AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Oceans

RESEARCH ARTICLE

10.1002/2015JC011408

Key Points:
- A neural network is developed to infer the vertical distribution of the backscattering coefficient
- The neural network requires as input Argo T/S profiles and ocean color remote sensing products
- Validation of the method is highly satisfactory which supports its application to the global ocean

Supporting Information:
- Supporting Information S1–S6

Correspondence to:
R. Sauzède,
sauzede@obs-vlfr.fr

Citation:
Sauzède, R., H. Claustre, J. Uitz, C. Jamet, G. Dall'Olmo, F. D'Ortenzio, B. Gentili, A. Poteau, and C. Schmechtig (2016), A neural network-based method for merging ocean color and Argo data to extend surface bio-optical properties to depth: Retrieval of the particulate backscattering coefficient, *J. Geophys. Res. Oceans*, *121*, 2552–2571, doi:10.1002/2015JC011408.

Received 27 OCT 2015
Accepted 15 MAR 2016
Accepted article online 21 MAR 2016
Published online 13 APR 2016

A neural network-based method for merging ocean color and Argo data to extend surface bio-optical properties to depth: Retrieval of the particulate backscattering coefficient

R. Sauzède[1], H. Claustre[1], J. Uitz[1], C. Jamet[2], G. Dall'Olmo[3,4], F. D'Ortenzio[1], B. Gentili[1], A. Poteau[1], and C. Schmechtig[1]

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, Observatoire Océanologique de Villefranche, Laboratoire d'Océanographie de Villefranche, Villefranche-Sur-Mer, France, [2]Laboratoire d'Océanologie et de Géosciences, UMR8187, ULCO/CNRS, Wimereux, France, [3]Plymouth Marine Laboratory, Plymouth, UK, [4]National Centre for Earth Observation, Plymouth, UK

Abstract The present study proposes a novel method that merges satellite ocean color bio-optical products with Argo temperature-salinity profiles to infer the vertical distribution of the particulate backscattering coefficient (b_{bp}). This neural network-based method (SOCA-BBP for Satellite Ocean-Color merged with Argo data to infer the vertical distribution of the Particulate Backscattering coefficient) uses three main input components: (1) satellite-based surface estimates of b_{bp} and chlorophyll *a* concentration matched up in space and time with (2) depth-resolved physical properties derived from temperature-salinity profiles measured by Argo profiling floats and (3) the day of the year of the considered satellite-Argo matchup. The neural network is trained and validated using a database including 4725 simultaneous profiles of temperature-salinity and bio-optical properties collected by Bio-Argo floats, with concomitant satellite-derived products. The Bio-Argo profiles are representative of the global open-ocean in terms of oceanographic conditions, making the proposed method applicable to most open-ocean environments. SOCA-BBP is validated using 20% of the entire database (global error of 21%). We present additional validation results based on two other independent data sets acquired (1) by four Bio-Argo floats deployed in major oceanic basins, not represented in the database used to train the method; and (2) during an AMT (Atlantic Meridional Transect) field cruise in 2009. These validation tests based on two fully independent data sets indicate the robustness of the predicted vertical distribution of b_{bp}. To illustrate the potential of the method, we merged monthly climatological Argo profiles with ocean color products to produce a depth-resolved climatology of b_{bp} for the global ocean.

1. Introduction

The ocean plays an important role in the regulation of the climate of our planet by influencing the amount of carbon dioxide (CO_2) in the atmosphere. An important part of this regulation takes place through the so-called biological carbon pump, which results from the sinking and sequestration to the deep oceans of part of the stock of Particulate Organic Carbon (POC) produced by phytoplankton photosynthesis [*Falkowski et al.*, 1998; *Volk and Hoffert*, 1985]. Despite their importance to the global carbon cycle, these processes are still poorly constrained. This is largely caused by a lack of observations of key biogeochemical properties and associated processes on relevant space and time scales. Traditional ship-based sampling and measurement methods provide direct, detailed information on biogeochemical properties of the water column, but with insufficient space-time coverage.

