




Article

# Machine Learning Regression Approaches for Colored Dissolved Organic Matter (CDOM) Retrieval with S2-MSI and S3-OLCI Simulated Data

Ana Belen Ruescas <sup>1,\*</sup>,<sup>†</sup> , Martin Hieronymi <sup>2</sup>, Gonzalo Mateo-Garcia <sup>1</sup> , Sampsa Koponen <sup>3</sup>, Kari Kallio <sup>3</sup> and Gustau Camps-Valls <sup>1</sup> 

<sup>1</sup> Image Processing Laboratory, Universitat de València, 46980 València, Spain; Gonzalo.Mateo-Garcia@uv.es (G.M.-G.); Gustau.Camps@uv.es (G.-C.V.)

<sup>2</sup> Institute for Coastal Research, Helmholtz-Zentrum Geesthacht, 21502 Geestacht, Germany; martin.hieronymi@hzg.de

<sup>3</sup> Finnish Environment Institute (Suomen ympäristökeskus, SYKE), FI-00251 Helsinki, Finland; Sampsa.Koponen@ymparisto.fi (S.K.); kari.y.kallio@ymparisto.fi (K.K.)

\* Correspondence: bruescas@uv.es; Tel.: +34-963-544-061

† Calle Catedrático José Beltran 2 46980 Paterna (València), Spain.

Received: 17 April 2018; Accepted: 17 May 2018; Published: 19 May 2018



**Abstract:** The colored dissolved organic matter (CDOM) variable is the standard measure of humic substance in waters optics. CDOM is optically characterized by its spectral absorption coefficient,  $a_{CDOM}$  at at reference wavelength (e.g.,  $\approx 440$  nm). Retrieval of CDOM is traditionally done using bio-optical models. As an alternative, this paper presents a comparison of five machine learning methods applied to Sentinel-2 and Sentinel-3 simulated reflectance ( $R_{rs}$ ) data for the retrieval of CDOM: regularized linear regression (RLR), random forest regression (RFR), kernel ridge regression (KRR), Gaussian process regression (GPR) and support vector machines (SVR). Two different datasets of radiative transfer simulations are used for the development and training of the machine learning regression approaches. Statistics comparison with well-established polynomial regression algorithms shows optimistic results for all models and band combinations, highlighting the good performance of the methods, especially the GPR approach, when all bands are used as input. Application to an atmospheric corrected OLCI image using the reflectance derived from the alternative neural network (Case 2 Regional) is also shown. Python scripts and notebooks are provided to interested users.

**Keywords:** remote sensing; CDOM; optically complex waters; linear regression; machine learning; Sentinel 2; Sentinel 3

## 1. Introduction

Ocean color retrievals from remote sensing have been produced regularly in the last decades with more or less accuracy in different regions of the planet and for several water types. The theory of aquatic optics was explained by Preisendorfer [1] and Jerlov [2]; and the related radiance transfer equations have been adapted by Mobley [3] for the Hydrolight model [4]. The development and validation of water quality algorithms, many of them empirically developed and implemented using in situ data from very specific locations, are the main topic of many of the published investigations. First more focused on open ocean or Case-1 waters- where the optical properties are determined mainly by phytoplankton-, later with further development of algorithms for more complex or Case-2 waters (coastal and lakes), the water quality variables reported as able to be estimated by remote sensing are: concentration of total suspended matter (TSM), turbidity, colored dissolved organic matter

(CDOM), concentration of chlorophyll\_a (Chl-a), occurrence of surface accumulating algal blooms, concentration of phycocyanin, and Secchi depth, e.g., [5–10]. The development of algorithms that do not require extensive in situ sampling for training is a central aim in remote sensing of water quality [11]. The atmospheric correction, previous step to derive water leaving reflectance, has turned out to be demanding particularly for non-oligotrophic waters [12,13], and especially complicated for darker CDOM-rich waters so called Case-2 absorbing (C2A) or extreme absorbing waters (C2AX) [14]. The water leaving signal is very low and the proportion of atmospheric noise can be very high (until 95% of the signal [15]). This atmospheric correction issue is, however, not part of the developments of this research, even though is a factor to take into account.

C2A and C2AX waters dominated by dissolved organic matter (DOM) are the focus of this research. DOM in the ocean has a relevant role and impact on the global carbon cycle, in concrete the colored component of DOM, which absorbs light exponentially decreasing from the ultra-violet (UV) to the visible parts of the spectrum. The estimation of CDOM from remote sensing data, as a proxy for dissolved organic carbon (DOC), requires of accurate algorithms [16]. In the boreal temperate and cold regions like Finland, Sweden and Estonia, humic waters in lakes and some coastal zones are abundant. These waters typically have fairly low TSM and Chl-a concentrations, even though some cases of “black lakes” with high Chl-a and TSM values have been reported too [17]. In these cases, the reflectance is negligible in the visible, and only in the red-near infrared is some times possible to detect the Chl-a signal. Within ESA’s C2X project, extreme absorbing waters were characterized by CDOM absorption  $a_{CDOM}(440) > 1 \text{ m}^{-1}$ , which results in very low reflectance, typically with maximum below  $0.005 \text{ sr}^{-1}$  [14]. In Finnish lakes, for instance, the median absorption coefficient of CDOM at 443 nm is around  $3.7 \text{ m}^{-1}$  [18,19]. In Finland the humic matter concentration of lakes correlates with the share of peat land in the drainage area [20]. Humic lakes can also originate from peat dredging, e.g., in the Netherlands. Information on humic substances is utilized in the application of official directives, lake management and climate change studies. The  $a_{CDOM}$  parameter is considered to be a measure of dissolve organic carbon, which could help to estimate  $\text{CO}_2$  efflux and to assess the carbon pool in carbon budget studies [19]. An accurate measurement of the  $a_{CDOM}$  parameter from remote sensing seems crucial in these types of water. However, it is known that CDOM is one of the most critical and uncertain ocean color (OC) product [21,22].

In the work presented here, we focus on the CDOM estimation, showing results of the application of several machine learning (ML) algorithms in ‘typical boreal waters’, with medium to extreme CDOM absorption. The main objective is the retrieval of the CDOM variable using the sensors developed by the European Space Agency (ESA) as part of the Copernicus Earth Observation Programme: Sentinel-2 Multi-Spectral Instrument (S2-MSI) and Sentinel-3 Ocean and Land Colour Instrument (S3-OLCI). The S2-MSI sensor and other high-spatial-resolution instruments have the drawback that, since they are designed initially for terrestrial applications, their spectral resolution, measurement frequency, and radiometric characteristics are not optimized for water quality mapping. Even though, S2-MSI gives more accurate water quality estimates through its enhanced channel configuration and better temporal resolution than other high resolution satellites like Landsat (see Section 2 for details). S3-OLCI is designed to measure ocean color over ocean and coastal zones with 300 m of spatial resolution (see Section 2 for specifications). With a very good signal-to-noise ratio, good radiometric stability, mitigation of sun-glint contamination and excellent cover of the global ocean, S3-OLCI images the spectral distribution of the radiance at top-of-atmosphere. After atmospheric correction, the upwelling radiance just above the sea surface (the water-leaving radiance) is retrieved and used to estimate a number of geophysical parameters through the application of specific bio-optical algorithms. S3-OLCI provides information on the atmosphere too, especially on the aerosols characterization necessary for the atmospheric correction process.

Concerning the retrieval of CDOM absorption, several band ratios have been proposed as predictive models for estimating CDOM from spectral data [11,23,24]. These parametric approaches only take into account a few spectral bands and thus they disregard the information contained in

other bands. Furthermore, they are usually based on models determined by training data, like in situ measurements that correlate with certain band combinations or ratios. Non-parametric regression algorithms can alternatively exploit the information contained in all spectral bands. For instance, semi-analytical or even neural network (NN) algorithms, i.e., MERIS Case II [25], Case 2 Regional Coast Color (C2RCC) [26,27], Boreal Lakes Processor [19], S3-OLCI Neural Network Swarm (ONNS) [28], are used to parametrize the inverse relationship between inherent optical properties (IOPs) and reflectance, allowing the retrieval of certain concentrations like Chl-a, TSM and  $a_{CDOM}$ . We assess here the performance of alternative non-parametric machine learning (ML) algorithms for CDOM estimation over typical boreal waters. The specific goals of the study are: (i) to test five different ML regression methods: multivariate regression (RLR), random forests (RFR), kernel ridge regression (KKR), Gaussian process regression (GPR) and support vector regression (SVR); (ii) to apply the methods on two datasets coming from two separate sources to demonstrate a wide range of applicability; (iii) to evaluate the validity by comparing with established methods, like empirically derived polynomial algorithms or neural networks (C2RCC and ONNS).

The remainder of the paper is organized as follows. §2 describes the materials and methods used, giving details about the distribution and components of the datasets; explaining details and giving references on the traditional band ratio methods; and finally introducing the ML approaches tested and associated statistics. §3 gives an empirical evidence of the performance of the proposed methods in comparison to standard bio-optical models for the particular datasets used; and it shows the validation results for the S3-OLCI-Case2Xtreme (C2X) dataset using an independent validation dataset. In §4 we focus the analysis on the C2X dataset, comparing the results of the S3-OLCI ML approaches with the results of the ONNS; finally, an application of the ML methods is tested on an S3-OLCI scene and the best method is compared with the C2RCC NN products included by default in the S3-OLCI L2. We conclude in §5 with remarks about the analysis done and an outline of the future work.

## 2. Datasets and Methods

### 2.1. Datasets

The datasets used are based on simulated remote sensing reflectance ( $R_{rs}$ ) with Hydrolight 5.2 [4].  $R_{rs}$  is the ratio of water-leaving radiance to downwelling irradiance above the sea surface. Here,  $R_{rs}$  refers to the simulations of clear atmosphere reflectance with Sun at zenith and viewing angle exactly perpendicular. We apply the ML approaches on two datasets: the first dataset comes from the Finnish Environmental Institute in Helsinki (Suomen ympäristökeskus, SYKE), and it is used to test and establish the methodology. The SYKE dataset is derived from routine monitoring measurements, with a broad representation of the water types and water status that can be found through the year in many of Finland's lakes and coastal waters. Specific inherent optical properties (SIOPs) were derived from literature [11,29]. The SYKE dataset is divided into two configurations: the first configuration simulates the Multi-Spectral Instrument carried in the Sentinel-2 satellite (S2-MSI). The S2-MSI samples 13 spectral bands. The wavelengths and bandwidths of the S2-MSI bands are shown in Table 1. The second SYKE configuration is related to the Ocean and Land Colour Instrument (S3-OLCI). S3-OLCI is based on the opto-mechanical and imaging design of ENVISAT MERIS. The S3-OLCI bands have been optimized to measure ocean color over open ocean and coastal zones and the sensor consists of 21 spectral bands with characteristics summarized in Table 1 as well.

The second dataset consists of S3-OLCI simulations calculated for the ESA's C2X project. Within the framework of the C2X project, in-water radiative transfer simulations for S3-OLCI were likewise carried out with the commercial software Hydrolight [4]. The source of the simulations is described in [30,31]. In the C2X project, the results of the simulations were grouped in five subcategories: Case 1, C2A, C2AX, Case 2 extreme scattering (C2S) and C2SX. Each subcategory comprehends 20,000 individual combinations of concentration of water constituents, inherent optical properties (IOPs), and sun positions. This large dataset is used for training and testing of the S3-OLCI ONNS

in-water processor [28]. One part of the S3-OLCI simulated dataset is put aside for validation purposes, with more than 4000 spectra per sub-category reserved exclusively for it. The C2X dataset contains simulations in 21 bands from where a subset of 15 bands is used here for water quality parameter estimation.

**Table 1.** Sentinel-2 Multi-Spectral Instrument(S2-MSI) and Sentinel-3 Ocean and Land Colour Instrument (S3-OLCI) spectral and spatial specifications (bands used in the experiments appear in blue).

