivities s_j will be positive for all bands and features. The empirical estimate of the sensitivity for the jth feature can be written as

$$s_j=\frac{1}{N}\sum_{n=1}^N\left(\frac{\partial\boldsymbol{\phi}(\mathbf{x}_n)}{\partial x_n^j}\right)^2 \tag{7}$$

where N denotes the number of training samples.

In our study, $\phi(\mathbf{x})$ represents the conditional mean function $\mu_{\text{GP}*}$. The resulting empirical estimate of the GP mean sensitivity is therefore obtained as follows:

$$\begin{aligned}s^j_{\mu_{\text{GP}\star}}&=\frac{1}{N}\sum_{q=1}^N\left(\frac{\partial\boldsymbol{\phi}(\mathbf{x}_q)}{\partial x_q^j}\right)^2\\&=\frac{1}{N}\sum_{q=1}^N\left(\frac{\partial\sum_{p=1}^N\alpha_pk(\mathbf{x}_p,\mathbf{x}_q)}{\partial x_q^j}\right)^2\\&=\frac{1}{N}\sum_{q=1}^N\left(\sum_{p=1}^N\frac{\alpha_p(x_p^j-x_q^j)}{\lambda_j^2}k(\mathbf{x}_p,\mathbf{x}_q)\right)^2.\end{aligned} \tag{8}$$

Note that the calculation of the empirical sensitivity is computed in closed form using the training data points and the inferred $\boldsymbol{\alpha}$.

C. Partial Least Squares Regression

Assume once again the $insitu$ Chl-a ($\mathbf{X}$) and Rrs ($\mathbf{y}$) training dataset $D\equiv\{\mathbf{X},\mathbf{y}\}$, where now the observations are collected in matrices, such that $\mathbf{X}$ is an $N\times D$ input data matrix consisting of $d=1,\ldots,D$ dimensions (spectral bands) and $n=1,\ldots,N$ observations, and let $\mathbf{y}$ be the corresponding $N\times1$ output vector (Chl-a measurements), holding $n=1,\ldots,N$ observations.

The partial least squares (PLS) model is based on introducing so-called latent variables, or X-scores, denoted by $\mathbf{T}$ ($N\times H$). $\mathbf{T}$ is relating $\mathbf{X}$ and $\mathbf{y}$, and H is the number of latent variables

(PLS components) [28], [39]. These latent variables are usually fewer than the number of features ($H < D$) and they are representing both $\mathbf{X}$ and $\mathbf{y}$ in the latent $\mathbf{T}$-space, such that the covariance between the projection of $\mathbf{X}$ and $\mathbf{y}$ in the $\mathbf{T}$-space is maximized. Then, the PLS model can be formally written by

$$\begin{aligned}\mathbf{X} &= \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}\\ \mathbf{y} &= \mathbf{T}\mathbf{c} + \mathbf{f}\\ \mathbf{T} &= \mathbf{X}\mathbf{W}^{\star}\\ \mathbf{W}^{\star} &= \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\end{aligned} \tag{9}$$

where $\mathbf{P}$ ($D \times H$) is a matrix of the *X*-loadings and $\mathbf{c}$ ($H \times 1$) is the *y*-loadings, and they are good representations (also referred to "summaries" in [28]) of $\mathbf{X}$ and $\mathbf{y}$, respectively. The term $\mathbf{W}^{\star}$ ($D \times H$) holds the weights of $\mathbf{X}$, and defines the common latent variable space (*X*-scores). The error terms, $\mathbf{E}$ ($N \times D$) and $\mathbf{f}$ ($N \times 1$), are assumed to be independent identically distributed $\sim \mathcal{N}(0, \sigma^2)$. In order to impose orthogonal latent variables ($\mathbf{T}$), the weight matrix $\mathbf{W}$ ($D \times H$) is introduced, and $\mathbf{W}$ holds the eigenvectors of the variance-covariance matrix $\mathbf{X}^{\mathbf{T}}\mathbf{Y}\mathbf{Y}^{\mathbf{T}}\mathbf{X}$. Thus, the vectors of $\mathbf{W}$ are orthonormal, and the row vectors of $\mathbf{T}$ are orthogonal to each other.

Then, the PLS model can be used for regression by expressing $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{X}\mathbf{W}^{\star}\mathbf{c} + \mathbf{f} = \mathbf{X}\mathbf{b} + \mathbf{f} \tag{10}$$

where $\mathbf{b} = \mathbf{W}^{\star}\mathbf{c}$. This way $\mathbf{y}$ can be estimated from $\mathbf{X}$, obtaining a meaningful relationship between $\mathbf{X}$ and $\mathbf{y}$. The best fit is achieved by minimizing the error term $\mathbf{f}$ in the PLSR model.

The *X*-scores, *X*- and *y*-loadings and the weights can be computed by using a PLS algorithm (an example of a PLS algorithm can be seen in Appendix A). For further details on the PLS model and the various PLS algorithms, we refer to [40]–[43] and [44] and [45].

The number of latent variables can be determined by using cross validation. However, in this paper, the training data are a multispectral dataset, where the maximum number of bands is 8 and $N >> D$, and we keep $H = D$ in the training process.

D. Feature Ranking for PLSR

Feature relevance in the PLSR model can be accessed directly from the regression coefficients $\mathbf{b}$ ($D \times 1$) in (10). However, here we focus another way to assign relevance to the input features, called the VIP method.

1) Variable Importance in Projection: The VIP_j measures the contribution to the total variance of the jth input feature ($j = 1, \ldots, D$), which is reflected by the weights (w_{hj}) from each component [46], [47]. It can be written by (note, the dataset is centered and scaled)

$$\mathrm{VIP}_j = \sqrt{D \sum_{h=1}^{H} \left(c_h^2 t_h^{\mathrm{T}} t_h\right) \left(w_{hj} / \parallel w_j \parallel^2\right) / \sum_{h=1}^{H} \left(c_h^2 t_h^{\mathrm{T}} t_h\right)}. \tag{11}$$

VIP is a measure of the contribution of each feature through the variance explained by each latent variable. The term $\left(c_h^2 t_h^{\mathrm{T}} t_h\right)$ is the variance of y explained by the hth latent variable. Thus, the VIP measure can also be expressed in term of sum of squares [48] by

$$\mathrm{VIP}_j = \sqrt{D \sum_{h=1}^{H} SS_h \left(w_{hj} / \parallel w_j \parallel^2\right) / \sum_{h=1}^{H} SS_h} \tag{12}$$

where SS_h is the percentage of y explained by the *h*th latent variable. Intuitively, the VIP value is a sum of squares, weighted by the PLS weights w_j, which takes into account the explained variance in the PLSR model. The average of the $(\mathrm{VIP}_j)^2$ is equal to one, therefore features with $\mathrm{VIP}_j > 1$ are picked as the most relevant feature [39].

III. Illustrating the Concept of the Feature Ranking Methods

In this section, we demonstrate the performance of the feature ranking methods and regression models on two controlled datasets. We simulate two cases: one simple low-dimensional and one complicated, very high-dimensional example. In both the cases, the relationship between the input and output is known, and the output is constructed to be a function of both relevant and irrelevant input features. These experiments give us some insight into the performance of the methods, and provide potentially more confidence in the results obtained, when they are applied to real data, where no ground-truth information is available.

A. Description of the Data

In the first experiment, we try to predict the response variable $\mathbf{y}$ from input vectors $\mathbf{x}_n = [x_n^1, x_n^2]$, where $x_n^1 \sim \mathcal{N}(0, 0.1)$, $x_n^2 \sim \mathcal{U}(0, 1)$, and $y_n = 2x_n^2, n = 1, \ldots, 1000$. The output (predicted mean) changes only in the second dimension, whereas it is fairly constant in the first. We expect that the feature ranking methods would identify the second dimension to be important in the prediction of y_n.

For the second experiment, we use a high-dimensional dataset $D = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$, where $\mathbf{x}_n = [x_n^1, \ldots, x_n^{1600}]$ is the input, and y_n is a scalar output, for $n = 1, \ldots, 200$. Let R^{1600} define a 1600-dimensional feature space, and let $R_i^{\kappa_i}$, for $i = 1, 2, 3, 4$, be four 121-dimensional subspaces of R^{1600}, where κ_i denote the sets of feature indices of these subspaces. Let furthermore for each $\mathbf{x}'_n \in R^{1600}$ be independent random variables distributed according to a Gaussian distribution with zero mean and 0.5 variance, $N(0, 0.5)$. Let z_n^i for $i = 1, 2, 3, 4$ be four random variables, also distributed by $N(0, 0.5)$, and let $y_n = z_n^1 z_n^2$. Define

$$x_n^j = \begin{cases} iz_n^i & \text{for } j \in \kappa_i \text{ and } i = 1, 2, 3, 4\\ x_n^j & \text{otherwise}\end{cases} \tag{13}$$

for $j = 1, \ldots, 1600$. The output y_n is hence only the product of z_n^1 and z_n^2, and we expect that the feature ranking methods (note, ARD for feature ranking could not be computed for the toy example because the optimization of the hyperparameters (λ) failed due to the high-dimensional dataset) assign relevance only to the corresponding subspaces R^{κ_1} and R^{κ_2}.

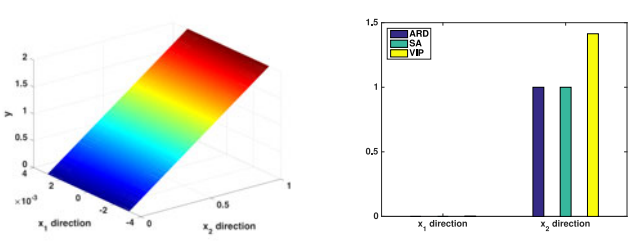


Fig. 1. Data for the first experiment (left) and the result of the feature ranking methods (right).

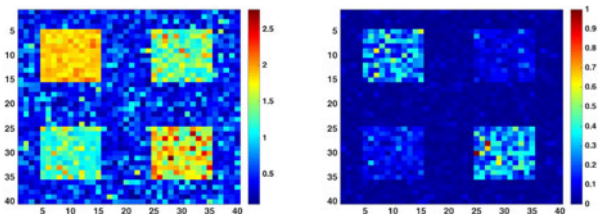


Fig. 2. Sensitivity map of the VIP (left) and the SA (right).

In this example, we also performed regression to compare the performances of the GPR and PLSR models. The toy example used here was inspired by [49].

B. Experiment 1

Fig. 1 shows the results obtained in the first example for the ARD, SA, and VIP methods. The left panel shows that the output (predicted mean) changes only in the second dimension. All the three methods can identify the feature that contributes the most to the prediction of y_n (right panel of Fig. 1).

C. Experiment 2

First, applying the SA and the VIP to the second experimental dataset returns a sensitivity map s_d, where $d = 1600$. Furthermore, transforming s_d into a matrix (image) allows the visualization of the performance of the feature ranking methods. Fig. 2 shows the sensitivity map for the VIP and the SA. The two important features z_n^1 and z_n^2 correspond to the squares in the left-top and right-bottom part, respectively, in Fig. 2. It can be observed, that both feature ranking methods could successfully identify the relevant features. However, in the case of the VIP, all features above the value 1 count as important features [47]. The computed sensitivity map (left part of Fig. 2) reveals that inputs at the top-right and bottom-left area, corresponding to z_n^3 and z_n^4, respectively, which are not relevant in the prediction of the output, were also assigned to have a sensitivity above the value of 1. Overall, VIP seems to show higher sensitivity to the irrelevant inputs than the SA.

In addition, we performed regression on this toy dataset by using the PLSR and GPR models. Fig. 3 shows the targets and the predicted values. In order to assess the strength of the regression, we computed several regression performance measures (see Table I): bias, accuracy by the normalized-root-mean-square error (NRMSE) and goodness of fit as measured by squared Pearson's correlation coefficient (R^2). It can be observed in Table I that the GPR model has the lowest bias, NRMSE values, and the highest correlation, $R^2 = 1$. Hence, the GPR model shows a better regression performance than the PLSR model. (The description of the computation of the regression performance measures can be seen in Section IV-C2.)

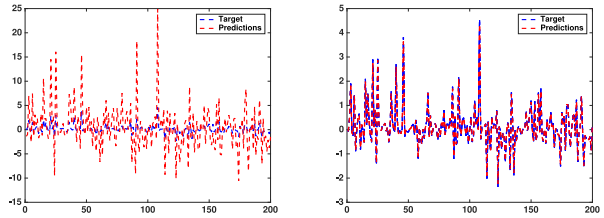


Fig. 3. Target values and predicted values for the PLSR model (left) and the GPR (right).

TABLE I
SUMMARY OF THE COMPUTED REGRESSION PERFORMANCE MEASURES FOR THE PLSR AND GPR MODEL FOR THE TOY DATA

Method	Bias	NRMSE	R^2
PLSR	2.7023	0.5208	0.5836
GPR	0.0271	0.0062	1.0000

D. Concluding Remarks

From these simulations, we may draw the following conclusions.

The first example showed that all the three feature ranking methods were sensitive to the relevant feature in the case of the low-dimensional controlled dataset.

The second experiment revealed that both the SA and VIP methods could successfully identify the important features in a very high dimensional dataset. The GPR resulted in more accurate regression than the PLSR model for this example.

Based on these experiments, we find it reasonable to apply the presented ranking methodologies to multispectral data in order to find the most relevant spectral bands in Chl-a estimation.

IV. EXPERIMENTAL SETUP

Next, we describe the experimental setup and show the results of the three ranking algorithms, the SA, ARD, and VIP, when applied to a Chl-a/Rrs matchup dataset, acquired by the ESA's MERIS sensor. (Note, we also performed the same analysis as presented below for a SeaWiFS (NASA) and a MODIS-Aqua (NASA) mathcup dataset, which have different spectral resolutions, and therefore may give slightly different conclusions [17]. The results of these analyses can be found in Appendix B.) The datasets can be obtained from the SeaBASS database (http://seabass.gsfc.nasa.gov and https://oceancolor.gsfc.nasa.gov/).

TABLE II
SUMMARY OF THE MERIS DATASET

	MERIS
Chl-a range (mgm^{-3})	0.017–40.23
No. of samples	567
Bands (λ_c (nm))	413 443 490 510 560 620 665 681
Bandwidth	10 and 7.5 nm

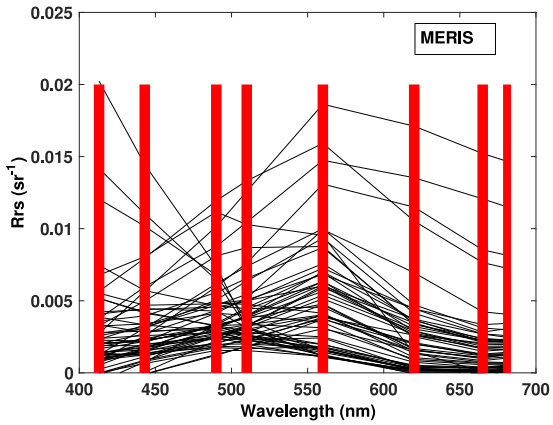


Fig. 4. Rrs (sr^{-1}) spectrum of the MERIS dataset. The red bars indicate the location of the spectral bands.

A. Description of the Dataset

Table II summarizes the MERIS dataset with respect to the center wavelength (λ_c), bandwidth, the range of the Chl-a contents, and the total number of samples.

The MERIS dataset consists of 567 measurements, measured between April 2002 and March 2012. It can be seen that the Chl-a content spans a wide range of concentration with values in the range between 0.017 and 40.23 mgm^{-3}. The bandwidth is here 10 nm for bands 1–7, and 7.5 nm for band 8.

Fig. 4 shows a few of the measured Rrs values for the MERIS dataset. The red bars indicate the position of the bands, and the width of the bars illustrates the band widths. In the following, we will number the bands chronologically 1, 2,...,8, where 1 corresponds to the smallest and 8 to the longest wavelengths.

The Rrs values show large variations across the dataset, corresponding to both Cases 1 and 2 conditions [50], which is bound to cause randomness in the estimated Chl-a contents. By definition [51] Case 1 conditions refer to waters, dominated by phytoplankton, and phytoplankton associated products, whereas Case 2 conditions can contain other constituents, and usually correspond to optically complex waters.

B. Feature Sets

1) Set A: Spectral Band Feature Set: This feature set contains eight features, the spectral bands of the MERIS dataset, ordered chronologically as noted above. Feature 1 is the band centered at 413 nm, and feature 8 corresponds to the spectral band at 681 nm.

2) Set B: Extended Spectral Band Feature Set: We extended the spectral band feature set by adding three additional features. These features are the band ratios from the OC2, OC3, and OC4 state-of-the-art models [11], [12]–[14] and [15]. These band ratios are the ratios of the measured Rrs in the blue and the green regions. The bands included in the band ratios are determined from the optical properties of the Chl-a absorption spectrum [13]. The three additional features in Set B are defined by

$$R_{\mathrm{OC2}} = \frac{\mathrm{Rrs}(490\ \mathrm{nm})}{\mathrm{Rrs}(560\ \mathrm{nm})} \tag{14}$$

$$R_{\mathrm{OC3}} = \frac{\max(\mathrm{Rrs}(443, 490\mathrm{nm}))}{\mathrm{Rrs}(560\ \mathrm{nm})} \tag{15}$$

$$R_{\mathrm{OC3}} = \frac{\max(\mathrm{Rrs}(443, 490, 510\ \mathrm{nm}))}{\mathrm{Rrs}(560\ \mathrm{nm})}. \tag{16}$$

Hence, Set B consists of 11 features, features 1–8 are the spectral bands chronologically ordered, and features 9–11 are the band ratios, corresponding to R_{OC2}, R_{OC3}, and R_{OC4}, respectively.

C. Test Setup

The test setup consists of a feature ranking analysis, and three regression performance tests.

1) Feature Ranking: First, we used Set B for ranking the relevance of the features using the SA, ARD, and VIP methods.[2] We also performed feature ranking on Set A. This was to help determine spectral bands, in the absence of the band ratio features, that are important for Chl-a retrieval, and to possibly add some insight to the physics of the problem.[3]

2) Regression: We carried out regression by splitting the dataset into 50% for training and 50% for testing. This was done by sorting the dataset based on the increasing Chl-a content. Then, we split the dataset, with odd numbers forming the training set and even numbers forming the test set, respectively. This allowed us to have approximately similar statistical variations in the training and test datasets.

Regression strength was evaluated by computing the following regression performance measures: the bias, the NRMSEs, and the squared correlation coefficient (R^2). These measures are expressed by

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}|(y_i - \hat{y}_i)| \tag{17}$$

$$\mathrm{NRMSE} = \frac{1}{y_{\max} - y_{\min}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{18}$$

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \tag{19}$$

[2]Note, in the SA method, we assume that all features can be treated as independent variables, although the band ratio features are functionally made up of other spectral bands in the feature set. Despite this fact, our results indicate that this has had no practical impact on the results.

[3]The GPR model has a computational load of $O(n^3)$. However, there are several techniques that can increase computational efficiency based on dimensionality reduction [52], and feature ranking for GPR can be an important tool in this regard.

TABLE III
COEFFICIENTS IN THE OC MODELS FOR THE MERIS DATASET

MERIS							
Model	Blue band	Green band	a_0	a_1	a_2	a_3	a_4
OC2	490	560	0.2389	− 1.9369	1.7627	− 3.0777	− 0.1054
OC3	443 > 490	560	0.2521	− 2.2146	1.5193	− 0.7702	− 0.4291
OC4	443 > 490 > 510	560	0.3255	− 2.7677	2.4409	− 1.1288	− 0.4990

TABLE IV
RANKED FEATURES FOR SET B

MERIS			
Ranked features	SA	ARD	VIP
1	R_{OC3}	665 nm	R_{OC2}
2	R_{OC2}	620 nm	R_{OC3}
3	R_{OC4}	681 nm	R_{OC4}
4	560 nm	443 nm	443 nm
5	620 nm	R_{OC2}	665 nm
6	413 nm	R_{OC3}	681 nm
7	665 nm	510 nm	413 nm
8	681 nm	490 nm	560 nm
9	443 nm	413 nm	510 nm
10	510 nm	R_{OC4}	490 nm
11	490 nm	560 nm	620 nm

TABLE V
RANKED FEATURES FOR SET A

MERIS			
Ranked bands	SA	ARD	VIP
1	560	560	560
2	413	490	413
3	620	510	510
4	443	620	620
5	665	665	490
6	681	681	443
7	510	443	665
8	490	413	681

where N is the number of observations in the test set, y is the true Chl-a content, $\hat{y}$ is the predicted Chl-a, $y_{\max}$ is the maximum observed value, $y_{\min}$ is the minimum observed value, and $\overline{y}$ is the mean of the observed Chl-a contents in the test set.

We performed regression studies in three test setups.

Test 1: First, we used Set B to evaluate the GPR and PLSR models, when only one feature was used in the regression models. For each feature, we computed the regression performance measures, and the study would hence find which single feature would result in the strongest regression.

Test 2: In the next step, we used features from Set B and gradually extended the number of features input to the regression models by sequentially adding one more feature at a time, following the order of importance determined by the SA, ARD, and VIP methods, respectively. In each case, we computed the bias, NRMSE, and R^2 values. This revealed how the number of features affected the regression performance, and how many and which features that would produce the best values for the regression performance measures. Furthermore, the three ranking methods could also be comparatively evaluated. Here, we assigned numbers to the ranked features from 1 to 11 according to the SA, ARD, and VIP. Feature number 1 corresponds to the most important feature for the given ranking method, whereas the number 11 is the least relevant feature. Hence, since the ranking methods evaluate the importance of the features differently, the actual feature associated with a given number and ranking method needs to be looked up in Table IV.

Test 3: Finally, we used Set A to perform the same sequential procedure as in Test 2. Using only the ranked spectral bands for regression, allows to determine which bands and the minimum number needed, without having a significantly decrease in regression strength. Here, the number 1 is assigned to the most relevant spectral band according to the three ranking methods, whereas the number 8 represents the least important band. The actual feature associated with a given number and ranking method needs to be looked up in Table V.

By comparing the results of Tests 2 and 3 with the results of Test 1, we can assess the increase in regression strength, when an increasing number of the ranked features are used in the regression. Hence, Test 1 can be seen as a reference performance level.

3) Comparison to the OC Models: The OC models are empirical fourth-order models. They use band ratios. The estimated Chl-a content can be expressed by

$$\text{Chl-a} = 10^{a_0 + \sum_{i=1}^{4} a_i \left(\log_{10} \left(\frac{\text{Rrs}(\lambda_{\text{blue}})}{\text{Rrs}(\lambda_{\text{green}})} \right) \right)^i} \tag{20}$$

where a_0 and a_i are the (polynomial) coefficients, $\text{Rrs}(\lambda_{\text{blue}})$ is the maximum of the measured Rrs values in the blue region, and $\text{Rrs}(\lambda_{\text{green}})$ is the measured Rrs on the green band. The sensor specific coefficients and bands used for the band ratios are listed in Table III. For further details on the OC models, we refer to NASA's OC website.[4]

4) Uncertainty of the Estimates: Last but not least, we illustrate the uncertainty of the estimates of the GPR model by choosing the strongest model, and comparing the uncertainty of the estimates with the *best* (lowest number of features and still strong regression performance) model. This shows how the uncertainty level changes when we reduce the number of bands in the GPR model.

V. RESULTS

A. Feature Ranking

Fig. 5 and Table IV summarize the results of the SA, ARD, and VIP feature ranking on the MERIS dataset, when all the

[4] *oceancolor.gsfc.nasa.gov*

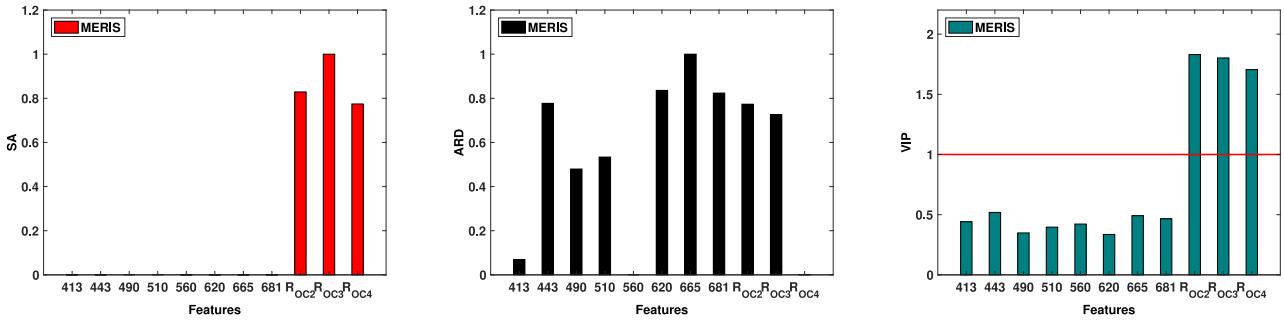


Fig. 5. SA of the GP mean (left), ARD (middle), and VIP (right) for Set B. For the VIP method, features above the red line are important in the estimation of Chl-a, whereas bands below are not likely to contribute to the prediction.

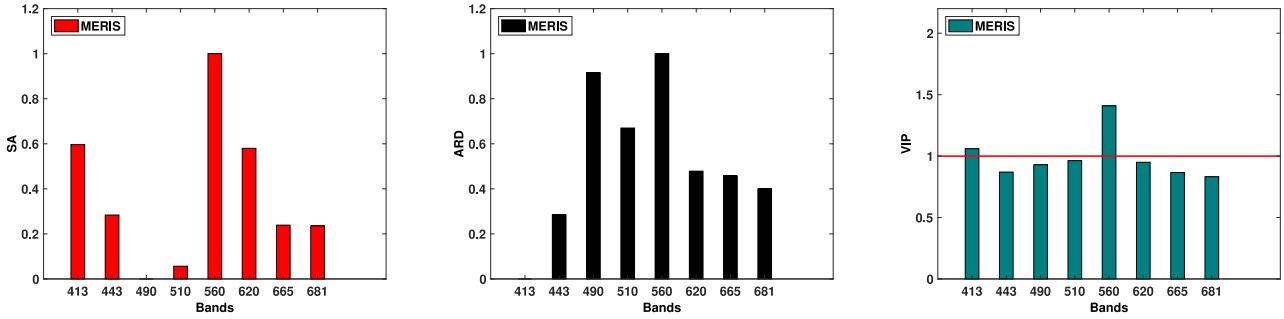


Fig. 6. SA of the GP mean (right), ARD (middle), and VIP (left) for Set A. For the VIP method, features above the red line are important in the estimation of Chl-a, whereas bands below are not likely to contribute to the prediction.

features were used (Set B). As can be seen, both the SA and the VIP methods assigned the highest relevance to the three band ratios, but they ranked their individual relevance differently. SA gave highest relevance to R_{OC3}, R_{OC2}, and R_{OC4}, in that order, and compared to these features; the relevance of the eight band features is more or less negligible. In the VIP method, only features with score above 1 are considered important. As seen in the left panel of Fig. 5, all band ratio features are scored above 1, whereas the band features are below, and hence, less important.

The ARD method (middle panel of Fig. 5) ranked the features differently. It gave highest relevance to the band centered at 665 nm, and high relevance to the bands at 443 and 620 nm, in addition to the band ratio features. However, except for the bands 560 and 413 nm, both of which have very low scoring, the relative differences in importance for ARD are not as pronounced as for the other two ranking methods.

Fig. 6 and Table V show the results of the ranking methods, when only the spectral bands are used (Set A). As can be seen, all the ranking methods assigned high relevance to the band positioned at 560 nm. This band is the denominator in all band ratio features, since this is a reference band because there is little or no absorption by Chl-a in this region [53], and the results reconfirm its importance. SA gave high importance also to the bands at 413 and 620 nm, whereas VIP, in addition to 560 nm, only gave the band at 413 nm a score above 1. The other bands are scored slightly below 1. ARD also puts high relevance to the bands at 490 and 510 nm. Both these bands are included in the R_{OC4} band ratio. In summary, these results suggest that the bands used in the band ratios (560 and 490 nm) are important. The high relevance of the band at 413 nm, as suggested by both SA and VIP, may be explained by the fact that the dataset also includes samples from eutrophic waters.

B. Regression Experiments

Test 1: Fig. 7 shows the three regression measures, bias, NRMSE, and R^2, for the single feature regression setup for the GPR model (upper panel) and for the PLSR model (lower panel). The numbers ticked on the x-axis are denoting band features in increasing order of wavelength, and the solid red line horizontally across each figure is inserted as a reference to ease the visual comparisons. As noted, using only one feature at a time, both the GPR and PLSR models resulted in the best performance for the three band ratio features, and according to the measures, all three have approximately similar regression strength (i.e., lowest bias and NRMSE, and the highest R^2). We also note that the four bands with the longest wavelengths, i.e., those centered at 560, 620, 665, and 681 nm, showed good regression performance, especially for the GPR model. For this regression model, the four band features with the shortest wavelengths showed a significantly weaker individual regression strength. On the other hand, for the PLSR model, this difference in performance between the short and the long wavelengths is less pronounced.