Recent advances in optical sensors implemented on in situ and remote-sensing platforms allow the study of biogeochemical variables and processes in the open ocean over a broad range of temporal and spatial scales. The increasing use of such optical tools has led the scientific community to develop optical proxies for estimating key biogeochemical parameters. Specifically the particulate backscattering coefficient and the particulate beam attenuation coefficient are widely used as proxies of POC [*Bishop and Wood*, 2009; *Bishop*, 2009; *Gardner et al.*, 2006]. The particulate backscattering coefficient (b_{bp}) has received much attention in the recent years because it can be continuously measured in situ from autonomous platforms [e.g., *Boss and Behrenfeld*, 2010; *Boss et al.*, 2008; *Dall'Olmo and Mork*, 2014] or retrieved from satellite remote

sensing of ocean color [*Behrenfeld et al*., 2005; *Siegel et al*., 2005; *Westberry et al*., 2008]. Aside from being a relevant proxy of POC [*Balch et al*., 2001; *Cetinić et al*., 2012; *Loisel et al*., 2001, 2002; *Stramski et al*., 1999, 2008], this bio-optical property can be used as an index of the particulate load, and its spectral dependence as an index of particle size [e.g., *Dall'Olmo and Mork*, 2014; *Loisel et al*., 2006] and phytoplankton size structure [*Kostadinov et al*., 2010]. Although still debated, several recent studies have shown that b_{bp} could also be considered as an indicator of phytoplankton carbon [*Behrenfeld et al*., 2005; *Graff et al*., 2015; *Martinez-Vicente et al*., 2013]. This would make b_{bp} an interesting alternative to chlorophyll *a* concentration for monitoring phytoplankton biomass in situ or from space. Therefore, b_{bp} appears as a key bio-optical property to study the space-time dynamics of POC and possibly of phytoplankton biomass, a prerequisite for ultimately improving the characterization and quantitative assessment of biologically mediated carbon fluxes in the global open ocean.

Satellite remote sensing of ocean color, coupled to relevant algorithms, has the potential to provide a quasi-synoptic view of b_{bp} which, in turn, can be interpreted in terms of POC [*Loisel et al*., 2001, 2002; *Stramski et al*., 1999, 2008]. We note that satellite-derived products of POC may also be obtained from reflectance-based or beam attenuation-based algorithms [e.g., *Gardner et al*., 2006; *Stramski et al*., 2008]. Several studies have used this potential to examine the spatial and temporal distribution of POC in the open ocean [*Gardner et al*., 2006; *Loisel et al*., 2002; *Stramska*, 2009]. However, such satellite-based estimates are restricted to the ocean surface layer and, in the context of global carbon cycle studies including carbon production and export, are insufficient. In fact, the photosynthetic activities of phytoplankton are not restricted to the near-surface layer but also to deeper layers in the water column. Moreover, POC, which is vector of carbon export, is also composed of biogenic detrital particles, microzooplankton, heterotrophic bacteria, viruses, and aggregates that are present within the entire water column in various proportions. So the vertical distribution of POC is important for understanding both pelagic ecosystems and carbon flux. The high spatial and temporal variability of the vertical distribution of POC makes the extension of surface POC to depth complex. To our knowledge, this has been attempted only by *Duforêt-Gaurier et al*. [2010] who based their study on a relatively small database of POC vertical profiles.

Because b_{bp} is tightly linked to the stock of biologically derived carbon (POC), its vertical distribution must be in some way driven by nutrient availability and light regime, which are in turn influenced by the physical forcing of the water column. Hence, one may expect that combining the satellite-derived surface data of b_{bp} with available information on the physical state of the water column will help extending surface b_{bp} to depth and constraining its vertical distribution. Since the launch of the Argo program, temperature, and salinity profiles are measured continuously with high spatiotemporal resolution throughout the world's oceans [*Roemmich et al*., 2009]. Now mature, with more than 3800 active floats, the Argo array provides a unique high-resolution view of hydrological properties in the upper 2000 m of the ocean. These data represent an ideal candidate for merging with satellite ocean color products. Therefore, in this study, we propose to develop and examine the potential of a new global method for merging satellite ocean color and physical Argo data to infer the vertical distribution of b_{bp} with a relatively high spatiotemporal resolution, i.e., the resolution of Argo-to-satellite matchup data.

In the past few years, the number of concurrent in situ observations of the vertical distributions of temperature, salinity, and b_{bp} has dramatically increased. This results from the integration of optical sensors on autonomous platforms, especially Bio-Argo profiling floats which almost all measure b_{bp} [*Boss et al*., 2008; *Claustre et al*., 2010a, 2010b; *Mignot et al*., 2014; *Xing et al*., 2014] in addition to physical vertical profiles of temperature and salinity. Hence, the numerous vertical profiles collected by Bio-Argo floats offer a new path for developing a global parameterization of the vertical distribution of this key bio-optical property. Our study aims to use the large database of physical and bio-optical vertical profiles collected by the Bio-Argo fleet within the global open ocean to establish the proposed method.