Sensor	Spatial Resolution (m)	Band Number	Central Wavelength (nm)	Bandwidth (nm)
S2-MSI <sup>1</sup>	10	2	496.6	98
		3	560.0	45
		4	664.5	38
		8	835.1	145
	20	5	703.9	19
		6	740.2	18
		7	782.5	28
		8a	864.8	33
		11	1613.7	143
	60	12	2202.4	242
		1	443.9	27
		9	945.0	26
S3-OLCI <sup>2</sup>	300	10	1373.5	75
		Oa1	400.0	15
		Oa2	412.5	10
		Oa3	442.5	10
		Oa4	490.0	10
		Oa5	510.0	10
		Oa6	560.0	10
		Oa7	620.0	10
		Oa8	665.0	10
		Oa9	673.75	7.5
		Oa10	681.25	7.5
		Oa11	708.75	10
		Oa12	753.75	7.5
		Oa13	761.25	2.5
		Oa14	764.375	3.75
Oa15	767.5	2.5		
Oa16	778.75	15		
Oa17	865.0	20		
Oa18	885.0	10		
Oa19	900.0	10		
Oa20	940.0	20		
Oa21	1020	40		

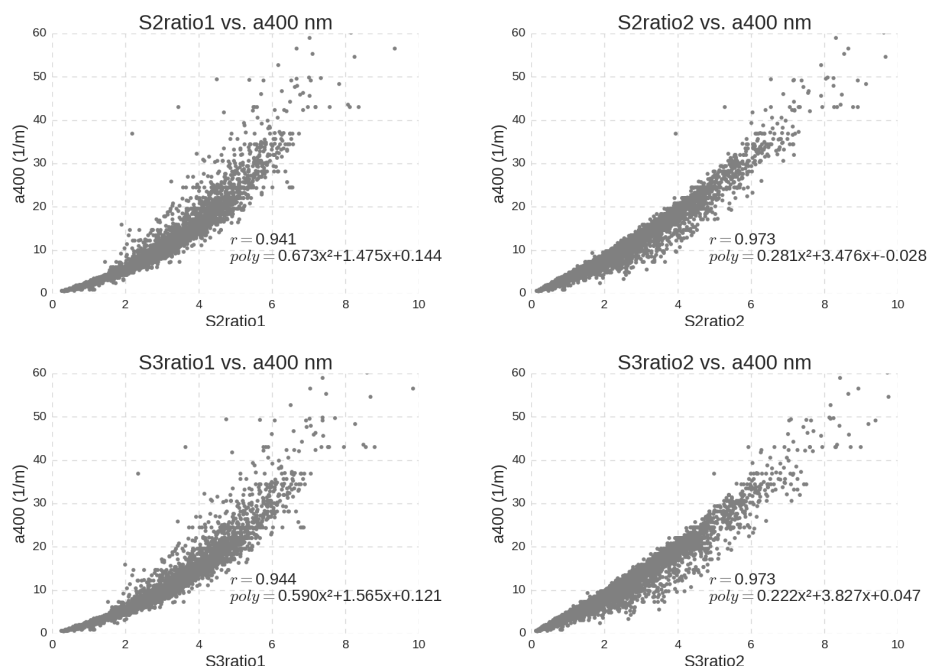
<sup>1</sup> <https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi>; <sup>2</sup> <https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-3-olci>.

## 2.2. Established Approaches Using Band Ratio Algorithms

Reports on established band ratio algorithms are mostly based on airborne and field measurements with negligible or only small atmospheric influence, like the simulated datasets used here. Important wavelength regions for the band ratio algorithms for CDOM retrieval are the ones between 400–600 nm, having 660–720 nm as reference. Following [11,23], CDOM can be estimated by a ratio of reflectance at wavelengths > 600 nm to reflectance in the 400–550 nm range. This ratio is valid for a wide range of water constituent combinations. In the work of [24] and [11] in situ measured data is used, and derived algorithms work well for  $a_{CDOM}$  and TSM. The TSM is estimated using a single band at 709 nm. Kallio [18] compared two band ratios using three wavebands at 490, 665 and 709 nm. Following that research, the optimized regression in our dataset has a polynomial form with coefficients varying depending on the ratio used:

$$y = p_1x^2 + p_2x + p_3 \quad (1)$$

Two band ratios are calculated with these simulated datasets and their correlation with the  $a_{400}$  nm value is shown in Figure 1. The 400 nm band is used here because its availability and because the CDOM absorption values at this ultraviolet (UV) wavelength are expected to be very high [18]. The first ratio combines the bands 665 nm with the 490 nm (*Ratio 1*); the second ratio combines the band in 705 or 708.75 nm with the 490 nm (*Ratio 2*). Both ratios are compared with the in situ absorption measured at 400 nm ( $a_{400} \text{ m}^{-1}$ ) and four new polynomial relations are then derived, two per each sensor configuration. The coefficients vary slightly from the ones in Tables 3 and 4 because all data available is used to study the ratio-parameter relationship. When applying the machine learning models, the dataset is divided randomly into a training set (75%) and a test set (25%). The S2-MSI and S3-OLCI configurations of the SYKE dataset include the whole range of Chl-a ( $0\text{--}120 \text{ mg m}^{-3}$ ), with variable TSM ( $0\text{--}123 \text{ g m}^{-3}$ ) and extreme CDOM at 400 nm ( $1\text{--}86 \text{ m}^{-1}$ ).



**Figure 1.** Polynomial regressions using the simulated Sentinel-2 Multi-Spectral Instrument (S2-MSI) and Sentinel-3 Ocean and Land Colour Instrument (S3-OLCI) configuration datasets.

### 2.3. Machine Learning Approaches

The methods based on ratios are parametric models, that is, an explicit relation is assumed between a subset of reflectance bands and the parameter of interest (CDOM in our case). Despite the good performance in general of such approximation, parametric models are generally limited because they do not exploit the wealth of spectral information and can only cope with (often too) simplistic and rigid relationships. An alternative approach is provided by nonparametric regression. In this case, an explicit relationship between reflectance and the parameter of interest is not assumed, and the functional form is actually learned (adjusted, inferred, fitted) from the data. Nonlinear very flexible relations can be accommodated, and very often improved performance in accuracy and bias is obtained. In addition, the trained models offer very low computational cost in the production (test) phase. On top of these advantages, the fields of machine learning and statistics offer a solid mathematical background, which allows to obtain bounds of performance and in some case confidence intervals for the predictions.

Currently there are a great many machine learning methods available for tackling nonlinear regression. They can be categorized in several families: multivariate linear regression, tree-based algorithms, neural networks, and kernel methods. In this work, we explore and compare

experimentally a representative method from each family. In particular, five machine learning algorithms for linear and non-linear regression are tested and compared to the polynomial regression explained above: (multivariate) RLR, RFR [32], KRR [33], GPR [34], and SVR [35]. The selected methods cover all the machine learning families: decision trees, randomized methods, kernel methods and probabilistic nonparametric approaches.

### 2.3.1. Multivariate Linear Regression

In RLR the output  $y$  (CDOM) is assumed to be a weighted sum of  $B$  input variables,  $\mathbf{x} = [x_1, \dots, x_B]^\top$ , that is  $\hat{y} = \mathbf{x}^\top \mathbf{w}$ . Maximizing the likelihood is equivalent to minimizing the sum of squared errors, and hence one can estimate the weights  $\mathbf{w} = [w_1, \dots, w_B]^\top$  by least squares minimization. Very often one imposes some smoothness constraints to the model and also minimizes the weights power,  $\|\mathbf{w}\|^2$ , thus leading to the regularized linear regression (RLR) method.

### 2.3.2. Decision Trees and Random Forests

An alternative method to linear regression is that of RFR [32]. An RFR model is an ensemble learning method for regression that operates by constructing a multitude of decision trees at training time and outputting the mean prediction of the individual trees. They combine many decision trees working with different subsets of features. The RFR strategy is very beneficial by alleviating the often reported over-fitting problem of simple decision trees. In addition, random forests (RFs) are quite robust to including a large number of input variables, excel in the presence of missing entries, heterogeneous variables, and can be easily parallelized to tackle large scale problems. RFs classification and regression have been applied in different areas of concern in forest ecology, such as modeling the gradient of coniferous species [36], the occurrence of fire in Mediterranean regions [37], the classification of species or land cover type [38,39], and the analysis of the relative importance of the proposed drivers [39] or the selection of drivers [38,40,41].

### 2.3.3. Kernel Methods

Kernel methods constitute a family of successful methods for regression [42]. We explore the performance of three instantiations: (i) the KRR is considered to be the (non-linear) version of the RLR [33]; (ii) GPR is a probabilistic approximation to non-parametric kernel-based regression, where both a predictive mean and predictive variance can be derived [43]; and (iii) the support vector regression (SVR) is the regression counterpart of the traditional support vector classifier, and delivers sparse solutions.

Kernel methods offer the same explicit form of the predictive model, which establishes a relation between the input (e.g., spectral data)  $x \in \mathbb{R}^B$  and the output variable (CDOM) is denoted as  $y \in \mathbb{R}$ . The prediction for a new input (radiance or reflectance) vector  $x_*$  can be obtained as:

$$\hat{y} = f(x) = \sum_{i=1}^N \alpha_i \mathbf{K}_\theta(\mathbf{x}_i, \mathbf{x}_*) + \alpha_o, \quad (2)$$

where  $\{\mathbf{x}_i\}_{i=1}^N$  are the spectra used in the training phase,  $\alpha_i$  is the weight assigned to each one of them,  $\alpha_o$  is the bias in the regression function, and  $K_\theta$  is a kernel or covariance function (parametrized by a set of hyper-parameters  $\theta$ ) that evaluates the similarity between the test spectrum  $x_*$  and all  $N$  training spectra. Depending on the functional to be minimized and the constraints imposed, different kernel methods can be derived giving rise to different values of the weights  $\alpha_i$ .

#### Kernel ridge regression

KRR is the kernel version of the regularized least squares linear regression [33,44]. The KRR essentially performs a regularized linear least squares regression in a feature space where the samples have been transformed to by a nonlinear mapping. It can be shown that the solution is analytical

and reduces to solving a matrix inversion. As in any kernel method, one has to select the form of the similarity kernel function. In our case, we selected the standard radial basis function (RBF) kernel, defined as

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2 / (2\sigma^2)), \quad (3)$$

where  $\sigma$  is the lengthscale parameter which is typically adjusted by cross-validation.

#### Gaussian process regression

GPR model assumes that the observed CDOM content is a function of the remote sensing reflectance using a multivariate joint Gaussian distribution of the available observations with zero mean and covariance matrix  $K$ :

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{x}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{K} + \sigma^2 I_n & \kappa_* \\ \kappa_*^T & \kappa_{**} + \sigma^2 \end{bmatrix}\right), \quad (4)$$

where  $\kappa_*$  is the covariance between the training vector and the test point,  $\kappa_{**}$  is the covariance between the test point with itself, and  $K + \sigma^2 I_n$  is the noisy covariance matrix of the training inputs. Applying Bayesian inversion is possible to compute the posterior distribution over the output  $x_*$  given the new input and the training dataset.

In our GPR implementation, we used the automatic relevance determination (ARD) kernel function,

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \nu \exp\left(-\sum_{b=1}^B (\mathbf{x}_b - \mathbf{x}'_b)^2 / (2\sigma_b^2)\right) + \sigma_n^2 \delta_{ij}, \quad (5)$$

and learned the hyper-parameters  $\theta = [\nu, \sigma_1, \dots, \sigma_B, \sigma_n]$  by marginal likelihood maximization. GP models [34] has recently provided excellent results in vegetation monitoring [43,45–47], as well as ocean chlorophyll content estimation [48].

#### Support vector regression

The SVR is the SVM implementation for regression and function approximation [35,49]. Unlike the previous KRR and GPR, the standard SVR formulation uses Vapnik's  $\varepsilon$ -insensitive cost function

$$\mathcal{L}_\varepsilon(e) = C \max(0, |e| - \varepsilon), \quad C > 0, \quad (6)$$

in which an error  $e = y - \hat{y}$  up to  $\varepsilon$  is not penalized, otherwise will incur in a linear penalization. This gives rise to a solution that is sparse, thus many  $\alpha_i$  become zero, meaning that the associated data points  $\mathbf{x}_i$  are irrelevant in the model. The SVR model has three hyper-parameters to be tuned: the  $C$  penalization factor, the  $\varepsilon$ -insensitive zone, and the kernel parameter  $\sigma$ , which as for KRR, we used the RBF kernel. Tuning all these hyper-parameters together is challenging. In our implementation, we followed a cross-validation procedure to do this.

#### 2.4. Comparison, Implementation and Reproducibility

A summary compilation of the main characteristics of the methods is given in Table 2. The running time of the training of the algorithms varies from model to model and it also depends on the inputs and amount of fine-tuned hyper-parameters. With these two particular databases the training time in a standard workstation with 16 GB of RAM is around 3 days for all the combinations of bands and models. In addition to this summary, we show a thorough comparison to a state-of-the-art neural network in §4. The code reproducing the results of this paper can be found in <https://github.com/IPL-UV/mlregocean>. An operational Matlab toolbox implementing all machine learning methods is available too at <https://github.com/IPL-UV/simpleR>.