Test 2: In this experiment, we gradually extend the number of features input to the regression models by sequentially adding one more feature at a time, following the order of importance

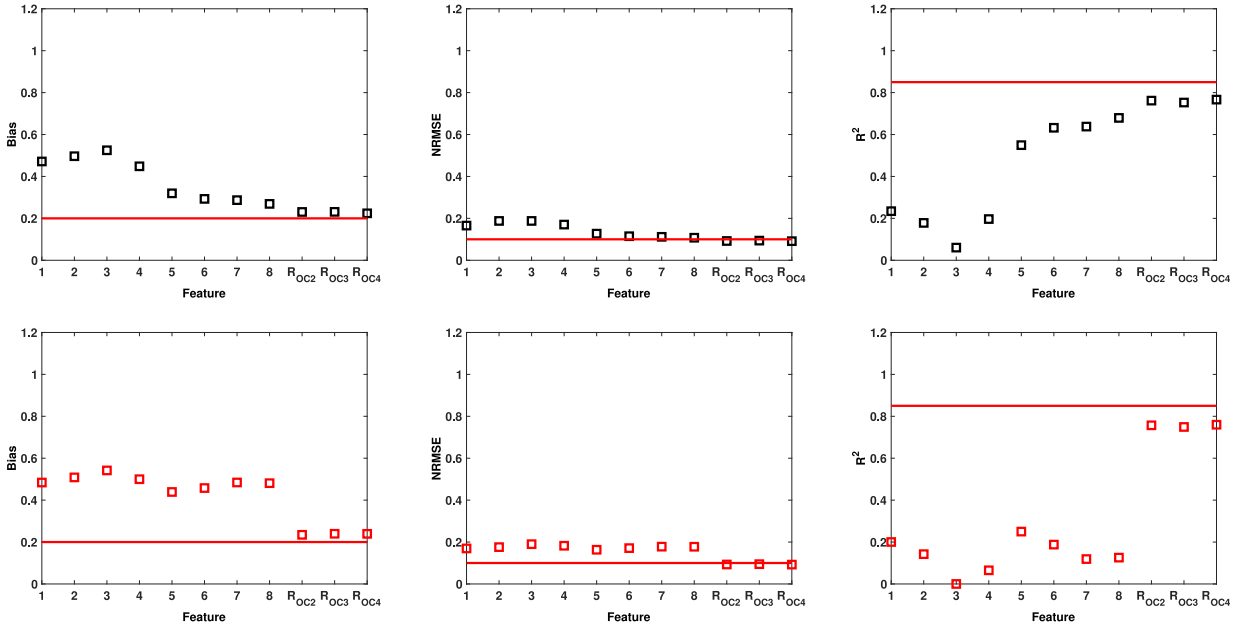


Fig. 7. Computed measures for the MERIS dataset for the GPR (top row) and PLSR (bottom row) model. Regression performance measures were computed by performing regression with only one feature at the time for Set B. (The red line is a reference line, allowing an easier comparison of the performance of the GPR and PLSR models.)

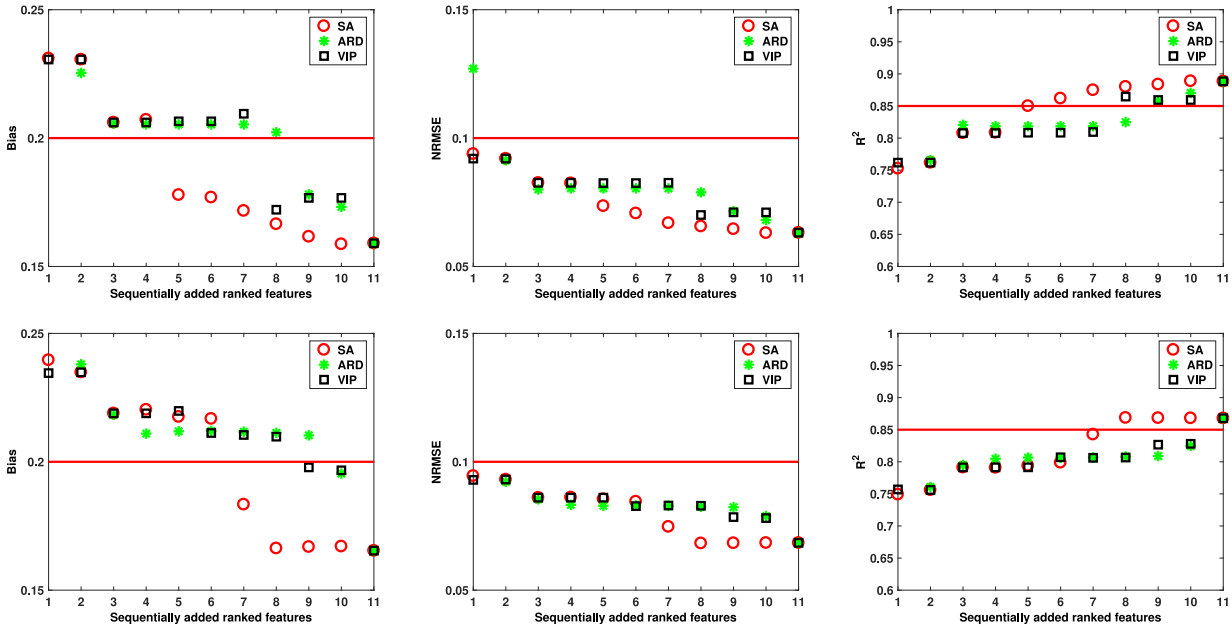


Fig. 8. Computed measures for Set B for the GPR (top row) and PLSR (bottom row) model. Here, the numbers represent the ranked features of the SA (red circle), ARD (green star), and VIP (black square) methods. The ranked features were added sequentially as inputs to the GPR and PLSR models. (The red line is a reference line, allowing an easier comparison of the performance of the GPR and PLSR models.)

determined by the SA, ARD, and VIP methods. The resulting regression performance measures as function of the number of ranked input features are summarized in Fig. 8 for GPR (upper panel) and PLSR (lower panel). The figures show that for both regression models, the regression performance improves as more and more input features are used. The best regression performance is achieved when as many as ten features, ranked by the SA method, were applied to the GPR model. This only excludes the band positioned at 490 nm, but this band is already contributing to the regression, as it is included in the band ratios. We also note that the *improvement curves*, i.e., the reduction in bias and NRSME and increase in R^2 from left to right, vary with ranking method and also with the regression model. We note that the curves associated with the SA ranking in general provide the best regression performance for both GPR and PLSR. For the SA ranking, we observe that the GPR

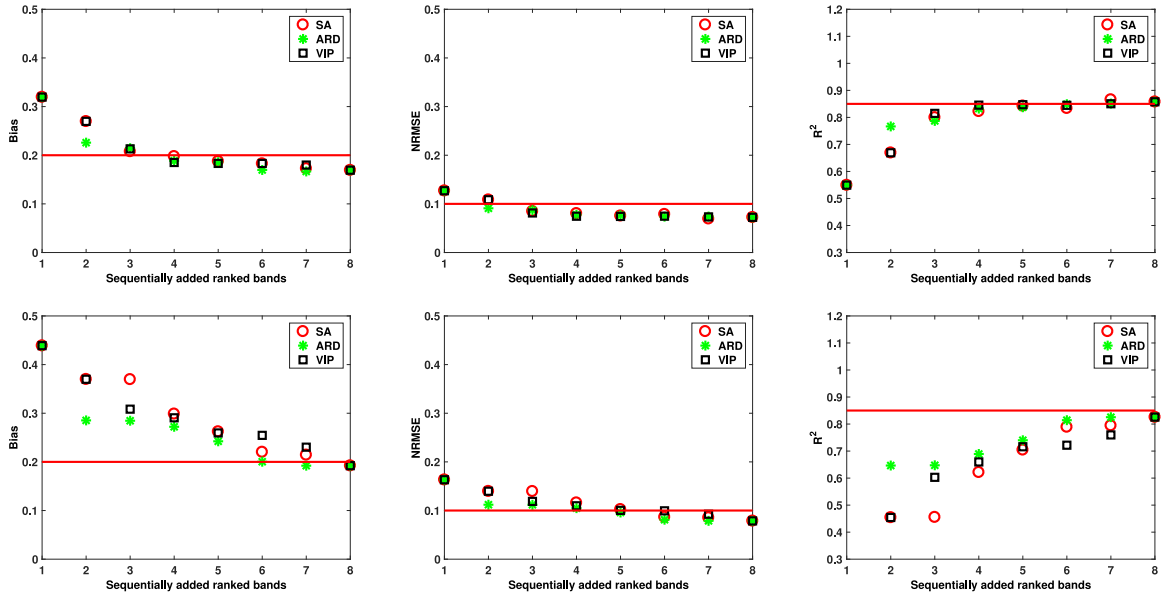


Fig. 9. Computed measures for Set A for the GPR (top row) and PLSR (bottom row) model. Here, the numbers represent the ranked features of the SA (red circle), ARD (green star), and VIP (black square) methods. The ranked features were added sequentially as inputs to the GPR and PLSR models. (The red line is a reference line, allowing an easier comparison of the performance of the GPR and PLSR models.)

curves have a clear stepwise trend, with big improvement steps at 3 and 5 feature inputs. There is basically no improvement of adding R_{OC2} to R_{OC3}, but significant improvements when also R_{OC4} is used. Similarly, there is little change in the measures by adding band feature 560 nm, but big improvement when band 620 nm is added. The curves for the PLSR model also have stepwise appearance, but the steps are at different numbers of feature inputs, and the curves seem to achieve the optimal performance with eight number of features following the SA ranking.

The GPR and PLSR models showed similar trends in performance when extending the feature sets. Also, the values for the performance measures were quite similar. The most noticeable difference occurred in the NRMSE value, where the GPR model showed a slightly lower value.

Test 3: Adding sequentially the ranked band features to the GPR and PLSR models revealed improvements, already when the second most important band was added (see Fig. 9). We observe that in general, the band features ranked by the ARD method showed the best regression performance measures as we extended the input sets, both for the GPR and PLSR models. Again we note that the ARD ranked the bands at 560 and 490 nm as the most relevant bands, and these bands correspond to the bands used in the R_{OC2} band ratio. However, the GPR model converged to a higher R^2 value, when many features were used, and the overall best performance was achieved with the GPR model using all features.

C. Comparison to the OC Models

Finally, we compared the regression performance of the GPR and PLSR models with the OC2, OC3, and OC4 models. These comparisons are summarized in Table VI in terms of resulting performance measures associated with some selected input features. The first nine rows display the performance measures associated with the three OC models and the GPR and the PLSR models using R_{OC2}, R_{OC3}, and R_{OC4} as inputs, respectively. Note that both the GPR and PLSR perform better than all the OC models. The best result is obtained with GPR using R_{OC4} as input feature. In the next rows, we present the numerical results for the GPR and PLSR using 1, 2, 8, and 10 input features, as described below.

TABLE VI
COMPARISON OF THE OC MODELS WITH GPR AND PLSR MODELS FOR THE MERIS DATASET

Regression model	Bias	NRMSE	R^2
OC2	0.2715	0.1114	0.7101
GPR with R_{OC2}	0.2306	0.0920	0.7618
PLSR with R_{OC2}	0.2345	0.0929	0.7570
OC3	0.2676	0.1090	0.7241
GPR with R_{OC3}	0.2311	0.0938	0.7526
PLSR with R_{OC3}	0.2391	0.0945	0.7491
OC4	0.2347	0.0949	0.7671
GPR with R_{OC4}	0.2241	0.0912	0.7666
PLSR with R_{OC4}	0.2392	0.0924	0.7598
GPR with band centered at 665 nm	0.2861	0.1143	0.6521
PLSR with band centered at 665 nm	0.4842	0.1780	0.1189
GPR with band centered at 681 nm	0.2691	0.1076	0.6793
PLSR with band centered at 681 nm	0.4809	0.1776	0.1257
GPR with bands centered at 490 and 560 nm	**0.2259**	**0.0909**	**0.7666**
PLSR with bands centered at 490 and 560 nm	0.2853	0.1120	0.6465
GPR with all bands	0.1692	0.0723	0.8578
PLSR with all bands	0.1922	0.0787	0.8257
GPR with all features, except band centered at 490 nm	**0.1587**	**0.0631**	**0.8889**
PLSR with all features, except band centered at 490 nm	0.1670	0.0684	0.8680

1) *Single band feature*: We display the regression performance of band features 7 (665 nm) and 8 (681 nm) when used as single feature inputs. These correspond to the band features with best performance, in the single feature experiment (see Fig. 7).
2) *Two band features*: Here, we display the regression performance when two bands define the input vector. We have chosen the bands at 560 and 490 nm, which according to Fig. 9 would give the best performance (ARD ranking).
3) *All band features*: For comparison, we also include the results when using all band features as input.
4) *Ten input features*: We finally combine the 3 band ratio features with 7 bands according to the SA ranking. This resulted in the overall best regression performance (see Fig. 8).

The GPR model with ten input features showed the strongest regression strength, and it actually outperforms all the OC models. The second strongest model was the PLSR with the same features. Both GPR and PLSR performed well in comparison to the OC models, also with few input bands. Hence, the results of Table VI suggest that Chl-a content retrieval can be improved in comparison to the OC models by using the GPR model with only two bands.

D. Uncertainty Level of the GPR Model

Based on our results in Section V-C, we illustrate the advantageous property of the GPR model, i.e., its ability to assign an uncertainty level to the estimates. Fig. 10 (top) shows the estimated Chl-a values by using the obtained strongest GPR model (see Table VI), the actual measured Chl-a values and the uncertainty level of the estimates for the MERIS dataset. Fig. 10 (bottom) shows how the uncertainty level changes when the GPR model uses only two spectral bands, 490 and 560 nm.

For some of the estimated values, the uncertainty level increases slightly when fewer features are used. This is in good correspondence with the computed regression performance measures.

The interesting observation to note is that the uncertainty level does not reveal a significant increase when only the two most important spectral bands are used in model, compared to the strongest GPR model, with ten input features.

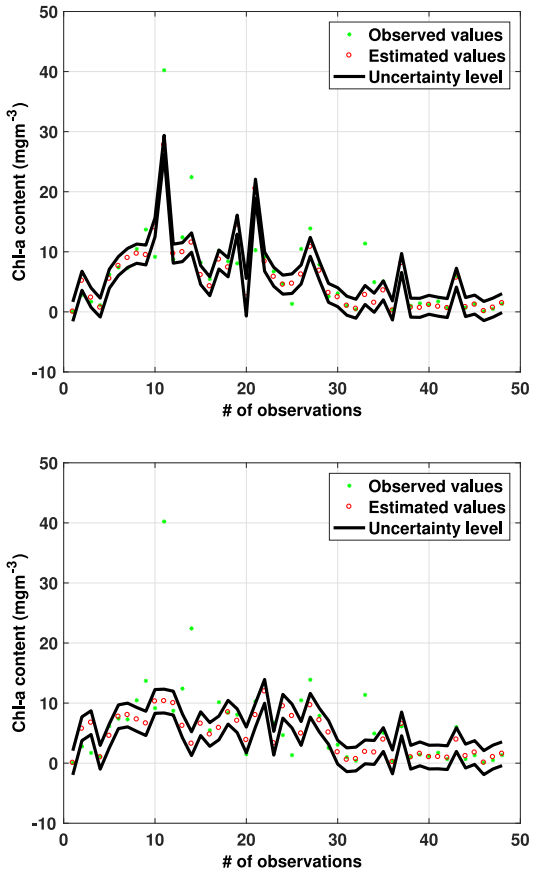


Fig. 10. Observed and estimated Chl-a contents by using the GPR model for the MERIS dataset. The black solid lines indicate the uncertainty level of the estimates. The top figure shows some of the observations by using the GPR model with the features that resulted the best values for the regression performance measures. The bottom panel illustrates the GPR model by using only two bands for regression.

VI. Illustrative Example

We illustrate the effect of using different algorithms and bands for Chl-a estimation on a test image acquired in July 2015 by MODIS-Aqua over high-latitude Arctic oceans (N = 89.9931°, S = 65.7186°, W = −174.2612°, E = 5.236°). Sea ice concentration was estimated to 33.7383% and the cloud coverage was 51.908%. The quasi-true color image can be seen in Fig. 11.

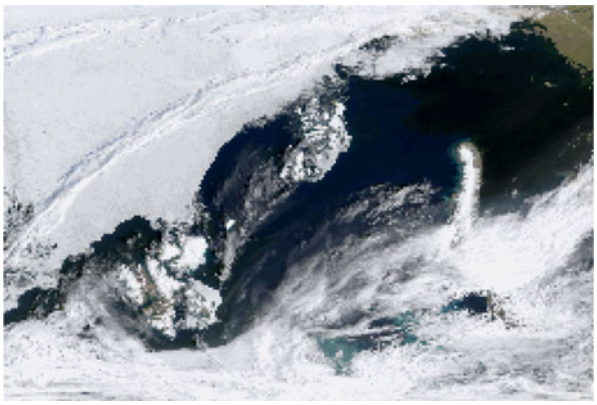


Fig. 11. Quasi-true color image of the test site.

Fig. 12 shows the Chl-a content maps estimated by the OC3 algorithm (top), by the GPR model with all the spectral bands (middle), and by the GPR model with bands centered at 488 and 678 nm (bottom). We observe that the estimated Chl-a maps have some differences. The GPR maps show lower concentration values, and reveal more details than the map of the OC3 algorithm. Both GPR estimates illustrate how the model captures internal structures, presumably associated to ocean current eddies. It is expected that phytoplankton blooms follow the pattern of the current eddies.

The corresponding computed regression performance measures for these cases were (see Table XII in Appendix B): bias = 0.2272, NRMSE = 0.1057, R^2 = 0.7868 (OC3); bias = 0.1628, NRMSE = 0.0702, R^2 = 0.8844 (GPR, all bands); and bias = 0.1774, NRMSE = 0.0746, R^2 = 0.8684 (GPR, 488 and

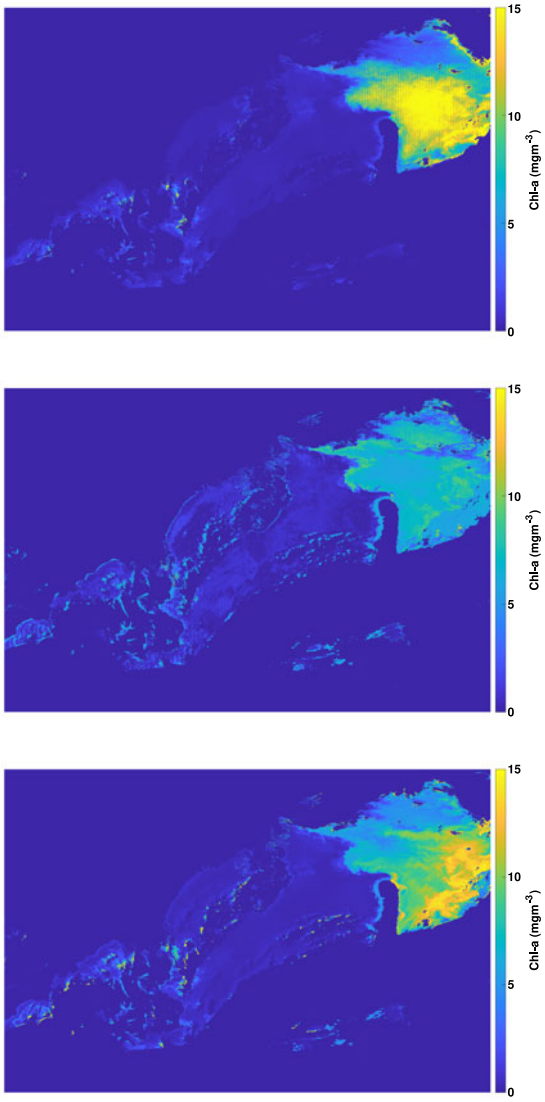


Fig. 12. Chl-a content estimates by using the OC3 algorithm (top), GPR model with all bands (middle), and the GPR model with bands centered at 488 and 678 nm (bottom).

678 nm). Based on these measures, the best results should be expected to be achieved by using the GPR model with all bands, followed by the GPR model, and with bands centered at 488 and 678 nm. The most pronounced difference between the maps computed by the GPR model with all bands (middle panel) and GPR with bands centered at 488 and 678 nm (bottom panel) is in the amount of the assigned Chl-a content. Which one is most correct, cannot be concluded without *in situ* information. This example, together with the computed statistics from the training set, shows that both the regression model and the input feature vector are important in OC applications, and that more research is needed to select the most reliable methodology.

VII. Conclusion and Future Work

In this paper, we studied feature ranking and regression performances of two regression methods, namely the GPR and the PLSR models, when applied for Chl-a content estimation based on a global MERIS dataset. In the GPR model, we use a Bayesian approach to learn the nonlinear functional relationship between the input feature vectors and the output Chl-a measurements, and the feature ranking was conducted using the ARD and the SA. The PLSR is a well-known linear regression model, which uses a so-called latent variable space to relate the input features to the Chl-a measurement. In PLSR, feature relevance was analyzed using a ranking method called VIP.

From the eight spectral bands of a MERIS matchup data set, we created two input feature sets. One (Set A) consisting of all the spectral bands, and the another extended feature set (Set B), which in addition to all bands, also consisted of the three band ratio features, denoted by R_{OC2}, R_{OC3}, and R_{OC4}, which are the inputs used in the state-of-the-art OC2, OC3, and OC4 regression models. The relevance of features were analyzed by all the ranking methods, and subsequently input to the two regression models in a test setup consisting of three tests. Using three measures, the bias, NRSME, and R^2, the individual regression strength of each feature as single input was computed. Next, we evaluated the regression strength of sets of features by gradually extending the number of features, adding one more feature at a time, following the order of importance determined by the SA, ARD, and VIP methods, respectively. We did this analysis first for Set B, and then for Set A.

Our results show that the all feature ranking methods can successfully assign sensible relevance to the features. Since the methods operate according to different ranking criteria, it is expected that they might assess the features differently. Both SA and VIP assigned the highest relevance to the three band ratio features, whereas ARD gave highest scores to the spectral bands centered at 665, 443, and 620 nm. ARD also found the band ratio features to be important. When applied to the features in Set A, i.e., only the band features, all the three methods agreed to give highest relevance to the band at 560 nm, but the order of the next relevant bands was somewhat different.

ARD ranked the spectral band at 490 nm as second most important. We note that the bands at 490 and 560 nm are included in all the band ratio features. The spectral region at 490 nm corresponds to the shifted Chl-a absorption peak, and has been used to avoid contribution from CDOM [54]. Even though this wavelength mostly represents accessory pigments, due to the correlation of these accessory pigments with Chl-a, the spectral band at 490 nm can successfully be used to derive Chl-a concentration. The band centered at 560 nm is a reference wavelength, since phytoplankton absorption is at the minimum around this green band [54].

In regression Test 1 (single feature input), we found that the three band ratio features achieved far the best regression performance in both regression models. This is not surprising, given the fact that these features are composed of two spectral bands, carefully selected. We also found that the spectral bands with the longest wavelengths, i.e., 560, 620, 665, and 681 nm, were performing significantly better than the shorter wavelengths, especially for the GPR model. It has previously been shown that bands in the red part of the visible region of the electromagnetic spectrum can successfully be applied for Chl-a retrieval due to the second absorption maximum of the Chl-a molecule [13]. Our results support this finding.

Not surprisingly, Test 2 showed that for both regression models, the regression performance improved as more and more input features are used as input. The best regression performance is achieved when ten features, ranked by the SA method, were applied to the GPR model. We note that the curves associated with the SA ranking in general provided the best regression performance for both GPR and PLSR. We also found that the improvement curves of GPR associated with the SA ranking had a clear stepwise trend, with big improvement steps at 3 and 5 feature inputs. These jumps would intuitively give clues to which features to select first, if dimensionality reduction were to be applied to the input feature space.

Test 3 also showed that the regression performance gradually improved when applying more and more input spectral bands, and the overall best performance was obtained with all eight spectral bands. However, we note that the gain in performance for the GPR model by increasing the number of input spectral bands from 4 to 8 is only minor (see Fig. 9).

Our comparisons between the GPR and PLSR regression models and the three OC models clearly demonstrated that both GPR and PLSR performed better than the state-of-the-art models for several different sets of input features. Based on the performance measures we have used, we also find that the GPR model in all cases has the strongest regression performance.

Note that we also performed the same study for two additional global matchup datasets: the SeaWiFS and MODIS-Aqua datasets. Due to lack of space, we did not include a detailed description of these studies here, but some important results have been tabulated in Appendix B. We found similar results for the SeaWiFS and MODIS-Aqua datasets as those reported in this paper for the MERIS dataset.

Based on the current studies, we conclude that there is a big potential for improvements in Chl-a retrieval from satellite-based observations by selecting the most appropriate regression model in combination with an optimal set of input features.

For future work, we plan to perform extensive validation studies of the GPR model, and compare its performance with the state-of-the-art Chl-a retrieval algorithms, and other algorithms, e.g., neural networks, on optically complex aquatic environments, such as coastal and Arctic waters, and midlatitude shallow lakes.

APPENDIX A

PLSR Algorithm

Here, we present the so-called nonlinear iterative PLS (NIPALS) algorithm introduced by [55]. The NIPALS algorithm can be written by

$$\begin{aligned}
&\textbf{for } h = 1, \ldots, H \\
&\quad \mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \\
&\quad \mathbf{t} = \mathbf{X}\mathbf{w} \\
&\quad c = \mathbf{y}^{\mathrm{T}}\mathbf{t}/\mathbf{t}^{\mathrm{T}}\mathbf{t} \\
&\quad \mathbf{p} = \mathbf{X}^{\mathrm{T}}\mathbf{t}/\mathbf{t}^{\mathrm{T}}\mathbf{t} \\
&\quad \mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}^{\mathrm{T}} \\
&\quad \mathbf{y} = \mathbf{y} - \mathbf{t}c \\
&\textbf{end for} \\
&\mathbf{W} = \mathbf{w}_1, \ldots, \mathbf{W}_H \\
&\mathbf{T} = \mathbf{t}_1, \ldots, \mathbf{t}_H \\
&\mathbf{P} = \mathbf{p}_1, \ldots, \mathbf{p}_H \\
&\mathbf{c} = c_1, \ldots, c_H . \qquad (21)
\end{aligned}$$

APPENDIX B

We performed the same experiments as in Sections IV and V for two additional datasets: the SeaWiFS and MODIS-Aqua datasets. We found that the results were in good correspondence with our findings for the MERIS dataset, namely a satisfactory regression can be already achieved by using the spectral bands centered at 490 and 555 nm for the SeaWiFS dataset, and 488 and 678 nm for the MODIS-Aqua dataset.

In the following, we present the description of the datasets and features, and the results of the feature ranking methods, regression models, and comparisons for these additional datasets.

The SeaWiFS and MODIS-Aqua Datasets

The SeaWiFS and MODIS-Aqua datasets are summarized in Table VII. These datasets represent both Cases 1 and 2 conditions. Table VIII shows the coefficients in the OC models for the SeaWiFS ans MODIS-Aqua datasets.

TABLE VII
SUMMARY OF THE SEAWIFS AND MODIS-AQUA DATASETS

SeaWiFS	
Chl-a range (mgm^{-3})	0.024–129.332
No. of samples	1465
Bands (λ_c (nm))	421 443 490 510 555 670
Bandwidth	20 nm
MODIS-Aqua	
Chl-a range (mgm^{-3})	0.0153–25.4985
No. of samples	579
Bands (λ_c (nm))	412 443 488 531 547 667 678
Bandwidth	10 nm, 15 nm

The band ratio features for the SeaWiFS dataset can be written by

$$R_{\mathrm{OC2}} = \frac{\mathrm{Rrs}(490\ \mathrm{nm})}{\mathrm{Rrs}(555\ \mathrm{nm})} \qquad (22)$$

$$R_{\mathrm{OC3}} = \frac{\max(\mathrm{Rrs}(443, 490\ \mathrm{nm}))}{\mathrm{Rrs}(555\mathrm{nm})} \qquad (23)$$

$$R_{\mathrm{OC3}} = \frac{\max(\mathrm{Rrs}(443, 490, 510\ \mathrm{nm}))}{\mathrm{Rrs}(555\mathrm{nm})} \qquad (24)$$

TABLE VIII
COEFFICIENTS IN THE OC MODELS FOR THE SEAWIFS AND MODIS-AQUA DATASETS

Model	Blue band	Green band	a_0	a_1	a_2	a_3	a_4
SeaWiFS							
OC2	490	555	0.2511	−2.0853	1.5035	−3.1747	0.3383
OC3	443 > 490	555	0.2515	−2.3798	1.5823	−0.6372	−0.5692
OC4	443 > 490 > 510	555	0.3272	−2.9940	2.7218	−1.2259	−0.5683
MODIS-Aqua							
OC2	488	547	0.2500	−2.4752	1.4061	−2.8233	0.5405
OC3	443 > 488	547	0.2424	−2.7423	1.8017	0.0015	−1.2280

TABLE IX
RANKED FEATURES FOR THE SEAWIFS AND MODIS-AQUA DATASETS

SeaWiFS			
Ranked features	SA	ARD	VIP
1	R_{OC4}	490 nm	R_{OC4}
2	R_{OC2}	R_{OC2}	R_{OC3}
3	R_{OC3}	R_{OC4}	R_{OC2}
4	490 nm	443 nm	412 nm
5	510 nm	412 nm	555 nm
6	555 nm	670 nm	490 nm
7	412 nm	510 nm	510 nm
8	443 nm	R_{OC3}	443 nm
9	670 nm	555 nm	670 nm
MODIS-Aqua			
Ranked features	SA	ARD	VIP
1	R_{OC3}	678 nm	R_{OC3}
2	R_{OC2}	531 nm	R_{OC2}
3	412 nm	R_{OC2}	488 nm
4	443 nm	412 nm	547 nm
5	547 nm	547 nm	531 nm
6	488 nm	443 nm	678 nm
7	678 nm	667 nm	443 nm
8	667 nm	488 nm	412 nm
9	488 nm	R_{OC3}	667 nm

and for the MODIS-Aqua dataset by

$$R_{OC2} = \frac{\text{Rrs}(488\text{ nm})}{\text{Rrs}(547\text{ nm})} \tag{25}$$

$$R_{OC3} = \frac{\max(\text{Rrs}(443, 488\text{ nm}))}{\text{Rrs}(547\text{ nm})}. \tag{26}$$

Feature Ranking for the SeaWiFS and MODIS-Aqua Datsets

The ranked features can be seen in Table IX, and the ranked spectral band features are presented in Table X.

Regression

Tables XI and XII show the computed regression performance measures for the SeaWiFS and MODIS-Aqua datasets, respectively.