Artificial neural networks (ANNs) are very powerful methods for approximating any differentiable and continuous functions [*Hornik et al*., 1989] and have been widely used for biogeochemical, geophysical, and remote sensing applications [e.g., *Bricaud et al*., 2007; *Friedrich and Oschlies*, 2009; *Gross et al*., 2000; *Jamet et al*., 2012; *Krasnopolsky*, 2009; *Niang et al*., 2006; *Palacz et al*., 2013; *Raitsos et al*., 2008; *Sauzède et al*., 2015; *Telszewski et al*., 2009]. These methods have a large potential to model complex and nonlinear relationships that are characteristic of ecological data sets [*Lek and Guégan*, 1999]. Furthermore, one of the benefit of using ANNs is that uncertainties in input data are accounted for during the training process of the neural

network. Indeed, ANNs are relatively insensitive to reasonable uncertainties in input data. Therefore, we selected this method as the most appropriate for reaching our goal.

In summary, this study presents a new ANN-based method that uses merged satellite ocean color-based products and physical Argo data to retrieve the vertical distribution of b_{bp} at the global scale. Hereafter, the method is referred to as SOCA-BBP for Satellite Ocean Color merged with Argo data to infer the vertical distribution of the Particulate Backscattering coefficient. SOCA-BBP uses three main input components: (1) a surface component composed of satellite-based estimates of b_{bp} and chlorophyll *a* concentration, (2) vertically resolved physical quantities derived from Argo temperature and salinity profiles, and (3) the day of the year of the considered satellite-to-Argo matchup. Our analysis utilizes a large database of 4725 concurrent in situ vertical profiles of temperature, salinity, and b_{bp} collected by Bio-Argo profiling floats, matched up with satellite ocean color observations. The resulting database is representative of various trophic conditions, making the method largely applicable to the global open ocean.

2. Data Presentation and Processing

Below we present the Bio-Argo database used in this study, which is composed of concurrent vertical profiles of temperature, salinity, and particulate backscattering coefficient. Then we present the procedure for matching up the Bio-Argo vertical profiles with satellite-based bio-optical products. We finally describe the resulting database used to develop and validate the SOCA-BBP algorithm.

2.1. Database of Concurrent Vertical Profiles of Temperature, Salinity, and Particulate Backscattering Coefficient

In addition to the standard conductivity-temperature-depth (CTD) sensors mounted on physical Argo profiling floats, Bio-Argo floats are equipped with additional bio-optical sensors that can be used to measure proxies of major biogeochemical variables. Specifically, the Bio-Argo floats are fitted with a CTD (Seabird), a sensor package (Satlantic OCR) that measures downwelling irradiance at three wavelengths and PAR (Photosynthetically Available Radiation), and a sensor package (WET Labs ECO Puck Triplet) composed of a chlorophyll *a* fluorometer, a CDOM (Colored Dissolved Organic Matter) fluorometer, and a sensor measuring the particulate backscattering coefficient at a wavelength of 700 nm (79 floats) or 532 nm (4 floats). In the present study, we use exclusively measurements of temperature, salinity, and particulate backscattering coefficient to train and validate the method.

The Bio-Argo floats used in this study typically collect measurements from 1000 m to the surface with a $\sim$1 m resolution every 10 days, 3 days, or even 3 times per day depending on the float mission configuration. When the float surfaces, data are transmitted in real time using Iridium communication. Thanks to this communication system, the float mission parameters can also be modified in real time (e.g., time interval between two profiling cycles).