**Table 2.** Main characteristics of the machine learning methods considered in this work: regularized linear regression (RLR), random forest regression (RFR), kernel ridge regression (KRR), Gaussian process regression (GPR), and support vector regression (SVR), grouped by family. We give explicit descriptors, such as the order  $\mathcal{O}$  of the computational and memory cost for training and testing as a function of dimension  $d$  (spectral bands), number of samples  $n$  for training and number of nodes  $t$  in the RFR model. Also the qualitative characteristics of the considered methods (strengths and weaknesses) is shown.

Method	Train Cost	Test Cost	Train Memory	Test Memory	Pros	Cons
RLR	$\mathcal{O}(d^2n)$	$\mathcal{O}(d)$	$\mathcal{O}(nd)$	$\mathcal{O}(d)$	Simple, fast	Cannot cope with nonlinear relations
RFR	$\mathcal{O}(tdn \log(n))$	$\mathcal{O}(td \log(n))$	$\mathcal{O}(nd)$	$\mathcal{O}(td)$	Fast, parallelizable	Do not provide confidence intervals, heuristic parameters to tune, prone to outliers
KRR	$\mathcal{O}(n^3)$	$\mathcal{O}(nd)$	$\mathcal{O}(n^2)$	$\mathcal{O}(nd)$	Fast for moderate sample sizes ( $n < 10,000$ ), effective	Do not provide confidence intervals, prone to outliers
GPR	$\mathcal{O}(n^3)$	$\mathcal{O}(nd)$	$\mathcal{O}(n^2)$	$\mathcal{O}(nd)$	Accurate, confidence intervals for predictions, automatic estimation of spectral bands and noise variance	Slow in moderate sample sizes
SVR	$\mathcal{O}(n^3)$	$\mathcal{O}(nd)$	$\mathcal{O}(n^2)$	$\mathcal{O}(nd)$	Generally accurate and robust to outliers	Slow, three hyper-parameters to tune by cross-validation



### 3. Experimental Results

#### 3.1. Experiments Setup

The two datasets, SYKE and C2X, are used separately, each one with a different strategy to train and test the six statistical models. The SYKE dataset is randomly split into training and test datasets before fitting the models (test size = 25%, random state = 42). The C2X dataset has two separated sets for training and testing. The calculations are carried out keeping the independence of both datasets and results are specific for each one. All the models have specific hyper-parameters for controlling over-fitting. We adjusted this hyper-parameters using standard ten fold cross validation on the training data. The experiments include five configurations of the input bands, which are common for the two datasets (Tables 3 and 4): the two firsts use one single ratio as input (*S\*-Ratio 1* and *S\*-Ratio 2*); the third configuration uses the two ratios together as input (*S\* two ratios*); the fourth configuration takes the full spectral range found in the simulation dataset (*S\* all bands*); and the fifth is a mix of the third and fourth configurations using the two ratios as well as the full spectra (*S\* all bands + ratios*). The spectral range of the simulated Sentinel-2 data consisted of 6 bands corresponding to some of the S2-MSI sensor wavelengths. The Sentinel-3 OLCI simulations included a total of 12 to 15 bands from the visible to the NIR. These are the common set found in many of the ocean color sensors used for water quality retrievals (see bands marked in blue in Table 1). Statistics used to check the validity of the methods are: the coefficient of determination ( $R^2$ ), the bias, root mean squared error (RMSE) absolute (abs) and relative (rel), the mean absolute error (MAE), the residual error (RES) and the Pearson's coefficient of correlation (R).

#### 3.2. Analysis of the Models

##### 3.2.1. Sentinel 2 MSI Data

Table 3 offers an overview of the metrics for all six methods tested with the five different combinations of input variables for S2-MSI. When using only one input (*S2-Ratio 1* =  $R_{rs}(665)/R_{rs}(490 \text{ nm})$ , *S2-Ratio 2* =  $R_{rs}(709)/R_{rs}(490 \text{ nm})$ ) the five ML methods tested do not improve the results strikingly in comparison with the polynomial regression. The metrics are very similar to each other and the polynomial regression is especially close to the KRR and GPR methods. Both approaches give very good results, but computation time will then be the determining factor, and the polynomial regression requires less time and it still efficient. However, when using more than one variable, non-linear ML methods get more relevance, as it can be seen in the *S2 two ratios*, *S2 all bands* and *S2 all bands + ratios* sections on Table 3. GPR and SVR methods show the best results in terms of correlation coefficients ( $> 0.98$ ) and errors ( $< 0.12$ ). When the two ratios are used as inputs, the RLR approach could still be considered as the simplest solution and probably preferred over the more sophisticated and computational time consuming ML techniques. However, it is clear that using all bands available as inputs leads to a significant improvement on the statistics (*S2 all bands + ratios*), especially using GPRs and also SVRs as mentioned before, reducing considerably the residuals and bias and improving the RMSEs.

The distribution of residuals for *S2 two ratios* and *S2 all bands + ratios* are shown in Figure 2. The box-plots of the errors show the clear improvement of the non-linear ML models in comparison with the linear one (RLR) for all cases (residuals closer to zero). The use of all bands and band ratios available leads to an amelioration in the errors.

The regression plots in Figure 3 compare the performance of the models as the expression of the value of the ratio against the values of the test dataset. Data plotted are the training (gray dots) and test (pink dots) points. The *Polyfit* line is the polynomial regression performance of the single ratios, in dark red, and it will be taken as reference (see also Figure 1). The plot on the left shows the results for *S-2 Ratio 1*. Data behaves quite regularly until the absorption values are higher than  $10 \text{ m}^{-1}$ , where they start to show more dispersion. At this point, it is useful to see how the RFR model behaves (orange line), because the

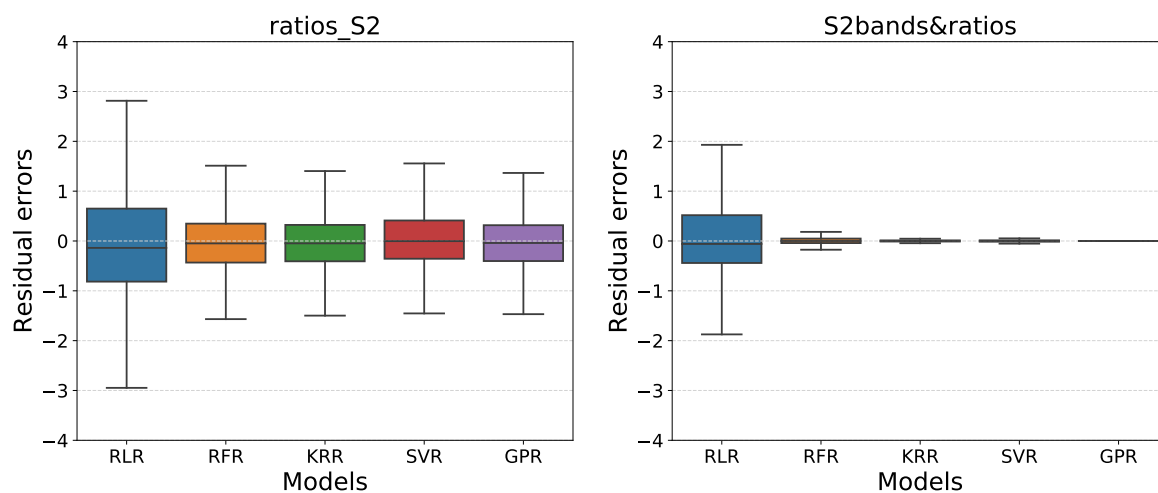
relative over-fitting of this model is indicating areas with outliers in the training data. The RLR is following a straight line (in blue), not fitting much the data at the beginning in the lower values and at the end of the relationship, although it has the same trend than the other models. The KRR (green line) and SVR (red line) reveal a better fit with the data and between them. The GPR (purple line) model seems similar to the KRR with the final portion closer to the SVR model, taking into account the higher  $a_{CDOM}(400)$  values. The area in light gray is the 95% confidence interval of the GPR prediction (see Section 2.3). The plot on the right shows the same regression fitting for the models compared with the polynomial fitting of the *S-2 Ratio 2*. In this case many of the models approach the *Polyfit* regression closer than the *the S-2 Ratio 1*. The regression line of the RFR seems to be smoother than the previous one and the uncertainty of the GPR tighter. This would probably mean that the *S-2 Ratio 2* is a better method to derive calibration coefficients for CDOM algorithm development with the SYKE dataset.

**Table 3.** Results obtained with empirical fitting and several machine learning methods on the simulated S2-MSI data ( $a_{CDOM}(400)$ ).

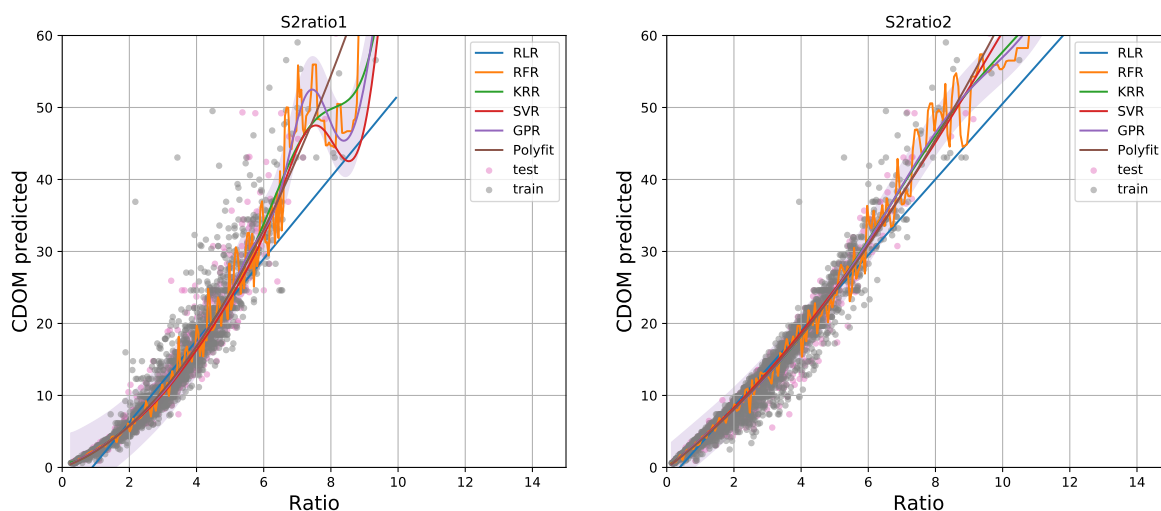
	R2	Bias	MAE	R	RMSE	RMSErel
<b>S2-Ratio 1: <math>x_1 = 665/490</math></b>						
Polyfit	0.934	0.004	1.068	0.966	2.054	0.133
RLR	0.887	0.657	1.747	0.942	2.684	0.768
RFR	0.895	-0.043	1.334	0.946	2.578	0.170
KRR	0.933	0.005	1.052	0.966	2.059	0.134
GPR	0.934	0.014	1.057	0.966	2.057	0.133
SVR	0.928	-0.281	1.026	0.966	2.136	0.130
<b>S2-Ratio 2: <math>x_2 = 705/490</math></b>						
Polyfit	0.966	-0.001	0.917	0.983	1.472	0.168
RLR	0.949	0.026	1.143	0.974	1.806	0.385
RFR	0.949	-0.035	1.138	0.974	1.811	0.189
KRR	0.967	-0.005	0.916	0.983	1.454	0.168
GPR	0.967	-0.004	0.916	0.983	1.456	0.168
SVR	0.964	0.224	0.893	0.982	1.550	0.177
<b>S2 two ratios: <math>x = [x_1, x_2]</math></b>						
RLR	0.949	0.026	1.145	0.974	1.806	0.388
RFR	0.969	-0.018	0.784	0.984	1.410	0.121
KRR	0.974	0.010	0.75	0.987	1.293	0.117
GPR	0.976	-0.005	0.748	0.986	1.346	0.115
SVR	0.974	0.065	0.745	0.987	1.283	0.118
<b>S2 all bands, <math>x \in \mathbb{R}^B</math></b>						
RLR	0.484	0.124	3.957	0.696	5.735	1.626
RFR	0.978	-0.049	0.295	0.990	1.164	0.074
KRR	0.997	0.006	0.210	0.999	0.408	0.06
GPR	0.998	-0.003	0.027	0.999	0.359	0.033
SVR	0.993	-0.002	0.092	0.996	0.670	0.068
<b>S2 all bands + ratios, <math>x \in \mathbb{R}^B</math></b>						
RLR	0.987	0.015	0.619	0.993	0.917	0.243
RFR	0.997	-0.009	0.145	0.998	0.447	0.038
KRR	0.998	-0.005	0.039	0.999	0.475	0.064
GPR	0.995	0.022	0.029	0.997	0.190	0.022
SVR	0.999	-0.012	0.044	0.999	0.182	0.022

An interesting tool for the analysis of the models is the permutation plots shown in Figure 4. These plots are the result of the permutation test, which calculates the probability of getting a value equal to or more extreme than an observed value of a test statistic under a specified null hypothesis by recalculating the test statistic after random re-orderings (shuffling) of the data [50]. The permutation plot can be seen as a mean of feature ranking. The statistic used is the MAE, which means that Figure 4 shows the MAE after shuffling each of the regressors  $P = 30$  times. According to our previous experience [51], permutation analysis results in robust estimates of variable relevance, and it is possible

to achieve stable feature rankings with a moderate number of runs. More than 30 realizations does not reduce much the variance of the estimates. By using the permutation plots, we are measuring the importance of each of the bands to the proposed model. The *S-2 all bands + ratios* composition permutation plots of 4 models are shown as example. The higher the MAE value when a particular band is removed, the more weight it has in the model: that is, for the RLR bands in the blue part of the spectrum (443 and 490 nm) caused a higher MAE when removed compared with the other bands, followed by the 560 and the 705 nm bands. Precisely the combination of the 705 and the 490 is the *S2-Ratio 2*, which is crucial for the RLR method as observed in the plot. For the KRR method bands seem to have more balanced distribution, highlighting the importance of band at 443 nm. The GPR approach shows major weight of band 740 nm, followed by 665 nm, diminishing the role of the blue and ratio bands compared with other models. In the SVR approach, with lower MAE values than the rest, the two ratios play a major role. In general all methods highlight the importance of the bands used for the ratios or the ratios themselves, which agrees with the physical knowledge about the bands contributing more to CDOM determination.



