Combining BGC-Argo floats and satellite observations for water column estimations of particulate backscattering coefficient

Jorge García-Jiménez[1], Ana B. Ruescas[1], Julia Amorós-López[1], and Raphaëlle Sauzède[2]

[1]Image Processing Laboratory, Universitat de València, Spain
[2]Institut de la Mer de Villefranche, FR3761, CNRS, Sorbonne Université, France

Correspondence: Jorge García-Jiménez (jorge.garcia-jimenez@uv.es)

Abstract. Monitoring carbon cycle processes is key to understanding climate system science. As the second largest carbon reservoir on Earth, the ocean regulates carbon balance through Particulate Organic Carbon (POC), which links surface biomass production, the deep ocean, and sedimentation. The degradation of POC in the deep ocean notably impacts atmospheric CO_2 levels. POC estimation is achieved by measuring proxies like the Particulate Backscattering Coefficient (b_{bp}), obtained from satellite observations and *in situ* sensors, such as the BioGeoChemical-Argo (BGC-Argo) floats. These floats provide global-scale profiles of ocean biogeochemical properties. Previous research has combined data from BGC-Argo floats and satellite sensors, demonstrating the potential of machine learning models to infer vertical bio-optical properties in the water column. By bridging the gap between surface optical properties and deep ocean processes, this approach enhances the estimation within the top 250 meters of the water column. This study focuses on such estimations, including remote sensing data from the Sentinel-3 Ocean and Land Colour Instrument (OLCI) sensor. The addition of optical information about absorption and scattering processes has improved the accuracy of the Random Forest models, which show promising results, especially within the first 50 meters in the Subtropical Gyres. However, in dynamic regions like the North Atlantic, results are less consistent, suggesting further research is needed to understand how the complexity of the water column's physical state modifies the b_{bp} vertical fluxes.

1 Introduction

The ocean covers approximately 70% of Earth's surface and plays a fundamental role in climate dynamics. It regulates and redistributes energy and carbon through various physical and biogeochemical processes. In the regulation of atmospheric carbon dioxide by the ocean, the so-called biological carbon pump is the process that enables the transfer of CO_2 from the atmosphere to the ocean floor. Photosynthetic organisms (phytoplankton) living in the upper layers where sunlight is available require carbon compounds to survive and reproduce (Falkowski et al., 1998; Siegel et al., 2023). The presence and amount of these organisms heavily depend on the availability of light and nutrients in the environment (Behrenfeld et al., 2006). The challenges of directly observing and quantifying POC at various depths, combined with the complex interaction between key variables (usually non-linear) and the low-resolution measurements in a highly dynamic environment, contribute to gaps in our understanding of specific marine processes, such as carbon sequestration, nutrient cycling, sedimentation, and ocean-

25 atmosphere CO_2 exchange. The carbon dioxide captured ends up forming ocean biomass, contributing to particulate organic carbon. Particulate carbon, both organic (POC) and inorganic (PIC), can also sink and become locked away in sediments. In the current global warming context, a more accurate representation of POC distribution is key for estimating the ocean's contribution to the Earth's carbon cycle.

Bio-optical sensors installed on autonomous platforms have become a reliable technology for acquiring *in situ* data about
30 water masses' ecological and physical status. These sensors measure the scattering of light in water, which provides information about radiative transfer conditions and the nature and dynamics of suspended particulate matter. The particulate backscattering coefficient (b_{bp}), an inherent optical property (IOP) of water, has proven to be a reliable bio-optical proxy for POC (Cetinić et al., 2012; Sullivan et al., 2013). IOPs are the intrinsic characteristics of water, determined solely by its composition and independent of the external light field or the geometrical angle conditions during observation. These properties include absorp-
35 tion, scattering, and attenuation processes, which describe how light behaves and propagates through water. IOPs are essential in studying light interactions in aquatic environments, as they reflect the presence of dissolved organic matter, phytoplankton, and suspended particles. On the contrary, Apparent Optical Properties (AOPs) are characteristics of light in water that depend not only on the water's Inherent Optical Properties (IOPs) but also on the geometric conditions of light, such as the azimuthal and zenithal angles of the sun. AOPs include measurements like reflectance, attenuation, and the diffuse attenuation coeffi-
40 cient, which are influenced by the water's composition (e.g., dissolved substances and particulates) as well as the surrounding environmental light conditions. The IOP b_{bp} can be measured by autonomous platforms spread out across the ocean, such as the Biogeochemical-Argo (BGC-Argo) profiling floats (Claustre et al., 2020); or estimated from onboard satellite sensors, such as the Sentinel-3 Ocean and Land Colour Instrument (OLCI)[1] (EUMETSAT, 2019; Jorge et al., 2021; Koestner et al., 2024). Designing observational strategies based on combining the two approaches constitutes a fundamental tool for improving
45 knowledge of ocean processes (BGC, 2016).

POC can be quantified from *in situ* filtered seawater samples. Conducting field campaigns is costly, and data availability is often limited. When using satellite ocean colour data, POC is quantified at a global scale daily. Several approaches have been developed to estimate POC from optical measurements of water leaving radiance (L_w), or by linking POC with remote sensing derived IOPs (Bisson et al., 2019; Evers-King et al., 2017; Loisel et al., 2002; Stramski et al., 2008). However, these
50 methods are designed to estimate parameters at the sea surface, which does not fully capture the complexities of carbon export in the ocean, as numerous vertical processes within the water column significantly influence the carbon cycle. Fusing satellite data with vertical profiles from BGC-Argo floats to extend the measurements of surface bio-optical properties (i.e. b_{bp}) to several depth layers is performed with the SOCA method in Sauzède et al. (2016, 2020). The initial SOCA2016 method consists of a neural network combining satellite surface estimates of b_{bp} and chlorophyll-a (chl-a) concentration, matched
55 up in space and time with depth-resolved physical properties derived from temperature-salinity profiles measured by BGC-Argo profiling floats. This method predicts b_{bp} for 10 different depths in the productive layer. In 2020, the availability of a larger database with new profiles -the under-sampling of many ocean regions in the SOCA2016 approach- and the opportunity to increase the vertical resolution of model outputs, enabling improved characterization and quantification of export carbon

---

[1]https://sentinel.esa.int/web/sentinel/user-guides/sentinel-3-olci/product-types/level-2-water

fluxes, led to the development of the SOCA2020 method. This approach includes additional Sea Level Anomaly (SLA) inputs
60 with information about sub-mesoscale processes; it replaces satellite-derived products (b_{bp} and chl-a) by simple reflectances
at several wavelengths and explores machine learning-based techniques that are efficient at estimating retrievals, in addition
to quantifying the uncertainty associated with the outputs. A significant improvement in the b_{bp} predictions was revealed,
especially near the surface layers.

Building on these results, this research proposes a more detailed analysis of estimating b_{bp} in the upper layers of the ocean
65 surface using Sentinel-3 Ocean and Land Colour Instrument (S3OLCI) data. We enhance spatial resolution from the 4 km
resolution of GlobColour level-3 merged products (1/24° at the equator) to the 300 m Full Resolution (FR) of Sentinel-3 OLCI.
Additionally, we evaluate model performance after incorporating OLCI spectral wavelengths as features for b_{bp} estimation and
compare these results with those obtained using GlobColour. Another key aspect of this study is determining whether adding
IOPs derived from satellite data improves the accuracy of b_{bp} estimation compared to using reflectances alone. These IOPs,
70 available from the Sentinel-3 OLCI processor, could significantly enhance regression models. The comparison is conducted
between BGC-Argo data and the various satellite datasets for two depth layers: from the surface to either 50 m or 250 m.

2 Data and methods

Data from *in situ* measurements collected by BGC-Argo floats, along with satellite data from various projects and missions
(GlobColour and Sentinel-3 OLCI) are utilized as inputs for the machine learning models. We employ three datasets for two
75 different maximum depths—50 m and 250 m: 1) Level-3 multi-sensor products from GlobColour; 2) Level-2 single-sensor
reflectances from Sentinel-3 OLCI processed with the Case 2 Regional Coast Colour (C2RCC) algorithm; and 3) The second
dataset (2) plus derived IOPs from OLCI using again the C2RCC processor.