The volume scattering function (VSF), $\beta(\theta,\lambda)$ (m^{-1} sr^{-1}), is defined as the angular distribution of scattering relative to the direction of light propagation θ at the optical wavelength λ. The backscattering sensor of Bio-Argo floats measures $\beta(124°, \lambda)$ with $\lambda = 700$ nm or 532 nm. The contribution of particles to the VSF, β_p, is calculated by subtracting the contribution of pure seawater, β_{sw}, from $\beta(124°, \lambda)$:

$$\beta_p(124^\circ,\lambda)=\beta_p(124^\circ,\lambda)-\beta_{sw}(124^\circ,\lambda). \quad (1)$$

with β_{sw} depending on temperature and salinity and computed using a depolarization ratio of 0.039 [*Zhang et al.*, 2009]. Then, the particulate backscattering coefficient at 700 or 532nm, $b_{bp}(\lambda)$, is determined from $\beta_p(124°, \lambda)$ and a conversion factor, χ, [*Boss and Pegau*, 2001; *Kokhanovsky*, 2012; *Sullivan and Twardowski*, 2009] as follows:

$$b_{bp}(\lambda)=2\pi\chi(\beta_p(124^\circ,\lambda)-\beta_{sw}(124^\circ,\lambda)). \quad (2)$$

The value of χ for an angle of 124° is 1.076 [*Sullivan and Twardowski*, 2009].

As the Bio-Argo database includes 10 times more $b_{bp}(700)$ profiles than $b_{bp}(532)$ profiles, in order to harmonize the b_{bp} profiles of the database, the profiles of $b_{bp}(532)$ were converted to $b_{bp}(700)$. The conversion was performed using a power law model of the particulate backscattering coefficient spectral dependency:

Table 1. Abbreviations Used in the Present Study and Their Significance

Abbreviations	Significance
b_{bp}	Optical particulate backscattering coefficient at 700 nm (m^{-1})
Chl	Chlorophyll *a* concentration (mg m^{-3})
b_{bp_MODIS}	MODIS-Aqua-derived b_{bp} using QAA algorithm [*Lee et al.*, 2002] (m^{-1})
b_{bp_VIIRS}	VIIRS-derived b_{bp} using QAA algorithm [*Lee et al.*, 2002] (m^{-1})
Chl_{MODIS}	MODIS-Aqua-derived Chl (mg m^{-3})
Chl_{VIIRS}	VIIRS-derived Chl (mg m^{-3})
b_{bp_SOCA}	Vertically resolved values of b_{bp} retrieved by SOCA-BBP (m^{-1})
b_{bp_Floats}	Vertically resolved values of b_{bp} collected by Bio-Argo profiling floats (m^{-1})
b_{bp_AMT}	Vertically resolved values of b_{bp} collected during the AMT cruise (m^{-1})
Z	Geometrical depth (m)
Z_{norm}	Depth at which the Chl profile returns to a constant background value at depth (m)
ζ	Depth normalized with respect to Z_{norm}, $\zeta = z/Z_{norm}$ (dimensionless)
Z_m	Mixed layer depth (m)
Z_e	Euphotic layer depth (m)
K_{d490}	Diffuse attenuation coefficient at 490 nm (m^{-1})
K_{PAR}	Diffuse attenuation coefficient for photosynthetically available radiation (m^{-1})
Day	Day of the year
Day_{rad}	Day transformed into radians
Z_{pd}	Penetration depth defined as $Z_{pd} = Z_e/4.6$ [*Morel and Berthon*, 1989] (m)
MAPD	Median Absolute Percent Difference (%)

$$b_{bp}(\lambda) = b_{bp}(\lambda_0) \cdot \left(\frac{\lambda}{\lambda_0}\right)^{-\gamma}. \quad (3)$$

We use a value of $\gamma = 2$ for the Bio-Argo profiles collected in the North Pacific subtropical gyre and a value of $\gamma = 3$ for those collected in the South Pacific Subtropical Gyre. These values are based on *Loisel et al.* [2006] who showed that the low chlorophyll waters of the subtropical gyres are typically associated with high γ values (between 2 and 3) whereas low or even negative γ values are found in the more productive areas of the ocean (between −1.5 and 1). The CTD data were quality controlled following the standard Argo protocol [*Wong et al.*, 2014]. A quality control procedure was applied to each profile of $b_{bp}(700)$ (hereafter b_{bp}; see Table 1 for a list of symbols): (1) the manufacturer-supplied offsets and scaling factors were applied to each raw profile; (2) high-frequency spikes were removed using a median filter; (3) b_{bp} values above 0.03 m^{-1} were discarded as considered outside of the sensor range of operation. The final Bio-Argo database of concurrent b_{bp} and temperature-salinity measurements is composed of 8330 vertical profiles collected by 83 Bio-Argo profiling floats.