**Figure 2.** Box-plots of the residuals *S2 two ratios* and *S2 All bands + ratios*. On each box, the central mark is the median, the edges of the box are the lower hinge (defined as the 25th percentile) and the upper hinge (the 75th percentile), the whiskers extend to the most extreme data points not considered outliers.



**Figure 3.** Comparison of the performance of the models using the linear regression representation, *S2-MSI*. On the y axis are the normalized CDOM values, on the x axis the value of the ratios.

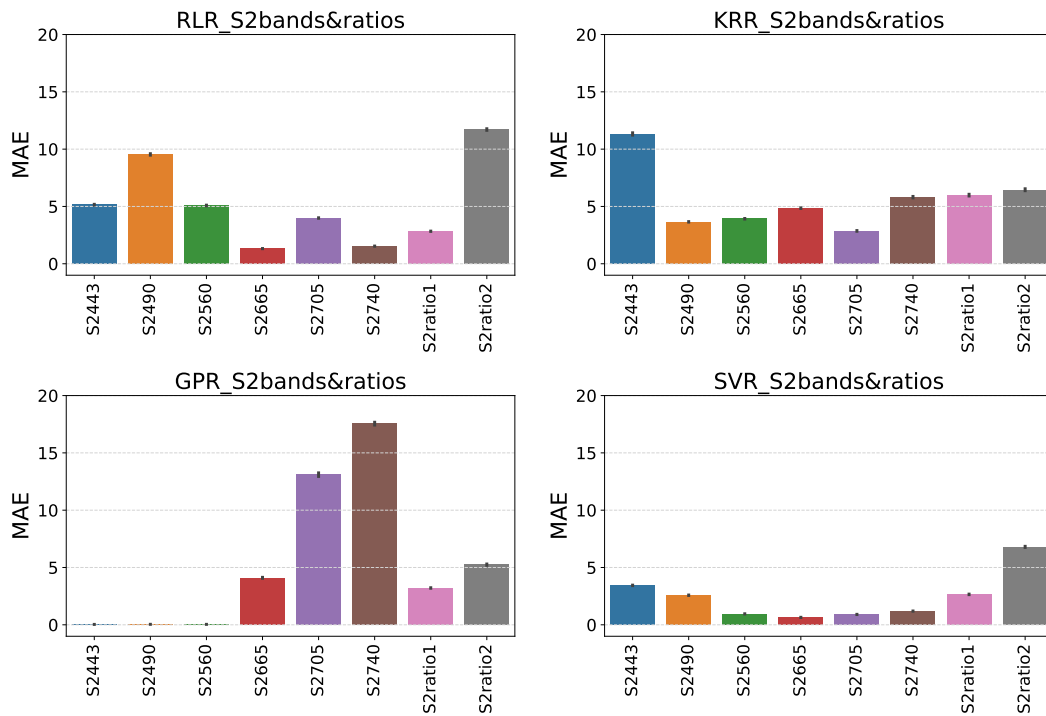


Figure 4. Permutation plots for four ML methods for the S2-MSI configuration.

3.2.2. Sentinel 3 OLCI Data

Statistic results are very similar for the S3-OLCI configuration (Table 4, left part), with in general good values for all approaches when using S3-Ratio 1 or S3-Ratio 2 as inputs ( $R > 0.94$  and  $RMSEs < 0.7$ ). The Polyfit and KRR methods stay within the same ranges, with very good performance in both cases. Increasing the number of inputs shows better results, especially with the more sophisticated ML approaches. The RLR fails when using S3 all bands, due to the non-linear behavior of the inputs as it happened with S2-MSI ( $R = 0.763$ ,  $RMSEr > 1.4$ ). The KRR, GPR and SVR approaches show the best statistics in S3 all bands and S3 all bands + ratios.

Figure 5 shows the residuals values in box plots for S3 two ratios and S3 all bands + ratios. Again the errors are reduced, in this case quite spectacularly, when using more sophisticated kernel-based methods (KRR, GPR and SVR) and they are especially low when using all inputs S3 all bands + ratios.

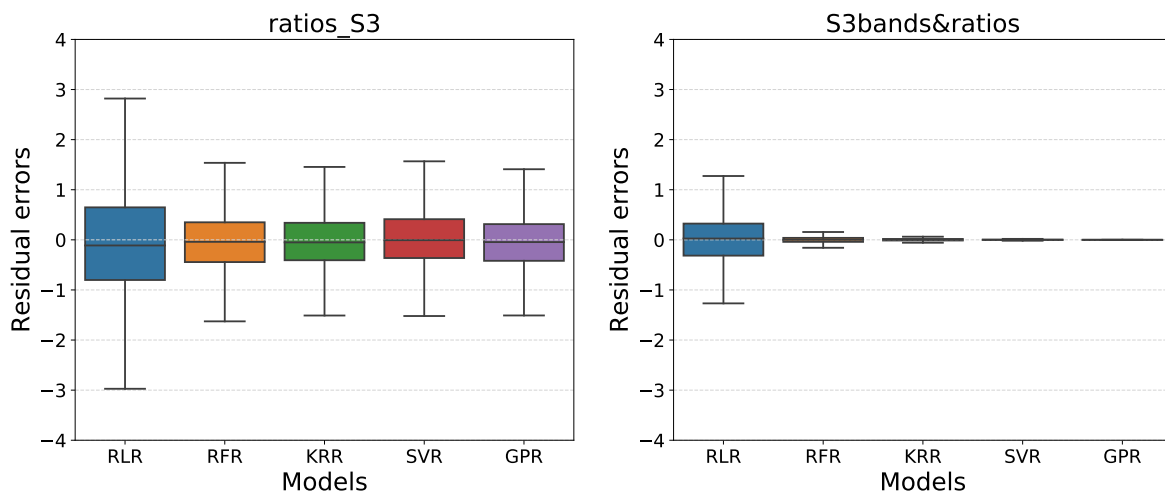


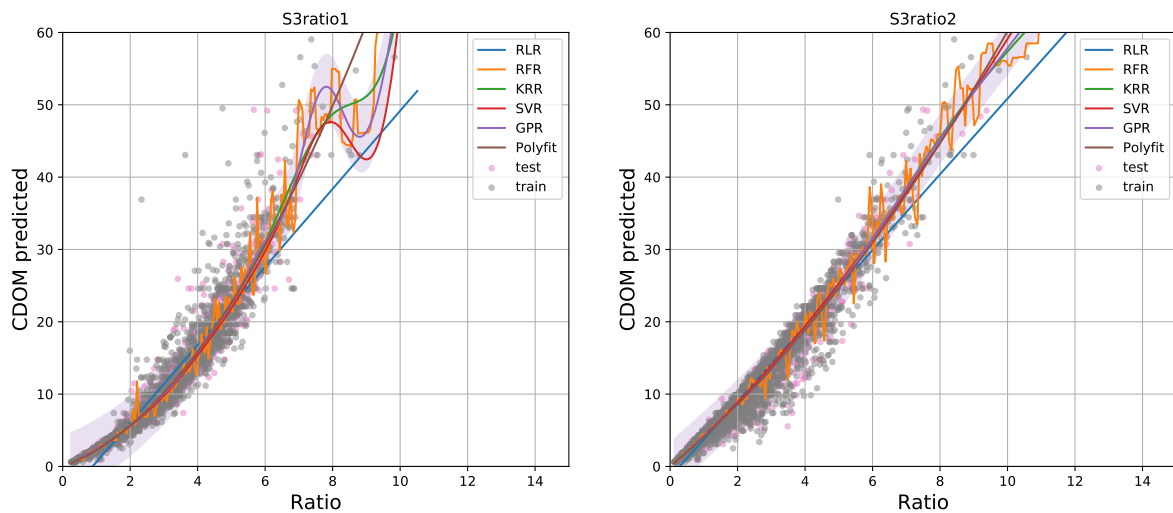
Figure 5. Box-plots of the residuals for S3 two ratios and S3 All bands + ratios.

**Table 4.** Results obtained with empirical fitting and the five machine learning methods on the simulated S3-OLCI data of the C2A(X) dataset.