2.1 Study Area

Two regions of the ocean are analyzed, the North Atlantic (NA), within latitudes 35º- 80ºN, and the Subtropical Gyres (STG),
80 within latitudes 15º- 40º North and South (see Figure 1). These two areas have very differentiated trophic states throughout
the year, experiencing great differences in terms of nutrients, light availability, minimum and maximum temperature regimes,
mixed layer depth (MLD) variations, thermocline levels and mesoscale dynamics. One of the main differences between these
two regions is the variability in the stratification of the upper ocean layers. This phenomenon determines the resistance of the
water to overturning, thus conditioning the supply of nutrients from deeper waters (Lozier et al., 2011). NA waters are season-
85 ally high in chl-*a* (mg.m^{-3}), which is a proxy of phytoplankton biomass. During winter, a weakly stratified upper ocean water
column overturns or mixes, facilitating the upwelling of nutrients needed to sustain surface productivity. In the STG region
spanning thousands of kilometers across the oceans, nutrients are in short supply, and waters range from ultra-oligotrophic (chl-
a ≤ 0.04 mg.m^{-3}) to oligotrophic (chl-a ≤ 0.07 mg.m^{-3}) (Letelier et al., 2004). During the summer and winter cycles, there is
expansion and contraction of their spatial coverage, respectively (Leonelli et al., 2022). Feucher et al. (2019) showed that both
90 Northern Hemisphere subtropical gyres have a qualitatively very similar stratification structure, with permanent pycnoclines in
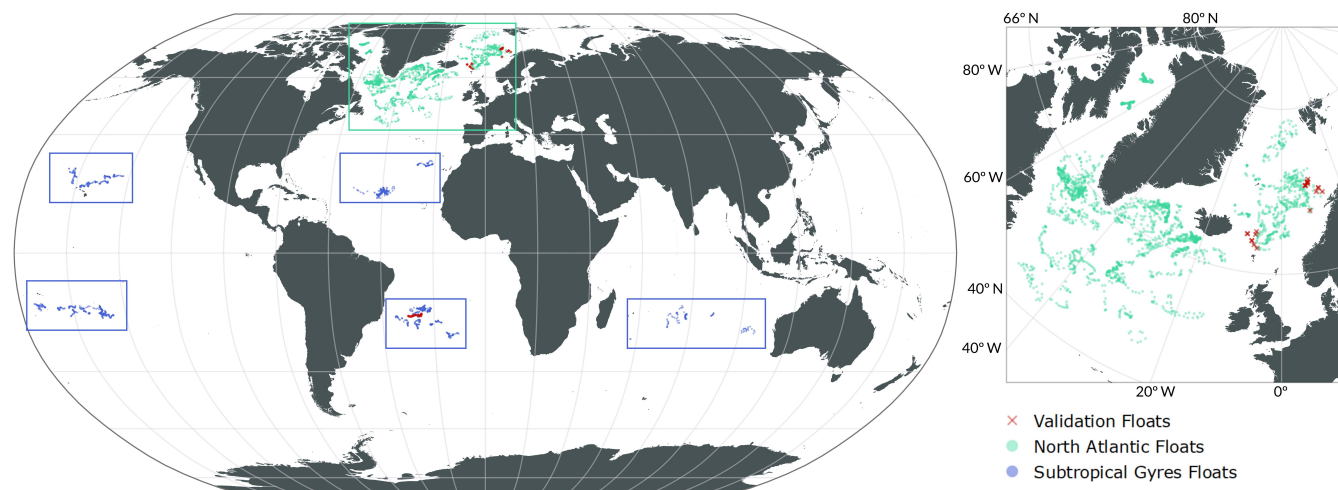
Figure 1. Spatial distribution of BGC-Argo and satellite data matchups.

the North Atlantic and North Pacific. The subtropical gyres of the South Atlantic and South Pacific Oceans are characterized by two different modes in the density and stratification space. Vertically homogeneous waters prevent deep-ocean nutrients from upwelling to the euphotic zone and control the biological pump, which plays an important role in carbon dioxide uptake. Despite the global coverage of the STG sampled regions, there is much more heterogeneity in the NA observations, which are
95 more complex and challenging for the models as can be observed in the results.

2.2 BGC-Argo Data

The international One-Argo program provides continuous ocean observations through an array of profiling floats, each equipped with sensors tailored to specific objectives: Core-Argo (for temperature and salinity measurements); BGC-Argo (for biogeo-chemical measurements); Deep-Argo (for measurements deeper than 2,000 m); and Polar-Argo (for measurements in polar
100 environments). Key bio-optical variables, such as chlorophyll-a, optical particulate backscattering, and irradiance, can be measured using BGC-Argo profiling floats. These variables are essential for generating products that support biogeochemical and ecosystem studies (Claustre et al., 2009, 2020). The BGC-Argo floats usually collect measurements from 1,000 m to the surface, with a depth resolution of ∼1 meter, every 10 days.

The lower boundary of the euphotic zone is defined as the depth where 1% of the Photosynthetically Available Radiation
105 (PAR) penetrates the water column. It varies in the global Ocean from ∼20 m to more than 120 m, depending on the region and season. The flux of sinking carbon that exits the euphotic zone due to gravity is a key component of the overall carbon sequestration budget (Siegel et al., 2014). In the experiments, a depth limit extending beyond the lower boundary of the euphotic zone (250 m depth) was selected. From 250 meters to the surface, measurements of temperature, salinity, density, and spiciness were taken every 2 meters from the floats, along with information about the MLD -calculated as the depth at which density

110 exceeds 0.03 kg m^{-3} relative to the density at 10 m (de Boyer Montégut et al., 2004). Vertical measurements of b_{bp} at the same vertical resolution are also available in the datasets. For the designed experiments, Table 1 shows the different types of variables used to train and validate the proposed models.

The b_{bp} value (Mignot et al., 2014) used here is calculated following the work of Sullivan and Twardowski (2009). The angular distribution of scattering relative to the direction of light propagation θ at the optical wavelength λ is known as the
115 volume scattering function (VSF), $\beta(\theta, \lambda)$ (m^{-1}, sr^{-1}). It is composed of the sum of pure sea water β_{sw} and particles β_p, where β_{sw} depends on temperature and salinity and computed using a depolarization ratio of 0.039 (Zhang et al., 2009). The contribution of β_p to the VSF is calculated subtracting the contribution of β_{sw} from $\beta(124º,\lambda)$:

$$\beta_p(124°, \lambda) = \beta(124°, \lambda) - \beta_{sw}(124°, \lambda) \tag{1}$$

Then, a conversion factor χ with a value 1.076 for an angle of 124º relates b_{bp} to β_p, making it possible to extrapolate the
120 measurement from a single angle (124º) to the total coefficient as follows (Boss and Pegau, 2001; Sullivan and Twardowski, 2009):

$$b_{bp}(\lambda) = 2\pi\chi(\beta(\theta, \lambda) - \beta_{sw}(\theta, \lambda)) \tag{2}$$

The backscattering sensor of the BGC-Argo floats measures $\beta(124º,\lambda)$ with λ = 700nm. The quality control procedure carried out is the one followed in the SOCA2016 method.

125 **2.3 BGC-Argo and Satellite Match-up Databases**

The match-up database created for the SOCA2020 experiments, which links BGC-Argo floats with GlobColour and GlobalOcean data, was utilized in this study. The GlobColour data consists of normalized water-leaving reflectances (*rho_wn*) at 5 wavelengths (412, 443, 490, 555 and 670 nm), as well as the Photosynthetically Active Radiation (PAR) product. This *rho_wn* are derived from a combination of sensors that constitute the GlobColour product: SeaWiFS, MERIS, MODIS Aqua,
130 VIIRS NPP and OLCI (ACRI-ST, 2020). The GlobalOcean set provides Sea Level Anomaly (SLA) data, calculated relative to a 20-year mean of sea surface height, generated with altimeter data from various missions (HY-2A, Saral/Altika, Cryosat-2, Jason-2, Jason-1 T/P, ENVISAT, GFO and ERS1/2) (CMEMS, 2022). In the cited work, the match-up with BGC-Argo floats was performed using the values from the closest available pixels within a $\pm$ 5-day window and on a 5x5 pixel grid. Further details about the procedure can be found in Sauzède et al. (2020).

135 The BGC-Argo measurements used here were matched with Sentinel-3 OLCI data using the Calvalus tool developed by Brockmann Consult GmbH (Fomferra, 2011). The spatio-temporal approach applied consists of a time window between the BGC-Argo profiles and the satellite measurements of $\pm$ 24 hours, and the spatial satellite coverage around the profile is 3x3 macro pixel on full-resolution imagery (300 m pixel). Once the match-up between satellite and float is performed, a baseline quality control is applied to guarantee that the geophysical parameters are under stable conditions and show certain

Table 1. Summary of the variables used in the study

Data	Description	Variable	Quantity	Variables processed	Type of pre-processing
BGC-Argo	*In situ* sensors	Temperature	26 (51m)/126 (250m)	5	PCA
		Salinity	26 (51m)/126 (250m)	5	PCA
		Density	26 (51m)/126 (250m)	5	PCA
		Spiciness	26 (51m)/126 (250m)	5	PCA
		MLD	1	1	Standardisation
		Lat/Lon	2	2	-
		DOY	1	1	-
		b_{bp}	26 (51m)/126 (250m)	26/126	Stand.+log10
GlobColour	Level-3 product	rho_wn	5	5	Standardisation
		PAR	1	1	Standardisation
GlobalOcean		SLA	1	1	Standardisation
Sentinel-3 OLCI	C2RCC L2 reflectance	rho_wn	12	12	Standardisation
	C2RCC L2 WQ products	IOPs	8	8	Standardisation

homogeneity around the measurement. First, a flag-based filter is applied, discarding pixels near or under probable cloudy conditions. This is followed by an outlier removal based on z score ($z = (x - \mu)/\sigma$), applied at macro pixel level band by band. Then, a coefficient of variation in the 560 nm band ($cv = \sigma/\mu$) is applied (Bailey and Werdell, 2006). Coefficient values under 0.2 assure a good spatial homogeneity (Ahmed et al., 2013; Hlaing et al., 2013; Zibordi et al., 2009). Finally, the median of the pixels left by macro-pixel is used (Hu et al., 2001), which is a standard procedure in studies focused on oceanic waters (Barnes et al., 2019). These criteria reduced the data set from the original 4115 to 763 match-ups. Specifically, 411 and 352 data points are available for the NA and STG regions. We excluded data from two floats to be used exclusively for validation purposes: in the NA, the float with unique WMO (for World Meteorological Organization Number) 6902545 -with 22 measurements- and in the STG region the float WMO 3902125 -with 28 measurements- constitute the independent dataset in the validation process.