TABLE X
RANKED SPECTRAL BANDS FOR THE SEAWIFS AND MODIS-AQUA DATASETS

SeaWiFS			
Ranked bands	SA	ARD	VIP
1	412	555	555
2	555	490	412
3	443	443	670
4	670	412	443
5	490	670	510
6	510	510	490
MODIS-Aqua			
Ranked bands	SA	ARD	VIP
1	488	488	547
2	678	678	412
3	547	412	531
4	667	531	443
5	412	547	488
6	443	667	678
7	531	443	667

TABLE XI
COMPARISON OF THE OC MODELS WITH GPR AND PLSR MODELS FOR THE SEAWIFS DATASET

Regression model	Bias	NRMSE	R^2
OC2	0.2319	0.0990	0.8128
GPR with R_{OC2}	0.2142	0.0908	0.8403
PLSR with R_{OC2}	0.2159	0.0915	0.8373
OC3	0.2275	0.0977	0.8180
GPR with R_{OC3}	0.2129	0.0914	0.8380
PLSR with R_{OC3}	0.2225	0.0939	0.8289
OC4	0.2123	0.0907	0.8406
GPR with R_{OC4}	0.2079	0.0890	0.8464
PLSR with R_{OC4}	0.2302	0.0952	0.8243
GPR with bands centered at 490 and 555 nm	**0.2101**	**0.0894**	**0.8450**
PLSR with bands centered at 490 and 555 nm	0.2592	0.1082	0.7724
GPR with all bands	**0.1792**	**0.0780**	**0.8820**
PLSR with all bands	0.2394	0.1019	0.7980
GPR with bands centered at 412, 443, 490, 670 nm, and features R_{OC2} and R_{OC4}	**0.1805**	**0.0780**	**0.8820**
PLSR with bands centered at 412, 443, 490, 670 nm, and features R_{OC2} and R_{OC4}	0.2021	0.0854	0.8583

TABLE XII
COMPARISON OF THE OC MODELS WITH GPR AND PLSR MODELS FOR THE MODIS-AQUA DATASET

Regression model	Bias	NRMSE	R^2
OC2	0.2270	0.1098	0.7697
GPR with R_{OC2}	0.2072	0.0894	0.8115
PLSR with R_{OC2}	0.2123	0.0910	0.8053
OC3	0.2272	0.1057	0.7868
GPR with R_{OC3}	0.2062	0.0880	0.8173
PLSR with R_{OC3}	0.2304	0.0958	0.7847
GPR with band centered at 678 nm	0.2713	0.1119	0.7042
PLSR with band centered at 678 nm	0.3760	0.1559	0.4403
GPR with bands centered at 488 and 678 nm	**0.1774**	**0.0746**	**0.8684**
PLSR with bands centered at 488 and 678 nm	0.2648	0.1064	0.7379
GPR with all bands	**0.1628**	**0.0702**	**0.8844**
PLSR with all bands	0.2049	0.0856	0.8271
GPR with bands centered at 412, 531 and 678 nm, and features R_{OC2}	**0.1697**	**0.0725**	**0.8771**
PLSR with bands centered at 412, 531 and 678 nm, and features R_{OC2}	0.1891	0.0837	0.8350

REFERENCES

[1] C. S. Reynolds, *The Ecology of Phytoplankton*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[2] Govindjee, *Bioenergetics of Photosynthesis*. New York, NY, USA: Academic, 1975.

[3] T. Volk and M. I. Hoffert, *Ocean Carbon Pumps: Analysis of Relative Strengths and Efficiencies in Ocean-Driven Atmospheric CO_2 Changes*. Washington, DC, USA: Amer. Geophys. Union, 2013. [Online]. Available: http://dx.doi.org/10.1029/GM032p0099

[4] K. R. Arrigo *et al.*, "Phytoplankton community structure and the drawdown of nutrients and CO_2 in the southern ocean," *Science*, vol. 283, no. 5400, pp. 365–367, 1999. [Online]. Available: http://science.sciencemag.org/content/283/5400/365

[5] M. Hein and K. Sand-Jensen, "CO_2 increases oceanic primary production," *Nature*, vol. 388, pp. 526–527, 1997.

[6] M. Hofmann, B. Worm, S. Rahmstorf, and H. J. Schellnhuber, "Declining ocean chlorophyll under unabated anthropogenic CO_2 emissions," *Environ. Res. Lett.*, vol. 6, no. 3, pp. 34–35, 2011. [Online]. Available: http://stacks.iop.org/1748-9326/6/i=3/a=034035

[7] N. T. T. Ha, K. Koike, and M. T. Nhuan, "Improved accuracy of chlorophyll-a concentration estimates from modis imagery using a two-band ratio algorithm and geostatistics: As applied to the monitoring of eutrophication processes over Tien Yen Bay (Northern Vietnam)," *Remote Sens.*, vol. 6, no. 1, pp. 421–442, 2014. [Online]. Available: http://www.mdpi.com/2072-4292/6/1/421

[8] X.-e. Yang, X. Wu, H.-l. Hao, and Z.-l. He, "Mechanisms and assessment of water eutrophication," *J. Zhejiang Univ. Sci. B*, vol. 9, no. 3, pp. 197–209, Mar. 2008. [Online]. Available: https://doi.org/10.1631/jzus.B0710626

[9] F. A. Al-Wassai, and N. V. Kalyankar, "Major limitations of satellite images," *J. Global Research Comput. Sci.*, vol. 4, no. 5, pp. 51–59, May 2013.

[10] J. Hill, C. Diemer, O. Stöver, and T. Udelhoven, "A local correlation approach for the fusion of remote sensing data with different spatial resolutions in forestry applications," in *Proc. Int. Archives Photogrammetry Remote Sens.*, 1999, vol. 32, pt. 7-4-3 W6, pp. 3–4.

[11] C. Hu, Z. Lee, and B. Franz, "Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference," *J. Geophys. Res.*, vol. 117, 2012, Art. no. C01011.

[12] A. Morel and S. Maritorena, "Bio-optical properties of oceanic waters: A reappraisal," *J. Geophys. Res., Oceans*, vol. 106, no. C4, pp. 7163–7180, 2001. [Online]. Available: http://dx.doi.org/10.1029/2000JC000319

[13] J. E. O'Reilly *et al.* "Ocean color chlorophyll algorithms for SeaWiFS," *J. Geophys. Res.*, vol. 103, pp. 24937–24953, 1998.

[14] J. E. O'Reilly *et al.*, "SeaWiFS postlaunch calibration and validation analyses, part 3," NASA Goddard Space Flight Center, Greenbelt, MD, USA, NASA Tech. Memo. 2000-206892, vol. 11, 2000.

[15] P. J. Werdell and S. W. Bailey, "An improved bio-optical data set for ocean color algorithm development and satellite data product validation," *Remote Sens. Environ.*, vol. 98, pp. 122–140, 2005.

[16] D. Blondeau-Patissier, J. F. Gower, A. G. Dekker, S. R. Phinn, and V. E. Brando, "A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans," *Prog. Oceanography*, vol. 123, pp. 123–144, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0079661114000020

[17] I. S. Robinson, *Measuring the Oceans From Space: The Principles and Methods of Satellite Oceanography*. Chichester, U.K.: Praxis Publ. Ltd., 2004.

[18] P. Cipollini, G. Corsini, M. Diani, and R. Grass, "Retrieval of sea water optically active parameters from hyperspectral data by means of generalized radial basis function neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1508–1524, Jul. 2001.

[19] R. Doerffer and H. Schiller, "The MERIS case 2 water algorithm," *Int. J. Remote Sens.*, vol. 28, no. 3–4, pp. 517–535, 2007. [Online]. Available: https://doi.org/10.1080/01431160600821127

[20] M. Hieronymi, D. Müller, and R. Doerffer, "The OLCI neural network swarm (ONNS): A bio-geo-optical algorithm for open ocean and coastal waters," *Frontiers Marine Sci.*, vol. 4, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmars.2017.00140

[21] H. Zhan, P. Shi, and C. Chen, "Retrieval of oceanic chlorophyll concentration using support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2947–2951, Dec. 2003.

[22] E. J. Kwiatkowska and G. S. Fargion, "Application of machine-learning techniques toward the creation of a consistent and calibrated global chlorophyll concentration baseline dataset using remotely sensed ocean color data," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 12, pp. 2844–2860, Dec. 2003.

[23] G. Camps-Valls, J. Muñoz-Marí, L. Gómez-Chova, K. Richter, and J. Calpe-Maravilla, "Biophysical parameter estimation with a semisupervised support vector machine," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 248–252, Apr. 2009.

[24] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, J. Amorós-López, and J. Calpe-Maravilla, "Retrieval of oceanic chlorophyll concentration with relevance vector machines," *Remote Sens. Environ.*, vol. 105, no. 1, pp. 23–33, 2006.

[25] L. Pasolli, F. Melgani, and E. Blanzieri, "Gaussian process regression for estimating chlorophyll concentration in subsurface waters from remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 3, pp. 464–468, Jul. 2010.

[26] C. E. Rasmussen and C. K. I. Williams, *Gaussian Process for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[27] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 8, pp. 514–520, 1996.

[28] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: A basic tool of chemometrics," *Chemometrics Intell. Lab. Syst.*, vol. 58, pp. 109–130, 2001.

[29] J. Verrelst, J. Muñoz, L. Alonso, J. P. Rivera, G. Camps-Valls, and J. Moreno, "Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for sentinel-2 and -3," *Remote Sens. Environ.*, vol. 118, pp. 127–139, 2012.

[30] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of vegetation biophysical parameters using Gaussian process techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5, pt. 2, pp. 1832–1843, May 2012. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84860332507&partnerI D=40&md5=8f89c24c1927827bf249a795b35098fc

[31] J. Verrelst, J. P. Rivera, A. Gitelson, J. Delegido, J. Moreno, and G. Camps-Valls, "Spectral band selection for vegetation properties retrieval using gaussian processes regression," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 52, pp. 554–567, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0303243416301234

[32] K. Blix, G. Camps-Valls, and R. Jenssen, "Gaussian process sensitivity analysis for oceanic chlorophyll estimation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1265–1277, Apr. 2017.

[33] H. Feilhauer, G. P. Asner, and R. E. Martin, "Multi-method ensemble selection of spectral bands related to leaf biochemistry," *Remote Sens. Environ.*, vol. 164, pp. 57–65, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425715001339

[34] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00401706.1993.10485033

[35] X. Wang, U. Kruger, and B. Lennox, "Recrusive partial least squares algorithms for monitoring complex industrial processes," *Control Eng. Pract.*, vol. 11, pp. 613–632, 2003.

[36] K. Ryan and K. Ali, "Application of a partial least-squares regression model to retrieve chlorophyll-a concentrations in coastal waters using hyper-spectral data," *Ocean Sci. J.*, vol. 51, no. 2, pp. 209–221, 2016. [Online]. Available: http://dx.doi.org/10.1007/s12601-016-0018-8

[37] P. J. Werdell and S. W. Bailey, "The seaWiFS bio-optical archive and storage system (seaBASS): Current architecture and implementation," NASA Goddard Space Flight Center, Greenbelt, MD, USA, NASA Tech. Memo. 2002-211617, p. 45, 2002.

[38] P. J. Werdell *et al.* "Unique data repository facilitates ocean color satellite validation," *EOS Trans. Amer. Geophys. Union*, vol. 84, no. 38, pp. 377–392, 2003.

[39] R. Gosselin, D. Rodrigue, and C. Duchesne, "A bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications," *Chemometrics Intell. Lab. Syst.*, vol. 100, pp. 12–21, 2010.

[40] N. L. Afanador, "Important variable selection in partial least squares for industrial process understanding and control," Ph.D. dissertation, Radboud Univ. Nijmegen, Nijmegen, The Netherlands, 2014.

[41] H. Abdi, "Partial least squares regression and projection on latent structure regression (PLS regression)," *Wiley Interdiscip. Rev., Comput. Statist.*, vol. 2, no. 1, pp. 97–106, 2010. [Online]. Available: http://dx.doi.org/10.1002/wics.51

[42] S. Rännar, F. Lindgren, P. Geladi, and S. Wold, "A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm," *J. Chemometrics*, vol. 8, pp. 111–125, 1994. [Online]. Available: http://dx.doi.org/10.1002/cem.1180080204

[43] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016974399385002X

[44] B. S. Dayal and J. F. MacGregor, "Improved PLS algorithms," *J. Chemometrics*, vol. 11, no. 1, pp. 73–85, 1997. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1099-128X(199701)11:1<73::AID-CEM435>3. 0.CO;2-

[45] K. Song, D. Lu, L. Li, S. Li, Z. Wang, and J. Du, "Remote sensing of chlorophyll-a concentration for drinking water source using genetic algorithms (GA)-partial least square (PLS) modeling," *Ecol. Informat.*, vol. 10, pp. 25–36, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574954111000756

[46] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold, "Multi- and megavariate data analysis. Principles and applications," *J. Chemometrics*, vol. 16, no. 5, pp. 261–262, 2001.

[47] P. Jonsson, "Surface status classification, utilizing image sensor technology and computer models," Ph.D. dissertation, , Mid Sweden Univ., Örnsköldsvik, Sweden, 2015.

[48] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods on partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 118, pp. 62–69, 2012.

[49] P. M. Rasmussen, K. H. Madsen, T. E. Lund, and L. K. Hansen, "Visualization of nonlinear kernel models in neuroimaging by sensitivity maps," *NeuroImage*, vol. 55, no. 3, pp. 1120–1131, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811910016198

[50] L. Gross-Colzy, S. Colzy, R. Frouin, and P. Henry, "A general ocean color atmospheric correction scheme based on principal components analysis: Part 1. Performance on case 1 and case 2 waters," *Proc. SPIE*, vol. 6680, 2007, Art. no. 668002. [Online]. Available: http://dx.doi.org/10.1117/12.738508

[51] C. Mobley, D. Stramski, W. Bissett, and E. Boss, "Optical modeling of ocean waters: Is the case 1- case 2 classification still useful?" *Oceanography*, vol. 17, pp. 60–67, 2004.

[52] E. L. Snelson, "Flexible and efficient Gaussian process models for machine learning," Ph.D. dissertation, Gatsby Comput. Neurosci. Unit, Univ. College London, London, U.K., 2007.

[53] H. R. Gordon *et al.* "A semianalytic radiance model of ocean color," *J. Geophys. Res.*, vol. 93, pp. 10909–10924, 1988.

[54] J. P. Cannizzaro and K. L. Carder, "Estimating chlorophyll a concentrations from remote-sensing reflectance in optically shallow waters," *Remote Sens. Environ.*, vol. 101, no. 1, pp. 13–24, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0034425705004025

[55] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III, "The partial least squares approach to generalized inverses," *SIAM J. Sci. Statist. Comput.*, vol. 5, pp. 735–743, 1984.

Katalin Blix received the B.S. degree in geosciences from the Sogn og Fjordane University College, Førde, Norway, in 2010, and the Civil Engineer/M.S. degree in technology with specialization to applied physics and mathematics, in 2014, from the University of Tromsø—The Arctic University of Norway, Tromsø, Norway, where she is currently working toward the Ph.D. degree in remote sensing.

Torbjørn Eltoft (M'92) received the degrees of Cand. Real. (M.S.) and Dr. Scient. (Ph.D.) from the University of Tromsø - The Arctic University of Norway, in 1981 and 1984, respectively.

He joined the Faculty of Science and Technology, University of Tromsø (UiT)—The Arctic University of Norway, Tromsø, Norway, in 1988. He is currently a Professor of remote sensing with the Department of Physics and Technology, UiT—The Arctic University of Norway. He is an Adjunct Professor of remote sensing with the Northern Research Institute Tromsø, Tromsø, Norway. He is also the Director of the Centre for Integrated Remote Sensing and Forecasting for Arctic Operations, a Centre for Research-based Innovation awarded by the Norwegian Research Council in 2014. His research interests include multidimensional signal and image analysis, statistical modeling, neural networks, and machine learning, with applications in multichannel synthetic aperture radar and ocean color remote sensing.

Dr. Eltoft was an Associate Editor for the Elsevier journal *Pattern Recognition* for the period 2005–2011, and was a Guest Editor for the journal *Remote Sensing* on the special issue for the POLinSAR 2017 conference. He has a significant publication record in the area of signal processing and remote sensing, and he was the recipient of the year 2000 Outstanding Paper Award in Neural Networks awarded by the IEEE Neural Networks Council, and of the Honorable Mention for the 2003 Pattern Recognition Journal Best Paper Award. He was the winner of the 2017 UiT Award for Research and Development.

Chapter 8

Paper 3:
Machine Learning Automatic Model Selection Algorithm for Oceanic Chlorophyll-a Content Retrieval

remote sensing

MDPI

Article

Machine Learning Automatic Model Selection Algorithm for Oceanic Chlorophyll-a Content Retrieval

Katalin Blix * and Torbjørn Eltoft

UiT the Arctic University of Norway, P.O. box 6050 Langnes, NO-9037 Tromsø, Norway; torbjorn.eltoft@uit.no

* Correspondence: katalin.blix@uit.no; Tel.: +47-483-49-399

Received: 13 March 2018; Accepted: 16 May 2018; Published: 17 May 2018

Abstract: Ocean Color remote sensing has a great importance in monitoring of aquatic environments. The number of optical imaging sensors onboard satellites has been increasing in the past decades, allowing to retrieve information about various water quality parameters of the world's oceans and inland waters. This is done by using various regression algorithms to retrieve water quality parameters from remotely sensed multi-spectral data for the given sensor and environment. There is a great number of such algorithms for estimating water quality parameters with different performances. Hence, choosing the most suitable model for a given purpose can be challenging. This is especially the fact for optically complex aquatic environments. In this paper, we present a concept to an Automatic Model Selection Algorithm (AMSA) aiming at determining the *best* model for a given matchup dataset. AMSA automatically chooses between regression models to estimate the parameter in interest. AMSA also determines the number and combination of features to use in order to obtain the *best* model. We show how AMSA can be built for a certain application. The example AMSA we present here is designed to estimate oceanic Chlorophyll-a for global and optically complex waters by using four Machine Learning (ML) feature ranking methods and three ML regression models. We use a synthetic and two real matchup datasets to find the *best* models. Finally, we use two images from optically complex waters to illustrate the predictive power of the *best* models. Our results indicate that AMSA has a great potential to be used for operational purposes. It can be a useful objective tool for finding the most suitable model for a given sensor, water quality parameter and environment.

Keywords: ocean color; remote sensing; model selection; feature ranking; regression

1. Introduction

Ocean Color (OC) monitoring from spaceborne and airborne platforms using remote sensing techniques has been receiving an increased focus in the past decades [1,2]. This is due to the fact that an ever-increasing amount of remote sensing data is getting available, but also, because of increased anthropogenic activity and climate change have resulted in changes in the water quality [3]. Coastal waters are one of the most sensitive areas due to their vulnerable ecosystems. Worsened water quality might endanger these ecosystems (such as fish's habitats [4]), which has both economical and ecological importance [3]. It is well-known that the eutrophication of coastal waters and inland waters has been increasing lately, leading to decreased water-quality [5,6]. Continuously monitoring of the water-bodies, with special focus to coastal waters is therefore important for various reasons. It can also contribute to improved understanding of the ongoing changes, and the impact of increased anthropogenic activities on the ecosystems [7].

The quality of water bodies, both globally and regionally, is most efficiently inferred from color using multi-spectral or hyper-spectral remote sensing. The color of the oceans is determined by the

different type, amount and distribution of water constituents. Being able to monitor these water constituents allows to retrieve information about the environmental state of the water [8]. The most common parameters used for monitoring water quality are Chlorophyll-a (Chl-a), Colored Dissolved Organic Matter (CDOM), Total Suspended Matter (TSM), Secchi Disk Depth (SDD), turbidity, Total Phosphorus (TP), to name some [3].

However, the retrieval of water quality contents from remote sensing data is not always strait forward. Algorithms are generally dependent on the sensors' characteristics, geographical location, and environmental conditions of the water body. The objective of this paper is to *present* and *demonstrate* a strategy for an Automatic Model Selection Algorithm (AMSA), for retrieval of water quality parameters from remote sensing data, given an appropriate matchup dataset. Since Chl-a is one of the most important and most studied of these water quality parameters [5], we will use Chl-a as an example parameter throughout the paper. Besides, estimating aquatic Chl-a concentration has several important applications, in addition to providing information about water-quality. Chl-a occurs in phytoplankton in aquatic environments. Phytoplankton uses photosynthesis in order to live and grow. Capturing of light, which is the driving of photosynthesis [9], takes place in the Chl-a molecule. Estimating Chl-a content allows to retrieve information about the aquatic biomass and several biophysical processes. During photosynthesis, phytoplankton takes up Carbon-Dioxid (CO_2) [10]. Therefore, monitoring phytoplankton through Chl-a might also contribute to the understanding of climate change [11–13].

Using Chl-a as an example, we will in the following give some rational and motivation for AMSA. Remote sensing of Chl-a content (and other water quality parameters) is done by optical imaging sensors onboard satellites, which have different spectral and spatial resolutions. Chl-a content is usually retrieved by relating the measured signal at the sensor, the remote sensing reflectance (Rrs), to coincident in-water Chl-a measurements (see for instance the National Aeronautics and Space Administration's (NASA) OC products [14–18]). This dataset is denoted a so-called matchup dataset, and forms the basis for most of the algorithms used for Chl-a content estimation from remotely sensed data. Since the various sensors have different number of bands at different central wavelengths (see Table 2), the matchup data has to be calibrated for each given sensor.

Furthermore, there are a manifold of retrieval algorithms available to the user [19–21]. Some of them are designed to estimate Chl-a globally, whereas others are region specific. These algorithms are in general sensor specific, this means they require a new or adjusted model for each sensor. For an untrained user, it is often challenging to establish or choose the most suitable Chl-a retrieval model. This is especially the fact for optically challenging aquatic environments, such as coastal waters [22]. Coastal waters are often dominated by other water constituents than Chl-a, such as CDOM, and CDOM and Chl-a are known to have their absorption peak in the same spectral region. This results in difficulties in distinguishing between the signals originating from Chl-a and CDOM, especially, when Chl-a content is estimated by algorithms that use the absorption peak of the Chl-a molecule.

As more datasets are collected, and computer processing power gets unlimited, machine learning (ML) algorithms have become more feasible in OC applications. ML models are not based on assumption about the Chl-a absorption spectrum. They learn the relationship between the in-situ Chl-a content and the available Rrs values, and use this learned functional relationship for prediction. These models use all the available spectral bands for learning and prediction, which results that the importance of the spectral bands in the regression process is kept hidden. It can be questioned whether all the bands are needed to obtain the best regression for a given model and region. Artificial Neural Networks (ANN) models have been lately successfully applied for Chl-a estimation [23–25], and to various other applications, such as for predicting the amount of generated electricity [26], suspendid sediment load in rivers [27] and rainfall and runoff predictions ahead in time [28]. For OC applications, satellite derived Chl-a in optically complex waters is also often estimated by using other ML algorithms [29–32].

Furthermore, complex waters show great regional variations, which leads to erroneous Chl-a estimates, when algorithms tuned on global datasets, are applied to a local region [19]. Therefore, it is often required to design local algorithms, which are trained on datasets from the given region [33,34]. However, choosing the most suitable model for a given region can still be challenging.

The above arguments suggest that an automatic model selection approach could be an important tool in choosing the optimum model to monitor a given aquatic environment. Comparisons of models for various OC applications have been carried out in [35–37], but to the best of our knowledge, a flexible and automatized model selection tool for OC application has not yet been proposed in the literature. Being able to objectively compare models and determine the most suitable one for the given data and purpose might be beneficial for the users.

The contribution of this paper is to present a strategy for an Automatic Model Selection Algorithm (AMSA), which outputs the most suitable water quality retrieval model, given the matchup dataset. The current AMSA model uses three ML models as input options. ML models usually rely on feature selection in prior to regression [38]. This is due to the fact that dimensionality reduction is often required to increase accuracy, robustness and computational time [39]. Using feature selection also helps to correctly interpret the data. The method for choosing the most optimal number and combination of features for the given model is model dependent, and needs to be developed in each case. AMSA uses feature ranking methods to assign relevance to the features, then it evaluates the number and combination of these ranked features in regression models using some quantitative regression performance measures.

Hence, AMSA is not only using feature selection prior to regression, but also feature ranking methods derived from regression models based on different principles. This means that the importance of the features is first determined by using several feature ranking approaches, one tailored to each regression model, then sequential forward selection is applied for comparison. Then the regression models are compared by computing regression performance measures. Finally, AMSA returns the best model for the given matchup dataset. Hence, AMSA is neither limited to a given water quality parameter nor to a feature ranking method/regression model/regression performance measure. The only input that it requires, is the matchup dataset.

For *demonstration* of the performance of AMSA, we use three sophisticated ML models for feature ranking and regression. These regression models are the Gaussian Process Regression (GPR), Support Vector Regression (SVR) and Partial Least Square Regression (PLSR) models. GPR has been shown to outperform empirical [31,40] and ML regression models [41] for biophysical parameter retrieval from remotely sensed data. GPR has several advantageous properties besides its excellent regression performance, for instance the certainty level of the estimates and the possibility to access feature relevance. Feature relevance for the GPR model can be accessed by the Sensitivity Analysis (SA) [31,32] and the Automatic Relevance Determination (ARD) [40,42] feature ranking methods.

The SVR model has also been shown to perform well for OC applications [29,43,44]. In this work, we applied the SA to the SVR model in order to access feature relevance. For classification in neuroimage applications, this has been done in [45]. Here, we introduced the methodology for regression in Chl-a content estimation.

The PLSR model was included in AMSA, because of the Variable Importance in Projection (VIP) feature ranking methods associated with it. PLSR is a strong regression model, which can handle high dimensional inputs, reduce noise and co-linearity in the data [46]. The PLSR model has been applied for OC applications in optically complex aquatic environments [47].

We have previously studied the SA of the GPR model, ARD and VIP feature ranking methods and the GPR and PLSR regression models for Chl-a content estimation in [32]. In [32], we used a MERIS matchup dataset and two additional matchups for the MODIS-Aqua and SeaWiFS sensors to evaluate the methodologies, and concluded that these feature ranking methods can be used to reduce the number of features, while still obtaining comparable estimates for Chl-a content, compared to the state-of-art algorithms.

In the current demonstration of AMSA, we show how the proposed strategy can be used to determine automatically a model for oceanic Chl-a content estimation for both global waters and optically complex waters. The matchup datasets we have used here include a synthetic dataset produced by the International Ocean-Colour Coordinating Group (IOCCG dataset) [48], plus two additional matchups, one for the MERIS sensor (MEdium Resolution Imaging Spectrometer) and one for the MODIS-Aqua (MODerate-resolution Imaging Spectroradiometer) sensor. The IOCCG dataset provides the possibility to threshold the data based on the absorption of the CDOM, and the amount of Chl-a concentrations. Hence, observations which are more likely to occur in complex aquatic environments, can be selected. Furthermore, we resample the IOCCG dataset to match the spectral resolution of the MERIS and MODIS-Aqua matchups.

An additional contribution of this work, is to further extend the feature ranking methods by the sensitivity analysis of the SVR model, which allows us to include the SVR regression model in the AMSA model library. We choose to use the IOCCG dataset to have better control over the optical properties of observations, and include the two matchups for the MERIS and MODIS-Aqua sensors to show that the approach work well on different data sets and for different environmental situations. We highlight that the goal here is to show how the AMSA approach can be used to perform an objective comparison and selection of an optimal model for the given dataset, according to the regression criteria used. AMSA automatically performs feature ranking and training and testing of the regression models. Hence, the output model is already validated. Finally, the demonstration includes two images acquired by MERIS over optically complex aquatic areas to visualize the predictions given by the selected optimal AMSA model.

The rest of this work is organized as follows. Section 2 introduces the general concept of the AMSA and explains the ML AMSA for oceanic Chl-a content estimation in details. Furthermore, the datasets used in this study are described. Section 3 presents the results. Section 4 discusses the results and approach, and highlights advantages and disadvantages of the methodology. Finally, Section 5 concludes this paper and outlines future work.

2. Materials and Methods

2.1. The Automatic Model Selection Algorithm

2.1.1. The Concept of the AMSA

The AMSA has two stages. In the first stage, relevance is assigned to all the available features by using feature ranking methods. The second stage is to perform regression by using the ranked features as inputs. The *best* regression model is determined by selecting the most optimal number and combination of features based on the selected goodness of fit criteria. Examples of goodness of fit measures are: Normalized Root Mean Squared Errors (NRMSE) and the Pearson's correlation coefficient (R^2).

Feature ranking: Assume a matchup dataset $D = \{\mathbf{x}_n; \mathbf{y}_n\}_{n=1}^{N}$, where $\mathbf{x}_n$ is the D dimensional input, D is the number of features, $\mathbf{y}_n$ is the corresponding output (ground-truth) and N is the number of measurements. This matchup dataset is used for ranking the D features in $\mathbf{x}$ by using feature ranking methods. Figure 1 shows the feature ranking stage of the algorithm.

The process starts by using all data in the matchup dataset to perform feature ranking. Assume, there are i feature ranking methods. Then the output of this step is i sets of ranked features, each ordered by decreasing relevance (i.e., the first feature in *Ranked feature set* is the most important, and the last is the least relevant).