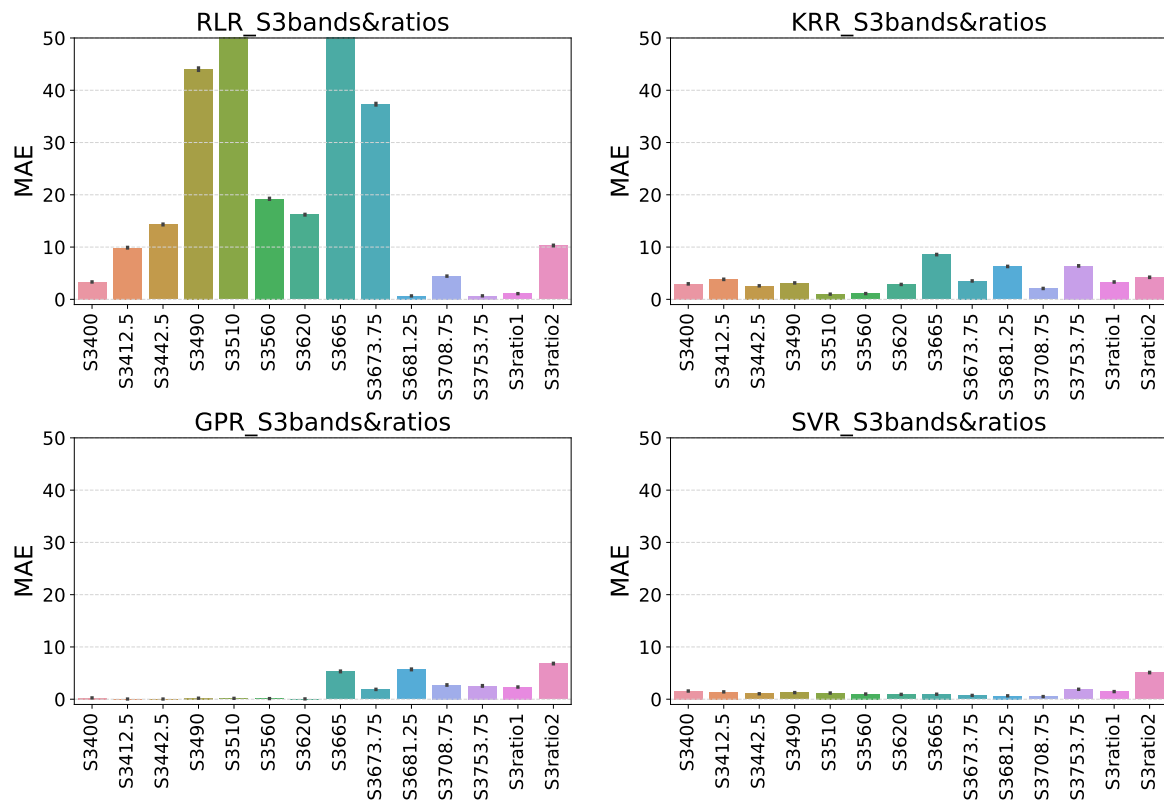
	SYKE				C2X			
	R2	Bias	MAE	RMSEr	R2	Bias	MAE	RMSEr
<b>S3-Ratio 1</b>								
Polyfit	0.936	0.004	1.048	0.133	0.931	−0.083	0.555	0.944
RLR	0.891	0.064	1.710	0.746	0.919	−0.059	0.695	1.547
RFR	0.899	0.029	1.330	0.174	0.901	−0.038	0.640	1.298
KRR	0.936	0.005	1.033	0.133	0.935	−0.067	0.523	0.587
GPR	0.932	0.013	1.037	0.132	0.935	−0.064	0.523	0.630
SVR	0.936	−0.254	1.010	0.129	0.926	−0.219	0.532	0.631
<b>S3-Ratio 2</b>								
Polyfit	0.959	−0.006	0.995	0.186	0.868	−0.115	0.712	5.110
RLR	0.948	0.015	1.139	0.325	0.646	−0.133	1.376	8.473
RFR	0.933	−0.048	1.228	0.206	0.861	−0.102	0.693	3.844
KRR	0.960	−0.010	0.996	0.185	0.898	−0.131	0.585	3.510
GPR	0.960	−0.009	0.996	0.196	0.893	−0.141	0.578	3.696
SVR	0.958	0.264	0.967	0.186	0.847	−0.238	0.689	4.021
<b>S3 two ratios</b>								
RLR	0.949	0.024	1.130	0.407	0.924	−0.076	0.689	2.306
RFR	0.968	−0.033	0.797	0.119	0.959	−0.080	0.392	0.529
KRR	0.973	0.007	0.750	0.117	0.957	−0.101	0.429	0.435
GPR	0.968	−0.014	0.760	0.114	0.953	−0.081	0.397	1.073
SVR	0.974	0.072	0.748	0.119	0.956	−0.180	0.393	0.477
<b>S3 all bands</b>								
RLR	0.580	0.036	3.100	1.483	0.605	0.001	1.781	7.145
RFR	0.981	−0.045	0.246	0.061	0.992	0.005	0.144	0.881
KRR	0.998	0.002	0.196	0.058	0.983	0.019	0.224	1.588
GPR	0.998	−0.002	0.029	0.068	0.996	0.008	0.113	1.104
SVR	0.993	−0.025	0.128	0.067	0.892	−0.245	0.383	0.845
<b>S3 all bands + ratios</b>								
RLR	0.991	0.006	0.474	0.164	0.962	−0.023	0.460	1.853
RFR	0.998	−0.007	0.126	0.038	0.992	−0.019	0.114	0.168
KRR	0.998	0.009	0.037	0.046	0.989	−0.008	0.131	2.264
GPR	0.999	0.012	0.014	0.035	0.996	−0.012	0.072	0.556
SVR	0.998	0.004	0.027	0.054	0.995	−0.023	0.095	0.334

In Figure 6, a similar pattern to the one described in the previous subsection with S2-MSI data (Figure 3) is found with the *S-3 Ratio 1* and *S-3 Ratio 2*, which leads to similar conclusions. The main difference between *S-2 Ratio 1* and *S-3 Ratio 1* is that the latter suffers less over-fitting in the smaller values of the RFR regression line.

Figure 7 shows the permutation plots of the S3-OLCI bands for four models and the *S3 all bands + ratios* configuration. They are, in general, very similar to the ones for S2-MSI. For RLR the importance of the bands in the blue and red spectrum seems to be more relevant. The RFR model stresses the importance of the ratios when those are used as inputs (not shown), behavior also observed in S2-MSI. For the KRR model, bands at 665, 681.25 and 753.75 nm point out, but generally the weight of the bands seems too be more distributed with lower MAE than the RLR. The GPR approach shows a major weight of the red and NIR bands (673.75, 681.25 and 753.75 nm), similar to the KRR. Finally, the bands for the SVR approach behaves in a similar way in S3-OLCI and S2-MSI, with the *S3-Ratio 2* pointing out. The higher weight of the blue bands in some of the approaches is expected, because the absorption maximum of CDOM takes place in these spectral ranges (400–490 nm), as it happens when the *S3 all bands* configuration is used (plots not shown). The contrast of these blue bands with the spectrum in the red and NIR (620–709 nm) are the inputs used in CDOM band-ratio algorithms in many studies. When the ratios are used as inputs as well, they gain relevance, as shown for S2-MSI.



**Figure 6.** Comparison of the performance of the models using the linear regression representation, S3-OLCI.



**Figure 7.** Permutation plots for four ML methods for the S3-OLCI configuration.

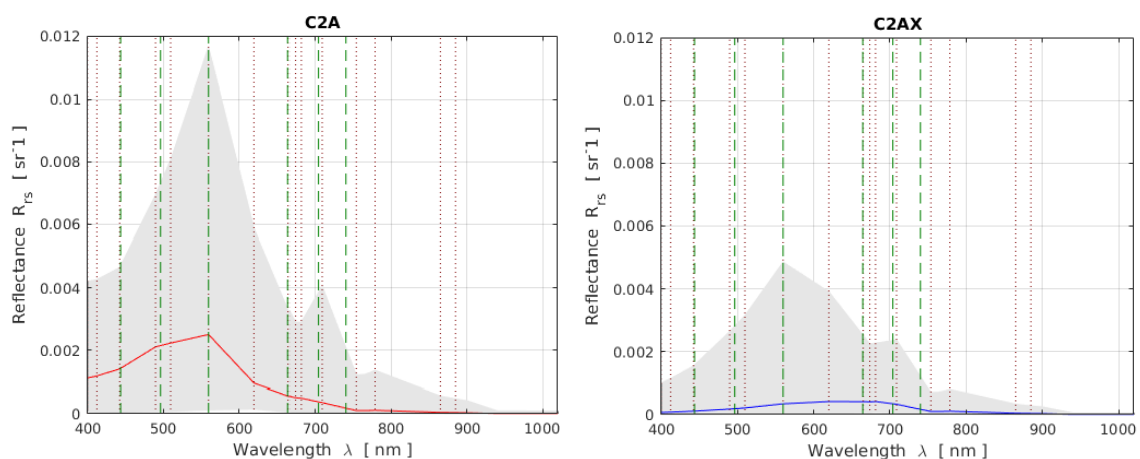
### 3.2.3. Application of the Models to the C2X Dataset

From the C2X dataset, we have selected all data classified as C2A and C2AX and in which chlorophyll is the main component of the specific absorption, that is only brown algae group dominates the signal. In total we have 5570 data that we use to train the models. Figure 8 shows the C2A data on the left and the C2AX on the right, both datasets are joined for the training of the models. Fifteen out of the 21 S3-OLCI wavebands are selected as inputs, including bands at 778.75, 865, 885 nm, which enlarge the experiment with new spectra. *S3-Ratio 1* and *S3-Ratio 2* are calculated to be included as inputs as

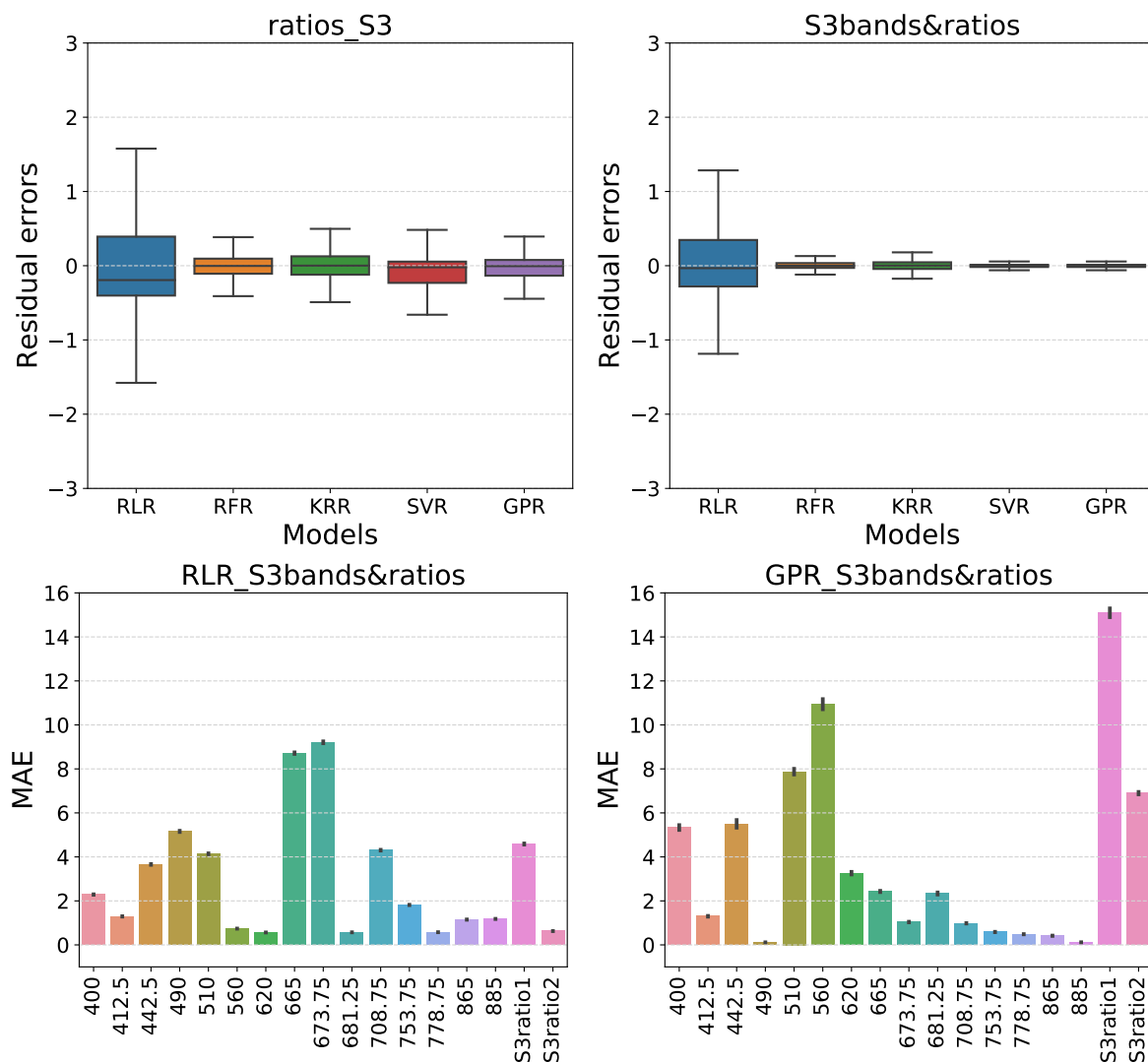
well. The  $a_{CDOM}(440)$  nm absorption coefficient, using both subgroups, has a range between 0.098 and  $20 \text{ m}^{-1}$ ; while the CHL content range from 0.03 rises until  $200 \text{ mg m}^{-3}$ ; TSM ranges from 0 to  $10 \text{ g m}^{-3}$ . It is worth noting that the absorption values are measure at a different wavelength than the SYKE dataset, which has to be taken into account when making comparisons. The C2X exclusive validation dataset is also filtered and 1783 spectra are left in an independent dataset, which is only used for validation purposes.

Table 4 (right part) shows the statistics of the model for comparison with the SYKE dataset. If the *S3-Ratio 1* is used as the only input, statistics show quite good results using the Polyfit model, outperformed slightly by the KRR and GPR methods, which raises the question of model complexity vs. simple models with fair results. *S3-Ratio 2* statistics show lower numbers than for *S3-Ratio 1*. The RLR fails compared with other approaches, having lower  $R^2$  (0.65) and higher RMSEr (8.47) and MAE (1.376). *S3-Ratio 2* seems to under-perform if we compare with the results with the SYKE dataset. When the *S3 two ratios* are used as inputs, the RLR method gives similar results than when using only the *S3-Ratio 1*. The influence of the *S3-Ratio 1* is then affecting the RLR and the rest of the models stronger than the *S3-Ratio 2*. However, if we compare the RLR and GPR permutation plots, the influence of *S3-Ratio 2* is still visible in the GPR model (Figure 9), though not higher than *S3-Ratio 1*. If all bands are used as input, generally good results are obtained for all models except the simplest one (RLR), with  $R > 0.9$  and  $MAE < 0.4$ . In terms of error, using all bands, the RFR does a very good job, with a high coefficient of determination (0.992) and one of the lowest RMSEr (0.881) and MAE (0.144) errors. If all reflectance bands and the ratios are used as inputs, improvements are clear with the RLR method in comparison with the other experiments. The best model is again the GPR, but all approaches have coefficients of determination and correlation  $> 0.96$  and low errors and biases.