The Sentinel-3 OLCI bands selected extend from 400 nm to 753 nm (bands 1 to 12) of normalized water-leaving reflectances (*rho_wn*). The extraction is done on level 2 data atmospherically corrected with the Case-2 Regional CoastColour (C2RCC) Processor (Brockmann et al., 2016). C2RCC relies on an extensive database of simulated water-leaving reflectances and related top-of-atmosphere radiances, with neural networks trained to perform inversions both for the atmospheric correction and the in-water quality parameter estimation. C2RCC provides parameters like the absorption and scattering of the different constituents

155 (IOPs) at 443 nm, that is: absorption of chlorophyll pigments (*apig*), yellow substances (*agelb*) and detritus (*adet*); scattering of particulate matter (*bpart*) and white scatterers (*bwit*), as well as the additive *atot* and *btot*. It also provides total suspended matter concentration, chlorophyll-a concentration, and alternative AOPs like K_d (diffuse attenuation coefficient). Each parameter have their associated error estimation. From the 25 parameters calculated by C2RCC, we selected the eight IOPs mentioned, plus the reflectance for bands 400 to 753 nm.

160 **2.4 Data Pre-processing**

Table 1 shows the number of parameters (measured or derived) available for the experiments. After excluding the measurements for validation, the two areas have a total of 713 inputs. The maximum number of input variables is 46. The size of the matrices can be seen in Table 2. Due to the different nature of the input variables (X) used to train the models and the high dimension and covariance of the variables measured along the water column by BGC-Argo floats, the data was preprocessed to reduce possible redundancy. The high-dimensional, non-independent variables (temperature, salinity, density, and spiciness) were the ones with the most significant number of features. Each variable had one measurement every 2 meters, which means 126 measurements in the first 250 m, or 26 measurements in the 50 m depth profiles.

To reduce the high dimensionality and simplify the regression models, a Principal Component Analysis (PCA) is applied to some of the input features. After this feature reduction on the high-dimensional variables, the 250 m and 50 m measurements with 126 and 26 inputs are reduced to 5 components for each variable, resulting in a total of 20 features. This method still retains 99% of the information. In addition, satellite-derived variables and the MLD were normalized using zscore standardization, i.e., removing the mean (μ_x) and dividing by the standard deviation (σ_x) of each feature.

A second preprocessing step consisted of a logarithmic transformation of the b_{bp} values coming from the floats. This step reduces the spread of the values that result from the significant differences between the surface and deeper depths, thus obtaining a Gaussian-like distribution that will help the model performance. Finally, variables that consider the spatio-temporal domain, like latitude, longitude and date (day of year) are also included.

Table 2. Matrix sizes for the different datasets depths. Dimensions specified as: $samples \times features \times outputs$

Depth	Region	GCGO+BGC	S3OLCI+BGC	S3OLCI+IOPs	S3OLCI
50 m	North Atlantic	$389 \times 32 \times 26$	$389 \times 46 \times 26$	$389 \times 26 \times 26$	$389 \times 15 \times 26$
	Subtropical Gyres	$324 \times 32 \times 26$	$324 \times 46 \times 26$	$324 \times 26 \times 26$	$324 \times 15 \times 26$
250 m	North Atlantic	$389 \times 32 \times 126$	$389 \times 46 \times 126$	$389 \times 26 \times 126$	$389 \times 15 \times 126$
	Subtropical Gyres	$324 \times 32 \times 126$	$324 \times 46 \times 126$	$324 \times 26 \times 126$	$324 \times 15 \times 126$

2.5 Multi-output machine learning models

There are two main approaches for dealing with multi-output regression problems. One way is to use univariate models, also known as problem transformation methods (Schmid et al., 2022; Borchani et al., 2015). These methods decompose the multi-output regression problem into multiple single-target problems, creating an independent model for each output. The predictions from these separate models are then combined. This approach ignores the relationships between the targets, which can adversely affect the prediction's overall accuracy. Alternatively, multivariate models are designed to capture dependencies and interactions between the outputs, potentially leading to more accurate predictions (Borchani et al., 2015). When and how to apply these two approaches depends on the nature of the data and the correlation between the targets. In our preprocessing results, PCA decomposition indicates a high covariance among measurements at different depths in the water column. Since our regression models estimate b_{bp} at different depths, it is logical to consider that nearby values in the water column are related to each other.

Different algorithms have been tested in previous works (see Sauzède et al. (2016, 2020)) to estimate b_{bp} at various depths. Both works are based on a multivariate model applied to all possible outputs. In SOCA16, a Multi-Layer Perceptron is developed, while in SOCA2020 a comparison between a linear model (Ridge) and an ensemble model (Random Forest) is done. The latter showed higher performance. The Multivariate Random Forest used in this study offers higher accuracy than the univariate Random Forest, especially when the outputs are highly correlated (Schmid et al., 2022) and when complex interactions demand structured inference to be effectively managed (Xu et al., 2019). All the previously mentioned algorithms, along with others such as Linear Regressor (LR), Ridge Linear Regressor (RLR), Random Forest Regressor (RFR), and Multi-Layer Perceptron (MLP), were tested for estimating b_{bp} during the dataset preparation phase. Based on these results, the Random Forest Regressor (RFR) was selected as the most suitable algorithm for this multi-input/multi-output problem.

Random Forest Regressor (Breiman, 2001) has been widely applied in geosciences and marine environmental studies for classification and regression tasks (Cutler et al., 2007; Ruescas et al., 2018). Regression trees are at the model's core, which effectively handles complex data when there are non-linear dependencies between a numerical response variable and a diverse set of predictors, whether qualitative or quantitative (D'Ambrosio et al., 2017). RFR is an ensemble method that combines many weak decision tree learners, which are grown in parallel to reduce the bias and variance of the model simultaneously, enhancing the model's predictive performance. Furthermore, RFR provides insights into the importance of the training features, which reveals the variables that have the most significant impact on the predictions. This capability makes the model's mechanisms and results easier to interpret and explain.

3 Performance of the Random Forest Regressor

Several dataset combinations were used as inputs for the RFR. The naming conventions and the data used as features are shown in Table 1. The dataset names in the table correspond to the specific features included: GCGO refers to the GlobColour-GlobOcean L3 satellite reflectance, combined with the PAR and SLA products (7 features); BGC denotes the Argo-BGC data

after pre-processing (27 features across 26 or 126 layers, depending on the depth of 50 or 250 m, respectively); S3OLCI
includes 12 reflectance bands (plus IOPs in the case of S3OLCI+BGC); and IOPs represent the eight C2RCC-derived IOPs.

The RFR was trained on 80% of the data, with the remaining 20% set aside for testing. Experiments were conducted in the
NA and STG regions across two depth layers: 0-50 m and 0-250 m. The test dataset was exclusively used to evaluate model
performance and was never exposed to the regressor during training. For each regression model, we analyzed the key features
that contributed to improving the estimation of b_{bp} in the different combinations. The final experiment validated the pre-trained
models using two independent floats in the NA and STG regions.

3.1 S3OLCI+BGC: results with BGC-Argo and OLCI data

A subset of the dataset utilized in the SOCA2020 experiment (GCGO+BGC) was included in the statistical analysis to facilitate
a comparison between our findings and the previous approach. Tables 1 and 2 present the input features and matrix sizes for
the different experiments. In the following sections, we analyze the results of the RFR model applied to these datasets, starting
with the GCGO+BGC and S3OLCI+BGC datasets to establish a baseline. In the NA region, 311 data points were used for
training and 78 for testing, while in the STG region, 259 data points were used for training and 65 for testing.