2.2. Bio-Argo and Satellite Ocean Color Matchup Database

For consistency with Bio-Argo b_{bp} data measured at (or converted to) 700 nm, the satellite-derived b_{bp} data were estimated for a wavelength of 700 nm using the Quasi-Analytical Algorithm (QAA) [*Lee et al.*, 2002]. Then, each profile of the Bio-Argo database described above was matched up with satellite data of surface $b_{bp}(700)$ and chlorophyll *a* concentration (Chl) using the closest pixel from standard level 3 eight day MODIS-Aqua composites (Reprocessing R2014.0) with a 9 km resolution (provided by the Ocean Color Web: http://oceancolor.gsfc.nasa.gov).

The matchup procedure led to discarding 43% of the profiles from the initial Bio-Argo database (see discarding rate for the major oceanic basins in Table 2). The geographic distribution of the 4725 remaining Bio-Argo profiles with concomitant MODIS-Aqua-derived products is presented in Figure 1. The database used in this study covers most of the major ocean basins (i.e., Southern Ocean, Indian Ocean, Mediterranean Sea, North Pacific, South Pacific, North Atlantic, and South Atlantic; see supporting information Figure S1 for details of basin boundaries). The underrepresentation of the Southern Hemisphere, due to undersampling (see Figure 1), is apparent in Figure 2. On a monthly basis, more profiles are available for spring and summer than for autumn and winter months for the Northern and Southern Hemisphere (Figure 2a). This temporal bias of data acquisition is mostly due to a lack of satellite images at high latitudes during winter and autumn. The annual distribution of the vertical profiles in the database covers 8 years from 2008 to 2015 (Figure 2b); most of the observations were nevertheless collected since 2013.

The resulting Bio-Argo and satellite matchup database appears to be representative of a broad variety of hydrological and biogeochemical conditions prevailing in the global open ocean (Figure 3). For instance, the values of mixed layer depth, Z_m, acquired by the Bio-Argo floats vary between 15 and 900 m (measurements from the North Atlantic Subtropical Gyre in spring and the North Atlantic in winter, respectively). Z_m is calculated from the density profiles using a density criterion of 0.03 kg m^{-3} as in *De Boyer Montégut et al.* [2004]. The database is also representative of most trophic conditions observed in the open ocean (i.e., from oligotrophic to eutrophic waters, see Figure 3b). The MODIS-Aqua-estimated Chl, Chl_{MODIS}, covers 3

Table 2. Summary of the Number of Profiles Rejected After the Satellite Versus Bio-Argo Matchup Procedure for the Different Bio-Argo Float Sampling Regions[a]

Area	Number of Profiles	Number of Profiles After Matchup	% of Profiles Removed
North Atlantic	3985	1791	56%
South Atlantic	1073	620	43%
North Pacific	364	173	53%
South Pacific	275	188	32%
Southern Ocean	203	47	77%
Mediterranean Sea	2235	1839	18%
Indian Ocean	195	67	66%
Global ocean (total)	8330	4725	56%

[a]The seven major oceanic basins boundaries used to compute this table are presented in supporting information Figure S1.

orders of magnitude (i.e., from 0.01 to 10 mg m^{-3}). The most oligotrophic conditions were found in the South Atlantic Subtropical Gyre in autumn and the most eutrophic in the North Atlantic, especially in the Labrador Sea during the spring bloom. Similarly, the MODIS-Aqua-derived b_{bp}, b_{bp_MODIS}, covers 3 orders of magnitude (from 0.00001 to 0.01 m^{-1}; Figure 3c).

Before splitting the Bio-Argo and satellite matchup database into two subsets for developing the neural network (i.e., the training and validation data sets), 314 profiles collected by four Bio-Argo floats were removed from the database to create an "independent data set" used for an additional validation of the method. These four floats were chosen in four major oceanic basins: the North Atlantic Subpolar Gyre, the North Atlantic Subtropical Gyre, the Southern Ocean, and the Mediterranean Sea.

The resulting matchup database was randomly split into two independent subsets, including 80% (3525 profiles used for training the MLP) and 20% (886 profiles used for validating the MLP) of the data. Similar to the training data set, the validation data set is representative of the hydrological and biogeochemical conditions prevailing in the global open ocean (see histograms in Figure 3).