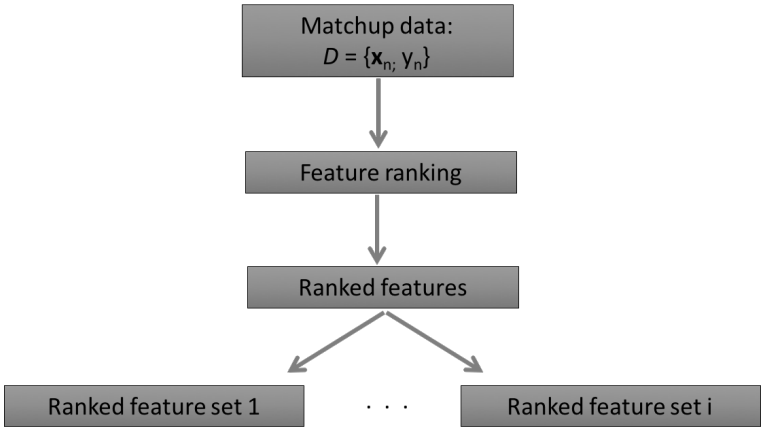


Figure 1. The feature ranking stage of the AMSA.

Regression and feature selection: Figures 2 and 3 show the flowcharts of the regression stage. In the regression stage, the dataset is split into two parts, 50% is used for training and 50% is used for testing. This partitioning ensures that both training and testing sets contain representative data. Assume j number of *Regression models* are available. Then an iterative process starts by training and testing *Regression model* $1, ..., j$ with the features in the *Ranked feature set* $1, ..., i$ by using a sequential forward selection approach.

For simplicity, let us assume using *Regression model 1* and *Ranked feature set 1*, containing D ranked features. *Regression model 1* starts the training on the training data by taking the most important feature in the *Ranked feature set 1*. When this model is trained, testing is performed on the test data by computing k *Regression performance measures*. The results of the computed *Regression performance measures* are saved (Figure 3).

Then *Regression model 1* adds the second most relevant feature of the *Ranked feature set 1*, in addition to the first one. The system trains and tests the model by computing k *Regression performance measures*, and saves the results. This procedure continues until the least important feature of the *Ranked feature set 1* has been included.

Regression model 1 repeats the same process with all the *Ranked feature sets* $(1, ..., i)$. The same procedure is done with all the *Regression models* $(1, ..., j)$. The k *Regression performance measures* are saved for all j *Regression models*, and for all i *Ranked feature sets* with all D number of ranked features.

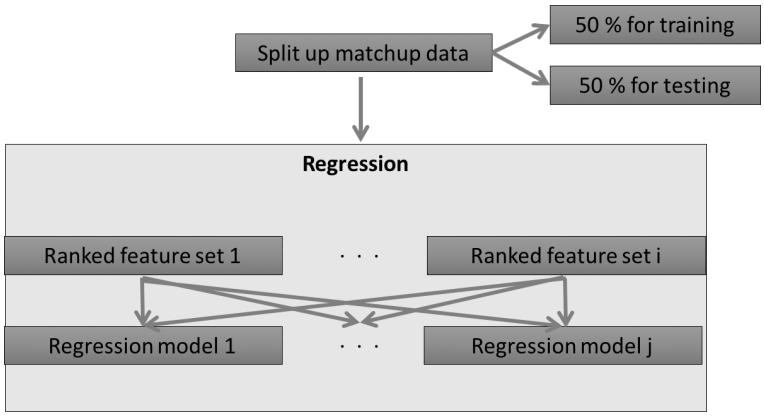


Figure 2. Regression stage of the AMSA (A).

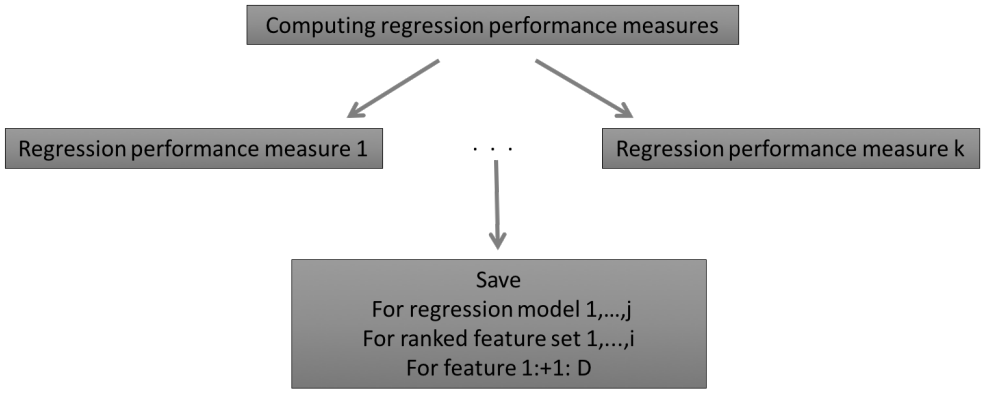


Figure 3. Regression stage of the AMSA (B).

Finally, AMSA searches in the stored *Regression performance measures* for the model, which resulted in the *best* performance. AMSA outputs: the *best* regression model based on the computed regression performance measures; the feature ranking method that resulted the best combination of features associated with the regression model; the number of features, which were needed to obtain the *best* model; the actual input-features of the *best* model and also the values of the regression performance measures. Table 1 shows the output of the algorithm.

Table 1. The output of the AMSA.

Regression model	Feature ranking method	The features	# of features	Value of the regression performance measures

There are obviously no limitations in the number of feature ranking methods, regression models and regression performance measures to be used in AMSA. Note, if feature ranking is not of interest, this stage can be turned off. In that case, only the most desirable regression model for the given dataset and predefined feature set is returned.

2.2. Demonstration of an AMSA Implementation

The AMSA concept can be used by the users to build an optimal model for her or his application. Any model can be selected, and it can be used for any water quality parameter estimation, as long as matchup data is available. Furthermore, user defined feature ranking methods, regression models and regression performance measures can be included. In this section we present the AMSA we designed for Chl-a estimation. It is based on the work and results presented in [32].

2.2.1. The Matchup Data

We focused on oceanic Chl-a content estimation from Rrs. Hence, the matchup data consists of Rrs measured on the wavelengths of the given sensor and corresponding in-situ Chl-a measurements.

For feature ranking, the complete available dataset was used, while for regression, the dataset was split up in 50% for training and 50% for testing. We chose to split up the data as it follows. The Chl-a values were sorted in an increasing order. The corresponding Rrs values were assigned to the sorted Chl-a values. Then we draw the even numbered observations for forming the training data, and the odd numbered measurements for testing purposes. Hence, both the training and test data was as representative as possible. Note, the way of splitting the data in AMSA can be defined differently. The dataset can be divided randomly and in a different proportion for training and testing, as well.

2.2.2. Regression Models

Assume a dataset consisting of *in situ* Chl-a values $y_{n=1}^{N}$ and corresponding input Rrs values $\{\mathbf{x}_n \in R^D\}_{n=1}^{N}$, where $n = 1, ..., N$ is the number of measurements and $d = 1, ..., D$ is the number of features (spectral bands) for all the regression models. We will use here regression models, namely Gaussian Process Regression, Support Vector Regression and Partial Least Squares Regression. These are briefly summarized below.

Gaussian Process Regression model: The Gaussian Process Regression (GPR) model assumes that the output (Chl-a) is a function of the input (Rrs) and some noise ε_n, which can be written by $y_n = f(\mathbf{x}_n) + \varepsilon_n$ for $n = 1, ..., N$, where the noise term is assumed to be additive, independently, identically Gaussian distributed with zero mean and constant variance, i.e., $\varepsilon_n \sim N(0, \sigma^2)$. The model learns this function by fitting a multivariate joint Gaussian distribution over the function values, $f(\mathbf{x}_1), ..., f(\mathbf{x}_N) \sim N(\mathbf{0}, \mathbf{K})$, with zero mean and covariance matrix $\mathbf{K}$. Then this can be used for predicting the unseen output Chl-a y_* for a new input Rrs $\mathbf{x}_*$ by defining a joint prior distribution between the available Chl-a $\mathbf{y} \equiv \{y_n\}_{n=1}^{N}$ and y_*. This can be mathematically expressed by

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} + \sigma^2\mathbf{I}_n & \mathbf{k}_* \\ \mathbf{k}_*^\top & k_{**} + \sigma^2 \end{bmatrix}\right), \tag{1}$$

where $\mathbf{k}_*$ is the covariance between the training vector and the test point, k_{**} is the covariance between the test point with itself, and $\mathbf{K} + \sigma^2\mathbf{I}_n$ is the $N \times N$ noisy covariance matrix of the training inputs. The posterior distribution over the output y_* can be analytically computed by using Bayes' formula: $p(y_*|\mathbf{x}_*, D) = N(y_*|\mu_{\mathrm{GP}*}, \sigma^2_{\mathrm{GP}*})$, where $\mu_{\mathrm{GP}*}$ is the predicted Chl-a and $\sigma^2_{\mathrm{GP}*}$ is the certainty level of the estimated Chl-a content (predictive variance). The predicted Chl-a content can be expressed by $\mu_{\mathrm{GP}*} = \mathbf{k}_*^\top(\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}$. Note, the predicted Chl-a content can also be written by $\mu_{\mathrm{GP}*} = \mathbf{k}_*^\top\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\mathbf{K} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}$ is the weight vector of the mean function of the GPR model. This allowed the application of the SA (Equation (11)). For further details on the GPR model we refer to [49].

Support Vector Regression model: The Support Vector Regression (SVR) model ([41,50–53]) estimates Chl-a value from Rrs values by $y_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b$, where $\mathbf{w}^{\mathrm{T}}$ is the transposed weight vector and b is the bias term. The SVR model uses the so-called ϵ-intensitive loss function to obtain estimates by penalizing errors exceeding an ϵ limit and at the same time obtaining a regression function as flat as possible. Hence the weights are estimated in the SVR model by minimizing the objective function $J = \frac{1}{\beta}\sum_{n=1}^{N}(\zeta_n^+ + \zeta_n^-) + \frac{1}{2}||\mathbf{w}||^2$ with respect to $\mathbf{w}$, ζ_n^+, ζ_n^- and constrained to

$$y_n - -\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - -b \leq \epsilon + \zeta_n^+ \quad \text{for} \quad n = 1, ..., N \tag{2}$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b - y_n \leq \epsilon + \zeta_n^- \quad \text{for} \quad n = 1, ..., N \tag{3}$$

$$\zeta_n^+, \zeta_n^- \geq 0 \quad \text{for} \quad n = 1, ..., N. \tag{4}$$

ζ_n^+ and ζ_n^- are called slack variables, and allow measurements to be larger than ϵ, and $\beta > 0$ is a constant controlling the trade-off between the flatness of the regression function and the magnitude of the deviations from ϵ.

Constructing a Lagrange function from the objective function allows to obtain the optimal solution for the weights: $\hat{\mathbf{w}} = \sum_{n=1}^{N}(\alpha_n^+ - -\alpha_n^-)\,\mathbf{x}_n$, where α_n^+ and α_n^- are the Lagrange multipliers, also referred to as support vectors. Define $a_n = \alpha_n^+ - -\alpha_n^-$, and collecting the estimated Chl-a values $\hat{y}_n$ into a vector $\hat{\mathbf{y}}$, the estimates can be written by

$$\hat{\mathbf{y}} = \hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x} + \hat{\mathbf{b}} = \sum_{n=1}^{N} a_n\mathbf{x}_n^{\mathrm{T}}\mathbf{x} + \hat{\mathbf{b}}. \tag{5}$$

Note, that a_n vanishes, when measurements do not exceed ϵ, which results that the solution for $\hat{\mathbf{w}}$ is sparse. Finally, applying the kernel function defined in Equation (13) to $\mathbf{x}_n^{\mathrm{T}}\mathbf{x}$, the estimated Chl-a value vector can be expressed by

$$\hat{\mathbf{y}} = \sum_{n=1}^{N} a_n k(\mathbf{x}_n, \mathbf{x}) + \hat{\mathbf{b}}. \tag{6}$$

Partial Least Square Regression model: Assume once again the *in-situ* Chl-a ($\mathbf{X}$) and Rrs ($\mathbf{y}$) training dataset $D \equiv \{\mathbf{X}, \mathbf{y}\}$, where now the observations are collected in matrices, such that $\mathbf{X}$ is an $N \times D$ input data-matrix consisting of $d = 1, ..., D$ features (spectral bands) and $n = 1, ..., N$ observations, and let $\mathbf{y}$ be the corresponding $N \times 1$ output-vector (Chl-a measurements), holding $n = 1, ..., N$ observations.

The Partial Least Square Regression (PLSR) model [46,54] relates the input Rrs $\mathbf{X}$ and the output Chl-a $\mathbf{y}$ through a latent-space. This is done by introducing so-called latent variables $\mathbf{T}$ ($N \times H$), which are representing both $\mathbf{X}$ and $\mathbf{y}$ in the latent-space, such that the covariance between the projection of $\mathbf{X}$ and $\mathbf{y}$ in this latent- space is maximized. The PLSR model can be written by

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \\ \mathbf{y} &= \mathbf{T}\mathbf{c} + \mathbf{f} \\ \mathbf{T} &= \mathbf{X}\mathbf{W}^{\star} \\ \mathbf{W}^{\star} &= \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}, \end{aligned} \tag{7}$$

where $\mathbf{P}$ ($D \times H$) is a matrix of the X-loadings and $\mathbf{c}$ ($H \times 1$) is the y-loadings, and they are good representations of $\mathbf{X}$ and $\mathbf{y}$, respectively. The term $\mathbf{W}^{\star}$ ($D \times H$) holds the weights of $\mathbf{X}$, and defines the common latent-space. The error terms, $\mathbf{E}$ ($N \times D$) and $\mathbf{f}$ ($N \times 1$), are assumed to be iid. $\sim N(0, \sigma^2)$. Then we estimate the output Chl-a $\mathbf{y}$ by

$$\mathbf{y} = \mathbf{X}\mathbf{W}^{\star}\mathbf{c} + \mathbf{f} = \mathbf{X}\mathbf{b} + \mathbf{f}, \tag{8}$$

where $\mathbf{b} = \mathbf{W}^{\star}\mathbf{c}$ and $\mathbf{W}$ ($D \times H$) is the weight matrix consisting of the eigenvectors of the variance-covariance matrix $\mathbf{X}^{\mathbf{T}}\mathbf{Y}\mathbf{Y}^{\mathbf{T}}\mathbf{X}$. Minimizing the error term $\mathbf{f}$ in the PLSR model results the most optimal regression. For further details on the PLSR model and algorithms we refer to [55–60].

2.2.3. Feature Ranking Methods

We chose four feature ranking methods to assign relevance to the features (in our case spectral bands). The four feature ranking methods are tailored to the regression models, and are the Sensitivity Analysis (SA) of the GPR model, Sensitivity Analysis (SA) of the SVR model, Automatic Relevance Determination (ARD) and Variable Importance in Projection (VIP).

SA of Kernel Machines (GPR and SVR): The SA feature ranking method for the SVR and GPR models are based on the same concept, but for different regression models. Although both the SVR and GPR are non-linear kernel machines, their underlying principles differ. The SA of the GPR model was introduced by [31,61], while the SA of the Support Vector Machine (SVM) for classification purposes was described in [45]. In this work, we extend the SA of the SVM to regression.

Let us define the sensitivity of feature j as

$$s_j = \int \left(\frac{\partial \boldsymbol{\phi}(\mathbf{x})}{\partial x_j} \right)^2 p(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{9}$$

where $p(\mathbf{x})$ is the probability density function of the D-dimensional input vector $\mathbf{x} = [x_1, \ldots, x_D]^\top$, and $\boldsymbol{\phi}(\mathbf{x})$ represents either the predictive mean function of the GPR model, $\mu_{\mathrm{GP}*}$ or the estimated output $\hat{y}$ of the SVR model. The empirical estimate of the sensitivity for the jth feature can be written as

$$s_j = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial \boldsymbol{\phi}(\mathbf{x}_n)}{\partial x_n^j}\right)^2, \tag{10}$$

where N denotes the number of training samples.

Applying the SA (Equation (10)) to the GPR model yields:

$$\begin{aligned} s^j_{\mu_{\mathrm{GP}\star}} &= \frac{1}{N}\sum_{q=1}^{N}\left(\frac{\partial \boldsymbol{\phi}(\mathbf{x}_q)}{\partial x_q^j}\right)^2 \\ &= \frac{1}{N}\sum_{q=1}^{N}\left(\frac{\partial \sum_{p=1}^{N}\alpha_p k(\mathbf{x}_p,\mathbf{x}_q)}{\partial x_q^j}\right)^2 \\ &= \frac{1}{N}\sum_{q=1}^{N}\left(\sum_{p=1}^{N}\frac{\alpha_p(x_p^j - x_q^j)}{\lambda_j^2}k(\mathbf{x}_p,\mathbf{x}_q)\right)^2, \end{aligned} \tag{11}$$

and to the SVR model gives

$$s^j_{SVR} = \frac{1}{N}\sum_{q=1}^{N}\left(\sum_{p=1}^{N}\frac{a_p(x_p^j - x_q^j)}{\lambda_j^2}k(\mathbf{x}_p,\mathbf{x}_q)\right)^2, \tag{12}$$

where the difference between Equations (11) and (12) is in the computation of α_p and a_p (Note that the calculation of the empirical sensitivity is computed in closed-form using the training data points and the inferred $\boldsymbol{\alpha}$ and $\mathbf{a}$).

ARD: Kernel Machines (GPR and SVR) use kernel functions to perform regression. The Squared Exponential (SE) kernel function is a widely used kernel function due to its advantageous properties, such as it has infinite derivatives and it is a universal kernel [62]. The SE kernel function can be written by

$$k(\mathbf{x}_p,\mathbf{x}_q) = \nu^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_p^d - x_q^d}{\lambda_d}\right)^2\right), \tag{13}$$

where λ_d is the length-scale for feature d, ν is the positive scale factor and σ^2 is the noise variance. The SE kernel also provides the possibility to access feature relevance. This can be achieved though the optimized length-scale hyperparameters in Equation (13) [40]. Small values of the length-scales indicate greater relevance, while larger values suggest less important features. Hence, the inverses of the optimized length-scale parameters allow the ranking of the features used in the SVR and GPR model.

VIP: The VIP feature ranking method is derived from the Partial Least Squares Regression (PLSR) model. VIP measures the contribution to the total variance of the jth input feature ($j = 1, ..., D$) [63,64]. The VIP can be written by [65]

$$\mathrm{VIP}_j = \sqrt{D\sum_{h=1}^{H} SS_h(w_{hj}/\parallel w_j \parallel^2)/\sum_{h=1}^{H} SS_h}, \tag{14}$$

where SS_h is the percentage of the output (Chl-a) explained by the *so-called* hth latent variable and w_j are the weights of the PLSR model.

2.2.4. Regression Performance Measures

We chose the Normalized Root Mean Squared Errors (NRMSE) and the Squared Correlation Coefficient (R^2) to evaluate regression strength. These measures are frequently used for model evaluation in remote sensing [66,67]. Using these measures might be appropriate, when comparison is in interest. These regression performance measures can be expressed by

$$\text{NRMSE} = \frac{1}{y_{\max} - -y_{\min}} \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - -\hat{y}_i)^2} \tag{15}$$

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - -\overline{y})^2}{\sum_{i=1}^{N}(y_i - -\overline{y}_i)^2}, \tag{16}$$

where N is the number of observations in the test set, y is the true Chl-a content, $\hat{y}$ is the predicted Chl-a, $y_{\max}$ is the maximum observed value, $y_{\min}$ is the minimum observed value, and $\overline{y}$ is the mean of the observed Chl-a contents in the test set.

2.2.5. Summary of the AMSA Approach

Figure 4 shows the summary of the ML AMSA for oceanic Chl-a content estimation. The ML AMSA uses in Stage 1 the Chl-a/Rrs matchup dataset to rank the features by using the SA GPR, SA SVR, ARD and VIP feature ranking methods. Then in the Stage 2, the dataset is split to perform regression by the GPR, SVR and PLSR models. Finally, the model with lowest NRMSE and highest R^2 is returned. This is the *best* model between the available possibilities. Figure 5 shows an illustrative example, how AMSA can be used for applications.

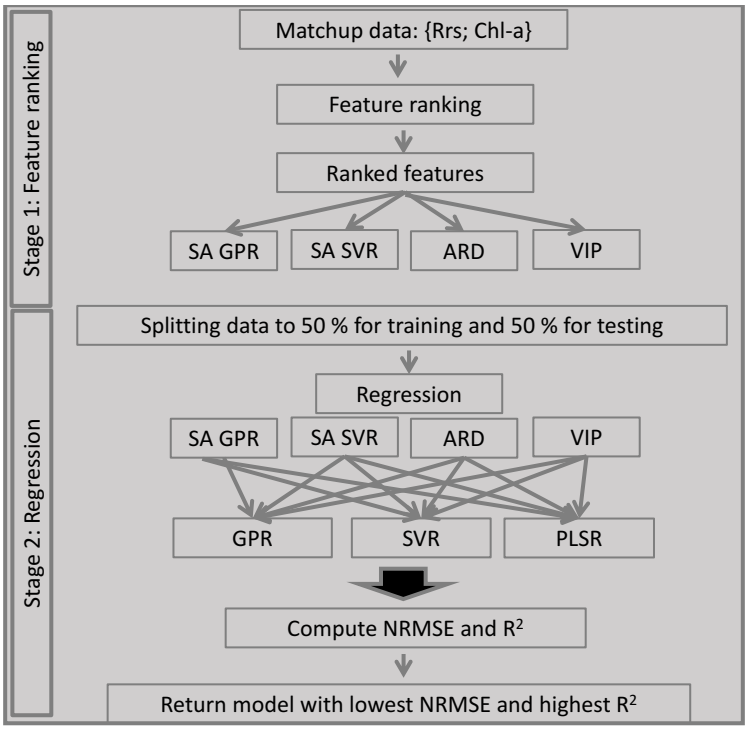


Figure 4. The ML AMSA for oceanic Chl-a content estiamtion.

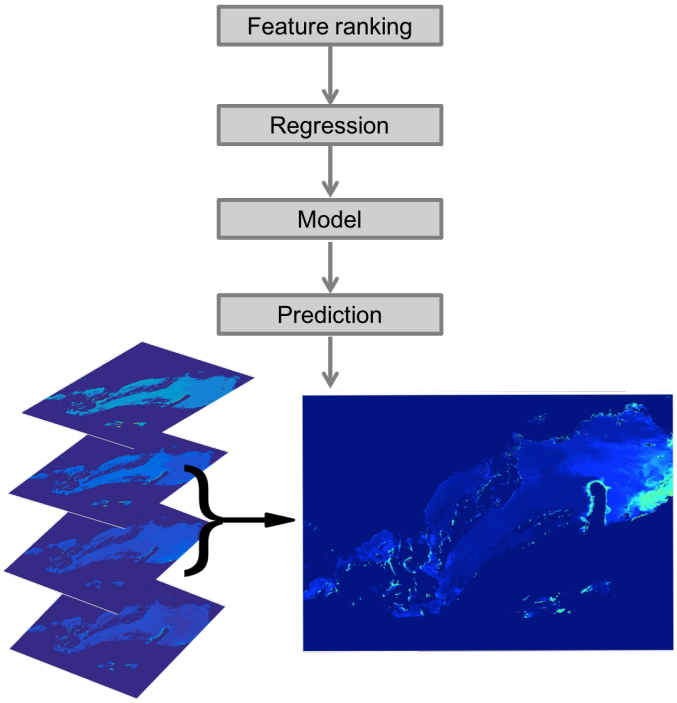


Figure 5. Illustration of the AMSA for application.

2.3. Data

We evaluated the AMSA algorithm on the IOCCG synthesized dataset [48] and a MERIS (MEdium Resolution Imaging Spectrometer) and MODIS-Aqua (MODerate-resolution Imaging Spectroradiometer) dataset obtained from SeaBASS database [68,69]. Table 2 summarizes the datasets we used for demonstrating the AMSA algorithm.

2.3.1. Training Data

The synthetic IOCCG dataset has a spectral region ranging from 400 to 800 nm on a 10 nm bandwidth, and containing both inherent (IOPs) and apparent optical properties (AOPs). We resampled the dataset to match the positions and bandwidths of the spectral bands of MERIS and MODIS-Aqua used for OC applications.

The summary of the synthetic resampled dataset can be seen in Table 2. We used the Rrs values with the corresponding Chl-a values. This dataset allows to mimic *eutrophic* conditions by defining a threshold based on the absorption coefficient for CDOM(a_{CDOM}) and Chl-a value. We partitioned the resampled data to *eutrophic* oceanic waters, for $a_{CDOM} > 0.06$ m^{-1} and Chl-a > 0.7 mgm^{-3}.

The MERIS dataset consists of 567 measurements, measured between April 2002 and March 2012. It can be seen that the Chl-a content spans a wide range of concentration with values in the range between 0.017 and 40.23 mgm^{-3}. The bandwidth is here 10 nm for bands 1–7, and 7.5 nm for band 8.

The MODIS-Aqua dataset has seven channels ranging from 405 nm to 683 nm. The spectral resolution is 10 nm, except for the first band, which has a bandwidth of 15 nm. The data we used here has 579 measurements between July 2002 and November 2012, and the Chl-a concentrations are between 0.0153 and 25.4985 mgm^{-3}.

In case of the MERIS and MODIS-Aqua datasets, only the Rrs and the corresponding Chl-a values were available, thus the division of the data was based on the Chl-a content only. The geographic locations of the measurements can be seen in Figure 6. The red dots indicate measurements for Chl-a value below 0.7 mgm^{-3}, and the black ones for Chl-a above 0.7 mgm^{-3}. It can be seen that measurements corresponding to *eutrophic* conditions are usually located in the coastal regions.

Table 2. Summary of the training datasets we used for model selection.

Synthetic resampled MERIS global (MS 1a)	
Bands (λ_c (nm))	413 443 490 510 560 620 665 681
Band width	10 nm and 7.5 nm
Spatial resolution	300 m
Chl-a range (mgm^{-3})	0.03–30
a_{CDOM} (m^{-1})	0.0025–2.3677
Nr. of samples	478
Synthetic resampled MERIS *eutrophic* (MS 1b)	
Chl-a range (mgm^{-3})	0.7–30
a_{CDOM} (m^{-1})	0.06–2.3677
Nr. of samples	300
MERIS global (MS 2a)	
Chl-a range (mgm^{-3})	0.017–40.23
Nr. of samples	557
MERIS *eutrophic* (MS 2b)	
Chl-a range (mgm^{-3})	0.7076–40.23
Nr. of samples	247
Synthetic resampled MODIS-Aqua global (MS 3a)	
Bands (λ_c (nm))	412 443 488 531 551 667 678
Band width	10 nm, 15 nm
Spatial resolution	1000 m
Chl-a range (mgm^{-3})	0.03–30
a_{CDOM} (m^{-1})	0.0025–2.3677
Nr. of samples	478
Synthetic resampled MODIS-Aqua *eutrophic* (MS 3b)	
Chl-a range (mgm^{-3})	0.03–30
a_{CDOM} (m^{-1})	0.06–2.3677
Nr. of samples	300
MODIS-Aqua global (MS 4a)	
Bands (λ_c (nm))	412 443 488 531 551 667 678
Band width	10 nm, 15 nm
Spatial resolution	1000 m
Chl-a range (mgm^{-3})	0.0153–25.4985
Nr. of samples	579
MODIS-Aqua *eutrophic* (MS 4b)	
Chl-a range (mgm^{-3})	0.703–25.4985
Nr. of samples	392

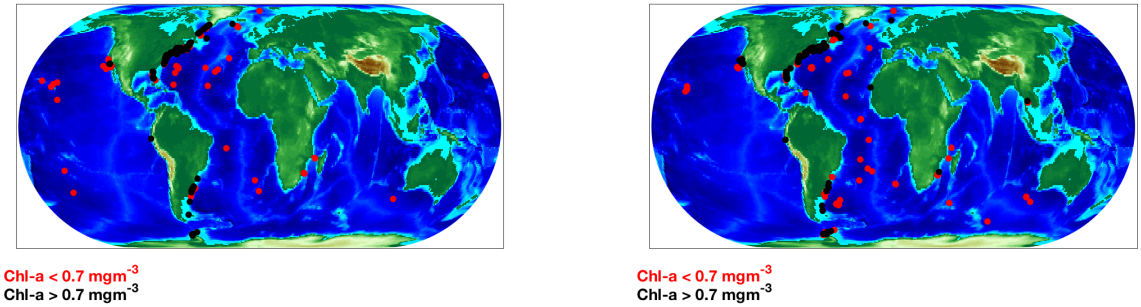


Figure 6. Position of the data for the MERIS (**left**) and MODIS-Aqua (**right**) global dataset. The red and black markers indicate oligotrophic and eutrophic conditions, respectively.

2.3.2. Test Data

We illustrate the results of the AMSA algorithm for *eutrophic* conditions on two full resolution images acquired by MERIS (We obtained the Rrs data from https://oceancolor.gsfc.nasa.gov/cgi/browse.pl?sen=am). The chosen areas are assumed to represent optically complex aquatic environments. One of the images is taken over the eastern coast of USA, and the other image is from the southern part of the Baltic sea. For better visualization purposes, we enlarged a part of the image.

3. Results

We applied AMSA to the eight datasets. For each dataset the total combination of models being evaluated by AMSA is (feature ranking) · (number of spectral bands) · (regression models). The total number of model evaluation are 84 and 96 for the MODIS-Aqua (7 bands) and MERIS (8 bands) datasets, respectively. This means that by using feature ranking methods, the total number of model evaluations are reduced, which speeds up the computational time required to return the most optimal model. Feature ranking reduces the total number of possible model-combinations by assigning relevance to the features. After the spectral bands were ranked, the sequential forward selection approach automatically trained and tested all the possible model combinations, and output the *best* model based on the computed regression performance measures. Table 3 shows the results of the AMSA algorithm for all the datasets. Note that the NMRSE and R^2 values in Table 3 are calculated from the test data.

Table 3. Selected models for the datasets.