The permutation analysis shows that the GPR model has comparable results with the SYKE dataset (Figure 9, right plot), with major weight of the blue bands when no ratios are taken as inputs (not shown), changing the weights to bands 510 and 560 nm followed by the blue bands, and highlighting the importance of the ratios when they are taken into account.



**Figure 8.** Ranges and mean of reflectance spectra at S3-OLCI wavebands from the C2X dataset for C2A and C2AX subsets. The utilized S2-MSI and S3-OLCI bands are highlighted for convenience.



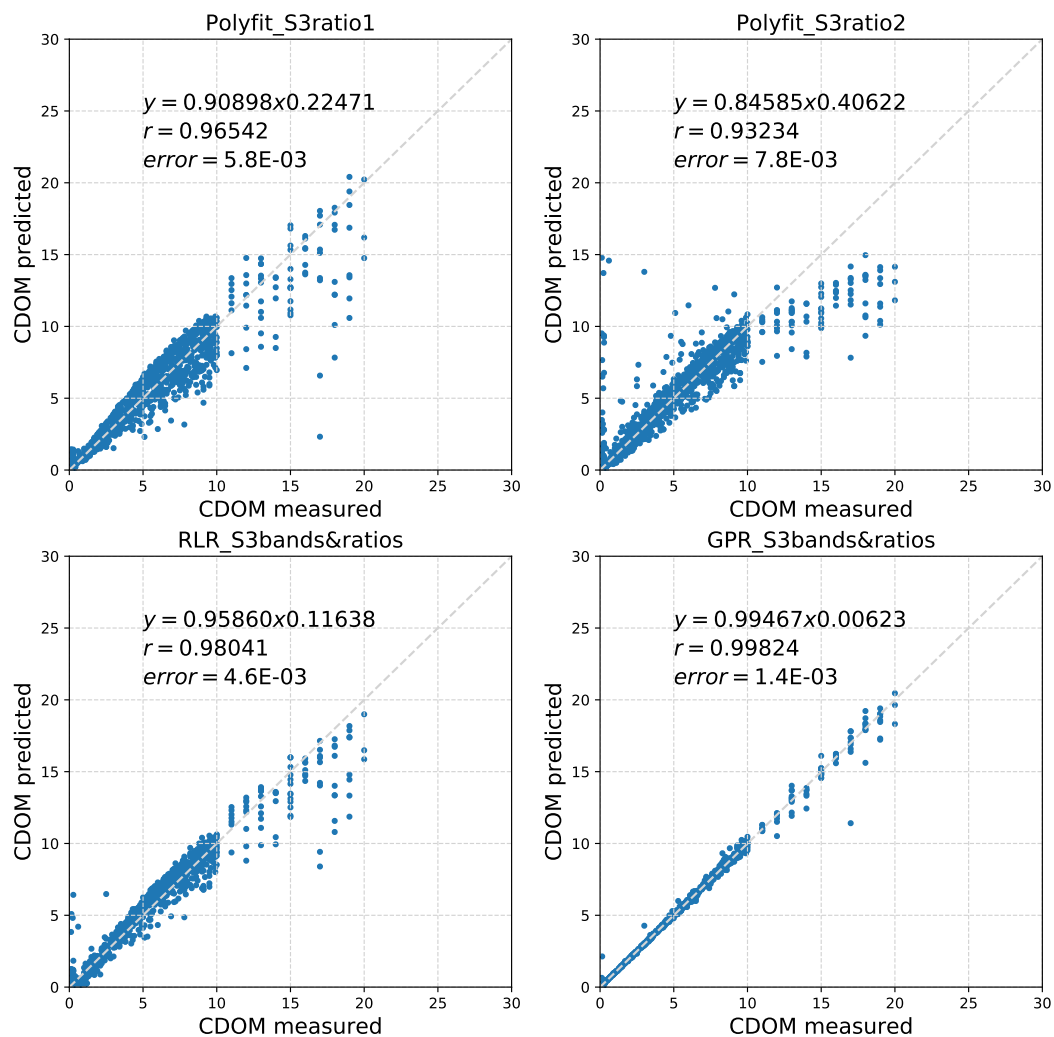
**Figure 9.** Statistic box-plots of the residual errors on the top; permutation plots of the RLR and GPR model on the bottom—C2X S3-OLCI.

### 3.2.4. Predicted vs. Measured CDOM with the C2X Dataset

As an example, plots and statistics comparing CDOM predicted (or retrieved) and CDOM measured (or simulated) for the six models and all possible band combinations have been generated. Figure 10 summarizes the main statistics using a linear regression of the two variables for some of models and band combinations. In general we obtain coefficients of correlation  $> 0.9$  for all cases, except for the RLR method in *S3 all bands* with 0.782, and the RLR for the *S3-Ratio 2*. These two correlations also have the biggest standard error and deviate from the 1:1 line in the regression plots showing a negative bias. The best correlation results uses the *S3 all bands + ratios* combination, as pointed out before, with  $r$  values  $> 0.98$  in all cases. The GPR and the SVR methods have similar results, with a slight lower error of the GPR.

Figure 10 (bottom row) shows an example of the best result -best band combination and best model-, which is the GPR *S3 all bands + ratios*, together with the RLR model with the same input. This comparison is made to observe the better adjustment of the results using the GPR approach, with a distribution of the data around the 1:1 line almost perfect when compared with the RLR results. The same figure shows the results for the simple Polyfit method on the two separated ratios on the top row, for a better understanding of the overall statistics and data distribution.





**Figure 10.** Retrieval performance with respect to the C2A(X) CDOM simulations: on the top left the scatter plot of the Polyfit method with the *Ratio1*; on the top right the scatter plot of the Polyfit method with the *Ratio2*. On the bottom left RLR method using all available bands and the two ratios as input; on the bottom right the GPR method with all available bands and the two ratios.

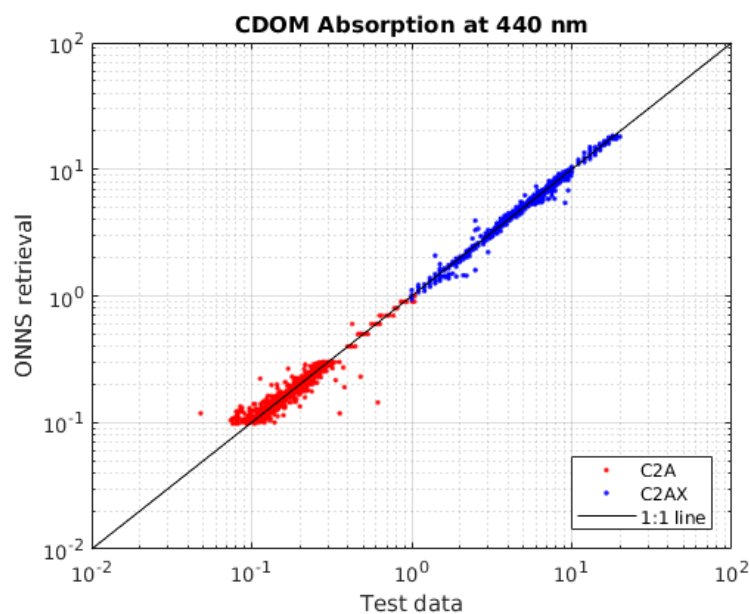
#### 4. Comparison of the C2X Experiment with Neural Nets

##### 4.1. Comparison of the Results Extracted with OLCI Neural Net Swarm

As it is the case with the machine learning methods, ONNS is an in-water algorithm, especially designed to retrieve water quality parameters from S3-OLCI satellite data. The aim of ONNS is to provide one algorithm that is suitable to all natural waters, from clearest oceanic waters to very turbid coastal or highly absorbing inland waters. For this purpose, a fuzzy logic optical water type classification is applied in conjunction with a set of specific neural networks. Input to the algorithm are normalized remote sensing reflectance at 11 OLCI bands, i.e., atmospheric corrected satellite data. ONNS retrieves 12 ocean color products with corresponding uncertainties, concentrations and optical properties, which form a system-inherent optical closure; CDOM absorption at 440 nm is one of the outputs. ONNS is based on the C2X dataset. The CDOM retrieval capabilities of ONNS are shown in comparison with the validation data in [28]; they state that lowest CDOM concentrations (in Case-1 waters) vanish in the noise, whereas high concentrations (C2AX) can be retrieved accurately. For C2A waters, the  $a_{CDOM}(440)$  obtains an absolute RMSE (evaluated in the linear space) equal to 0.023; the coefficient of correlation is 0.985; and the bias is slightly negative with  $-0.005$ . CDOM

concentration is higher in C2AX waters, thus, RMSE and bias are larger, 0.435 and  $-0.021$  respectively, but the correlation keeps high at 0.993.

The C2X dataset used in the present work is a small subset of the original database, only the ‘brown spectral phytoplankton group’. As an aside, in both cases C2A and C2AX, approximately 1.2 of cases are non-classifiable with ONNS, which results in NaNs after the application of the NN. Attending at the statistics of the linear regression between predicted and C2A(X) test data (see Figure 11 in log10 scale), with slope of 1.003, intercept of 0.004, correlation coefficient of 0.998, and standard error of  $1.6 \times 10^{-3}$ , we can conclude that ONNS shows similar good results, with slightly higher standard errors but same correlation coefficients than the GPR-S3 *all bands + ratios*, the method that we have considered the best one in the previous section.



**Figure 11.** Absorption coefficient of CDOM at 440 nm retrieved by ONNS (IOP NNs) for the C2A (red dots) and C2AX (blue dots) spectral types. Plot is in log10 scale.

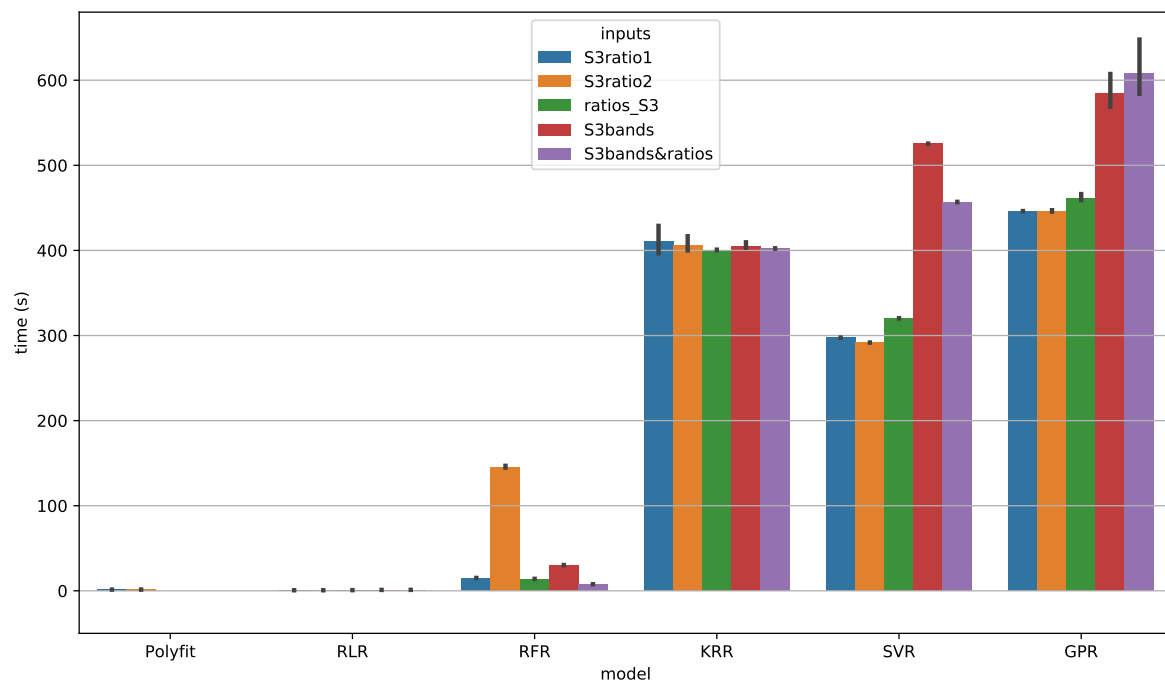
#### 4.2. Comparison Against the Standard S3-OLCI Product

ML approaches have statistically demonstrated to be very promising techniques for bio-geophysical land and water quality parameter retrieval [46,48]. To set the applicability of these methods for satellite image processing, here there is an example of the processing of an OLCI scene using the GPR model. This exercise cannot be considered a proper validation, since there are too many differences in the approaches, starting with the source of the reflectance bands. The ML methods rely on atmospherically corrected data, the training of the models is done with the C2X dataset, and then apply to the reflectance on the scene. The standard OLCI ocean processing module provides water-leaving reflectance in several steps (<https://sentinel.esa.int/web/sentinel/user-guides/sentinel-3-olci/processing-levels/level-2>). The first step is performed using two algorithms, selectable by a dedicated switch:

- The Baseline Atmospheric Correction (BPAC) removes all the contributions to TOA reflectance, including glint correction and white cap effects. Later in the processing it estimates the near-infra-red water-leaving reflectance to perform the atmospheric correction.
- The Alternate Atmospheric Correction (AAC) uses a neural network approach to provide water-leaving reflectance, but these products are not written by default in the standard L2 provided by EUMETSAT.