3.1.1 Shallow waters: from 0 to 50 meters depth

In Figure 2 (A), R^2 value profiles represent 20% of the total NA (green lines) and STG (blue lines) dataset used for model
testing. The total R^2 value for the NA region using the S3OLCI+BGC is 0.751, which is slightly higher than the 0.721 obtained
with the GCGO+BGC (Table 3). Changes in precision depend on the layer's depth. The S3OLCI+BGC set performs better both
at superficial and deeper layers. A weaker performance is observed below 20 meters depth, being coincident with the location
of the MLD, where more variations in b_{bp} values can be found. This is due to the strong baroclinic component and the density
dependence on the high thermal gradient of these waters. Figure 2 (B) illustrates the corresponding error statistics, and Figure
2 (C) compares observed and predicted b_{bp} profiles for this region. The predicted profiles exhibit greater homogeneity and
less variability compared to the observations. Note that spikes along the water column are difficult to predict by the models.
Figure 3 (*left*) shows the feature importance for the RFR models in the NA region. The most relevant feature in Sentinel-3
OLCI is the reflectance at 560 nm wavelength (band 6), demonstrating that satellite data provide good quality information
for estimating b_{bp} at both sea surface and, at least, the first 50 m depth. As demonstrated in previous studies, the relationship
between the reflectance at 555 nm and POC is well known. Stramski et al. (1999) and Oubelkheir et al. (2005) noted that
backscattering and absorption coefficients in the 510 to 555 nm spectral region seem to be well correlated with a broad range of
particle concentrations and compositions. The Photosynthetically Active Radiation (PAR) feature shows relative importance,
as it describes the mean daily photon flux density that can be used for photosynthesis at the moment of the observation.
Other relevant features are the day-of-year (*DOY*) and the longitude (*lon*), which are related to the significant spatio-temporal
differences across the North Atlantic region and the differences in the physical and biogeochemical characteristics of water. The
feature *DOY*, which accounts for the temporal component, reflects the seasonality that affects the phytoplankton cycles and,
thus, the POC dynamics on these regions. The importance of longitude indicates that several water types must be within the
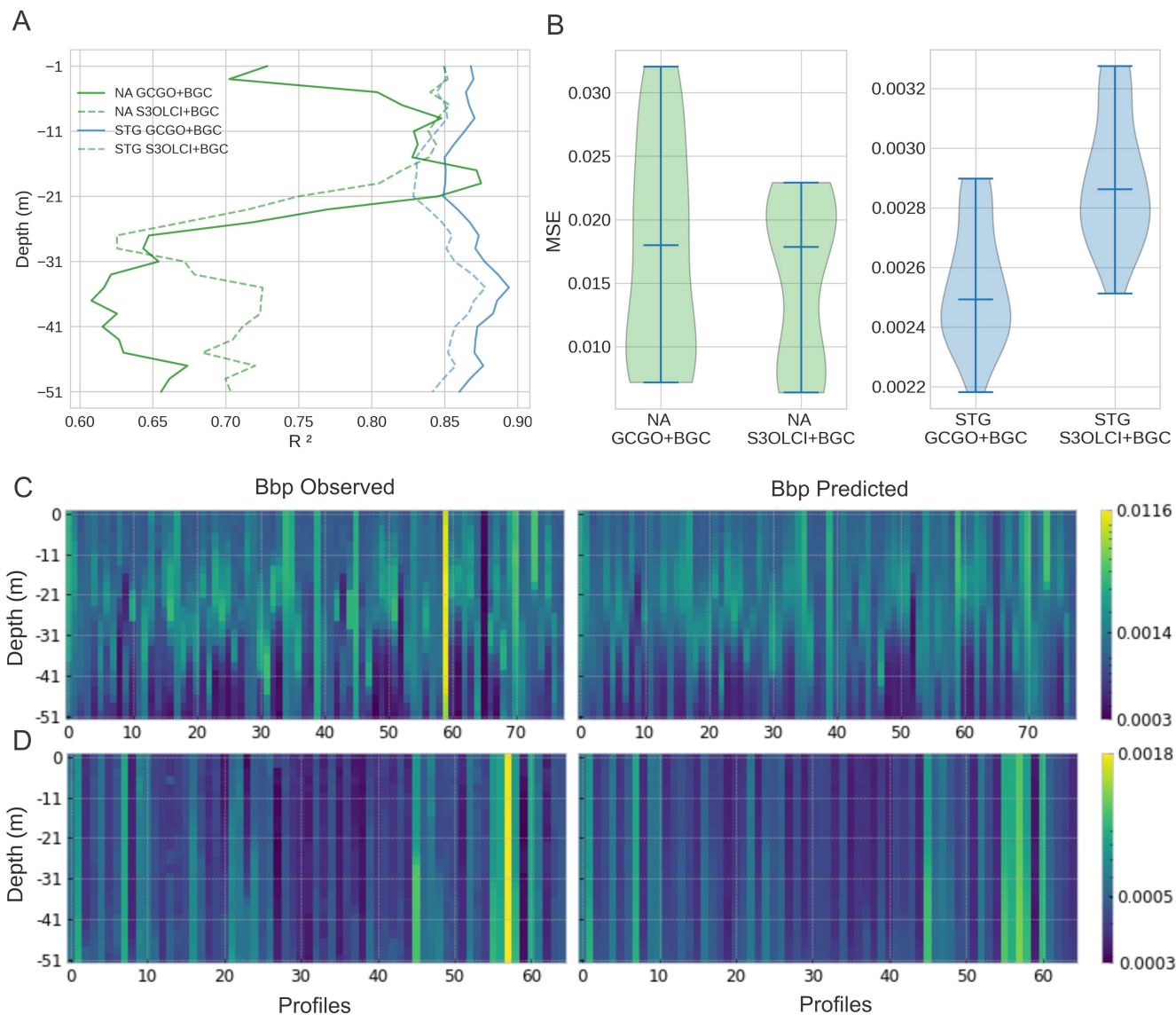
Figure 2. (A) R^2 obtained for the 50 m depth models with BGC-Argo and satellite data (GCGO+BGC and S3OLCI+BGC).(B) Violin plots with Mean Squared Error distribution by model. Observed and predicted b_{bp} (700 nm, m^{-1}) profiles with S3OLCI+BGC in the North Atlantic (C) and Subtropical Gyres (D).

same region. Another feature of importance is the first principal component of temperature measurements (*Temp_pc1*), related to the baroclinic dynamics and highlighting the importance of temperature in the POC cycle at high latitudes. The maximum phytoplankton blooms usually occur from June to August, when the water temperature is higher (Yang et al., 2020). Particular

245   attention is needed for the feature iop_bwit, introduced by the C2RCC processor. Due to the broad range of particles in water,

Table 3. Statistics by region at 50 m and 250 m depth models with satellite and BGC-Argo data.

Depth	Region		GCGO+BGC	S3OLCI+BGC	S3OLCI+IOPs	S3OLCI
50 m	North Atlantic	R^2	0.721	0.751	0.732	0.753
		MAE	0.075	0.074	0.081	0.076
		MSE	0.017	0.015	0.016	0.015
	Subtropical Gyres	R^2	0.868	0.849	0.881	0.848
		MAE	0.035	0.038	0.034	0.035
		MSE	0.0025	0.0028	0.0023	0.0029
250 m	North Atlantic	R^2	0.844	0.813	0.794	0.805
		MAE	0.039	0.046	0.052	0.049
		MSE	0.0047	0.0055	0.0064	0.0061
	Subtropical Gyres	R^2	0.904	0.893	0.881	0.884
		MAE	0.030	0.031	0.035	0.034
		MSE	0.0019	0.0021	0.0023	0.0023

particle scattering in this processor is parametrized by two components. One component (iop_bwit) represents white particles (coccolithophorids, foam and air bubbles), while small particles (iop_bpart) represent the other. This distinction enables the model to better characterize the scattering source. For instance, it can identify whether the measured parameter originates from a source that produces b_{bp}, like phytoplankton, or if it is produced by white caps or foam, which are very common in NA
250 waters. This characterization is responsible for the better accuracy in the most superficial layers, and it's also noticeable in the validation results (see 3.3 Section), where the results with this data are better.

In the STG region, the model with GCGO+BGC performs slightly better (total R^2 is 0.868) than the S3OLCI+BGC model (see Table 3 and Figure 2 (A)). Accuracy at different depths does not show significant changes, remaining constant and even improving as depth increases. There are no gradients along the water column, but b_{bp} values are much lower than in the NA
255 case, which indicates a low presence of organic matter due to the scarcity of nutrients and high incidence of sunlight on this region. This homogeneity facilitates the model estimations and the results are significantly better than in the NA region. In Figure 3 (*right*), we see the feature importance values for the STG RFR models. Latitude is the most relevant feature in this case. This suggests that conditions in the STG are very similar throughout the year, the DOY is less important because of the low seasonality in these areas (Mignot et al., 2014; Cornec et al., 2021). However, values vary according to their position relative to
260 the vortex around which the gyres revolve. The central areas of the subtropical gyres are under almost permanent atmospheric and geostrophic high pressure, making it difficult for different water masses to mix. A high stratification is expected to inhibit productivity. However, at the edges of the gyres, close to the eastern, western, sub-polar and equatorial currents, there is usually a greater influx of nutrient-laden due to the divergence of waters, upwelling, and therefore more favorable ecological conditions for phytoplankton. In these areas, seasonality is also more noticeable, and winter mixing due to a weaker stratification is clearly
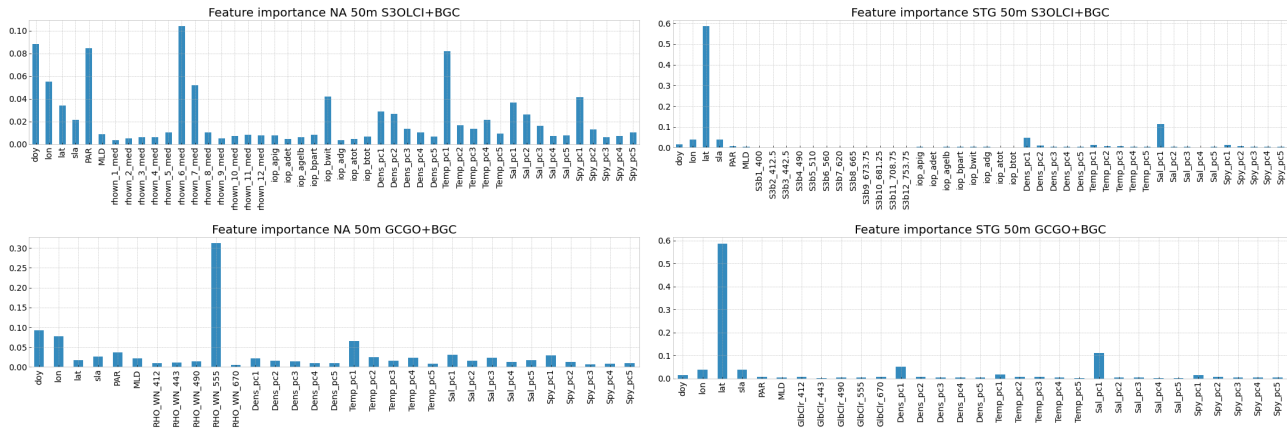
Figure 3. Feature importance for the models with GCGO+BGC and S3OLCI+BGC Argo data for 50 m depth in the North Atlantic (*left*) and in Subtropical Gyres (*right*).

265 associated with phytoplankton blooms and primary productivity (Lozier et al., 2011). However, this stratification variability may not be enough to lead to perceptible changes in large parts of the subtropical ocean. The importance of the density and salinity features (*Dens_pc1* and *Sal_pc1*) reflects the barotropic dynamics of these oceanic regions, where isobars and isopycnals are stratified parallel to the ocean surface and vary together as depth is gained (Leonelli et al., 2022).