Data Label	Model	Spectral Bands	# of Bands	NRMSE	R^2
MS 1a	GPR by VIP	1,...,7	7	0.0983	0.9463
MS 1b	GPR by VIP	4, 5 and 6	3	0.1363	0.9157
MS 2a	GPR by SA GP	1, 2, 5, 6 and 7	5	0.0764	0.9159
MS 2b	SVR by VIP	4, 5 and 6	3	0.1305	0.8332
MS 3a	GPR by ARD	1, 3 and 7	3	0.1082	0.9353
MS 3b	GPR by ARD	1, 3, 5 and 7	4	0.144	0.9068
MS 4a	SVR by VIP	1, 2, 3, 4, 5 and 7	6	0.1094	0.8402
MS 4b	SVR by ARD	1, 2, 3 and 7	4	0.1180	0.7540

In case of all the synthetic datasets (MS 1a, MS 1b, MS 3a and MS 3b) the *best* regression model was found to be the GPR, while for most of the real datasets (MS 2b, MS 4a and MS 4b) the strongest regression was obtained by the SVR model. This can be due to the fact, that the synthetic dataset

has low noise level in comparison to the real dataset (The parameter that handles noise in the GPR model, should have been tuned for the real datasets. However, in order to make the AMSA as robust as possible, we chose to compute the initial noise parameter by following the same formula).

For the MERIS datasets the *best* regression was achieved by using the spectral bands ranked by the VIP ranking method for most of the cases (MS 1a, MS 1b and MS 2b). In case of the MODIS-Aqua datasets, the ARD ranking method seemed to result in the best ranking (MS 3a, MS 3b and MS 4b).

For global monitoring, the *best* model was obtained by using most of the available spectral bands for almost all cases (MS 1a, MS 2a and MS 4a). The only exception was the synthetic MODIS-Aqua dataset, where the *best* model was already achieved by using only 3 spectral bands.

For *eutrophic* conditions AMSA resulted in the *best* regression, when only three or four bands were used. In case of the MERIS datasets (MS 1b and MS 2b), these bands are centered at 510, 560 and 620 nm. For the MODIS-Aqua datasets bands centered at 412, 488 and 678 nm were included in the regression models for both the synthetic (MS 3b) and real (MS 4b) dataset to achieve the strongest regression model.

The regression performance measures show, that the lowest NRMSE and highest R^2 were achieved for the synthetic global datasets (MS 1a and MS 3a), while the models resulting in highest NRMSE and lowest R^2 were for the *eutrophic* real datasets (MS 2b and MS 4b). These results also confirm the challenges of Chl-a content estimation from optically complex waters.

3.1. Chlorophyll-a Maps

In order to illustrate the performance of the *best* models for *eutrophic* conditions, we chose two full resolution MERIS images acquired over areas, which are assumed to be optically complex waters.

3.2. Cross Validation

The outputs of AMSA for the MERIS datasets (MS1b and MS2b), were the GPR and SVR models with bands centered at 510, 560 and 620 nm. We used cross validation to assess the robustness of the models. This was done by randomly dividing the datasets (MS1b and MS2b) into 80% for training and 20% for testing. Then training and testing of the models was performed by computing the NRMSE and R^2 measures. This was done in 500 iterations. The mean values of the computed measures for the cross validations can be seen in Table 4.

Table 4. Results of the cross validation.

	MS1b	
	NRMSE	**R^2**
GPR	0.1497	0.8973
SVR	0.1527	0.836
	MS2b	
	NRMSE	**R^2**
GPR	0.1464	0.824
SVR	0.1438	0.831

The cross validation resulted in very similar computed measures for both models. In case of the MS1b dataset, the GPR model resulted in slightly better values, while for the MS2b data the SVR model showed some improvements. This is in good agreement with the measures output by AMSA (Table 3). The cross validation results also indicate, that in case of the MS1b dataset the difference between the computed regression performance measures for the two models is larger, than is case of the MS2b dataset.

3.3. Visual Illustrations

We applied the AMSA selected GPR and SVR models to the test images. Figures 7 and 8 show the estimated Chl-a content. Figure 7 shows the estimated Chl-a content for the coastal water of East USA by using the GPR (left-column) and SVR (right-column) model with bands centered at 510, 560 and 620 nm. The overall Chl-a maps show that the GPR model predicts higher Chl-a content than the SVR model (top-row). It can be seen in the enlarged area (bottom-row), that there are regions where the SVR model assigns higher values to the Chl-a contents.

Figure 8 shows the estimated Chl-a content maps for the southern part of the Baltic sea. In this case, the overall predicted Chl-a content values (top-row) seem to be more similar for the GPR and SVR models. There are some regional variations in this case as well. The bottom-row in Figure 8 shows the enlarged area. Both models seem to capture the eddies in fine details.

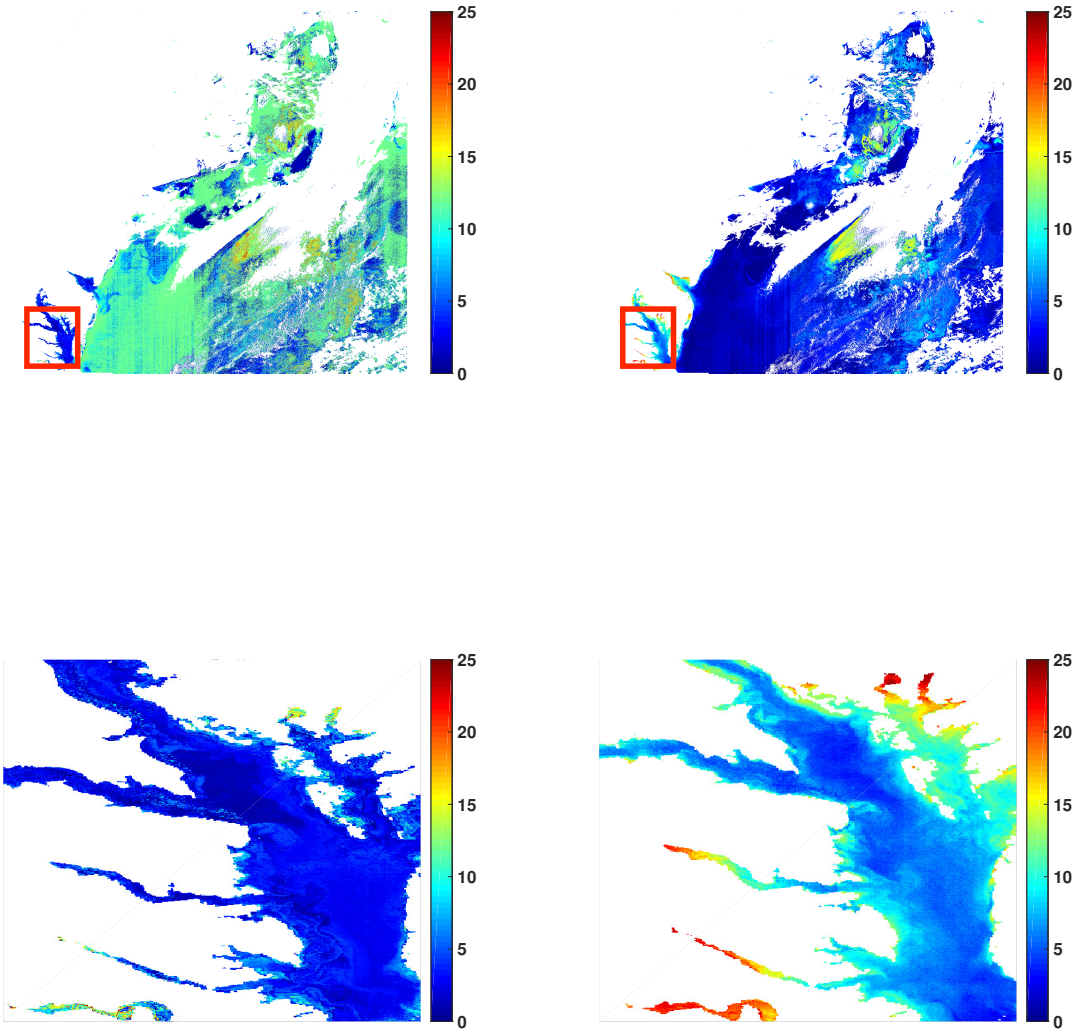


Figure 7. Estimated Chl-a map for the coast of East USA by using the GPR (**left-column**) and SVR (**right-column**) model with bands centered at 510, 560 and 620 nm. The bottom row shows the enlarged area indicated by the red squares.

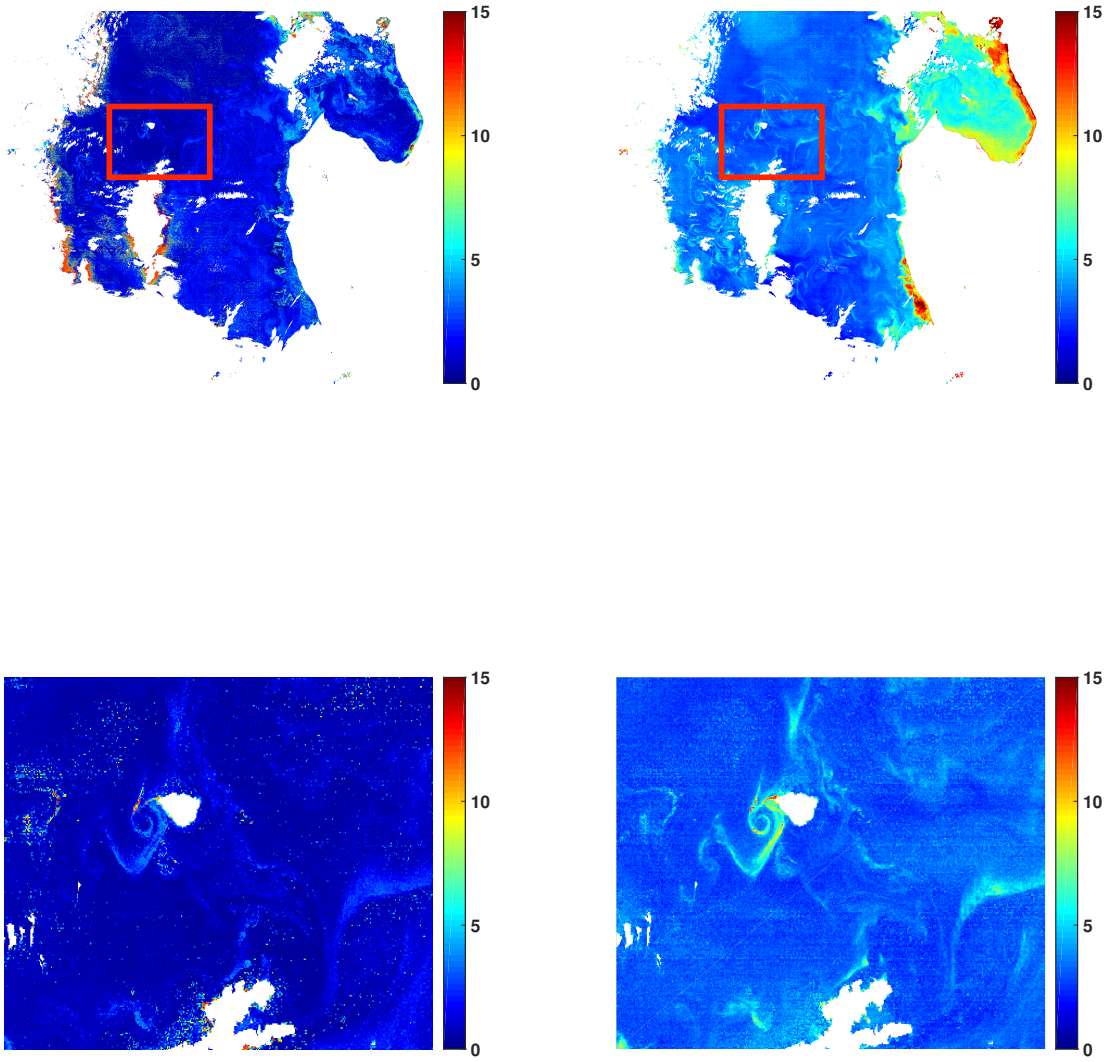


Figure 8. Estimated Chl-a map for the southern Baltic sea by using the GPR (**left-column**) and SVR (**right-column**) model with bands centered at 510, 560 and 620 nm. The bottom row shows the enlarged area indicated by the red squares.

4. Discussion

In this work, we presented a strategy to automatically determine the most suitable model for a given dataset for OC applications. The AMSA approach chooses the *best* model to estimate any water quality parameter from remotely sensed data. AMSA can determine the most suitable model for any regions and sensors. The input to AMSA is the matchup data, and the output is the *best* model. AMSA also outputs the number and combination of features needed to obtain the output model, and the regression performance measures for the *best* model.

We presented the AMSA for oceanic Chl-a content estimation by using ML methods. The AMSA we built here, has four feature ranking methods, the SA GPR, SA SVR, ARD and VIP methods, three regression models, the GPR, SVR and PLSR models, and two regression performance measures, the NRMSE and R^2 to evaluate the regression models. The four feature ranking methods are associated with the three sophisticated regression models, therefore it was a natural choice to include them in the AMSA we chose here.

Both the GPR and SVR models have been shown to be strong regression models for OC applications. They are flexible non-linear kernel methods, using kernel functions in the regression stage.

The choice of the kernel function is strongly dependent on the nature of the data. Here we used the most common kernel, the squared exponential kernel function, which has several advantageous properties. It is a universal kernel [62], and infinitely differentiable. This is a very important property with regard to the SA feature ranking methods, which uses the partial derivatives of the mean function in the GPR and SVR models. The squared exponential kernel function also allows to assess feature relevance by using the length-scale parameter in the function. The ARD feature ranking method uses the inverse of the optimized length-scale parameter to assign relevance. Optimization is done numerically through the maximum likelihood function, which in some cases can be trapped in a local minimum. This might result in erroneous ranking. Furthermore, the initialization of the parameters in the kernel function also have an impact on the optimization, and hence on the regression as well. Therefore, developing the robustness of initializing these parameters from the data should be prioritized in future methodological development.

Despite the PLSR model differs from the kernel machines in its underlying fundamental principles, it also provides the possibility to assess feature relevance through the VIP method. In this work, the AMSA has not output the PLSR model as the most suitable algorithm for Chl-a content estimation. However, in many cases (see Table 3) the VIP method seemed to rank the spectral bands such that the strongest regression was achieved by using the kernel machines. Thus, using the VIP method for feature ranking and kernel machines for regression might be a good combination of methods.

In this work we showed how these ML methods can be used to build an AMSA to estimate Chl-a content in different water conditions and for different sensors. The chosen matchup datasets (MERIS, MODIS-Aqua and the synthesized IOCCG dataset) allowed us to simulate water conditions with increased complexity. Note, although the Chl-a threshold we set here to 0.7 mgm^{-3} might be low for optically complex waters, the observations in the real *eutrophic* datasets above this value, still seem to originate from coastal environments (Figure 6).

AMSA gave as result that for the synthetic datasets the GPR performed *best*, but for most of the real dataset the *best* model was obtained with the SVR model. However, the cross validation results suggest that the SVR model might only have slightly better performance than the GPR model for these datasets.

Generally, for global Chl-a content estimation most of the spectral bands were needed to achieve the *best* regression with the chosen models. This might be due to the larger variety in the data. This result was in contrast to water conditions of increased complexity, where using only three or four of the available spectral bands as inputs resulted the strongest regression. In case of MERIS, these bands were centered at 510, 560 and 620 nm for both the synthetic and real datasets. The spectral band at 510 nm is used to estimate Chl-a content in CDOM rich waters [70]. This is due to the fact that both CDOM and Chl-a has absorption in the blue region of the visible part of the electromagnetic spectrum. The spectral band at 510 nm is mostly representative for the accessory pigments. However, since these pigments are strongly correlated with Chl-a, this band has been widely used to estimate Chl-a content from optically complex waters [16,17,70]. Furthermore, the green spectral band, centered at 560 nm is commonly used for Chl-a estimation, since there is little or no absorption due to phytoplankton in this region. Therefore, this is an important band to use as a reference wavelength in many Chl-a content retrieval algorithms [70]. Using red bands, included the band centered at 620 nm, to estimate Chl-a content has also been commonly used for optically complex waters due to the second absorption peak of the Chl-a [16].

For the *eutrophic* MODIS-Aqua datasets, the spectral bands centered at 412, 488 and 678 nm were found to have importance in the estimation of Chl-a for both the synthesized and real datasets. The bands centered at 488 and 678 nm are in good correspondence with the results for the MERIS datasets. The spectral band centered at 412 nm has also been suggested for Chl-a estimation in complex waters due to the deviation in the absorption between CDOM and Chl-a in this spectral region [71].

Since we used a synthetic resampled dataset to present the performance of AMSA, the model outputs differed from the real datasets. Therefore, we chose to illustrate the *best* models for both the

synthetic and real datasets for *eutrophic* conditions for the MERIS sensor for applications. In case of the coastal part of eastern USA, the GPR assigned in general higher values to the Chl-a content than the SVR model. However, enlarging a region close to shore revealed that the SVR model estimated higher Chl-a than the GPR model. This was also observable for the southern part of the Baltic sea, with less pronounced differences. The illustrative example also showed that both models could capture the same patterns and reveal fine details. Most probably there is a systematic bias occurring in the models. This can be adjusted by tuning the initial parameters in the kernel function, once the model for a given purpose is determined.

5. Conclusions

We conclude, based on this illustrative study, that the AMSA can be a helpful tool for water quality analysis from remote sensing data. It may also be useful in further development of new algorithms. AMSA can be used to objectively compare models with newly introduced algorithms. Furthermore, AMSA might also contribute to improved understanding of the underlying physical processes for various water conditions due to the inclusion of the feature ranking methods.

We have shown that combining ML feature ranking and regression methods in AMSA can reduce computational time and result in improved regression. Furthermore, kernel machines, such as the GPR and SVR models are confirmed to show strong regression power.

For future work, we plan to generalize AMSA by extending the methodology and applying it to different complex aquatic environments and sensors. We also plan to design a flexible AMSA so that user defined models can be added.

Author Contributions: K.B. conceived the idea; K.B. and T.E. developed the strategy, the demonstration of the AMSA model, the statistical analysis, cross validation of the results and application to satellite images. K.B. performed the implementations and prepared the representative datasets from the matchups. K.B. and T.E. analyzed and interpreted the results. K.B. wrote the article with significant contribution from T.E.

Funding: This research received no external funding.

Acknowledgments: This research is partly funded by CIRFA partners and the Research Council of Norway (grant number 237906).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kahru, M.; Mitchell, B.G. Ocean Color Reveals Increased Blooms in Various Parts of the World. *Eos Trans. Am. Geophys. Union* **2008**, *89*, 170. [CrossRef]
2. McClain, C.R. A Decade of Satellite Ocean Color Observations. *Ann. Rev. Mar. Sci.* **2009**, *1*, 19–42. [CrossRef] [PubMed]
3. Gholizadeh, M.H.; Melesse, A.M.; Reddi, L. A Comprehensive Review on Water Quality Parameters Estimation Using Remote Sensing Techniques. *Sensors* **2016**, *16*, 1298. [CrossRef] [PubMed]
4. Wilson, C. The rocky road from research to operations for satellite ocean-colour data in fishery management. *ICES J. Mar. Sci.* **2011**, *68*, 677–686. [CrossRef]
5. Ha, N.T.T.; Koike, K.; Nhuan, M.T. Improved Accuracy of Chlorophyll-a Concentration Estimates from MODIS Imagery Using a Two-Band Ratio Algorithm and Geostatistics: As Applied to the Monitoring of Eutrophication Processes over Tien Yen Bay (Northern Vietnam). *Remote Sens.* **2014**, *6*, 421–442. [CrossRef]
6. Yang, X.E.; Wu, X.; Hao, H.L.; He, Z.L. Mechanisms and assessment of water eutrophication. *J. Zhejiang Univ. Sci. B* **2008**, *9*, 197–209. [CrossRef] [PubMed]
7. Behrenfeld, M.J.; O'Malley, R.T.; Siegel, D.A.; McClain, C.R.; Sarmiento, J.L.; Feldman, G.C.; Milligan, A.J.; Falkowski, P.G.; Letelier, R.M.; Boss, E.S. Climate-driven trends in contemporary ocean productivity. *Nature* **2006**, *444*, 752–755. [CrossRef] [PubMed]
8. Ritchie, J.C.; Zimba, P.V.; Everitt, J.H. Remote Sensing Techniques to Assess Water Quality. *Photogramm. Eng. Remote Sens.* **2003**, *69*, 695–704. [CrossRef]
9. Govindjee. *Bioenergetics of Photosynthesis*; Academic Press: Cambridge, MA, USA, 1975.

10. Volk, T.; Hoffert, M.I. *Ocean Carbon Pumps: Analysis of Relative Strengths and Efficiencies in Ocean-Driven Atmospheric CO_2 Changes*; American Geophysical Union: Washington, DC, USA, 2013; pp. 99–110.
11. Arrigo, K.R.; Robinson, D.H.; Worthen, D.L.; Dunbar, R.B.; DiTullio, G.R.; VanWoert, M.; Lizotte, M.P. Phytoplankton Community Structure and the Drawdown of Nutrients and CO_2 in the Southern Ocean. *Science* **1999**, *283*, 365–367. [CrossRef] [PubMed]
12. Hein, M.; Sand-Jensen, K. CO_2 increases oceanic primary production. *Nature* **1997**, *388*, 526–527. [CrossRef]
13. Hofmann, M.; Worm, B.; Rahmstorf, S.; Schellnhuber, H.J. Declining ocean chlorophyll under unabated anthropogenic CO_2 emissions. *Environ. Res. Lett.* **2011**, *6*, 34–35. [CrossRef]
14. Hu, C.; Lee, Z.; Franz, B. Chlorophyll a algoritms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *J. Geophys. Res.* **2012**, *117*. [CrossRef]
15. Morel, A.; Maritorena, S. Bio-optical properties of oceanic waters: A reappraisal. *J. Geophys. Res. Ocean.* **2001**, *106*, 7163–7180. [CrossRef]
16. O'Reilly, J.E.; Maritirena, S.; Mitchell, B.G.; Siegel, D.A.; Carder, K.L.; Garver, S.A.; Kahru, M.; McClain, C. Ocean color chlorophyll algorithms for SeaWiFS. *J. Geophys. Res.* **1998**, *103*, 24937–24953. [CrossRef]
17. O'Reilly, J.E.; Maritorena, S.; O'Brien, M.C.; Siegel, D.A.; Toole, D.; Menzies, D.; Smith, R.C.; Mueller, J.L.; Mitchell, B.G.; Kahru, M.; et al. SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3. *Nasa Tech. Memo.* **2000**, *11*, 3–8.
18. Werdell, P.J.; Bailey, S.W. An improved bio-optical data set for ocean color algorithm development and satellite data product validation. *Remote Sens. Environ.* **2005**, *98*, 122–140. [CrossRef]
19. Blondeau-Patissier, D.; Gower, J.F.; Dekker, A.G.; Phinn, S.R.; Brando, V.E. A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Prog. Oceanogr.* **2014**, *123*, 123–144. [CrossRef]
20. Matthews, M.W. A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters. *Int. J. Remote Sens.* **2011**, *32*, 6855–6899. [CrossRef]
21. Odermatt, D.; Gitelson, A.; Brando, V.E.; Schaepman, M. Review of constituent retrieval in optically deep and complex waters from satellite imagery. *Remote Sens. Environ.* **2012**, *118*, 116–126. [CrossRef]
22. Gitelson, A.A.; Schalles, J.F.; Hladik, C.M. Remote chlorophyll-a retrieval in turbid, productive estuaries: Chesapeake Bay case study. *Remote Sens. Environ.* **2007**, *109*, 464–472. [CrossRef]
23. Wang, T.S.; Tan, C.H.; Chen, L.; Tsai, Y.C. Applying Artificial Neural Networks and Remote Sensing to Estimate Chlorophyll-a Concentration in Water Body. *Int. Symp. Intell. Inf. Technol. Appl.* **2008**, *1*, 540–544.
24. Canziani, G.; Ferrati, R.; Marinelli, C.; Dukatz, F. Artificial neural networks and remote sensing in the analysis of the highly variable Pampean shallow lakes. *Math. Biosci. Eng.* **2008**, *5*, 691–711. [CrossRef] [PubMed]
25. Gross, L.; Thiria, S.; Frouin, R. Applying artificial neural network methodology to ocean color remote sensing. *Ecol. Modell.* **1999**, *120*, 237–246. [CrossRef]
26. Nabavi-Pelesaraei, A.; Bayat, R.; Hosseinzadeh-Bandbafha, H.; Afrasyabi, H.; Chau, K.-W. Modeling of energy consumption and environmental life cycle assessment for incineration and landfill systems of municipal solid waste management—A case study in Tehran Metropolis of Iran. *J. Clean. Prod.* **2017**, *148*, 427–440. [CrossRef]
27. Chen, X.Y.; Chau, K.W. A Hybrid Double Feedforward Neural Network for Suspended Sediment Load Estimation. *Water Resour. Manag.* **2016**, *30*, 2179–2194. [CrossRef]
28. Alizadeh, M.J.; Kavianpour, M.R.; Kisi, O.; Nourani, V. A new approach for simulating and forecasting the rainfall-runoff process within the next two months. *J. Hydrol.* **2017**, *548*, 588–597. [CrossRef]
29. Zhan, H.; Shi, P.; Chen, C. Retrieval of Oceanic Chlorophyll Concentration Using Support Vector Machines. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 2947–2951. [CrossRef]
30. Camps-Valls, G.; Gómez-Chova, L.; Muñoz-Marí, J.; Vila-Francés, J.; Amorós-López, J.; Calpe-Maravilla, J. Retrieval of oceanic chlorophyll concentration with relevance vector machines. *Remote Sens. Environ.* **2006**, *105*, 23–33. [CrossRef]
31. Blix, K.; Camps-Valls, G.; Jenssen, R. Gaussian Process Sensitivity Analysis for Oceanic Chlorophyll Estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 1265–1277. [CrossRef]
32. Blix, K.; Eltoft, T. Evaluation of Feature Ranking and Regression Methods for Oceanic Chlorophyll-a Estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1403–1418. [CrossRef]

33. Sawaya, K.E.; Olmanson, L.G.; Heinert, N.J.; Brezonik, P.L.; Bauer, M.E. Extending satellite remote sensing to local scales: Land and water resource monitoring using high-resolution imagery. *Remote Sens. Environ.* **2003**, *88*, 144–156. [CrossRef]
34. Brando, V.E.; Dekker, A.G. Satellite hyperspectral remote sensing for estimating estuarine and coastal water quality. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 1378–1387. [CrossRef]
35. Vargas, M.; Brown, C.W.; Sapiano, M.R.P. Phenology of marine phytoplankton from satellite ocean color measurements. *Geophys. Res. Lett.* **2009**, *36*. [CrossRef]
36. D'Alimonte, D.; Melin, F.; Zibordi, G.; Berthon, J.F. Use of the novelty detection technique to identify the range of applicability of empirical ocean color algorithms. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 2833–2843. [CrossRef]
37. Fukushima, H.; Higurashi, A.; Mitomi, Y.; Nakajima, T.; Noguchi, T.; Tanaka, T.; Toratani, M. Correction of atmospheric effect on ADEOS/OCTS ocean color data: Algorithm description and evaluation of its performance. *J. Oceanogr.* **1998**, *54*, 417–430. [CrossRef]
38. Guyon, I.; Elisseeff, A. An Introduction to Feature Extraction. In *Feature Extraction: Foundations and Applications*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–25.
39. Ferreira, E. Model Selection in Time Series Machine Learning Applications. Ph.D. Thesis, University of Oulu: Oulu, Finland, 2015.
40. Verrelst, J.; Alonso, L.; Camps-Valls, G.; Delegido, J.; Moreno, J. Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1832–1843. [CrossRef]
41. Verrelst, J.; Muñoz, J.; Alonso, L.; Rivera, J.P.; Camps-Valls, G.; Moreno, J. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sens. Environ.* **2012**, *118*, 127–139. [CrossRef]
42. Verrelst, J.; Rivera, J.P.; Gitelson, A.; Delegido, J.; Moreno, J.; Camps-Valls, G. Spectral band selection for vegetation properties retrieval using Gaussian processes regression. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 554–567. [CrossRef]
43. Kwiatkowska, E.J.; Fargion, G.S. Application of Machine-Learning Techniques Toward the Creation of a Consistent and Calibrated Global Chlorophyll Concentration Baseline Dataset Using Remotely Sensed Ocean Color Data. *IEEE Trans. Geosci. Remote Sens.* **2003**, *41*, 2844–2860. [CrossRef]
44. Camps-Valls, G.; Muñoz-Marí, J.; Gómez-Chova, L.; Richter, K.; Calpe-Maravilla, J. Biophysical Parameter Estimation With a Semisupervised Support Vector Machine. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 248–252. [CrossRef]
45. Rasmussen, P.M.; Madsen, K.H.; Lund, T.E.; Hansen, L.K. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *NeuroImage* **2011**, *55*, 1120–1131. [CrossRef] [PubMed]
46. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chem. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
47. Ryan, K.; Ali, K. Application of a partial least-squares regression model to retrieve chlorophyll-a concentrations in coastal waters using hyper-spectral data. *Ocean Sci. J.* **2016**, *51*, 209–221. [CrossRef]
48. Lee, Z.P. *Remote Sensing of Inherent Optical Properties: Fundamentals, Test of Algorithms, and Applications*; Technical Report; International Ocean-Colour Coordinating Group, IOCCG: Busan, Korea, 2006.
49. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Process for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
50. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
51. Schölkopf, B.; Smola, A. *Learning with Kernels-Support Vector Machines, Regularization, Optimization and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
52. Murphy, K.P. *Machine Learning A probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012; pp. 496–498.
53. Kung, S.Y. *Kernel Methods and Machine Learning*; Cambridge University Press: Cambridge, UK, 2014; pp. 381–383.
54. Gosselin, R.; Rodrigue, D.; Duchesne, C. A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chem. Intell. Lab. Syst.* **2010**, *100*, 12–21. [CrossRef]
55. Afanador, N.L. Important Variable Selection in Partial Least Squares for Industrial Process Understanding and Control. Ph.D. Thesis, Radboud University Nijmegen: Nijmegen, The Netherlands, 2014.