Finally, a totally independent data set from an AMT (Atlantic Meridional Transect) field cruise in 2009 is also used to validate independently the method (i.e., different b_{bp} sensor, different time and location of b_{bp} acquisition). This addition validation is done in order to demonstrate the good generalization of the method

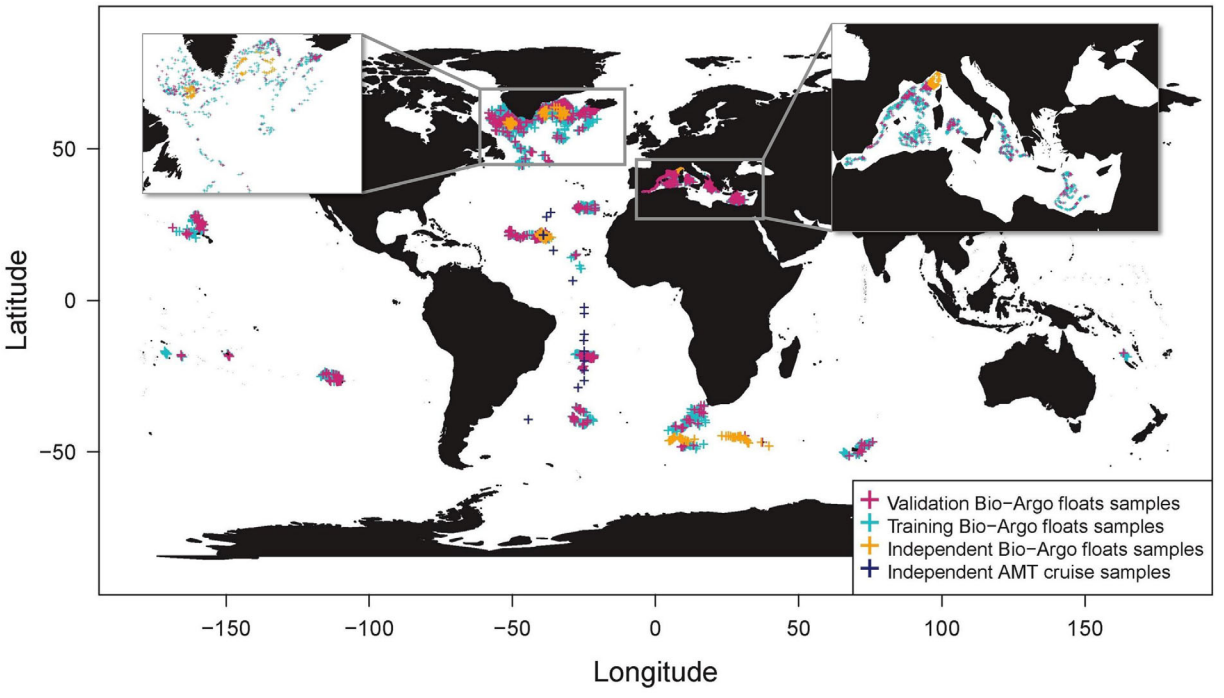

Figure 1. Geographic distribution of the 4725 stations used in the present study. For each station, concurrent profiles of temperature, salinity, and b_{bp} collected by Bio-Argo floats were matched up with concomitant MODIS-Aqua-derived products. Turquoise and purple crosses indicate the location of the profiles of the so-called "training" and "validation" data sets, respectively (see text). The vertical profiles collected by the four independent Bio-Argo floats are shown as orange crosses. The vertical profiles collected during the AMT oceanographic cruise (also used for an independent validation of the method) are shown as dark blue crosses.