Several other processors compute all the needed products from the water-leaving reflectance (OC4ME Chlorophyll, IMT Neural Net, Transparency Product, PAR Product). The AAC atmospheric correction scheme originally designed for MERIS (Medium Resolution Imaging Spectrometer), the *MERIS Case-2* neural network or C2R [25], is used as basis for the OLCI NN development [26]. Some of the products retrieved using the C2R neural net are concentrations (CHL\_NN, TSM\_NN, ADG443\_NN) together with atmospheric data (aerosol optical thickness and angstrom exponent at 865, integrated water column, etc.).

The scene processed here is downloaded from the CODA EUMETSAT service (<https://eoportal.eumetsat.int>), with the date of acquisition on the 24 May 2016 (09:09:53–09:11:53 am). Processor version at time of downloading is IPF-OL-2 06.11. It is difficult, due to the short operational time of the S3 satellite and the complex atmospheric situation of the area, to find perfectly clear scenes over the Baltic Sea area. Other issues regarding the atmospheric correction and its potential effect on CDOM retrieval is the overcorrection of naturally low radiances in the Baltic area, which can lead to estimations of negative reflectance when using the BPAC processing. This bias can affect the CDOM retrieval, with extremely high and unrealistic values. In order to overcome some of these potential problems, we downloaded the Level-1 image and processed it with the C2RCC algorithm, available in the Sentinel-3 Toolbox of the SeNtinel Application Platform (SNAP) (<http://step.esa.int/main/toolboxes/snap/>). We used the remote sensing reflectance bands for the GPR model previously trained with C2X dataset. The OLCI scene has dimensions  $2728 \times 4865$ , from where 2,625,816 is the number of valid pixels left after application of the C2RCC flags (Cloud\_risk, Rtos\_a\_OOS, Rtos\_a\_OOR, Rhov\_OOS, Rhov\_OOR). Figure 12 shows in a bar plot how long it takes for each model and band combination to compute the results. As the model begins to be more complex, the computation time increases. We observe a minimum of a few seconds when we apply the RLR until approximately 10 min of the GPR using all bands and ratios.

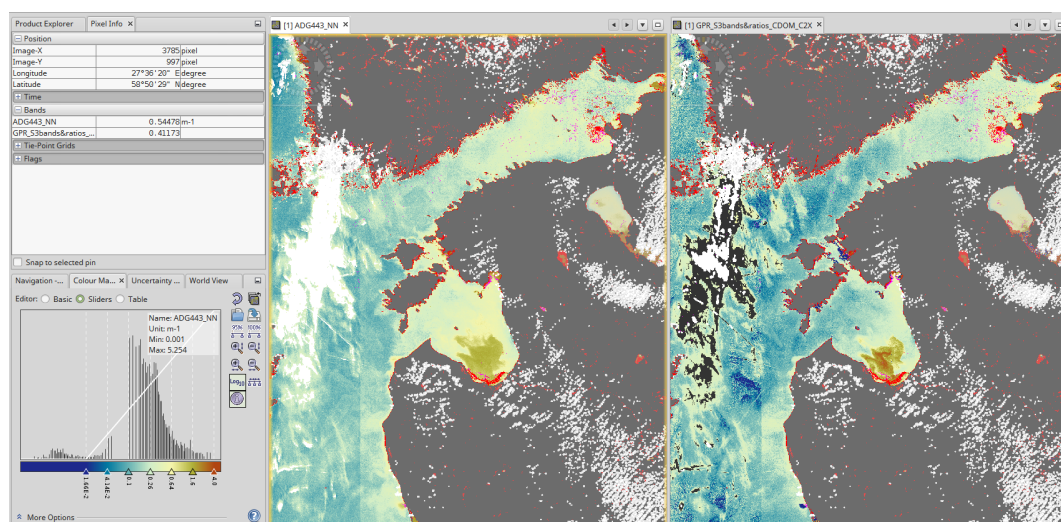


**Figure 12.** Computational time in seconds of each model and band combination on an standard OLCI scene.

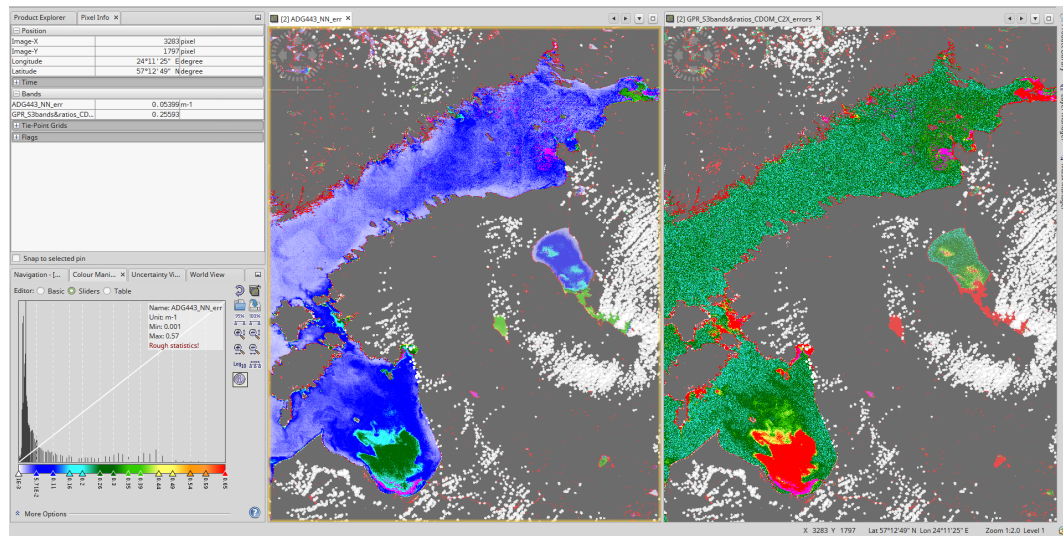
The comparison of the ADG443\_NN product and the CDOM product calculated with the the GPR model is shown within a snapshot of the SNAP software in Figure 13. Both images show similar patterns: the ADG443\_NN is shown in the left, in the right the GPR\_CDOM output. The area in the

map shows the Gulf of Finland and the eastern part of the Baltic Sea. Only valid sea pixels are shown, with land pixels masked in gray and clouds masked in white. Red pixels are the invalid pixels flagged by WQSF\_1sb\_OCNN\_fail, Rtosa\_OOS and Rhov\_OOS (BPAC and C2R NN flags). A cloudy area in the center of the Baltic Sea is masked by the Cloud\_risk flag and appears in white in the ADG443\_NN product. This flag is applied previously in the GPR\_CDOM processing and that is the reason the pixels of this area appear as black (Not A Number) in the derived CDOM product. The uncertainty maps of both products are shown in Figure 14. ADG443\_NN\_err is shown in the left (higher uncertainties in cyan and green); in the right the GPR associated uncertainties (higher uncertainties in green and red). The ADG443\_NN and associated uncertainties include the CDOM absorption plus the absorption of the detritus (minerals), which makes the quantitative comparison of both products a bit more difficult. However, the patterns of the products and the map of the uncertainties look very similar, with higher uncertainties in the same areas, for instance in the southern part of Lake Peipus or the Gulf of Riga.

To better understand the differences, it is important to highlight that the C2R NN has been trained with a large amount of Hydrolight simulations generated by the forward model [27], plus a bio-optical model relating scattering and absorption coefficients to concentrations. The bio-optical model is also based on a large dataset of in situ measurements of inherent optical properties [26]. The C2R NN converts the directional water leaving radiance reflectance (which recalculates within the NN) into a number of inherent optical properties. These are then converted into concentrations of water constituents by simple regression [26]. The complexity of this NN, including the improved atmospheric correction, is being compared here with less complex, but still sophisticated algorithms, with results that are quite optimized. ML approaches can help to reduce the training phase -and the generation of simulations- and processing time of satellite data in complex waters giving more than fair results.



**Figure 13.** Comparison of ADG443\_NN vs. GPR model ( $m^{-1}$ ): left ADG443\_NN product; right GPR CDOM output.



**Figure 14.** Comparison of ADG443\_NN uncertainties vs. the uncertainties of the GPR model ( $m^{-1}$ ): left ADG443 uncertainties; right GPR CDOM uncertainties.

Other models, including SVR and KRR are also applied, and they show very similar patterns to the ADG443\_NN output (not shown here) and quantitative values in the Gulf of Finland and northern part of Lake Peipus closer to each other than the GPR output. However, these methods do not allow to derive the uncertainty maps. In any case, if we consider the ADG443\_NN as a reference, we could say that the SVR, KRR and the GPR underestimate CDOM in the western part of this scene. In the Gulf of Finland the images are very similar, with a slight underestimation going further to the west too; and a slight overestimation of the GPR (not in KRR or SVR) in the Gulf of Riga, the area showing higher uncertainties in the standard product.

## 5. Conclusions

This study presents the statistical evaluation of five machine learning approaches for the retrieval of ocean color remote sensing products, focused in the CDOM parameter. The advanced statistical approaches used are classical and well-established machine learning methods: multivariate regression (RLR), random forests regression (RFR), kernel ridge regression (KRR), Gaussian process regression (GPR) and support vector regression (SVR). They are applied on remote sensing reflectance simulating S2-MSI and S3-OLCI bands. The experiments are made using two different datasets: the SYKE dataset (S2-MSI and S3-OLCI) and the C2X dataset (S3-OLCI). The datasets are representative of water types that can be classified as Case 2 absorbing and extreme absorbing waters (C2A and C2AX respectively). The results of these advanced statistic approaches are compared with more traditional empirical algorithms (second grade polynomial); and the inputs of each method is studied through permutation analysis to determine their weight in the models. The analysis tries to determine the relationship of the results of each model with observed natural behaviors and with the results of simpler algorithms.

The main message after analyzing the statistics is that the more input bands in the model, the better, independently of the instrument (MSI or OLCI). However, this would assume that the input values are perfect and no errors after the atmospheric corrections are included. In the real cases, errors are larger in some bands than in others, which makes the option of not including all of them well-founded. Furthermore, if the inputs contain already some features (ratios) that correlate with the output products (CDOM in this case), the gains are clearly visible since they help to improve the statistics significantly. The complexity of the models has a relevant role when all the available inputs are used; but with the proper inputs it is possible to rely in simpler models and obtain still quite good results.

Part of the validation of the models is made with an independent dataset derived from the C2X dataset and the results shown in this paper are limited to the OLCI sensor. Comparison between

predicted values and measured/simulated ones is made by linear regression statistics. Basically all methods give quite good coefficients of correlation ( $0.782 < r < 0.99$ ), with greater success of the GPR and SVR methods in all the experiments. Results are also compared with other machine learning approaches like neural nets. The OLCI Neural Net Swarm (ONNS) is an algorithm specifically design with the complete C2X dataset for all five main water types (C1, C2A, C2S, C2AX, C2SX). Only results for C2A and C2AX water types are compared here, with the GPR model (followed by SVR) and the ONNS offering similar good performance for both water types. This evaluation increase the confidence in ML approaches utilization for OC retrievals.

An application to a real OLCI scene is also made for verification of the applicability of the methods on atmospherically corrected satellite data. The  $R_{rs}$  data used from the image is derived with the C2RCC processor. This fact can introduce already important discrepancies since the input reflectance data come from different sources. The standard proxy for CDOM (ADG443\_NN) is found in the OLCI L2 product derived by ESA/EUMETSAT and calculated using the C2R NN. The C2R NN relies on the derivation of IOPs and calculates concentrations applying a bio-optical model, which introduces another important source of uncertainty in the comparison. However, results with the GPR, SVR and KRR outputs show similar patters and small differences in those areas with lower uncertainties, which gives hope on the good performance of a faster to train and simpler mathematical algorithms for operational production.