56. Abdi, H. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Int. Rev. Comput. Stat.* **2010**, *2*, 97–106. [CrossRef]
57. Rännar, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *J. Chem.* **1994**, *8*, 111–125. [CrossRef]
58. De Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chem. Intell. Lab. Syst.* **1993**, *18*, 251–263. [CrossRef]
59. Dayal, B.S.; MacGregor, J.F. Improved PLS algorithms. *J. Chem.* **1997**, *11*, 73–85. [CrossRef]
60. Song, K.; Lu, D.; Li, L.; Li, S.; Wang, Z.; Du, J. Remote sensing of chlorophyll-a concentration for drinking water source using genetic algorithms (GA)-partial least square (PLS) modeling. *Ecol. Inf.* **2012**, *10*, 25–36. [CrossRef]
61. Blix, K.; Camps-Valls, G.; Jenssen, R. Sensitivity Analysis of Gaussian Processes for Oceanic Chlorophyll Prediction. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IGARSS, Milan, Italy, 26–31 July 2015; pp. 996–999.
62. Micchelli, C.A.; Xu, Y.; Zhang, H. Universal Kernels. *J. Mach. Learn. Res.* **2006**, *7*, 2651–2667.
63. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. Multi- and Megavariate Data Analysis. Principles and Applications. *J. Chem.* **2001**, *16*, 261–262.
64. Jonsson, P. Surface Status Classification, Utilizing Image Sensor Technology and Computer Models. Ph.D. Thesis, Mid Sweden University: Sundsvall, Sweden, 2015.
65. Mehmood, T.; Liland, K.H.; Snipen, L.; Sæbø, S. A review of variable selection methods on Partial Least Squares Regression. *Chem. Intell. Lab. Syst.* **2012**, *118*, 62–69. [CrossRef]
66. Liang, L.; Qin, Z.; Zhao, S.; Di, L.; Zhang, C.; Deng, M.; Lin, H.; Zhang, L.; Wang, L.; Liu, Z. Estimating crop chlorophyll content with hyperspectral vegetation indices and the hybrid inversion method. *Int. J. Remote Sens.* **2016**, *37*, 2923–2949. [CrossRef]
67. Sayuri, F.; Watanabe, F.; Alcantara, E.; Walesza, T.; Rodrigues, T.; Imai, N.; Clemente, C.; Barbosa, C.; Luiz, H.; Da, S.; et al. Estimation of Chlorophyll-a Concentration and the Trophic State of the Barra Bonita Hydroelectric Reservoir Using OLI/Landsat-8 Images. *Int. J. Environ. Res. Public Health* **2015**. *12*, 10391–10417.
68. Werdell, P.J.; Bailey, S.W. The SeaWIFS Bio-optical Archive and Storage System (SeaBASS): Current architeture and implementation. In *NASA Technical Memoranda 2002-211617*; Fargion, G.S., McClain, C.R., Eds.; NASA Goddard Space Flight Center: Greenbelt, MD, USA, 2002; p. 45.
69. Werdell, P.J.; Bailey, S.W.; Fargion, G.S.; Pietras, C.; Knobelspiesse, K.D.; Feldman, G.C.; McClain, C.R. Unique data repository facilitates ocean color satellite validation. *EOS Trans. AGU* **2003**, *84*, 387. [CrossRef]
70. Cannizzaro, J.P.; Carder, K.L. Estimating chlorophyll a concentrations from remote-sensing reflectance in optically shallow waters. *Remote Sens. Environ.* **2006**, *101*, 13–24. [CrossRef]
71. Wei, J.; Lee, Z. Retrieval of phytoplankton and colored detrital matter absorption coefficients with remote sensing reflectance in an ultraviolet band. *Appl. Opt.* **2015**, *54*, 636–649. [CrossRef] [PubMed]

Chapter 9

Paper 4:
Remote Sensing of Water Quality Parameters over Lake Balaton by Using Sentinel-3 OLCI

water

MDPI

Article

Remote Sensing of Water Quality Parameters over Lake Balaton by Using Sentinel-3 OLCI

Katalin Blix [1,*], Károly Pálffy [2], Viktor R. Tóth [2] and Torbjørn Eltoft [1]

[1] Department of Physics and Technology, UiT the Arctic University of Norway, P.O. Box 6050 Langnes, NO-9037 Tromsø, Norway; torbjorn.eltoft@uit.no

[2] Balaton Limnological Institute, Hungarian Academy of Science, Centre for Ecological Research, Klebelsberg K. street 3, 8237 Tihany, Hungary; palffy.karoly@okologia.mta.hu (K.P.); toth.viktor@okologia.mta.hu (V.R.T.)

* Correspondence: katalin.blix@uit.no; Tel.: +47-483-49-399

Received: 23 July 2018; Accepted: 4 October 2018; Published: 11 October 2018

Abstract: The Ocean and Land Color Instrument (OLCI) onboard Sentinel 3A satellite was launched in February 2016. Level 2 (L2) products have been available for the public since July 2017. OLCI provides the possibility to monitor aquatic environments on 300 m spatial resolution on 9 spectral bands, which allows to retrieve detailed information about the water quality of various type of waters. It has only been a short time since L2 data became accessible, therefore validation of these products from different aquatic environments are required. In this work we study the possibility to use S3 OLCI L2 products to monitor an optically highly complex shallow lake. We test S3 OLCI-derived Chlorophyll-a (Chl-a), Colored Dissolved Organic Matter (CDOM) and Total Suspended Matter (TSM) for complex waters against in situ measurements over Lake Balaton in 2017. In addition, we tested the machine learning Gaussian process regression model, trained locally as a potential candidate to retrieve water quality parameters. We applied the automatic model selection algorithm to select the combination and number of spectral bands for the given water quality parameter to train the Gaussian Process Regression model. Lake Balaton represents different types of aquatic environments (eutrophic, mesotrophic and oligotrophic), hence being able to establish a model to monitor water quality by using S3 OLCI products might allow the generalization of the methodology.

Keywords: shallow lake; Chl-a; CDOM; TSM; Gaussian process regression; automatic model selection algorithm

1. Introduction

Large freshwater lakes play an important role in the earth's ecosystems, not only because they contain 68% of the global fresh water reservoir, but also because of their economic, social and biological importance as they provide habitats for wildlife, irrigation for agriculture, energy, transport and most importantly water for drinking [1]. The large areal extent of some of these lakes makes traditional water monitoring time and resource consuming, hence inefficient, yet continuous water quality monitoring of lakes is of great importance in detecting environmental changes [2].

Lake Balaton, which covers an area of 596 km^2, is the largest lake in Central Europe and one the most important natural and tourist attractions in Hungary and Central Europe. It provides recreational facilities, and is an aesthetics and cultural resort, which attracts the largest tourist industry in the country [3]. There are several ongoing ecosystem monitoring programs at Lake Balaton. These programs aim to monitor important biological and ecological aspects of biodiversity and food web interactions in the lake. Examples for former monitoring programs for Lake Balaton can be found in [4,5].

The lake has gone through significant changes in the past decades, and only lately were these changes experienced as advantageous. In the 1970s, increased nutrient loads of anthropogenic origin, such as inadequate wastewater management and agricultural runoff, and abiotic factors resulted in degradation of water quality of Balaton. Anthropogenic impacts, i.e., intensification of agricultural activities and increase in the number of settlements along the shore, caused eutrophication of the lake. The eutrophication process was successfully stopped and reversed by introducing a combination of technological and management solutions [6,7]. Recent unpublished data suggests that the lake has recovered and returned to the pre-eutrophic conditions.

As a result of these past events, there is an increasing demand for continuous monitoring of biotic and abiotic changes of the lake. Advances in remote sensing technology allow for the use of satellites for monitoring water constituents. The European Space Agency's (ESA) Ocean and Land Color Instrument (OLCI) onboard the Sentinel 3A and 3B satellites collects data of high spectral and spatial resolutions, and due to the frequent revisit time, they provide the possibility to monitor the water quality of Lake Balaton. In this work, we will study the water monitoring capabilities of Sentinel 3 (S3) for this lake, focusing on three important water quality parameters that affect the lake's water color through scattering and/or absorption: Chlorophyll-a (Chl-a), Colored Dissolved Organic Matter (CDOM) and Total Suspended Matter (TSM).

Chl-a is a major photosynthetic pigment which occurs in phytoplankton, i.e., in the ubiquitous, microscopic, free-floating and suspended organisms found in the illuminated (euphotic) layer of the lakes. The amount of phytoplankton in the water collectively accounts for the trophic state of the lake. Although these organisms are the base of the aquatic food web, their excess could be harmful. Phytoplankton face a great number of abiotic and biotic limitations (light, temperature, other algae, herbivores, etc.), which influence the phytoplankton growth [8]. Nutrient enrichment is very important, since it leads to the eutrophication of lakes, which can lead to alternate states [9].

CDOM is the colored (optically active) fraction of the dissolved organic matter (DOM) of waters, consisting mostly of humic and fulvic acids. Although CDOM is considered as an indicator of DOM [10,11], its origin can vary, as the amount of CDOM is affected by external factors and diffuse sources from the catchment. CDOM in waters is autochtonous, i.e., coming from degradation of algae or macrophytes in the given water body, and/or allochtonous, i.e., coming from the catchment area.

TSM includes a wide range of particulate material for the given water column. The origin of TSM can be local, such as wind induced resuspension and/or distant, for instance from tributaries [12]. TSM contains both organic and inorganic matter, and has a significant impact on the spatial and temporal aspects of the optical properties of the water bodies [13].

Ocean color remote sensing methodology could potentially be a useful tool to track the variability and monitor these water quality parameters [14–16]. In situ observations have documented that Lake Balaton shows a large spatial and temporal variation in the amount and the distribution of Chl-a, CDOM and TSM. This, and the fact that Lake Balaton is regularly monitored by field sampling and measurements, makes the lake particularly well suited for validating retrieval of water quality products for complex aquatic environments from the Copernicus S3 OLCI instrument. The computation of the standard Chl-a, CDOM and TSM maps from OLCI is generally performed by using a Neural Network (NN) method [17,18].

However, optical properties of local environments might show large deviations from the data used for training state-of-the-art models. This can lead to erroneous retrieval of water quality parameters [19]. Therefore, it is often required to use a local model, adjusted to the given area. An alternative powerful regression approach, the Gaussian Process Regression (GPR) model, has lately been investigated for biophysical parameter retrieval from remotely sensed data. The GPR model has been shown to outperform some other parameteric and non-parameteric machine learning methods, such as NNs, in the estimation of these biophysical parameters [20–24]. Hence, the GPR model can be an alternative candidate for estimating water quality parameters from data acquired by S3 OLCI in Lake Balaton.

In this work, our primary objective is to investigate the quality of the global S3 OLCI complex water products for Lake Balaton. For this, we compare the OLCI Level 2 (L2) water quality products (Chl-a, CDOM and TSM) against in situ measurements collected at six fixed stations in the lake in 2017. Hence, the first part of the work is a preliminary study, which aims to investigate the possibility of using S3 OLCI L2 water quality products to monitor Lake Balaton, and at the same time evaluate the performance of S3 OLCI L2 products for this highly complex aquatic environment.

Our secondary objective is to investigate the performance of the Machine Learning GPR approach, tuned locally for Lake Balaton. The GPR model is noted to have several advantageous properties. In addition to it's powerful regression strength, it also provides the possibility to access feature relevance, through feature ranking. As shown in [24,25], the regression strength and the efficiency of the model can be improved by using features selected by using ranking methods. In order to select the most suitable number and combination of spectral bands to be used in the GPR model for estimating Chl-a content of Lake Balaton, we applied the recently published Automatic Model Selection Algorithm (AMSA) [25] to data from the lake, extended with synthesised data of the same Chl-a ranges.

Finally, we visually compare the estimates for S3 OLCI L2 Chl-a products with the locally trained GPR model. Note, we do not specifically aim to compare the estimates of the NN with the locally trained GPR model, since the NN was trained on a dataset which differs in optical properties and size from the matchup data we used to train the local GPR model. Hence, our contribution in this work is to test S3 OLCI L2 water quality products for the diverse Lake Balaton conditions, and to comparatively assess the value of using a locally tuned Machine Learning GPR model.

2. Materials and Methods

2.1. Study Area

Lake Balaton is the largest shallow lake in Central Europe, situated in western Hungary (46°50′ N, 17°40′ E, Figure 1). The surface area of the lake is 596 km^2 with an average depth of 3.5 m, and the volume is about 2×10^9 m^3. Geomorphologically, the lake could be divided into four basins. One half to two thirds of the inflow is discharged by the main tributary, the Zala River, that enters the lake at the westernmost, Keszthelyi Basin. In past decades, the Zala River has carried a great amount of nutrients into Lake Balaton [26]. This resulted in the deterioration of water quality, mostly in the westernmost, Keszthelyi Basin, which led to a prominent trophic gradient in the lake in the 70s–90s [27]. Although phytoplankton biomass in Lake Balaton has significantly decreased during the last two decades, the trophic gradient along the SW-NE axis still exists.

The northern shore of Lake Balaton is steeper than in the south, which results in a difference in depth between the northern and southern shore. This can allow light to reach the bottom near the southern shore in particular. The bottom of the lake is dominated by fine grain size magnesite-bearing calcareous sediments [28]. This can be easily re-suspended under windy weather conditions, resulting in high turbidity. The spatial variability of algal biomass, bathymetry and bottom sediment content lead to high complexity of the optical properties of Lake Balaton.

In situ measurements are collected monthly in ice free periods. Six stations are visited, from the westernmost part of the lake, at the outflow of Zala River (Station 1), ending with Station 6 at the easternmost part of the lake (Figure 1 and Table 1). Usually, the data collection is performed at positions assumed to represent typical characteristics of the lake in those areas.

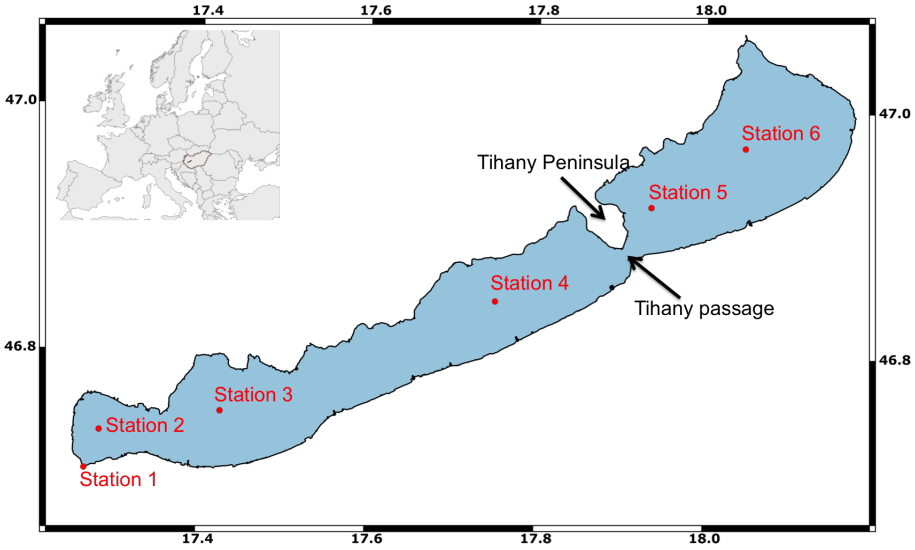


Figure 1. Location of Lake Balaton and the investigated stations.

Table 1. Geographical information of the investigated stations in Lake Balaton.

Station	Basin	Latitude	Longitude	Depth (m)	Area (km^2)
1	Mouth of Zala river	46°42′15.65″ N	17°15′39.16″ E	2.0	0.14
2	Keszthely basin	46°44′09.5″ N	17°16′58.3″ E	2.5	38.00
3	Szigliget basin	46°45′11.2″ N	17°25′14.5″ E	3.5	145.42
4	Szemesi basin	46°50′66.1″ N	17°44′59.5″ E	3.5	185.36
5	Siófok basin (T)	46°55′32.7″ N	17°55′64.9″ E	3.5	75.69
6	Siófok basin (Bf)	46°59′15.7″ N	18°04′74.7″ E	4.0	151.38

2.2. Data

2.2.1. Water Sampling

Chlorophyll-a concentration was determined from integrated water samples, which were collected from the whole water column. Water samples of known volume in replicates of 3 were filtered into GF-C filter (Whatman). Chl-a was spectrophotometrically measured after hot methanol extraction [29].

The concentration of CDOM was measured in Pt (platina) units (mg Pt L^{-1}). Water samples of known volume were filtered through a 0.45 µm pore size cellulose acetate filter, buffered with borate buffer and measured against a blank of buffered Milli-Q water at 440 nm and 750 nm using a Shimadzu UV 160A spectrophotometer. Pt units were calculated from the absorbance values according to [30].

TSM content was determined gravimetrically after sample filtration through a 0.4 µm pore size cellulose acetate filters.

2.2.2. Sentinel-3A OLCI Level-2 Products

Water Quality Products

We used the latest reprocessed (14 February 2018) Sentinel-3A OLCI Full Resolution (FR) Level-2 water quality products for complex waters for validation. These products include Chl-a, CDOM and TSM, retrieved from the spectral measurements by using NN techniques. Even though some part of Lake Balaton seems to show oligotrophic conditions, most of the lake is highly complex. Hence, it is reasonable to use water quality products for complex waters retrieved by NN. For further details on the NN retrieval algorithm we refer to [17,18,31].

There were six cloud free images available for the validation study. We located the coordinates of the six stations in the images, and used a 3 × 3 pixel matrix as described in [32], and applied the

recommended flags. Images were acquired at the days of the in situ measurements or one of the neighboring days. We assume weather conditions were similar. We used the Sentinel Application Platform (SNAP) version 5.0 for processing and preparing the matchups. In total, we could obtain 36 matchups for Chl-a, CDOM and TSM.

We converted the S3 OLCI retrieved CDOM absorbance (m^{-1}) to color (Pt units) by using the expression: $Color_{440}$(g Pt m^{-3}) = 18.216 × a_{440} − 0.209 [30,33].

Remote Sensing Reflectance (Rrs)

We have also extracted the Level-2 Rrs for the spectral bands summarized in Table 2, by following the same procedure as described above. This data was included in the dataset used for training and testing the alternative GPR approach to retrieve the Chl-a water quality parameter.

Table 2. Summary of the Sentinel 3A OLCI spectral bands.

Nr. of Band	Center Wavelength (nm)	Bandwidth (nm)
1	412.5	15
2	442.5	10
3	490	10
4	510	10
5	560	10
6	620	10
7	665	10
8	673.75	7.5
9	681.25	7.5

2.2.3. Synthetic Dataset

An additional synthetic dataset was generated by using HydroLight simulation. The dataset includes Chl-a concentrations over a wide range, with corresponding Rrs values of the S3 OLCI bands. We extracted the values corresponding to the ranges of in situ Chl-a measurements from Lake Balaton. This dataset was used for evaluating the alternative model to estimate Chl-a concentration in Lake Balaton.

2.3. Methodology

2.3.1. Statistical Analysis

We evaluated the S3 OLCI products by comparing the retrieved values to in situ measurements of Chl-a, CDOM and TSM, respectively. For each water quality parameter, we quantified the correspondence in terms of three statistical measure. These measures are the Bias, the Normalized Root Mean Squared Errors (NRMSE), and the Squared Correlation Coefficient (r^2). They are defined by:

$$\text{Bias} = \frac{1}{N}\sum_{i=1}^{N}|(y_i - \hat{y}_i)|, \tag{1}$$

$$\text{NRMSE} = \frac{1}{y_{\max} - y_{\min}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{2}$$

$$r^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2}, \tag{3}$$

where N is the number of observations, y is the in situ measurement, $\hat{y}$ is the S3 OLCI product, $y_{\max}$ is the maximum observed value, $y_{\min}$ is the minimum observed value, and $\overline{y}$ is the mean of the in situ measurements. We have also computed the p-value for assessing the level of significance. The p-value ranges between 0 and 1. A low p-value indicates that the null-hypothesis, which states there is no

relationship between the results and the data, can be rejected. The cut off value is user-defined, and usually set to 0.05. Hence a p-value < 0.05, means that the results are significant, while a p-value > 0.05 indicate little or no significance.

2.3.2. Machine Learning Gaussian Process Regression for Water Quality Estimation

GPR Model

Machine Learning by Gaussian Process Regression (GPR) has been demonstrated to perform excellently in the prediction of water quality parameters from remotely sensed data [20,21,23,24]. Therefore, we have chosen to evaluate this methodology on the matchup data obtained for Lake Balaton in 2017.

The GPR model is a flexible, non-linear kernel method, which learns the functional relationship between the input and output by using a Bayesian framework [34]. In this work, the input data ($\{\mathbf{x}_n \in R^D\}_{n=1}^N$) is formed by using the spectral bands from S3 OLCI Rrs (Table 2), where $n = 1, ..., N$ is the number of measurements, and $d = 1, ..., D$ is the number of spectral bands. The output ($y_{n=1}^N$) is the *in situ* and synthetic measurements for Chl-a.

The functional relationship between the input and output can be written by $y_n = \mathrm{f}(\mathbf{x}_n) + \varepsilon_n$, for $n = 1, ..., N$, where the noise term, ε_n, is assumed to be additive, independently, identically Gaussian distributed, with zero mean and constant variance, i.e., $\varepsilon_n \sim N(0, \sigma^2)$. The GPR model fits a multivariate joint Gaussian distribution over the function values $f(\mathbf{x}_1), ..., f(\mathbf{x}_N) \sim N(\mathbf{0}, \mathbf{K})$, with zero mean and covariance matrix $\mathbf{K}$. Using a Bayesian inversion, the posterior distribution can be analytically computed for the predicted output (y_*) for the corresponding new input ($\mathbf{x}_*$). This can be written by $p(y_*|\mathbf{x}_*, D) = N(y_*|\mu_{\mathrm{GP}*}, \sigma^2_{\mathrm{GP}*})$, where $\mu_{\mathrm{GP}*}$ is the predicted Chl-a, $\sigma^2_{\mathrm{GP}*}$ is the certainty level of the estimate, and D is the training data. The predicted Chl-a can be expressed by $\mu_{\mathrm{GP}*} = \mathbf{k}_*^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y}$, where $\mathbf{k}_*^\top$ is the transposed covariance between the training vector and the test point. For further details on the GPR model we refer to [34].

Automatic Model Selection Algorithm

We used the Automatic Model Selection Algorithm (AMSA), described in [25], to determine the most suitable Chl-a retrieval GPR model for Lake Balaton. AMSA uses feature ranking methods to select the combination of features that results in the strongest regression, based on some predefined quantitative regression performance measures.

Since different ranking methods, may rank the features differently, we used four feature ranking methods here. These are the Sensitivity Analysis (SA) of the GPR and Support Vector Regression (SVR) models, the Automatic Relevance Determination (ARD), and the Variable Importance in Projection (VIP).

For each station, the spectral bands were ranked by these four methods. Then the ranked bands were fed into the GPR model to perform regression, starting with the most relevant band, then the second most important band, and subsequently, the next ranked bands in decreasing order of importance. At each iteration, regression performance measures are computed, and used for evaluating the strength of the GPR with the combination of features. The computation is done until no further improvement is achieved, and is repeated for all the four sets of ranked spectral bands resulting from the SA GPR, SA SVR, ARD and VIP feature ranking methods. This process was done for each station.

Machine Learning GPR for Lake Balaton

We had six matchups available for each of the stations. These matchups were merged with synthetic data of the corresponding Chl-a contents. This allowed us to obtain a larger representative dataset. We used the procedure described above to determine a 'best' GPR model, i.e., a best spectral combination for each station. The purpose of this exercise was to assess if the GPR model is spectrally sensitive to the observed changes in the water conditions.

We also wanted to find a 'best' GPR model for the whole Lake Balaton. Hence, in order to find a GPR model that generalizes best for the whole lake, each of the station-wise 'best' models was next trained and tested on the whole data set. The training and testing were done by carrying out cross validation in 500 iterations. We also evaluated the GPR model using all spectral bands in the input vector.

3. Results

3.1. Data Acquisition

The optical properties of the stations show great spatial and temporal variation. Station 1 is rich in CDOM, hence the color of the water appears dark-brown, while stations 5 and 6 are usually oligotrophic, resulting in blue water color, similarly to open oceans. Figure 2 shows an RGB image acquired in August 2017 by S3 OLCI, supplemented by photos taken at the stations, when the corresponding sampling was carried out. As can be clearly observed the color of the water is changing from station to station.

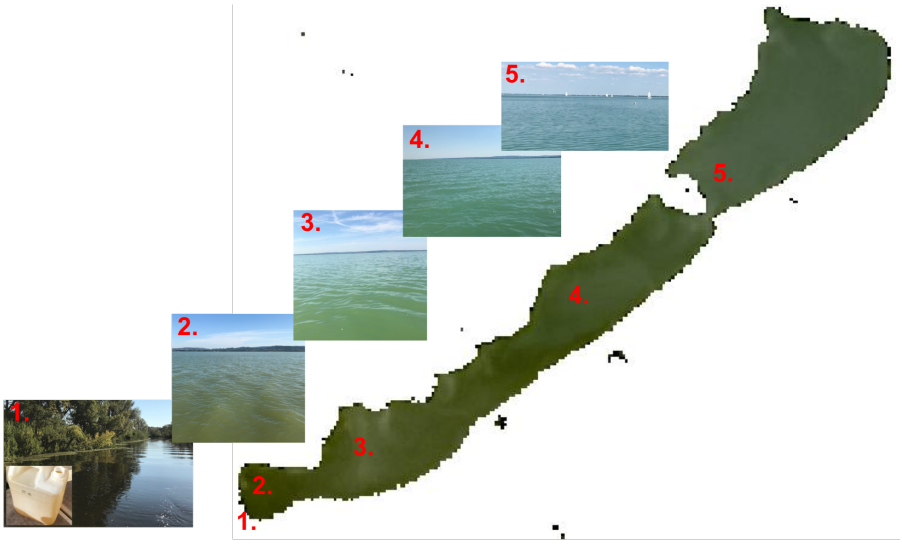


Figure 2. Color gradient in Lake Balaton. The RGB image was acquired by S3 OLCI at 18 August 2017, and the photos were taken at the stations, while the corresponding is situ measurements were collected.

3.1.1. In Situ Measurements

Table 3 summarizes the results of the in situ measurements for every month and station. It can be observed that every month shows large spatial variation in all water quality parameters. More details of these variations are depicted in Figure 6, where the temporal variations of the water quality parameters at each station, together with the S3 OLCI L2 products, are presented. Note that the temporal variations at the stations seem to show differences between the measured parameters. In case of Chl-a, stations 1, 2 and 3 have the largest variations, while stations 4, 5 and 6 have quite steady values. The range of CDOM concentration decreases from station 1 to 6, following the trophic gradient of the lake.

For most of the measurements, we can disregard the contribution of bottom reflectance to the measured signal, since the depth of the euphotic zone does not reach the bottom. However, there were three measurements (in June at station 5 and 6, and in August at station 5), which might include contribution from bottom reflectance. This presumption based on evaluation of the respective computed light extinction coefficients.

Table 3. Summary of the range of the *in situ* measured water quality parameters in 2017. See also Figure 6 for further representation of the variablility of the water quality parameters for every station.

Monthly Range			
Month	**Chl-a (mg m^{-3})**	**CDOM (g Pt m^{-3})**	**TSM (g m^{-3})**
March	2–20	5–64	5–11
May	3–6	3–100	4–21
June	2–25	4–103	2–12
July	5–46	2–95	12–61
August	3–55	5–124	7–21
September	5–33	2–84	4–60
Station Wise Range			
Station	**Chl-a (mg m^{-3})**	**CDOM (g Pt m^{-3})**	**TSM (g m^{-3})**
1	4–55	64–124	4–14
2	5–38	9–19	6–60
3	6–38	6–14	9–51
4	3–6	4–7	8–14
5	2–5	2–7	2–12
6	2–5	2–5	5–15

3.1.2. Satellite Products

Figure 3 shows the Rrs values of the six stations. It can be observed that the CDOM rich stations show a greater variation in the spectrum (Figure 3 top-row) than stations with low CDOM concentrations (Figure 3 bottom-row). This may be explained by the overlapping absorption spectra of Chl-a and CDOM. It might also be a result of the higher Chl-a concentration in itself, since stations with higher CDOM also have higher Chl-a in general. Station 1 and 2 have similar spectra, they are comparable in terms of Chl-a, but they significantly differ in CDOM (and in TSM too) concentration.