3.1.2 Deep waters: from 0 to 250 meters depth

270 Table 3 and Figure 4 (A) and (B) present the total and in-depth layer results for models trained with data up to 250 meters depth. Compared to the model trained on data up to 50 meters depth, the enhanced performance in the NA could be due to the inherent complexity of the region's shallow waters. Despite the valuable insights from *in situ* data, the upper 50 meters of the NA are highly dynamic. However, as depth increases, the relevance of BGC-Argo measurements becomes more significant. A decrease in accuracy is observed around 30 meters, coinciding with the MLD (Figure 4 (A)). However, as depth increases,

275 an upward trend in R^2 values is noticeable until approximately 150 meters. In the observed-predicted profiles (Figure 4 (C)), high b_{bp} values are occasionally observed at depths exceeding 100 meters. When predicting the NA results at 250 m, the model relies on different features as compared to the 50 m depth model. The most relevant features are the PCA 1 of the density, the DOY and the longitude (*lon*). These differences can be attributed to the variety of physical processes in this region, such as the mixing of distinct water masses through ocean eddies or ventilation processes, which affect the physicochemical properties

280 of the water layers and, consequently, the b_{bp} dynamics. The complexity and diversity of the region's dynamics are further emphasized by the importance of variables like sea level anomaly (*Sla*) and the first PCA of spiciness (*"Spy_pc1"*), which relate to vertical changes in the water column. Additionally, the variable *"rhown6_nm"* (at a 560 nm wavelength) remains a key contributor, highlighting the continued relevance of satellite data in estimating b_{bp} in the euphotic zone, even at depths as great as 250 meters.
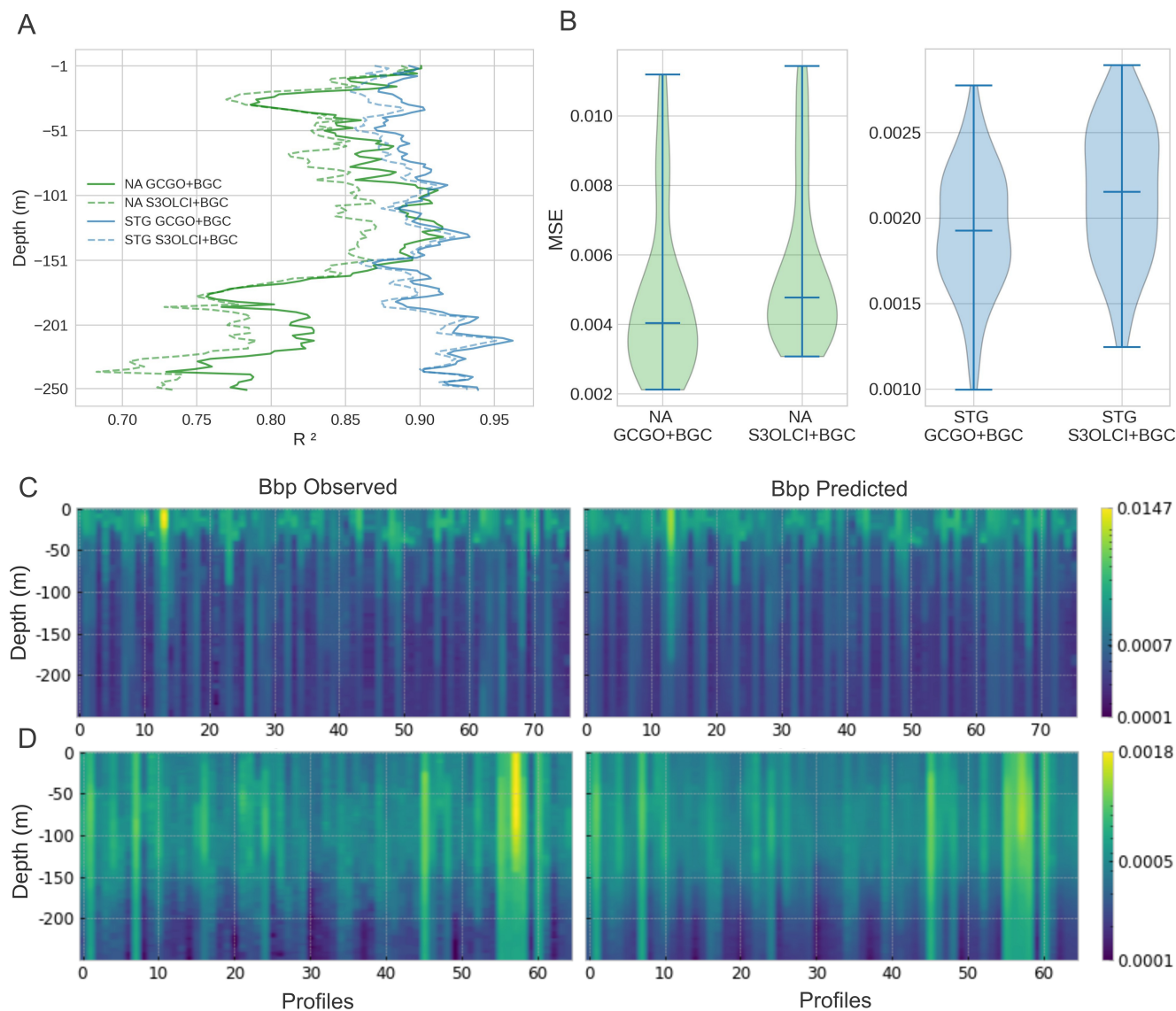
Figure 4. (A) R^2 obtained for the 250 m depth models with BGC-Argo and satellite data (GCGO+BGC and S3OLCI+BGC).(B) Violin plots with Mean Squared Error distribution by model. (C) Observed and predicted b_{bp} (700 nm, m^{-1}) profiles with GCGO+BGC in the North Atlantic and (D) Subtropical Gyres.

285   In the STG region, the results are better than in the NA region, likely due to the more homogeneous water conditions. Although the MLD is typically around 50 m depth, in this case, no significant increase in error is observed until the 120-meter layer. This zone corresponds to the region of effective biomass accumulation, known as the *deep biomass maxima*. It marks the interface between two distinct water masses: the nutrient-limited upper layer and the light-limited lower layer (Cornec et al.,

2021). This interface is typically found in oligotrophic waters at depths between 150 and 200 meters (Mignot et al., 2014).
290   Despite the lower productivity of these waters, their vast expanse makes them a significant contributor at the planetary scale
(Jenkins and Doney, 2003). As noted in the results for the 50-meter model, latitude remains the most important feature for the
Random Forest Regression (RFR) model.

3.2   S3OLCI: results of Sentinel-3 OLCI without BGC-ARGO data

As demonstrated in the previous experiments, satellite-derived features play a significant role in the models when profile depths
295   reach 50 meters, thus answering the initial hypothesis of this study. It is clear that sea surface signals aid in estimating b_{bp} at
subsurface levels. However, the extent of this contribution across different depth layers became evident only when comparing
models trained with different depth limits. The feature importance of the 50 m depth models shows that, at least in the NA
region, the importance of the features measured by satellite sensors are just as relevant as the inputs from the floats. For this
reason, we designed one last experiment with only satellite data (S3OLCI) to check how the models perform in-depth with the
300   normalized water-leaving reflectance bands. Additionally, we investigated the contribution of the satellite-derived IOPs from
the C2RCC processor, that is, adding the absorption and scattering variables as input features (S3OLCI+IOPs).

In the NA region, the model using only reflectance data (S3OLCI) outperforms the model that includes both reflectance and
the absorption and scattering products derived from the C2RCC processor (S3OLCI+IOPs) (see Table 3 and Figure 5 (A, B)).
While the MLD is still a barrier, accuracy improves beyond this depth for approximately another 10 meters. In the b_{bp} profiles
305   (Figure 5 (C)), despite the noted errors in deeper estimations, the model is capable of predicting significant contrast events
using only surface data from 36 meters onward, except an extreme case (profile 59). In the feature importance ranking, the
620 nm band is the most relevant of the spectrum. However, the spatio-temporal features (day of year, longitude and latitude)
seem to have greater weight than the results obtained with the data sets that include BGC-Argo data at the same depth (see
Section 3.2.1).

310   In the STG region, the S3OLCI+IOPs model achieves better results (Table 3). However, it is possible to see how the model
is not able to predict some spikes along the water column (Figure5 (D)). In the feature importance ranking (not shown here),
latitude remains the most relevant feature. The improved performance of the S3OLCI+IOPs model, compared to the S3OLCI
(reflectance-only) model, could be attributed to the contribution of marine particle scattering at 443 nm (*iop_bpart*) provided
by the C2RCC processor.

3.3   Validation with independent floats

The previously trained RFR models were applied to predict b_{bp} values using independent float data that was not included in the
training or testing sets. Statistical metrics and corresponding scatter plots are provided in Table 4 and Figure 6.