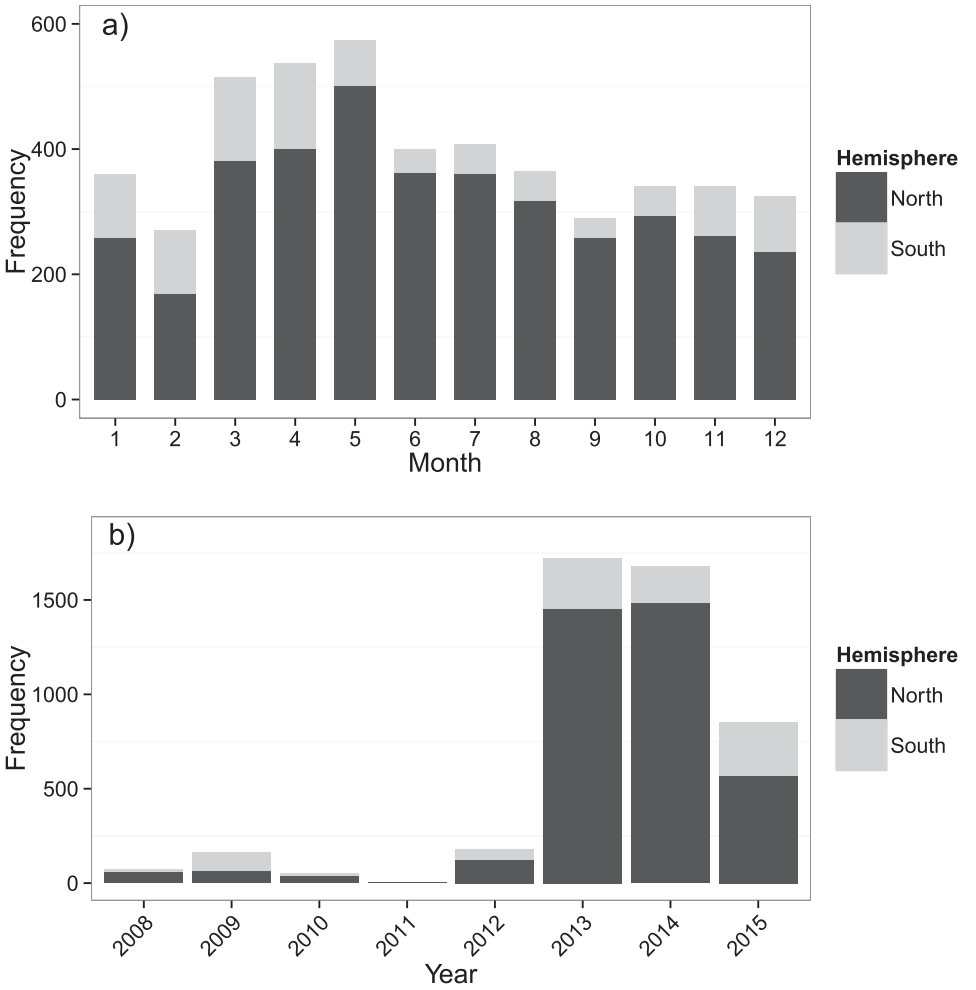

Figure 2. Temporal distribution of the 4725 stations for which both Bio-Argo and satellite data were simultaneously available as a function of (a) months and (b) years with black and grey colors indicating the hemisphere of data acquisition.

(i.e., good performance of the method in other conditions as used for the neural network training). During this AMT cruise, the continuous b_{bp} measurements at 470 and 526 nm were made using a WET Labs ECO-BB3 sensor. The AMT $b_{bp}(700)$ profiles were then computed by linearly combining b_{bp} measurements at 470 and 556 nm. After matchup, this subset is composed of 16 matchup satellite and in situ profiles of temperature, salinity, and b_{bp}.

To summarize, the geographical distribution of the sampling stations included in the training, validation, independent four float, and totally independent-AMT subsets is shown in Figure 1.

2.3. Normalization of the Vertical Profiles of the Particulate Backscattering Coefficient

SOCA-BBP is designed to predict the vertical distribution of b_{bp} within the so-called productive layer. This corresponds essentially to the layer where most particle and phytoplankton stocks are confined. In mixed conditions, the thickness of the productive layer roughly coincides with that of the mixed layer. In stratified conditions (typically associated with the presence of a deep chlorophyll maximum), the productive layer is more linked to the euphotic layer. Here the productive layer is described through the introduction of a dimensionless depth, ζ [*Sauzède et al.*, 2015], with ζ defined as the geometrical depth, z, divided by a normalization depth, Z_{norm}:

$$\zeta = {}^{z}/_{Z_{norm}}. \tag{4}$$

with Z_{norm} defined as the depth at which the Chl vertical profile returns to a constant background value (depth of the bottom of the productive layer). As fluorescence profiles are always collected simultaneously