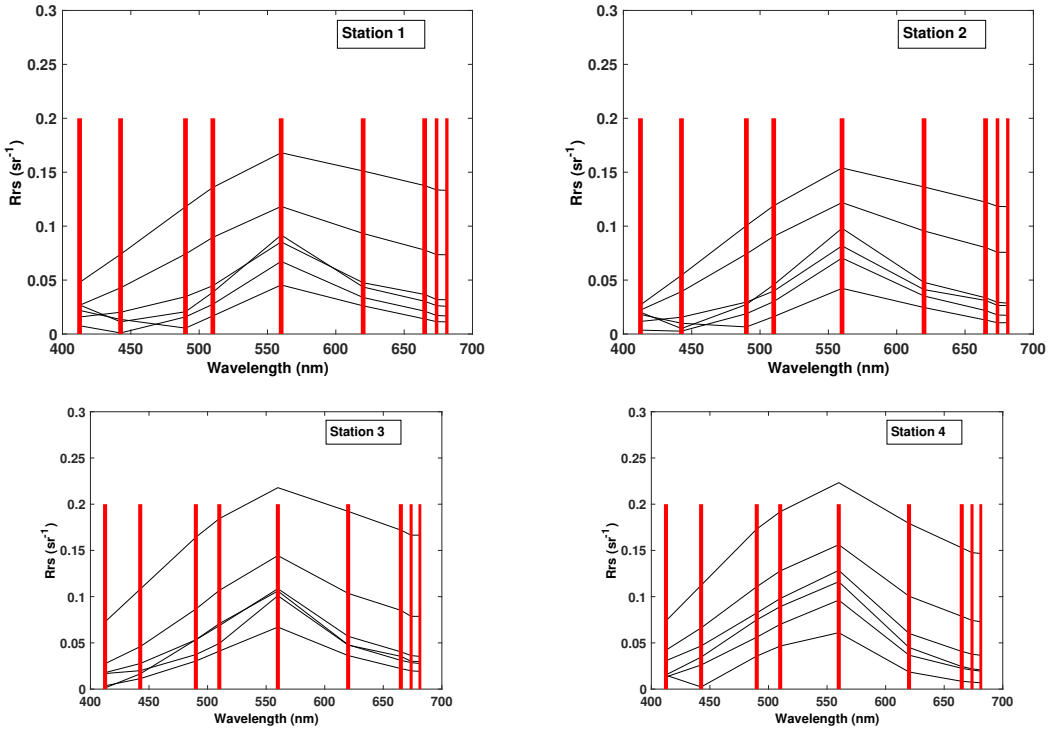


Figure 3. *Cont.*

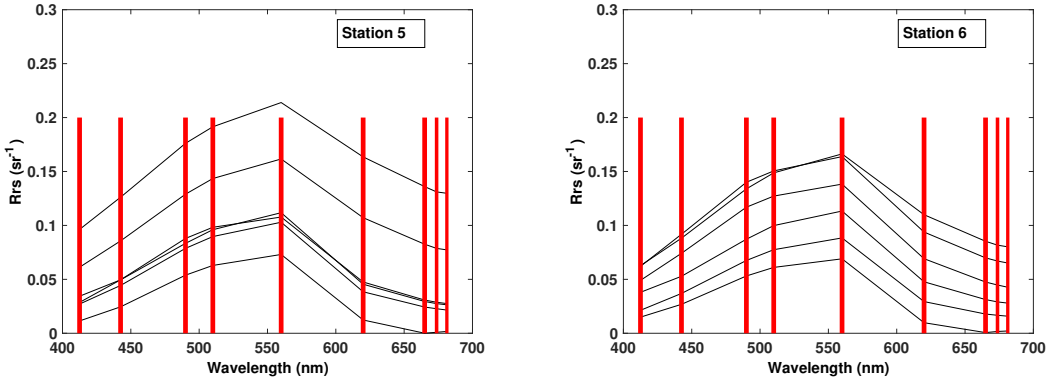


Figure 3. S3 retrieved Rrs values for the 9 spectral bands at the 6 stations. The red bars indicate the position of the bands, and their widths illustrate the relative proportion of the width of the bands.

3.2. Validation

First, we compared the in situ measurements with the S3 OLCI-derived products for all the available data. This allowed us to have an overall understanding about the accuracy of the estimation of the parameters.

Figure 4 shows the correspondence between the histograms of the S3 products and the in situ measurements. It can be observed that for the Chl-a (Figure 4 top) the histograms show similar and overlapping distributions of the estimated values. However, there are no satellite-derived estimates above 30 mg m^{-3}. In case of CDOM (Figure 4 bottom-left), the histograms also reveal similar distributions, although the satellite estimates are shifted to higher values. Furthermore, the satellite estimates could not capture values above 50 g Pt m^{-3}. The histograms of the TSM concentrations (Figure 4 bottom-right) show little agreement. Satellite estimates have a more uniform spread, with a significant shift towards higher values, compared to the in situ measurements.

Figure 5 shows scatter-plots of the measured in situ water quality parameters versus the corresponding satellite-derived products for all stations. It can be observed that in case of the Chl-a content (Figure 5 top), the S3 OLCI Chl-a retrieval algorithm does not estimate concentrations above 30 mg m^{-3}. The opposite of this tendency can be seen for the CDOM (Figure 5 bottom-left) and TSM (Figure 5 bottom-right) concentrations. The satellite products show significantly higher values than the in situ measurements.

With reference to Figure 5, the corresponding r^2 measure showed no correlation for Chl-a, but some correlation for CDOM and TSM. However, the lowest bias was computed for Chl-a, while both for CDOM and TSM the bias were higher. Finally, the NRMSE values were similar for Chl-a and CDOM, and higher for TSM.

In order to detect both monthly and station wise variations in the estimation of water quality products by using S3 OLCI, we compared the in situ measurements with the L2 products for every station and month. The results of the computed statistical measures can be seen in Tables 4 and 5.

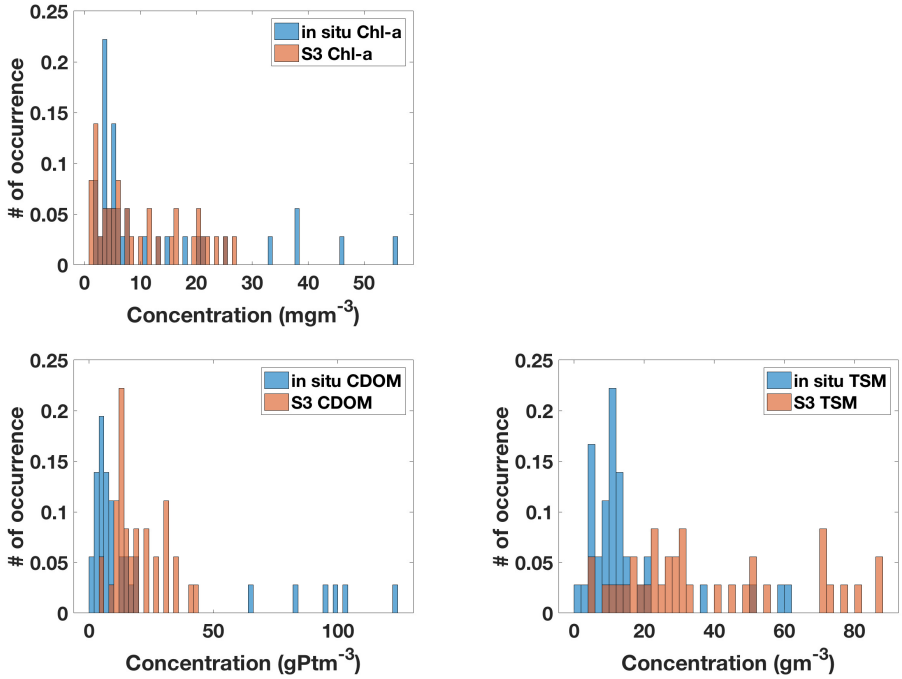


Figure 4. Histogram of the in situ and satellite-derived (S3) water quality concentrations: Chl-a (**top**), CDOM (**bottom-left**) and TSM (**bottom-right**).

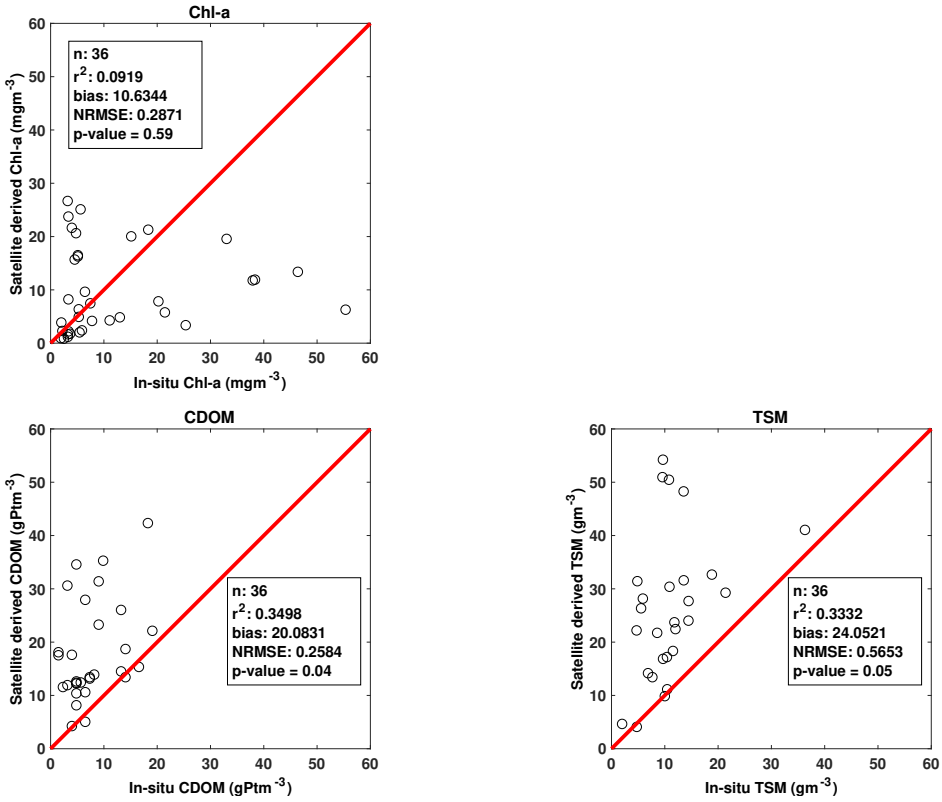


Figure 5. In situ versus satellite-derived water quality concentrations: Chl-a (**top**), CDOM (**bottom-left**) and TSM (**bottom-right**).

Station Wise Analysis

Analyzing the computed statistical measures station-wise revealed poor correspondence between the satellite retrievals and in situ measurements for all water quality parameters (Table 4). Stations 6 and 5 seemed to show the best values for S3 OLCI Chl-a and CDOM retrieval, respectively. These stations correspond to the area where both Chl-a and CDOM concentrations are low (Table 3). For the estimated TSM concentration, station 3 seemed to show the best computed statistical measures.

In order to visually assess the temporal variations of the water quality parameters at the stations, we have depicted the in situ measurements and the S3 OLCI-derived values for every station in Figure 6.

It can be seen that Chl-a is underestimated for stations 1, 2 and 3, with the exception of the May month. For stations 4, 5 and 6 S3 the OLCI algorithm both over- and underestimates Chl-a content. However, these biases seem to decrease as in situ Chl-a content decreases and shows less variations. CDOM is overestimated almost at all stations, with the exception of station 1, where it is underestimated for all months. The TSM concentration is also overestimated at all stations. The largest deviation seems to occur at station 1, while the smallest difference occurs at station 3. This is in good agreement with the computed statistical measures.

Table 4. Validation results: summary of the computed measures for the water quality parameters for every station.

Chl-a				
Station	**NRMSE**	**Bias**	$\mathbf{r^2}$	***p*-Value**
1	0.5405	24.5915	0.1538	0.442
2	0.4326	11.6173	0.0200	0.789
3	0.4311	10.6361	0.0104	0.847
4	3.0461	6.9700	0.0026	0.92
5	3.1640	6.1978	0.0863	0.571
6	1.6021	3.7935	0.4163	0.166
CDOM				
Month	**NRMSE**	**Bias**	$\mathbf{r^2}$	***p*-Value**
1	1.1985	68.0526	0.0290	0.747
2	1.4492	11.3669	7.3^{-4}	0.959
3	1.9824	11.4616	0.2600	0.301
4	4.2534	11.2679	0.1297	0.483
5	2.7703	10.9008	0.3768	0.195
6	2.6877	7.4488	0.2942	0.266
TSM				
Month	**NRMSE**	**Bias**	$\mathbf{r^2}$	***p*-Value**
1	5.0034	42.8140	0.1322	0.478
2	0.6156	25.4804	0.1138	0.51
3	0.6311	18.8408	0.4160	0.166
4	2.6384	20.5351	0.1662	0.42
5	3.2746	22.4133	0.3522	0.21
6	2.0486	14.2292	0.1731	0.41

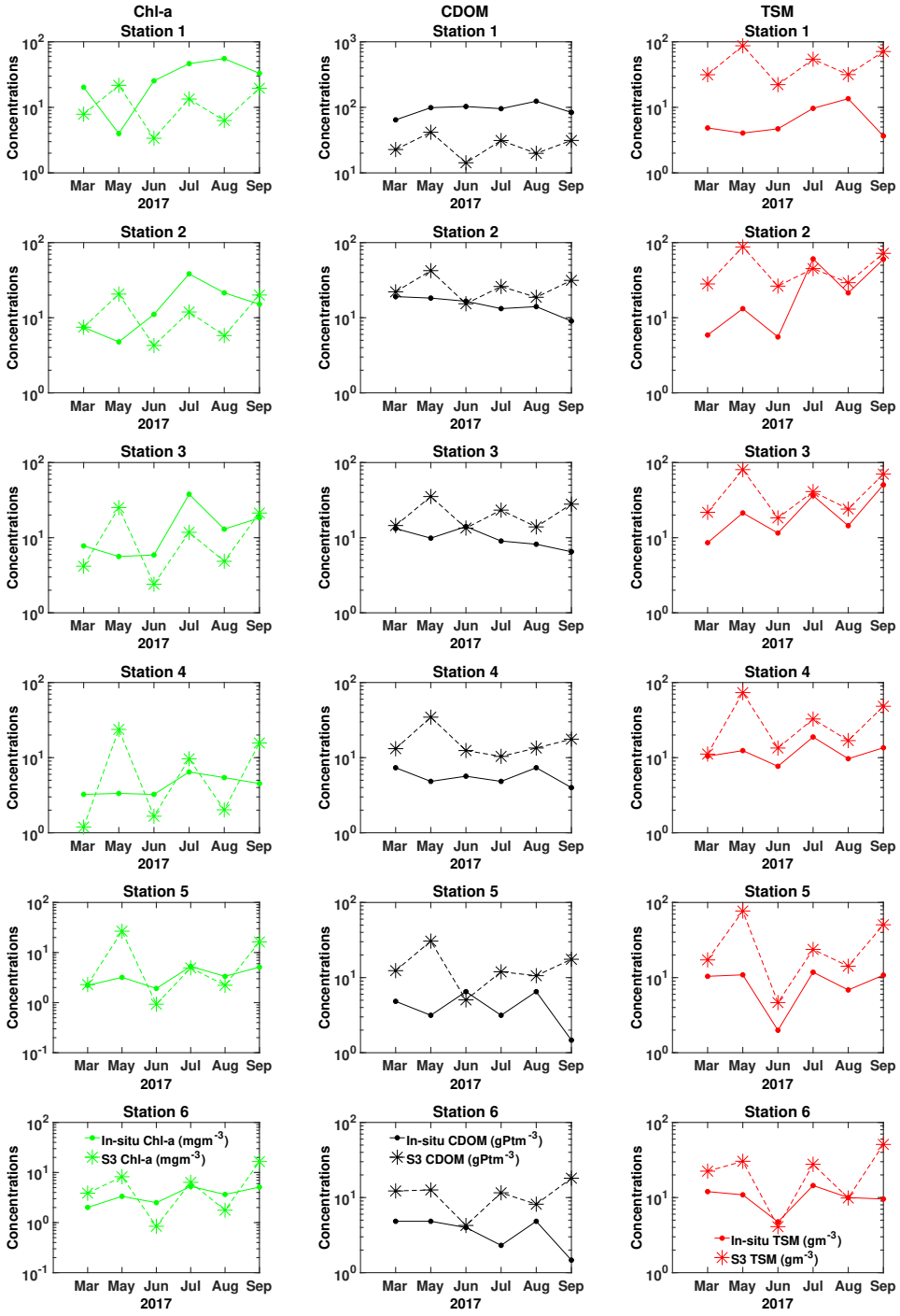


Figure 6. In situ versus satellite-derived water quality products for the stations. Chl-a is shown in the left panel, CDOM in the middle, and TSM in the right panel. Y-axis is presented on a logarithmic scale.

Monthly Analysis

Analyzing the data for each month revealed that the poorest performance was obtained in May for all the three parameters (Table 5). This might be related to the mixing of the water layers, which may cause the sensitivity of the NN algorithm to be biased towards the TSM. However, the computed biases were large for all months and parameters. The highest agreement between in situ observations and S3 OLCI products were found for the Chl-a concentration, with the exception for May. The computed correlation coefficients were found to be low for both the CDOM and TSM concentrations for most of the months.

Table 5. Validation results: summary of the computed measures for the water quality parameters for each month.

Chl-a				
Station	**NRMSE**	**Bias**	r^2	***p*-Value**
March	0.2955	3.3325	0.5923	0.074
May	7.421	16.9651	0.0684	0.617
June	0.4077	6.0809	0.5064	0.113
July	0.4949	15.0353	0.8098	0.015
August	0.4106	13.2152	0.6798	0.044
September	0.3489	9.1772	0.5459	0.093
CDOM				
Month	**NRMSE**	**Bias**	r^2	***p*-Value**
March	0.2984	11.1337	0.6453	0.054
May	0.3357	28.6014	0.2296	0.336
June	0.3670	16.5093	0.1972	0.378
July	0.3002	19.1573	0.5500	0.092
August	0.3586	21.3310	0.4514	0.144
September	0.3305	23.7658	0.3420	0.22
TSM				
Month	**NRMSE**	**Bias**	r^2	***p*-Value**
March	2.2354	13.3317	0.5928	0.073
May	3.6864	60.4756	0.0012	0.948
June	1.2296	9.0269	0.1444	0.4575
July	0.4206	17.3275	0.1049	0.5313
August	0.6787	8.3610	0.5797	0.079
September	0.7089	35.7899	0.3694	0.20

3.3. GPR for Lake Balaton Chlorofyll: A Content Retrieval

The validation results above indicate that there is a need for a local model in the estimation of water quality parameters over Lake Balaton based on S3 OLCI data. Therefore, in the following section we present the results of a locally tuned GPR model for Chl-a content.

3.3.1. AMSA for Improving the GPR Model for Chl: A Content Retrieval

We used AMSA to determine the number and positions of the most important spectral bands for the six stations for Chl-a. This was done by extracting the Chl-a and Rrs pairs from the synthetic dataset corresponding to the in situ Chl-a ranges for every station. Then the synthetic dataset was merged with the in situ data. This was used as input to AMSA. Then the first stage of AMSA, feature ranking, was done by using all the available samples (Table 6 Nr. of samples) for each station. The feature selection and evaluation part of AMSA were performed by splitting the data to training and testing samples. The test samples were formed by the in situ measurements, while the training samples held the rest of the samples. Table 6 summarizes the results for the stations. The *p*-value was below 0.0001 for all cases. Note, the results in Table 6 show the strongest models for the stations. However, using only few ranked bands as input to the GPR model already resulted in strong performance. The goal is to determine the 'best' model, therefore, these results are not reported here.

Table 6. Summary of the stationwise evaluation of AMSA for Chl-a for the merged dataset.

Station	Nr. of Samples	Nr. of Bands	NRMSE	r^2
1	769	8	0.0187	0.9995
2	675	6	0.0653	0.9984
3	646	9	0.0273	0.9988
4	784	5	0.0187	1.0000
5	695	6	0.0501	0.9979
6	745	3	0.0657	1.0000

The spectral bands needed to achieve the 'best' GPR model are summarized in Figure 7. It can be observed that for all stations, bands centered at 673.25 and 681.25 nm were needed to obtain the strongest regression for Chl-a content estimation in the GPR model. For station 6, using only three bands were already enough to determine the 'best' model. These three bands are centered at 442.5, 673.75 and 681.25 nm, which is in good correspondence with the Chl-a absorption and fluorescence spectrum. Station 6 is known to be less affected by CDOM, hence possibly the first absorption peak of Chl-a is not masked by CDOM.

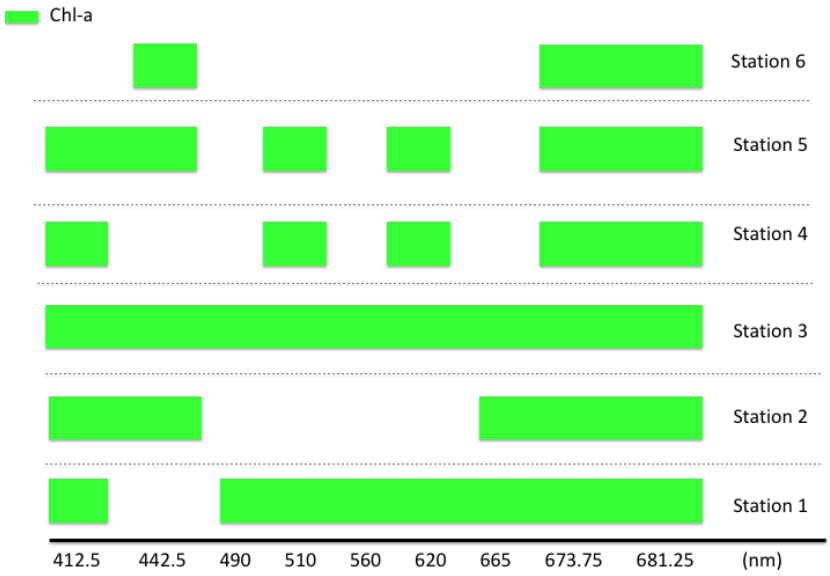


Figure 7. The most important spectral bands of Chl-a for each station.

3.3.2. Determining a General Model for Chl-a Content Retrieval

We used the results of the station-wise feature ranking from AMSA to determine a general GPR model tuned for the whole lake. Firstly, we used all the available spectral bands in the GPR model. This was defined as our reference model. Then we used the results of the ranking methods presented in Figure 7 for the stations to perform regression experiment involving the complete merged dataset.

Table 7 shows the computed statistics for the GPR models. Note that for Station 3, AMSA suggested that all bands were needed. All stations considered, the general observation was that the lowest bias was achieved by using bands centered at 412.5, 510, 620, 673.75 and 681.25 nm, and the lowest NRMSE was obtained with the bands centered at 442.5, 673.75 and 681.25 nm. Hereafter, we refer to these models as the all bands, the 5-band and the 3-band models, respectively, The *p*-value, which was very low in all cases, and r^2 measure could not reveal any differences between the models.

Table 7. Summary of the computed statistical measures for the six GPR models. The *p*-value was significantly below 0.0001 for all cases.

Station	Bands Used in the GPR Model	NRMSE	Bias	r^2
	All	0.00448	0.2056	1.0000
1	1, 3, 4, 5, 6, 7 and 8	0.0046	0.2047	1.0000
2	1, 2, 6, 7 and 8	0.0047	0.2037	1.0000
3	All	0.00448	0.2056	1.0000
4	1, 4, 6, 8 and 9	0.0031	0.1351	1.0000
5	1, 2, 4, 6, 8 and 9	0.0034	1.1414	1.0000
6	2, 8 and 9	0.003	0.1365	1.0000

3.3.3. Cross Validation

We used all bands, 5-band and 3-band models to perform cross-validation. For this purpose, we merged the synthetic and in situ data for all stations. In order to reduce computational time we used a subset of this merged dataset. This data was formed by sampling from the values from every station, hence the data was still representative for the whole lake. The total number of samples were 624.

We used this representative dataset to randomly draw samples from both the synthetic and in situ measurements for training the models, while the rest of the data was used for testing. The total number of samples used for training and testing, was 430 and 194, respectively. Then we computed the statistical measures on the test set. This was done for 500 times. The results are summarized in Table 8. It can be seen that both the 5-band and 3-band models resulted in improved performance in comparison to the all band model. The lowest NRMSE and bias were achieved by the 5-band model, and the highest r^2 was obtained with the 3-band model. The *p*-value were low in all cases. Note, both models include bands centered at 673.75 and 681.25 nm. These results confirm the importance of using these bands to estimate Chl-a in optically highly complex waters.

Table 8. Summary of the cross validation. The results show the mean values of the NRMSE, Bias, r^2 and *p*-value by using the GPR model with all bands, 5-bands and 3-band models for 500 iterations.

GPR Model	NRMSE	Bias	r^2	*p*-Value
All bands	0.1136	2.2532	0.7909	<0.0001
5-bands	0.1042	2.0563	0.8253	<0.0001
3-bands	0.1043	2.1247	0.8298	<0.0001

3.3.4. Chl-a Maps

By comparing the satellite products with the ground-truth measurements for all months, revealed that May had the largest deviations according to the statistical measures for all water quality parameters (Table 5).

The RGB image of Lake Balaton acquired at the 22 May 2017 can be seen in Figure 8. The yellowish pattern are most likely due to the mixing of the bottom layers. These patterns show good correspondence with the dominating wind direction, Northern winds, and the geography of the Northern shore of the lake. Note, the patches, which appear green in the image, are in areas well-known to be shadowed for the Northern winds.

Figure 9 shows the estimated Chl-a content by using S3 OLCI NNs (left) and the 5-band GPR model (right). It can be observed that the S3 OLCI product overestimates Chl-a content. This might be due to a too strong sensitivity to TSM. Comparing the RGB image and the Chl-a estimates-derived by S3 OLCI, we see that it follows the pattern of thoroughly mixed waters with higher TSM. the 5-band GPR model seem to show less (no) sensitivity to the TSM concentration. Chl-a estimates show higher values in the western basin, around the Tihany passage and also around the eastern basin. Fine details and patterns can also be observed in the image produced by the 5-band GPR model.

Patches with higher Chl-a content seem to appear in areas, where the primary productivity is assumed to be increased. The map (Figure 9 right) revealed regions with higher Chl-a values, in the western and eastern side of the Tihany passage. This is an interesting feature, which can be explained by the bathymetry of the lake. The water depth drops around the southern part of the passage [35,36], allowing benthic algae to appear in surface waters under suitable mixing conditions. The RGB image showed heavy mixing in the particular month we chose for this illustration. Favorable wind direction and speed might have caused the occurrence of a current in the Tihany passage, transporting Chl-a rich waters from the western part to the eastern side.

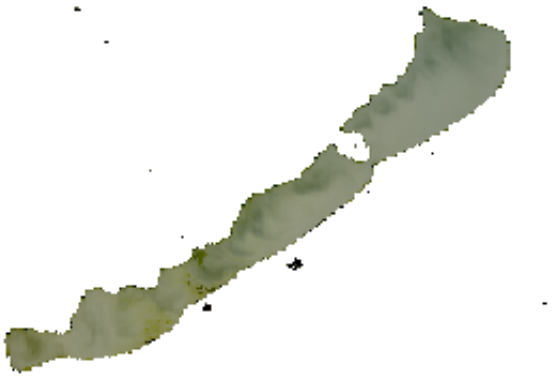


Figure 8. The RGB image of Lake Balaton at the 22 May 2017.

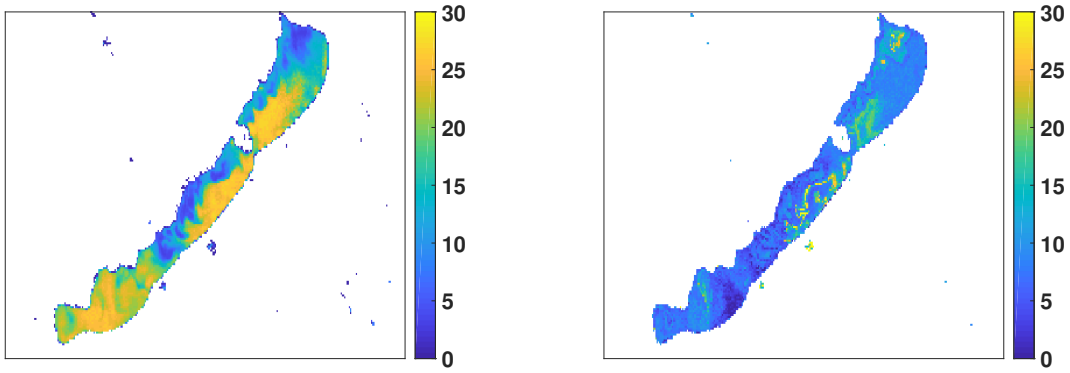


Figure 9. Chl-a content estimates by S3 OLCI (**left**) and the 5-band GPR (**right**). The units are in mg m^{-3}.

4. Discussion

In this work, we studied the possibility of using S3 OLCI L2 products to monitor water quality parameters in Lake Balaton. For this, we first used in situ measurements of Chl-a, CDOM and TSM to evaluate the performance of the state-of-the-the-art complex water algorithm for S3 OLCI. The overall finding was that the correlation between in situ measurements and the S3 OLCI L2 products was low and not significant. It was the lowest value for Chl-a content, and somewhat higher for CDOM and TSM. Note, there are few published validation results for S3 OLCI L2 water quality parameters for complex waters, since S3 OLCI data only lately has become available. However, for the MEdium Resolution Imaging Spectrometer (MERIS), which had similar spectral and spatial resolution as S3 OLCI, similar validation results have been documented using NN algorithms to retrieve water quality parameters. This includes the over and underestimation of Chl-a concentration [37], and large overestimation of TSM [31].

The station-wise study resulted in the best qualitative correspondence, i.e., lowest NRSME and bias, and highest correlation, for Chl-a and CDOM at stations representing oligotrophic waters (Stations 5 and 6). The range of the in situ measurements at these stations were between 2 and 5 mg m^{-3} for Chl-a and 2–7 g Pt m^{-3} for CDOM, which are the lowest of all stations. Here, the TSM concentrations were also in the lower ranges, in comparison to the other stations. The computed measures did not reveal any significant differences between the stations for TSM.

The monthly analyses showed that the S3 OLCI estimates were in quite good correspondence with the observations for Chl-a. CDOM and TSM estimates had less agreement with the in situ measurements. We found that May resulted in the poorest fit in terms the computed statistical measures. The in situ Chl-a ranges were lowest in May, but conversely, for this month the CDOM and TSM ranges were large.

These results might be related to inaccuracies in the atmospheric correction and water quality retrieval algorithms because of the lack of training data from Lake Balaton in the dataset used to establish the state-of-the-the-art models for complex waters [38].