In the NA region, the float identified as WMO 6902545 (see location in Figure 1) yields better estimates with the S3OLCI
models (R^2 ranging from 0.475 to 0.534) compared to the reference GCGO+BGC model, where the R^2 value drops to 0.266.
320   Figure 6(A) reveals the higher b_{bp} variability along the water depth in the NA region, as indicated by the colour scale. There is
an overestimation in the surface measurements (less than 30 m) and a slight underestimation at deeper depths. This validation
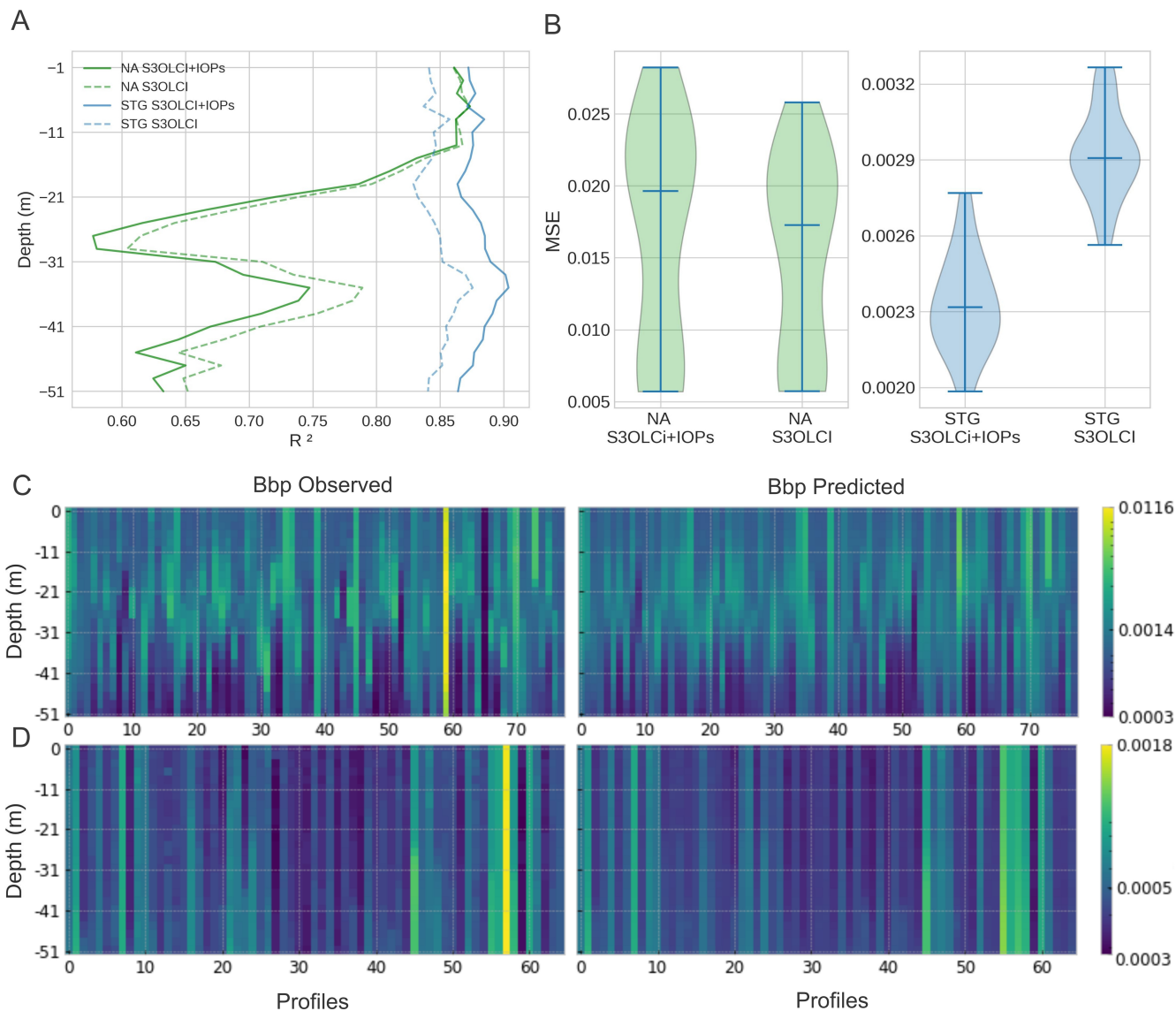
Figure 5. (A) R^2 obtained for the 50 m depth models using only satellite data.(B) Violin plots with Mean Squared Error distribution by model. (C) Observed and predicted b_{bp} (700 nm, m^{-1}) profiles with S3OLCI data in the North Atlantic and (D) with S3OLCI+IOPs in the Subtropical Gyres.

set includes data from several dates in 2017 and 2018, spanning from April to August. These temporal variations explain some of the observed drifts in the plots, where different float cycles (water depth profiles) are also evident.

The STG region's statistics and plots for the float with identifier 3902125 show better correlation coefficients and lower errors. The datasets incorporating S3OLCI data yielded better results. In Figure 6(B), two clusters of data are visible: one
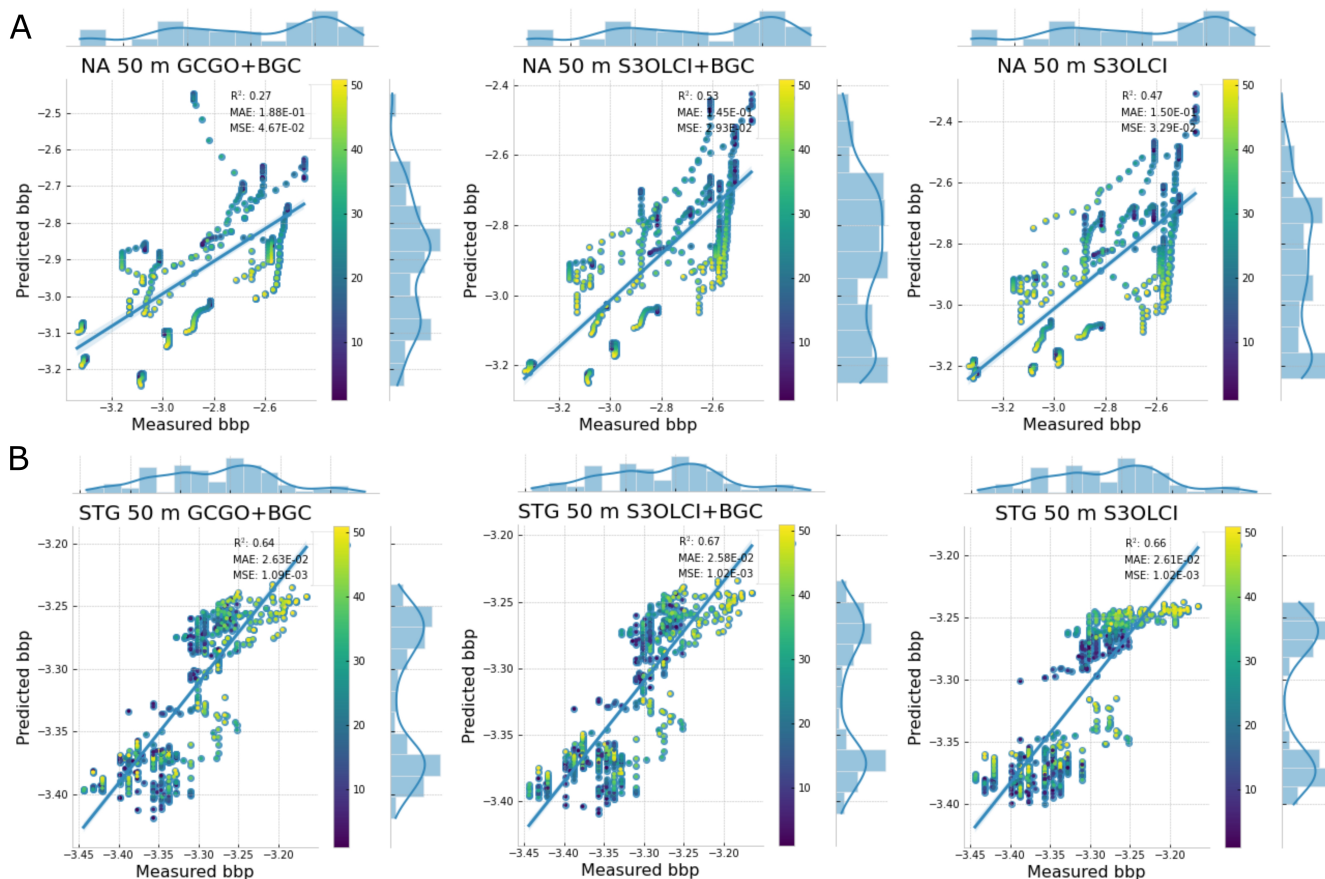
Figure 6. (A) Scatter plots with marginal histograms with the validation of the 50 m model performance on an independent float on the NA region (ID 6902545) and (B) on the STG region (ID 3902125). The color scales shows the depth of the measurements and b_{bp} values are in log10.

associated with low b_{bp} values and the other clustering around slightly higher values. The models tend to underestimate the lower b_{bp} values, while the higher values show a closer fit to the 1:1 line. However, in the model that uses only reflectance data (S3OLCI), a clear overestimation of higher values occurs. Unlike in the North Atlantic region, depth separation is not evident here, but the lower values correspond to measurements taken during the winter months in the South Atlantic Gyre, while the higher values were recorded during the summer months in the Southern Hemisphere, where the float was located.

These results reinforce the observations made in previous sections: models provide more accurate b_{bp} estimations in the STG region than in the NA, confirming the effectiveness of using the S3OLCI bands and derived C2RCC IOPs at shallow water depths.

Table 4. Validation with independent floats by region at 50 m and 250 m depth models with satellite and BGC-Argo data.

Depth	Region		GCGO+BGC	S3OLCI+BGC	S3OLCI+IOPs	S3OLCI
50 m	North Atlantic	R^2	0.266	0.534	0.525	0.475
		MAE	0.188	0.145	0.143	0.150
		MSE	0.0467	0.0293	0.0298	0.0329
	Subtropical Gyres	R^2	0.641	0.670	0.662	0.660
		MAE	0.026	0.025	0.026	0.026
		MSE	0.0054	0.0057	0.0054	0.0055
250 m	North Atlantic	R^2	0.331	0.315	0.290	0.276
		MAE	0.046	0.047	0.045	0.046
		MSE	0.0013	0.0010	0.0010	0.0010
	Subtropical Gyres	R^2	0.552	0.564	0.545	0.596
		MAE	0.041	0.040	0.042	0.039
		MSE	0.0028	0.0027	0.0026	0.0026

4 Discussion and Conclusion

335 Vertical estimates of the particulate backscattering coefficient (b_{bp}), used as a proxy for particulate organic carbon, were calculated along the water column. Estimations were performed in two distinct regions and across several layers within two maximum depth limits. Data from BGC-Argo profiles and ocean color satellite data (OLCI) were used. The Random Forest model was applied with different sets of variables for the b_{bp} estimations, following methodologies from previous works by Sauzède et al. (2016, 2020).