A rigorous validation procedure cannot be followed since there is still not a large match-up dataset available. Hopefully, the Sentinel-3 Validation Team and other public and private initiatives would help to solve this issue in the next few years, and more quantitatively validation of the products will be possible.

Future work will be focused on three main issues: (i) retrieval of other OC variables like chlorophyll-a or total suspended matter; (ii) adding new approaches like joint or deep Gaussian processes; (iii) extent and improve the quantitative validation with proper data and methods, including application to imagery that has been atmospherically corrected with the same or similar methods than the input data use for the training of the models.

**Supplementary Materials:** The code reproducing the results of this paper can be found in <https://github.com/IPL-UV/mlregoocean>. In addition, an operational Matlab toolbox implementing all machine learning methods is available at <https://github.com/IPL-UV/simpleR>.

**Author Contributions:** A.B.R. is the main author of the paper, designed, developed and implemented the experiments. She contributed in the Case2eXtreme project, main source of the C2X dataset. G.M.-G. designed and implemented the experiments with the different databases in Python. M.H. contributed with the development of the C2X dataset; he also built the ONNS and recalculated the partial results shown in this paper. M.-G. profusely wrote and reviewed this work. S.K. and K.K. contributed with the SYKE database and the writing and reviewing of the paper. G.C.-V. helped with the selection of the ML approaches, since the SimpleR code written in Matlab and authored by him is the basis of the current development. He also wrote the ML section and helped with the overall revision of the paper.

**Acknowledgments:** The research is funded by the European Research Council (ERC) under the ERC-CoG-2014 SEDAL project (grant agreement 647423). Special thanks to the Case2eXtreme project team (funded by ESA) for the availability of the C2X dataset. Thanks to SYKE for the availability and preparation of the SYKE dataset.

**Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

## References

1. Preisendorfer, R.W. *Hydrologic Optics*; U.S. Department of Commerce; National Oceanic and Atmospheric Administration; Environmental Research Laboratories; Pacific Marine Environmental Laboratory: Honolulu, HI, USA, 1976.
2. Jerlov, N. *Marine Optics*, 2nd ed.; Elsevier Science: New York, NY, USA, 1976; Volume14, p. 230.
3. Mobley, C. *Light and Water: Radiative Transfer in Natural Waters*; Academic Press: Cambridge, MA, USA, 1994.

4. Mobley, C.; Sundman, L.K. *HydroLight 5.2, Ecolight 5.2, Technical Documentation*; Technical Report; Sequoia Science Inc.: Bellevue, WA, USA, 2013.
5. Morel, M.; Prieur, L. Analysis of variations in ocean color. *Limnol. Oceanogr.* **1977**, *22*, 709–722. [[CrossRef](#)]
6. Dekker, A. Detection of Optical Water Quality Parameters for Eutrophic Waters by High Resolution Remote Sensing. Ph.D. Thesis, Vrije Universiteit, Amsterdam, The Netherlands, 1993.
7. Bukata, R.; Jerome, J.; Kondratyev, K.; Pozdnyakov, D. *Optical Properties and Remote Sensing of Inland and Coastal Waters*; CRC Press: Boca Raton, FL, USA, 1995.
8. Simis, S.; Tijdens, M.; Peters, S.; Gons, H. Optical characterization of cyanobacterial bloom termination. *Verhandlungen Internationale Vereinigung für Theoretische und Angewandte Limnologie* **2005**, *29*, 941–944. [[CrossRef](#)]
9. Moses, W.; Gitelson, A.; Berdnikov, S.; Povazhnyy, V. Estimation of chlorophyll- a concentration in case II waters using MODIS and MERIS data: successes and challenges. *Environ. Res. Lett.* **2009**, *4*, 045005. [[CrossRef](#)]
10. Giardino, C.; Bartoli, M.; Candianai, G.; Brescian, M.; Pellegrini, L. Recent changes in macrophyte colonisation patterns: an imaging spectrometry-based evaluation of southern Lake Garda (Northern Italy). *J. Appl. Remote Sens.* **2007**, *1*, 011509.
11. Kallio, K. Optical properties of Finnish lakes estimated with simple bio-optical models and water quality monitoring data. *Hydrol. Res.* **2006**, *37*, 183–204. [[CrossRef](#)]
12. 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](#)]
13. Guanter, L.; Ruiz-Verdu, A.; Odermatt, D.; Giardino, C.; Simis, S.; Estelles, V.; Heege, T.; Dominguez-Gomez, J.A.; Moreno, J. Atmospheric correction of ENVISAT/MERIS data over inland waters: Validation for European lakes. *Remote Sens. Environ.* **2010**, *114*, 467–480. [[CrossRef](#)]
14. Hieronymi, M.; Krasemann, H.; Mueller, D.; Brockmann, C.; Ruescas, A.; Stelzer, K.; Nechad, B.; Ruddick, K.; Simis, S.; Tislone, G.; Steinmetz, F.; Regner, P. Ocean Colour Remote Sensing of Extreme Case-2 Waters. In Proceedings of the 2016 ESA Living Planet Symposium, Prague, Czech Republic, 9–13 May 2016.
15. 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. [[CrossRef](#)]
16. Tehrani, N.; DSa, E.; Osburn, C.; Bianchi, T.; Schaeffer, B. Chromophoric Dissolved Organic Matter and Dissolved Organic Carbon from Sea-Viewing Wide Field-of-View Sensor (SeaWiFS), Moderate Resolution Imaging Spectroradiometer (MODIS) and MERIS Sensors: Case Study for the Northern Gulf of Mexico. *Remote Sens.* **2013**, *5*, 1439–1464. [[CrossRef](#)]
17. Kutser, T.; Paavel, B.; Verpoorter, C.; Ligi, M.; Soomets, T.; Toming, K.; Casal, G. Remote Sensing of Black Lakes and Using 810 nm Reflectance Peak for Retrieving Water Quality Parameters of Optically Complex Waters. *Remote Sens.* **2016**, *8*, 497. [[CrossRef](#)]
18. Kallio, K. *Water Quality Estimation by Optical Remote Sensing in Boreal Lakes*; Monographs of the Boreal Environment Research 39; The Finnish Environment Institute: Helsinki, Finland, 2012.
19. Kallio, K.; Koponen, S.; Ylöstalo, P.; Kervinen, M.; Pyhälähti, T.; Attila, J. Validation of MERIS spectral inversion processors using reflectance, IOP and water quality measurements in boreal lakes. *Remote Sens. Environ.* **2015**, *157*, 147–157. [[CrossRef](#)]
20. Kortelainen, P. Content of Total Organic Carbon in Finnish Lakes and Its Relationship to Catchment Characteristics. *Can. J. Fish. Aquat. Sci.* **1993**, *50*, 1477–1483. [[CrossRef](#)]
21. Attila, J.; Koponen, S.; Kallio, K.; Lindfors, A.; Kaitala, S.; Ylöstalo, P. MERIS Case II water processor comparison on coastal sites of the northern Baltic Sea. *Remote Sens. Environ.* **2013**, *128*, 138–149. [[CrossRef](#)]
22. Beltran-Abauza, J.M.; Kratzer, S.; Brockmann, C. Evaluation of MERIS products from Baltic Sea coastal waters rich in CDOM. *Ocean Sci.* **2014**, *10*, 377–396. [[CrossRef](#)]
23. Brezonik, P.; Olmanson, L.; Finlay, J.; Bauer, M. Factors affecting the measurement of CDOM by remote sensing of optically complex inland waters. *Remote Sens. Environ.* **2015**, *157*, 199–215. [[CrossRef](#)]
24. Alikas, K.; Lauth, S.; Reinart, A. *D3.4 Adapted Water Quality Algorithms*; Technical Report, GLaSS Project, H2020; European Union: Brussels, Belgium, 2014.
25. Doerffer, R.; Schiller, H. The MERIS Case 2 water algorithm. *Int. J. Remote Sens.* **2007**, *28*, 517–535. [[CrossRef](#)]
26. Doerffer, R. *OLCI L2 ATBD*; Technical Report; GKSS: Gothenburg, Sweden, 2011.

27. Brockmann, C.; Doerffer, R.; Peters, M.; Stelzer, K.; 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 Proceedings of the Living Planet Symposium, Prague, Czech Republic, 9–13 May 2016.
28. Hieronymi, M.; Mueller, D.; Doerffer, R. The OLCI Neural Network Swarm (ONNS): A Bio-Geo-Optical Algorithm for Open Ocean and Coastal Waters. *Front. Mar. Sci.* **2017**, *4*, 140. [[CrossRef](#)]
29. Ylöstalo, P.; Kallio, K.; Seppälä, J. Absorption properties of in-water constituents and their variation among various lake types in the boreal region. *Remote Sens. Environ.* **2014**, *148*, 190–205. [[CrossRef](#)]
30. Hieronymi, M.; Kraseman, H.; Ruescas, A.; Brockmann, C.; Steinmetz, F.; Tilstone, G.; Simis, S. *Algorithm Theoretical Basis Document; Technical Report, Case 2 eXtreme Project; ESA: Paris, France, 2015.*
31. Kraseman, H.; Hieronymi, M.; Simis, S.; Steinmetz, F.; Tilstone, G.; Nechad, B.; Kraemer, U. *Database for Task 2, Technical Note; Technical Report, Case 2 eXtreme Project; ESA: Paris, France, 2016.*
32. Breiman, L.; Friedman, J. Estimating Optimal Transformations for Multiple Regression and Correlation. *J. Am. Statist. Assoc.* **1985**, *80*, 1580–1598.
33. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004.
34. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; The MIT Press: New York, NY, USA, 2006.
35. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [[CrossRef](#)]
36. Evans, J.S.; Cushman, S.A. Gradient modeling of conifer species using random forests. *Landsc. Ecol.* **2009**, *24*, 673–683. [[CrossRef](#)]
37. Oliveira, S.; Oehler, F.; San-Miguel-Ayanz, J.; Camia, A.; Pereira, J.M. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *For. Ecol. Manage.* **2012**, *275*, 117–129. [[CrossRef](#)]
38. Gislason, P.O.; Benediktsson, J.A.; Sveinsson, J.R. Random Forests for land cover classification. *Pattern Recognit. Lett.* **2006**, *27*, 294–300. [[CrossRef](#)]
39. Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forest for classification in Ecology. *Ecology* **2007**, *88*, 2783–2792. [[CrossRef](#)] [[PubMed](#)]
40. Genuer, R.; Poggi, J.M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [[CrossRef](#)]
41. Jung, M.; Zscheischler, J. A Guided Hybrid Genetic Algorithm for Feature Selection with Expensive Cost Functions. *Procedia Comput. Sci.* **2013**, *18*, 2337–2346. [[CrossRef](#)]
42. Camps-Valls, G.; Bruzzone, L. *Kernel methods for Remote Sensing Data Analysis*; Wiley & Sons: Oxford, UK, 2009.
43. Camps-Valls, G.; Verrelst, J.; Munoz-Mari, J.; Laparra, V.; Mateo-Jimenez, F.; Gomez-Dans, J. A Survey on Gaussian Processes for Earth Observation Data Analysis. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 58–78. [[CrossRef](#)]
44. Suykens, J.A.K.; Gestel, T.V.; Brabanter, J.D.; Moor, B.D.; Vandewalle, J. (Eds.) *Least Squares Support Vector Machines*; World Scientific Publishing Co.: Singapore, 2002.
45. Furfaro, R.; Morris, R.D.; Kottas, A.; Taddy, M.; Ganapol, B.D. A Gaussian Process Approach to Quantifying the Uncertainty of Vegetation Parameters from Remote Sensing Observations. In Proceedings of the AGU Fall Meeting Abstracts, San Francisco, CA, USA, 11–15 December 2006.
46. Verrelst, J.; Muñoz, J.; Alonso, L.; Delegido, J.; Rivera, J.; 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](#)]
47. 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](#)]
48. Blix, K.; Jenssen, R.; Camps-Valls, G. Gaussian Process Sensitivity Analysis for Oceanic Chlorophyll Estimation. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2017**, *1*, 1–13. [[CrossRef](#)]
49. Schölkopf, B.; Smola, A. *Learning with Kernels—Support Vector Machines, Regularization, Optimization and Beyond*; MIT Press: Cambridge, MA, USA, 2002.



50. Anderson, M.; Robinson, J. Permutation Tests for Linear Models. *Aust. N. Z. J. Stat.* **2001**, *43*, 75–88. [[CrossRef](#)]
51. Camps-Valls, G.; Jung, M.; Ichii, K.; Papale, D.; Tramontana, G.; Bodesheim, P.; Schwalm, C.; Zscheischler, J.; Mahecha, M.; Reichstein, M. Ranking drivers of global carbon and energy fluxes over land. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).