The above results motivated us to investigate the capabilities of a locally trained GPR model for monitoring the complex environment of Lake Balaton. The overall findings for the S3 OLCI products showed the poorest performance for Chl-a content retrieval, which is the most important water quality parameter. Therefore, we studied the possibility of improving Chl-a content estimation in Lake Balaton by using the alternative approach. We obtained a larger, more representative dataset suitable for evaluating a locally tuned model by extending the in situ measurements with a synthetic dataset for S3 OLCI, generated for complex waters.

Using the AMSA approach to determine the most suitable number and combination of spectral bands to be used in the GPR model, we obtained significant improvements in regression strength. Even though the four feature ranking methods currently implemented in AMSA are-derived from different mathematical principles, the ranking showed high consistency. Our station-wise feature ranking experiment showed that the most relevant bands were highly dependent of the water properties and the water quality parameter in question. Our study suggested that for Chl-a estimation in Lake Balaton the bands 1, 4, 6, 8 and 9 are the most important in the GPR model. These bands have been previously shown to be sensitive to Chl-a in different datasets [24]. Bands positioned in the red part of the electromagnetic spectrum, corresponding to the longer wavelengths, might be important due to the second absorption peak of the Chl-a molecule [39]. Recent studies have presented the benefit of using S3 OLCI red bands to monitor Chl-a in optically complex environments [40,41]. Chl-a estimation can be improved by using models with these red bands. This is in good correspondence with our results. The station-wise analysis of AMSA showed that inclusion of red bands were necessary to obtain the 'best' GPR model for all cases. The 5-band model for Lake Balaton also was found to use these red bands as inputs to achieve improved Chl-a retrieval. The inclusion of additional blue-green bands has been shown to be advantageous, when the aquatic environment has large variation in Chl-a content [42]. Our results also indicated that bands corresponding to lower relative wavelengths are also required to optimize the GPR model for the lake.

We visually compared the predictive power of the locally tuned 5-band GPR model with S3 OLCI L2 Chl-a products for Chl-a estimation. The Chl-a map produced by using S3 OLCI L2 NN algorithm seemed to show high sensitivity to the TSM content. The estimated Chl-a contents were significantly above the in situ measurements, indicating overestimation. This is in good agreement with the validation results, which showed that S3 OLCI assigns high values to Chl-a content below about 10 mg m^{-3}. This is a surprising finding, since the state-of-the-the art NN was trained on samples containing values up to 30 mg m^{-3}. A possible explanation for this overestimation is that complex optical properties of the lake results in sensitivity to other water constituents, such as TSM. This might lead to erroneous Chl-a content estimates. This also suggests the importance of using an alternative flexible approach for local, highly complex aquatic environment. The Chl-a map produced by the 5-band GPR model seemed to show better correspondence with the measured Chl-a content range for

the particular month. The model could capture fine details and patches, which can be explained by the bathymetry and currents in the lake.

5. Conclusions

Our analysis showed that S3 OLCI provides the excellent possibility to monitor Lake Balaton, due to its spectral and spatial resolution and the good quality of the data. However, our validation results indicate the need of algorithm development for optically highly complex waters. We can conclude that based on the evaluation study of the alternative approach on the composite dataset, the GPR model seems to be able to improve the estimation of Chl-a concentration in Lake Balaton.

We believe that the development of an accurate, fast and robust water quality retrieval model for Lake Balaton would certainly be generally beneficial. This is due to the fact that Lake Balaton's optical properties represent different kinds of aquatic environments: eutrophic, mesotrophic, oligotrophic, turbid and clear waters, and possible contribution of bottom reflectance. Hence, the lake represents a unique test site for the development of retrieval models for water quality parameters for optically complex waters.

For future work, we will collect in situ radiometric data, which might allow to further exploit the optical properties of Lake Balaton and understand eventual challenges with regard to the atmospheric correction algorithm. Furthermore, we will further test and validate the alternative model presented here on data originating from various other water bodies. This might allow us to understand the generalization capabilities of the 5-band GPR model.

Author Contributions: K.B. conceived the idea, methodology, performed the implementations, validation, formal analysis, data processing and analysis, visualization and prepared the original draft. K. P., V. R. T. and T. E. contributed to the investigation, interpretation of the results, writing-review and editing. T. E. supervised the work.

Funding: This research received no external funding.

Acknowledgments: This research is partly funded by CIRFA partners and the Research Council of Norway (grant number 237906). We thank the Hungarian Academy of Science, Center for Ecological Research, Balaton Limnological Institute for providing the data, and to Balázs Németh for his useful comments and discussions. We thank EUMETSAT for producing and distributing the S3 OLCI L2 data. We thank Nima Pahlevan (National Aeronautics and Space Administration/Goddard Space Flight Center) for providing the synthetic dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Beeton, A.M. Large freshwater lakes: Present state, trends, and future. *Environ. Conserv.* **2002**, *29*, 21–38, doi:10.1017/S0376892902000036.
2. Palmer, S.; Zlinszky, A.; Balzter, H.; Perea, V.N.; Tóth, V. Copernicus Framework for Monitoring Lake Balaton Phytoplankton. In *Earth Observation for Land and Emergency Monitoring*; Wiley-Blackwell: Hoboken, NJ, USA, 2017; Chapter 10, pp. 173–191, doi:10.1002/9781118793787.ch10.
3. Rátz, T.; Michalkó, G.; Kovács, B. The influence of Lake Balaton's tourist milieu on visitors' quality of life. *Tourism Int. Interdiscip. J.* **2008**, *56*, 127–142.
4. Mózes, A.; Présing, M.; Vörös, L. Seasonal Dynamics of Picocyanobacteria and Picoeukaryotes in a Large Shallow Lake (Lake Balaton, Hungary). *Int. Rev. Hydrobiol.* **2006**, *91*, 38–50, doi:10.1002/iroh.200510844.
5. Riddick, C.A.L.; Hunter, P.D.; Tyler, A.N.; Martinez-Vicente, V.; Horváth, H.; Kovács, A.W.; Vörös, L.; Preston, T.; Présing, M. Spatial variability of absorption coefficients over a biogeochemical gradient in a large and optically complex shallow lake. *J. Geophys. Res. Oceans* **2015**, *120*, 7040–7066, doi:10.1002/2015JC011202.
6. Somlyódy, L.; van Straten, G. Background to the Lake Balaton Eutrophication Problem. In *Modeling and Managing Shallow Lake Eutrophication with Application to Lake Balaton*; Springer: Berlin, Germany, 1986; pp. 3–18.
7. Istvánovics, V.; Clement, A.; Somlyody, L.; Specziár, A.; G.-Tóth, L.; Padisak, J. Updating water quality targets for shallow Lake Balaton (Hungary), recovering from eutrophication. *Hydrobiologia* **2007**, *581*, 305–318, doi:10.1007/s10750-006-0509-1.

8. Reynolds, C.S.; Scheiz, Z. The response of phytoplankton communities to changing lake environments. *Swiss J. Hydrol.* **1987**, *49*, 220–236, doi:10.1007/BF02538504.
9. Scheffer, M.; Hosper, S.; Meijer, M.L.; Moss, B.; Jeppesen, E. Alternative equilibria in shallow lakes. *Trends Ecol. Evol.* **1993**, *8*, 275–279, doi:10.1016/0169-5347(93)90254-M.
10. Brezonik, P.L.; Olmanson, L.G.; Finlay, J.C.; Bauer, M.E. Factors affecting the measurement of CDOM by remote sensing of optically complex inland waters. *Remote Sens. Environ.* **2015**, *157*, 199–215, doi:10.1016/j.rse.2014.04.033.
11. Toming, K.; Kutser, T.; Tuvikene, L.; Viik, M.; Nõges, T. Dissolved organic carbon and its potential predictors in eutrophic lakes. *Water Res.* **2016**, *102*, 32–40, doi:10.1016/j.watres.2016.06.012.
12. Madsen, J.D.; Chambers, P.A.; James, W.F.; Koch, E.W.; Westlake, D.F. The interaction between water movement, sediment dynamics and submersed macrophytes. *Hydrobiologia* **2001**, *444*, 71–84, doi:10.1023/A:1017520800568.
13. Giardino, C.; Oggioni, A.; Bresciani, M.; Yan, H. Remote Sensing of Suspended Matter in Himalayan Lakes. *Mt. Res. Dev.* **2010**, *30*, 157–168, doi:10.1659/MRD-JOURNAL-D-09-00042.1.
14. Büttner, G.; Korándi, M.; Gyömörei, A.; Köte, Z.; Szabó, G. Satellite remote sensing of inland waters: Lake Balaton and reservoir Kisköre. *Acta Astronaut.* **1987**, *15*, 305–311.
15. Palmer, S.C.; Hunter, P.D.; Lankester, T.; Hubbard, S.; Spyrakos, E.; Tyler, A.N.; Présing, M.; Horváth, H.; Lamb, A.; Balzter, H.; et al. Validation of Envisat MERIS algorithms for chlorophyll retrieval in a large, turbid and optically-complex shallow lake. *Remote Sens. Environ.* **2015**, *157*, 158–169, doi:10.1016/j.rse.2014.07.024.
16. Tyler, A.N.; Svab, E.; Preston, T.; Présing, M.; Kovács, W.A. Remote sensing of the water quality of shallow lakes: A mixture modelling approach to quantifying phytoplankton in water characterized by high-suspended sediment. *Int. J. Remote Sens.* **2006**, *27*, 1521–1537, doi:10.1080/01431160500419311.
17. Doerffer, R.; Schiller, H. The MERIS Case 2 water algorithm. *Int. J. Remote Sens.* **2007**, *28*, 517–535, doi:10.1080/01431160600821127.
18. Brockmann, C.; Doerffer, R.; Peters, M.; Kerstin, S.; Embacher, S.; Ruescas, A. Evolution of the C2RCC Neural Network for Sentinel 2 and 3 for the Retrieval of Ocean Colour Products in Normal and Extreme Optically Complex Waters. In *Living Planet Symposium, Proceedings of the Conference, Prague, Czech Republic, 9–13 May 2016*; ESA Special Publication: Paris, France, 2016; Volume 740, pp. 54.
19. Blondeau-Patissier, D.; Gower, J.F.; Dekker, A.G.; Phinn, S.R.; Brando, V.E. A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Prog. Oceanogr.* **2014**, *123*, 123–144, doi:10.1016/j.pocean.2013.12.008.
20. Pasolli, L.; Melgani, F.; Blanzieri, E. Gaussian Process Regression for Estimating Chlorophyll Concentration in Subsurface Waters From Remote Sensing Data. *IEEE Geosci. Remote Sens. Lett.* **2010**, *7*, 464–468, doi:10.1109/LGRS.2009.2039191.
21. Verrelst, J.; Muñoz, J.; Alonso, L.; Rivera, J.P.; Camps-Valls, G.; Moreno, J. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sens. Environ.* **2012**, *118*, 127–139.
22. Verrelst, J.; Alonso, L.; Camps-Valls, G.; Delegido, J.; Moreno, J. Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1832–1843, doi:10.1109/TGRS.2011.2168962.
23. Blix, K.; Camps-Valls, G.; Jenssen, R. Gaussian Process Sensitivity Analysis for Oceanic Chlorophyll Estimation. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2017**, *10*, 1265–1277, doi:10.1109/JSTARS.2016.2641583.
24. Blix, K.; Eltoft, T. Evaluation of feature ranking and regression methods for oceanic chlorophyll-a estimation. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2018**, *11*, 1403–1418, doi:10.1109/JSTARS.2018.2810704.
25. Blix, K.; Eltoft, T. Machine Learning Automatic Model Selection Algorithm for Oceanic Chlorophyll-a Content Retrieval. *Remote Sens.* **2018**, *10*, 775, doi:10.3390/rs10050775.
26. Somlyódy, L.; Jolánkai, G. Nutrient loads. In *Modeling and Managing Shallow Lake Eutrophication with Application Toe Lake Balaton*; Springer: Berlin, Germany, 1986; pp. 125–156.
27. Herodek, S.; Laczkó, L.; Virág, A. *Lake Balaton: Research and Management*; NEXUS Press: Budapest, Hungary, 1988.
28. Tompa, E.; Nyirő-Kósa, I.; Rostási, A.; Cserny, T.; Pósfai, M. Distribution and composition og Mg-calcite and dolomite in the water and sediments of Lake Balaton. *Centrel Eur. Geol.* **2014**, *57*, 113–136, doi:10.1556/CEuGeol.57.2014.2.1.

29. Iwamura, T.; Nagai, H.; Ichimura, S.E. Improved methods for determining contents of chlorophyll, protein, ribonucleic acid, and deoxyribonucleic acid in planktonic populations. *Int. Revue ges. Hydrobiol.* **1970**, *55*, 131–147, doi:10.1002/iroh.19700550106.
30. Cuthbert, I.D.; del Giorgio, P. Toward a standard method of measuring color in freshwater. *Limnol. Oceanogr.* **1992**, *37*, 1319–1326.
31. Alikas, K.; Reinart, A. Validation of the MERIS products on large European lakes: Peipsi, Vänern and Vättern. *Hydrobiologia* **2008**, *599*, 161–168.
32. Cristina, S.C.V.; Moore, G.F.; Goela, P.R.F.C.; Icely, J.D.; Newton, A. In situ validation of MERIS marine reflectance off the southwest Iberian Peninsula: assessment of vicarious adjustment and corrections for near-land adjacency. *Int. J. Remote Sens.* **2014**, *35*, 2347–2377, doi:10.1080/01431161.2014.894657.
33. V.-Balogh, K.; Németh, B.; Vörös, L. Specific attenuation coefficients of optically active substances and their contribution to the underwater ultraviolet and visible light climate in shallow lakes and ponds. *Hydrobiologia* **2009**, *632*, 91–105.
34. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, MA, USA, 2005.
35. Torma, P.; Krámer, T. Modeling the effect of waves on the diurnal stratification of a shallow lake. *Period. Polytech. Civ. Eng.* **2016**, *61*, 165–175, doi:10.3311/PPci.8883.
36. Zlinszky, A.; Molnár, G. Georeferencing the first bathymetric maps of Lake Balaton, Hungary. *Acta Geod. Geoph. Hung.* **2009**, *44*, 79–94, doi:10.1556/AGeod.44.2009.1.8.
37. Nimit, K.; Lotlikar, A.; Kumar, T.S. Validation of MERIS sensor's CoastColour algorithm for waters off the west coast of India. *Int. J. Remote Sens.* **2016**, *37*, 2066–2076, doi:10.1080/01431161.2015.1129564.
38. Toming, K.; Kutser, T.; Uiboupin, R.; Arikas, A.; Vahter, K.; Paavel, B. Mapping water quality parameters with sentinel-3 ocean and land colour instrument imagery in the Baltic Sea. *Remote Sens.* **2017**, *9*, 1070, doi:10.3390/rs9101070.
39. O'Reilly, J.E.; Maritirena, S.; Mitchell, B.G.; Siegel, D.A.; Carder, K.L.; Garver, S.A.; Kahru, M.; McClain, C. Ocean color chlorophyll algorithms for SeaWiFS. *J. Geophys. Res.* **1998**, *103*, 24937–24953.
40. Watanabe, F.; Alcântara, E.; Imai, N.; Rodrigues, T.; Bernardo, N. Estimation of chlorophyll-a concentration from optimizing a semi-analytical algorithm in productive inland waters. *Remote Sens.* **2018**, *10*, 227, doi:10.3390/rs10020227.
41. Lins, R.C.; Martinez, J.M.; Motta Marques, D.d.; Cirilo, J.A.; Fragoso, C.R. Assessment of chlorophyll-a remote sensing algorithms in a productive tropical estuarine-lagoon System. *Remote Sens.* **2017**, *9*, 516, doi:10.3390/rs9060516.
42. Smith, M.E.; Lain, L.R.; Bernard, S. An optimized chlorophyll a switching algorithm for MERIS and OLCI in phytoplankton-dominated waters. *Remote Sens. Environ.* **2018**, *215*, 217–227, doi:10.1016/j.rse.2018.06.002.

Chapter 10

Conclusion and future work

In this thesis, the Sensitivity Analysis of the Gaussian Process Regression model's mean and variance functions was introduced and analyzed. The method measures the integrated squared gradient of these functions in all directions. Intuitively, the SA of the GPR thus quantifies how much the mean and variance functions change in the input dimensions, and assigns the most relative importance to the dimensions with highest variations.

The controlled simulated experimental setups revealed that the SA of the GPR mean could consistently assign high relevance to the important feature, and that the SA of the GPR variance was able to capture the spacing of the data in the input dimension. The performance of the methodology was evaluated on Chl-a/ Rrs matchups, with promising results.

When comparing the performance of the SA of the GPR mean function with other feature ranking methods for regression, for instance the SA of the SVR and VIP of the PLSR, it was found that also these methods give high importance to the input features with largest variation. (See Chapter 4 Figures 4.3, 4.4, 4.5 and 4.6.) This suggests that the introduced methodology is consistent with other methods.

In this work, an automatized model selection approach called AMSA, was introduced. It combines feature ranking and regression method selection to objectively determine the most suitable model for a given dataset. Evaluating AMSA on Chl-a/ Rrs matchups representing several different water conditions, showed that the GPR, with a certain set of features, in most cases is chosen as the *best* of the investigated methods. This is in good correspondence with other studies on biophysical parameter estimation using the GPR model (for example in [4] and [5]).

It is also shown in this thesis, that using feature selection in the GPR model can improve the method and result in Chl-a estimates comparable (or better than) the estimates of the state-of-the-art algorithms.

Furthermore, the features selected by the SA reflect the biophysical properties of the water bodies. This can also be expected, since the investigated feature ranking methods returns changes in the Rrs spectrum. The Rrs signal carries the biophysical signature of the illuminated part of the water body.

The AMSA approach was furthermore used for establishing a unified model for Chl-a monitoring using the S3 OLCI sensor. The chosen data originated from Lake Balaton, a lake in Hungary, which is known to represent different kinds of water conditions. Evaluating AMSA on the data from Lake Balaton resulted in a model that could successfully estimate Chl-a for the whole lake. This model was henceforth referred to as Balaton model, and is currently under evaluation in Arctic inland, coastal and open waters.

For future work, the Balaton model will be further tested on various other local and global aquatic environments. The goal will be to create a generalized model for Chl-a and other water quality parameter estimation, with specific focus on the S3 OLCI sensor. Having one Chl-a product available for all kinds of waters, allowing a wider range of users to utilize water quality data provided by S3 OLCI, would be a great achievement, and represent an important tool for understanding and monitoring the water quality of Earth's water reservoirs.

The studies conducted in this thesis show the strength of the GPR method. However, further studies of the GPR are required. Although the GPR model is a sophisticated method, it has certain disadvantages. For example, the choice of the initial hyper-parameters for the optimization of the kernel parameters has an impact on the GPR, and it might influence the SA of the GPR's mean and variance functions, as well. It is suggested that other strategies for the choice and optimization of the hyper-parameters should be investigated. The goal would be the have a more reliable and user - friendly approach, which would ensure that the optimized parameters correspond to the global maximum of the likelihood function of the hyper-parameters learned from the given training data. The computational efficiency of the method also requires improvements, although there are already several approaches, which can speed up the method.

If these issues are addressed and resolved, the GPR would have the potential of becoming a popular approach in a wide range of application. It would not only have an extraordinary learning strength, but it would also be an approach, which is trackable, and where the driving mechanisms of the method are fully understood. This thesis has contributed to this understanding.

Bibliography

[1] J. M. Wang, D. J. Fleet, and A. Hetzmann, "Gaussian Process Dynamical Models for Human Motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, 2008.

[2] N. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," *Advances in neural information processing systems*, vol. 16 (3), pp. 329–336, 2004.

[3] A. Krause, A. Singh, and C. Guestrin, "Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies," *Journal of Machine Learning Research*, vol. 9, pp. 235–284, 2008.

[4] J. Verrelst, J. Muñoz, L. Alonso, J. P. Rivera, G. Camps-Valls, and J. Moreno, "Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for sentinel-2 and -3," *Remote Sensing of Environment*, vol. 118, pp. 127–139, 2012.

[5] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50 (5), pp. 1832–1843, 2012.

[6] T. Volk and M. I. Hoffert, *Ocean Carbon Pumps: Analysis of Relative Strengths and Efficiencies in Ocean-Driven Atmospheric CO2 Changes*. American Geophysical Union, 2013.

[7] K. R. Arrigo, D. H. Robinson, D. L. Worthen, R. B. Dunbar, G. R. DiTullio, M. VanWoert, and M. P. Lizotte, "Phytoplankton Community Structure and the Drawdown of Nutrients and CO2 in the Southern Ocean," *Science*, vol. 283, no. 5400, pp. 365–367, 1999.

[8] M. Hein and K. Sand-Jensen, "CO2 increases oceanic primary production," *Nature*, vol. 388, pp. 526–527, 1997.

[9] M. Hofmann, B. Worm, S. Rahmstorf, and H. J. Schellnhuber, "Declining ocean chlorophyll under unabated anthropogenic CO2 emissions," *Environmental Research Letters*, vol. 6, no. 3, pp. 034–035, 2011.

[10] N. T. T. Ha, K. Koike, and M. T. Nhuan, "Improved accuracy of chlorophyll-a concentration estimates from modis imagery using a two-band ratio algorithm and geostatistics: As applied to the monitoring of eutrophication processes over tien yen bay (northern vietnam)," *Remote Sensing*, vol. 6, no. 1, pp. 421–442, 2014.

[11] X.-e. Yang, X. Wu, H.-l. Hao, and Z.-l. He, "Mechanisms and assessment of water eutrophication," *Journal of Zhejiang University SCIENCE B*, vol. 9, pp. 197–209, Mar 2008.

[12] I. S. Robinson, *Measuring the Oceans from Space: The principles and methods of satellite oceanography*. Praxis Publishing Ltd, 2004.

[13] H. R. Gordon, O. B. Brown, R. H. Evans, J. W. Brown, R. C. Smith, K. S. Baker, and D. K. Clark, "A Semianalytic Radiance Model of Ocean Color," *Journal of Geophysical Research*, vol. 93, pp. 10909–10924, 1988.

[14] J. E. O'Reilly, S. Maritorena, M. C. O'Brien, D. A. Siegel, D. Toole, D. Menzies, R. C. Smith, J. L. Mueller, B. G. Mitchell, M. Kahru, F. P. Chavez, P. Strutton, G. F. Cota, S. B. Hooker, C. R. Mcclain, K. L. Carder, F. Müller-Karger, L. Harding, A. Magnuson, D. Phinney, G. F. Moore, J. Aiken, K. R. Arrigo, R. Letelier, and M. Culver, "SeaWiFS postlaunch calibration and validation analyses, part 3," *Nasa Tech. Memo. 2000-206892*, vol. 11, 2000.

[15] J. E. O'Reilly, S. Maritirena, B. G. Mitchell, D. A. Siegel, K. L. Carder, S. A. Garver, M. Kahru, and C. McClain, "Ocean color chlorophyll algorithms for SeaWiFS," *Journal of Geophysical Research*, vol. 103, pp. 24937–24953, 1998.

[16] D. Blondeau-Patissier, J. F. Gower, A. G. Dekker, S. R. Phinn, and V. E. Brando, "A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans," *Progress in Oceanography*, vol. 123, pp. 123 – 144, 2014.

[17] H. Zhan, P. Shi, and C. Chen, "Retrieval of oceanic chlorophyll concentration using support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41 (12), pp. 2947–2951, 2003.

[18] E. J. Kwiatkowska and G. S. Fargion, "Application of machine-learning techniques toward the creation of a consistent and calibrated global chlorophyll concentration baseline dataset using remotely sensed ocean color data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41 (12), pp. 2844 – 2860, December 2003.

[19] G. Camps-Valls, J. Muñoz-Marí, K. R. L. Gómez-Chova, and J. Calpe-Maravilla, "Biophysical parameter estimation with a semisupervised support vector machine," *IEEE Geoscience and Remote Sensing Letters*, vol. 6 (2), pp. 248 – 252, 2009.

[20] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, J. Amorós-López, and J. Calpe-Maravilla, "Retrieval of oceanic chlorophyll concentration with relevance vector machines," *Remote Sensing of Environment*, vol. 105(1), pp. 23 – 33, 2006.

[21] P. Cipollini, G. Corsini, M. Diani, and R. Grass, "Retrieval of sea water optically active parameters from hyperspectral data by means of generalized radial basis function neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, pp. 1508–1524, 2001.

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

[23] C. Brockmann, R. Doerffer, M. Peters, S. Kerstin, S. Embacher, and A. Ruescas, "Evolution of the C2RCC Neural Network for Sentinel 2 and 3 for the Retrieval of Ocean Colour Products in Normal and Extreme Optically Complex Waters," in *Living Planet Symposium*, vol. 740 of *ESA Special Publication*, p. 54, August 2016.

[24] K. Nimit, A. Lotlikar, and T. S. Kumar, "Validation of meris sensor's coastcolour algorithm for waters off the west coast of india," *International Journal of Remote Sensing*, vol. 37 (9), pp. 2066–2076, 2016.

[25] K. Alikas and A. Reinart, "Validation of the meris products on large european lakes: Peipsi, vänern and vättern," *Hydrobiologia*, vol. 599, pp. 161–168, 02 2008.

[26] K. Blix, K. Pálffy, V. R. Tóth, and T. Eltoft, "Remote sensing of water quality parameters over lake balaton by using sentinel-3 olci," *Water*, vol. 10, no. 10, 2018.

[27] J. R. Jensen, *Remote Sensing of the Environment: An Earth Resource Perpective*. Pearson Prentice Hall, 2007.

[28] H. M. Dierssen and K. Randolph, *Remote Sensing of Ocean Color*, pp. 439–472. New York, NY: Springer New York, 2013.

[29] C. Mobley, D. Stramski, W. Bissett, and E. Boss, "Optical modeling of ocean waters: Is the case 1- case 2 classification still useful?," *Oceanography*, 2004.

[30] Z. P. Lee, "Remote sensing of inherent optical properties: Fundamentals, test of algorithms, and applications.," tech. rep., Reports of the International Ocean-Colour Coordinating Group, IOCCG, 2006.

[31] T. Iwamura, H. Nagai, and S.-E. Ichimura, "Improved methods for determining contents of chlorophyll, protein, ribonucleic acid, and deoxyribonucleic acid in planktonic populations," *Internationale Revue der gesamten Hydrobiologie und Hydrographie*, vol. 55 (1), pp. 131–147, 1970.

[32] I. D. Cuthbert and P. del Giorgio, "Toward a standard method of measuring color in freshwater," *Limnology and Oceanography*, vol. 37, pp. 1319–1326, 1992.

[33] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[34] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," *in Advances in Neural Information Processing Systems*, vol. 8, pp. 514 – 520, 1996.

[35] K. Ryan and K. Ali, "Application of a partial least-squares regression model to retrieve chlorophyll-a concentrations in coastal waters using hyper-spectral data," *Ocean Science Journal*, vol. 51, no. 2, pp. 209–221, 2016.

[36] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, 2004.

[37] B. Schölkopf and A. Smola, "Learning with kernels-support vector machines, regularization, optimization and beyond," *MIT Press*, 2002.

[38] K. P. Murphy, *Machine Learning A probabilistic Perspective*. The MIT Press, 2012.

[39] S. Y. Kung, *Kernel Methods and Machine Learning*. Cambridge University Press, 2014.

[40] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, pp. 109 – 130, 2001.

[41] R. Gosselin, D. Rodrigue, and C. Duchesne, "A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 100, pp. 12–21, 2010.

[42] N. L. Afanador, *Important Variable Selection in Partial Least Squares for Industrial Process Understanding and Control*. PhD thesis, Radboud University Nijmegen, 2014.

[43] H. Abdi, "Partial least squares regression and projection on latent structure regression (PLS Regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97 – 106, 2010.

[44] S. Rännar, F. Lindgren, P. Geladi, and S. Wold, "A PLS kernel algorithm for data sets with many variables and fewer objects. part 1: Theory and algorithm," *Journal of Chemometrics*, vol. 8, pp. 111 – 125, 1994.

[45] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251 – 263, 1993.

[46] B. S. Dayal and J. F. MacGregor, "Improved PLS algorithms," *Journal of Chemometrics*, vol. 11, no. 1, pp. 73 – 85, 1997.

[47] K. Song, D. Lu, L. Li, S. Li, Z. Wang, and J. Du, "Remote sensing of chlorophyll-a concentration for drinking water source using genetic algorithms (GA)-partial least square (PLS) modeling," *Ecological Informatics*, vol. 10, pp. 25 – 36, 2012. Ecological Informatics and Ecosystem Conservation.

[48] K. Blix, G. Camps-Valls, and R. Jenssen, "Sensitivity analysis of gaussian processes for oceanic chlorophyll prediction," in *2015 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2015, Milan, Italy, July 26-31, 2015*, pp. 996–999, 2015.

[49] K. Blix, G. Camps-Valls, and R. Jenssen, "Gaussian process sensitivity analysis for oceanic chlorophyll estimation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10 (4), pp. 1265–1277, April 2017.

[50] P. M. Rasmussen, K. H. Madsen, T. E. Lund, and L. K. Hansen, "Visualization of nonlinear kernel models in neuroimaging by sensitivity maps," *NeuroImage*, vol. 55, no. 3, pp. 1120 – 1131, 2011.

[51] K. Blix and T. Eltoft, "Machine learning automatic model selection algorithm for oceanic chlorophyll-a content retrieval," *Remote Sensing*, vol. 10 (5), p. 775, 2018.

[52] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold, "Multi- and Megavariate Data Analysis. principles and applications," *Journal of Chemometrics*, vol. 16, no. 5, pp. 261 – 262, 2001.

[53] P. Jonsson, *Surface Status Classification, Utilizing Image Sensor Technology and Computer Models*. PhD thesis, Mid Sweden University, 2015.

[54] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods on Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62 – 69, 2012.