340 Previous studies estimating b_{bp} from satellite-derived remote sensing reflectance (R_{rs}) have typically employed empirical or semi-analytical models, with most focused on surface layers. In Bisson et al. (2019), b_{bp} profiles from floats were processed by averaging b_{bp} values within the surface mixed layer, followed by a comparison between different sensors and b_{bp} retrieval inversion products by NASA. In that case, OLCI -with data from Reduced Resolution mode at 1.2 km pixel resolution- under-performed compared to MODIS (Moderate Resolution Imaging Spectroradiometer) with 1 km at nadir (r= 0.32 to 0.47 and r =

345 0.60 to 0.79 respectively). This difference was attributed to higher coefficients of variation (30% for OLCI and 5% for MODIS) across bands between 412 and 555 nm and aerosol optical thickness at 865 nm. In the present work, OLCI Full Resolution (FR) data, with a spatial resolution of 300 m, was used. Additionally, the most relevant wavelength in some of our models (620 nm) was not considered in Bisson et al. (2019).

Additional considerations regarding our work include the time criteria for match-ups, the higher spatial resolution, and the

350 macro pixel sizes (e.g., 3×3, 5×5), which differ significantly from those in other published studies. Sauzède et al. (2020) used profiles that reach up to 1000 m depth, where the contribution of surface satellite data was overshadowed by the valuable

information from BGC-Argo profiles. In this study, the focus was on the possible contribution of satellite-derived water-leaving reflectance to b_{bp} estimation within the first 250 meters, situated in the twilight zone. Satellite features have proved to be indeed relevant for the b_{bp} estimations, especially in the Subtropical Gyres region. These oligotrophic waters, characterized by high

355 stratification, rely heavily on nutrient injection from deeper zones, as the upper euphotic zone is typically nutrient-limited. In fact, Letelier et al. (2004) and Mignot et al. (2014) describe these gyres as a two-layer system: an upper layer nutrient limited but not light-limited, and a deeper layer that is light-limited but has greater nutrient access. These authors also highlight a seasonal distinction, with winter bringing greater water mixing than summer. During winter, average light intensity for PAR in the mixed layer decreases while turbulence increases. This seasonal variation may explain the two distinct clusters observed

360 in the validation exercise for the STG region. The inclusion of satellite surface data, along with derived parameters such as inherent optical properties (IOPs), in combination with *in situ* profile data, should be considered for estimating b_{bp}, and by extension, approximating particulate organic carbon (POC), at least for layers up to 250 meters depth. It is important to note that organic carbon fixation primarily occurs in the upper ocean layers. This organic matter is subsequently transformed through respiration, particle aggregation, zooplankton grazing, feces production, and microbial decomposition (Siegel et al.,

365 2014), before eventually sinking to deeper layers.

Concerning the results, the models that relied exclusively on satellite data (S3OLCI and S3OLCI+IOPs) produced reasonable estimations for the upper layers in both the North Atlantic and Subtropical Gyres regions. This is encouraging, as satellite data, with its synoptic spatial coverage and broad temporal scope, can efficiently complement Argo float measurements. Satellite observations provide valuable insights into mesoscale ocean processes over various temporal ranges, extending at least the

370 past three decades. However, remote sensing products are limited—only about 20% of the euphotic zone is directly observable by satellite sensors. This highlights the critical importance of extending surface observations to deeper layers (Claustre et al., 2010).

Future work should be focused on enlarging the database with new BGC-Argo profiles and satellite data; and extending the data to new areas of the global ocean. The role of the MLD on the different regions is also an issue that deserves attention

375 in order to further understand the effect that it has on biochemical parameter estimations. Planned sensors with extended capabilities, like the hyperspectral NASA PACE, might be also a path of research to follow, since we have seen that adding new wavelengths had a positive effect on the results of our models compared with previous works. Possible improvements in the detection of CDOM with the UV bands can be an important contribution to better estimating particulate organic material (POM) and, consequently, to POC.

380 *Data availability.* Both BGC-Argo measurements and OLCI data are open and freely available for the scientific and public community.

Author contributions. JGJ, JAL and ABR designed the experiments and prepared the match-up data set based on the BGC-Argo data pre-processed by RS. JGJ processed the data and run the models. ABR made the independent validation. All four author contributed with the revision and writing of the paper.

Competing interests. No competing interests are present

385 *Acknowledgements.* AI4CS - GVA PROMETEO Projecte "Artificial Intelligence for complex systems: Brain, Earth, Climate, Society", funded by Conselleria de Innovación, Universidades, Ciencia y Sociedad Digital, CIPROM/2021/56

References

ACRI-ST: GlobColour Product User Guide, Product user guide, https://www.globcolour.info/CDR_Docs/GlobCOLOUR_PUG.pdf, 2020.

Ahmed, S., Gilerson, A., Hlaing, S., II, A. W., Arnone, R., and Wang, M.: Evaluation of ocean color data processing schemes for VIIRS sensor using in-situ data of coastal AERONET-OC sites, in: Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2013, edited by Jr., C. R. B., Mertikas, S. P., Neyt, X., and Bruyant, J.-P., International Society for Optics and Photonics, SPIE, https://doi.org/10.1117/12.2028821, 2013.

Bailey, S. W. and Werdell, P. J.: A multi-sensor approach for the on-orbit validation of ocean color satellite data products, Remote Sensing of Environment, 102, 12–23, https://doi.org/10.1016/j.rse.2006.01.015, 2006.

Barnes, B. B., Cannizzaro, J. P., English, D. C., and Hu, C.: Validation of VIIRS and MODIS reflectance data in coastal and oceanic waters: An assessment of methods, Remote Sensing of Environment, 220, 110–123, https://doi.org/10.1016/j.rse.2018.10.034, 2019.

Behrenfeld, M. J., O'Malley, R. T., D.A.Siegel, McClain, C., Sarmiento, J., Feldman, G., P.G., A. M., Falkowski, Letelier, R., and Boss, E.: Climate-driven trends in contemporary ocean pro- ductivity, Nature, 44, 752–755, https://doi.org/10.1038/nature05317, 2006.

BGC, A. P.: The scientific rationale, design and implementation plan for a Biogeochemical-Argo float array, Report, https://doi.org/10.13155/46601, 2016.

Bisson, K. M., Boss, E., Westberry, T. K., and Behrenfeld, M. J.: Evaluating satellite estimates of particulate backscatter in the global open ocean using autonomous profiling floats, Optics Express, 27, 30 191–30 203, https://doi.org/10.1364/OE.27.030191, 2019.

Borchani, H., Varando, G., Bielza, C., and Larrañaga, P.: A survey on multi-output regression, WIREs Data Mining and Knowledge Discovery, 5, 216–233, https://doi.org/10.1002/widm.1157, 2015.

Boss, E. and Pegau, W. S.: Relationship of light scattering at an angle in the backward direction to the backscattering coefficient, Applied Optics, 40, 5503–5507, https://doi.org/10.1364/AO.40.005503, 2001.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, 2001.

Brockmann, C., Doerffer, R., Peters, M., Stelzer, K., Embacher, S., Ruescas, and Ana: 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, Living Planet Symposium, ESA, 2016.

Cetinić, I., Perry, M. J., Briggs, N. T., Kallin, E., D'Asaro, E. A., and Lee, C. M.: Particulate organic carbon and inherent optical properties during 2008 North Atlantic Bloom Experiment, Journal of Geophysical Research: Oceans, 117, https://doi.org/10.1029/2011JC007771, 2012.

Claustre, H., Bishop, J., Boss, E., Bernard, S., Berthon, J. F., Coatanoan, C., Johnson, K., Lotiker, A., Ulloa, O., Perry, M. J., D'Ortenzio, F., D'andon, O. H., and Uitz, J.: Bio-optical profiling floats as new observational tools for biogeochemical and ecosystem studies: Potential synergies with ocean color remote sensing, ESA Publication., https://doi.org/10.5270/OceanObs09.cwp.17, 2009.

Claustre, H., Bishop, J., Boss, E., Bernard, S., Berthon, J., Coatanoan, C., Johnson, K., Lotiker, A., Ulloa, O., M.J., M. P., D'ortenzio, F., D'andon, O. H. F., and Uitz, J.: Bio-Optical Profiling Floats as New Observational Tools for Biogeochemical and Ecosystem Studies: Potential Synergies with Ocean Color Remote Sensing, Tech. rep., https://doi.org/https://doi.org/0.5270/OceanObs09.cwp.17, 2010.

Claustre, H., Johnson, K. S., and Takeshita, Y.: Observing the Global Ocean with Biogeochemical-Argo, Annual Review of Marine Science, 12, 23–48, https://doi.org/10.1146/annurev-marine-010419-010956, 2020.

CMEMS: Product User Manual For Sea Level Altimeter Products, Product user guide, https://catalogue.marine.copernicus.eu/documents/PUM/CMEMS-SL-PUM-008-032-068.pdf, 2022.

Cornec, M., Claustre, H., Mignot, A., Guidi, L., Lacour, L., Poteau, A., D'Ortenzio, F., Gentili, B., and Schmechtig, C.: Deep Chloro-
phyll Maxima in the Global Ocean: Occurrences, Drivers and Characteristics, Global Biogeochemical Cycles, 35, e2020GB006 759,
https://doi.org/10.1029/2020GB006759, 2021.

Cutler, D. R., Edwards Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J.: Random Forest for classification in
Ecology, Ecology, 88, 2783–2792, https://doi.org/10.1890/07-0539.1, 2007.

de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., and Iudicone, D.: Mixed layer depth over the global ocean: An examination of
profile data and a profile-based climatology, Journal of Geophysical Research: Oceans, 109, https://doi.org/10.1029/2004JC002378, 2004.

D'Ambrosio, A., Aria, M., Iorio, C., and Siciliano, R.: Regression trees for multivalued numerical response variables, Expert Systems with
Applications, 69, 21–28, https://doi.org/10.1016/j.eswa.2016.10.021, 2017.

EUMETSAT: Sentinel-3 OLCI Inherent Optical Properties, Algorithm theoretical basis documents, https://www.eumetsat.int/media/44307,
2019.

Evers-King, H., Martinez-Vicente, V., Brewin, R. J. W., Dall'Olmo, G., Hickman, A. E., Jackson, T., Kostadinov, T. S., Krase-
mann, H., Loisel, H., Röttgers, R., Roy, S., Stramski, D., Thomalla, S., Platt, T., and Sathyendranath, S.: Validation and Inter-
comparison of Ocean Color Algorithms for Estimating Particulate Organic Carbon in the Oceans, Frontiers in Marine Science,, 4,
https://doi.org/10.3389/fmars.2017.00251, 2017.

Falkowski, P., Barber, R., and Smetacek, V.: Biogeochemical Controls and Feedbacks on Ocean Primary Production, Science, 281, 200–7,
1998.

Feucher, C., Maze, G., and Mercier, H.: Subtropical Mode Water and Permanent Pycnocline Properties in the World Ocean, Journal of
Geophysical Research: Oceans, 124, 1139–1154, https://doi.org/10.1029/2018JC014526, 2019.

Fomferra, N.: Cal/Val and User Service - Calvalus, Final report, https://doi.org/http://www.brockmann-
consult.de/calvalus/pub/docs/Calvalus-Final_Report-Public-1.0-20111031.pdf, 2011.

Hlaing, S., Harmel, T., Gilerson, A., Foster, R., Weidemann, A., Arnone, R., Wang, M., and Ahmed, S.: Evaluation of the VIIRS ocean color
monitoring performance in coastal regions, Remote Sensing of Environment, 139, 398–414, https://doi.org/10.1016/j.rse.2013.08.013,
2013.

Hu, C., Carder, K. L., and Muller-Karger, F. E.: How precise are SeaWiFS ocean color estimates? Implications of digitization-noise errors,
Remote Sensing of Environment, 76, 239–249, https://doi.org/10.1016/S0034-4257(00)00206-6, 2001.

Jenkins, W. J. and Doney, S. C.: The subtropical nutrient spiral, Global Biogeochemical Cycles, 17, https://doi.org/10.1029/2003GB002085,
2003.

Jorge, D. S., Loisel, H., Jamet, C., Dessailly, D., Demaria, J., Bricaud, A., Maritorena, S., Zhang, X., Antoine, D., Kutser, T., Bélanger, S.,
Brando, V. O., Werdell, J., Kwiatkowska, E., Mangin, A., and d'Andon, O. F.: A three-step semi analytical algorithm (3SAA) for estimating
inherent optical properties over oceanic, coastal, and inland waters from remote sensing reflectance, Remote Sensing of Environment, 263,
https://doi.org/10.1016/j.rse.2021.112537, 2021.

Koestner, D., Stramski, D., and Reynolds, R. A.: Improved multivariable algorithms for estimating oceanic particulate or-
ganic carbon concentration from optical backscattering and chlorophyll-a measurements, Frontiers in Marine Science, 10,
https://doi.org/10.3389/fmars.2023.1197953, 2024.

Leonelli, F. E., Bellacicco, M., Pitarch, J., Organelli, E., Buongiorno Nardelli, B., de Toma, V., Cammarota, C., Marullo, S., and Santoleri, R.:
Ultra-Oligotrophic Waters Expansion in the North Atlantic Subtropical Gyre Revealed by 21 Years of Satellite Observations, Geophysical
Research Letters, 49, e2021GL096 965, https://doi.org/10.1029/2021GL096965, 2022.

Letelier, R. M., Karl, D. M., Abbott, M. R., and Bidigare, R. R.: Light driven seasonal patterns of chlorophyll and ni-trate in the lower euphotic zone of the North Pacific Subtropical Gyre, Limnology and Oceanography, 49, 508–519, https://doi.org/https://doi.org/10.4319/lo.2004.49.2.0508, 2004.

465 Loisel, H., Nicolas, J.-M., Deschamps, P.-Y., and Frouin, R.: Seasonal and inter-annual variability of particulate organic matter in the global ocean, Geophysical Research Letters, 29, 491–494, https://doi.org/10.1029/2002GL015948, 2002.

Lozier, M. S., Dave, A. C., Palter, J. B., Gerber, L. M., and Barber, R. T.: On the relationship between stratification and primary productivity in the North Atlantic, Geophysical Research Letters, 38, https://doi.org/10.1029/2011GL049414, 2011.

Mignot, A., Claustre, H., Uitz, J., Poteau, A., D'Ortenzio, F., and Xing, X.: Understanding the seasonal dynamics of phytoplankton biomass
470 and the deep chlorophyll maximum in oligotrophic environments: A Bio-Argo float investigation, Global Biogeochemical Cycles, 28, 856–876, https://doi.org/https://doi.org/10.1002/2013GB004781, 2014.

Oubelkheir, K., Claustre, H., Sciandra, A., and Babin, M.: Bio-optical and biogeochemical properties of different trophic regimes in oceanic waters, Limnology and Oceanography, 50, 1795–1809, https://doi.org/10.4319/lo.2005.50.6.1795, 2005.

Ruescas, A., Hieronymi, M., Mateo-Garcia, G., Koponen, S., Kallio, K., and Camps-Valls, G.: Machine Learning Regression Approaches
475 for Colored Dissolved Organic Matter (CDOM) Retrieval with S2-MSI and S3-OLCI Simulated Data, Remote Sensing, 10, 786, https://doi.org/10.3390/rs10050786, 2018.

Sauzède, R., Claustre, H., Uitz, J., Jamet, C., Dall'Olmo, G., D'Ortenzio, F., Gentili, B., Poteau, A., and Schmechtig, C.: A neural network-based method for merging ocean color and Argo data to extend surface bio-optical properties to depth: Retrieval of the particulate backscat-tering coefficient, Journal of Geophysical Research: Oceans, 121, n/a–n/a, https://doi.org/10.1002/2015JC011408, 2016.

480 Sauzède, R., Johnson, J., Claustre, H., Camps-Valls, G., and Ruescas, A.: Estimation of oceanic POC with machine learning, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, V-2-2020, 949–956, https://doi.org/10.5194/isprs-annals-V-2-2020-949-2020, 2020.

Schmid, L., Gerharz, A., Groll, A., and Pauly, M.: Machine Learning for Multi-Output Regression: When should a holistic multivariate approach be preferred over separate univariate ones?, https://doi.org/10.48550/ARXIV.2201.05340, 2022.

485 Siegel, D. A., Buesseler, K. O., Doney, S. C., Sailley, S. F., Behrenfeld, M. J., and Boyd, P. W.: Global assessment of ocean carbon export by combining satellite observations and food-web models, Global Biogeochemical Cycles, 28, 181–196, https://doi.org/10.1002/2013GB004743, 2014.

Siegel, D. A., DeVries, T., Cetinić, I., and Bisson, K. M.: Quantifying the Ocean's Biological Pump and Its Carbon Cycle Impacts on Global Scales, Annual Review of Marine Science, 15, 329–356, https://doi.org/10.1146/annurev-marine-040722-115226, pMID: 36070554, 2023.

490 Stramski, D., Reynolds, R. A., Kahru, M., and Mitchell, B. G.: Estimation of Particulate Organic Carbon in the Ocean from Satellite Remote Sensing, Science, 285, 239–242, http://www.jstor.org/stable/2898690, 1999.

Stramski, D., Reynolds, R. A., Babin, M., Kaczmarek, S., Lewis, M. R., Röttgers, R., Sciandra, A., Stramska, M., Twardowski, M. S., Franz, B. A., and Claustre, H.: Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern South Pacific and eastern Atlantic Oceans, Biogeosciences, 5, 171–201, https://doi.org/10.5194/bg-5-171-2008, 2008.

495 Sullivan, J. and Twardowski, M.: Angular shape of the oceanic particulate volume scattering function in the backward direction, Applied optics, 48, 6811–9, https://doi.org/10.1364/AO.48.006811, 2009.

Sullivan, J., Twardowski, M., Zaneveld, J., and Moore, C.: Measuring optical backscattering in water, Light Scattering Reviews 7, pp. 189–224, https://doi.org/10.1007/978-3-642-21907-8_6, 2013.

Xu, D., Shi, Y., Tsang, I. W., Ong, Y.-S., Gong, C., and Shen, X.: A Survey on Multi-output Learning, https://doi.org/10.48550/ARXIV.1901.00248, 2019.

Yang, B., Boss, E. S., Haëntjens, N., Long, M. C., Behrenfeld, M. J., Eveleth, R., and Doney, S. C.: Phytoplankton Phenology in the North Atlantic: Insights From Profiling Float Measurements, Frontiers in Marine Science, 7, https://doi.org/10.3389/fmars.2020.00139, 2020.

Zhang, X., Hu, L., and He, M.-X.: Scattering by pure seawater: Effect of salinity, Optics Express, 17, 5698–5710, https://doi.org/10.1364/OE.17.005698, 2009.

Zibordi, G., Berthon, J.-F., Mélin, F., D'Alimonte, D., and Kaitala, S.: Validation of satellite ocean color primary products at optically complex coastal sites: Northern Adriatic Sea, Northern Baltic Proper and Gulf of Finland, Remote Sensing of Environment, 113, 2574–2591, https://doi.org/10.1016/j.rse.2009.07.013, 2009.