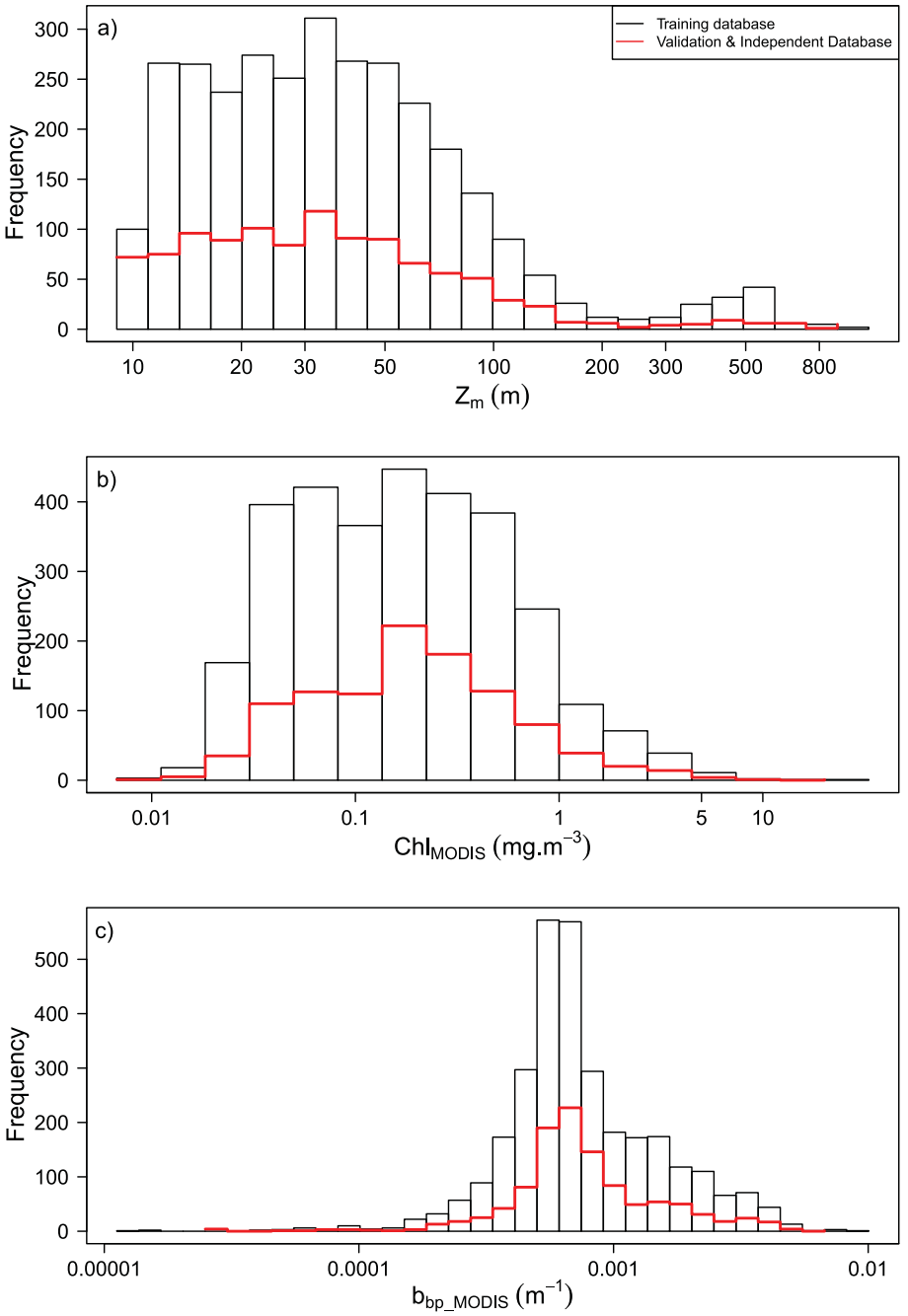

Figure 3. Histogram of frequency of (a) the mixed layer depth, Z_m (m), (b) the satellite-derived chlorophyll *a* concentration, Chl_{MODIS} (mg m^{-3}), and (c) the satellite-derived particulate backscattering coefficient, b_{bp_MODIS} (m^{-1}). The black histogram represents the distribution of the data used to train the method and the red histogram is for the independent data set (20% of the initial database) used to validate the method.

with temperature, salinity, and b_{bp} profiles by the Bio-Argo floats, Z_{norm} can be computed with precision for the Bio-Argo database using the fluorescence profiles (see *Sauzède et al.* [2015] for details).

As the main objective of the SOCA-BBP method is to merge satellite and Argo data without using vertical bio-optical profiles acquired by Bio-Argo floats, for application purposes, we developed a statistical relationship to estimate Z_{norm} from two parameters accessible or derivable from our input data set: (1) the euphotic layer depth, Z_e, the depth at which irradiance is reduced to 1% of its surface value, and (2) the mixed layer depth, Z_m. Z_e is computed with the following procedure: (1) the attenuation coefficient at 490 nm, K_{d490}, is

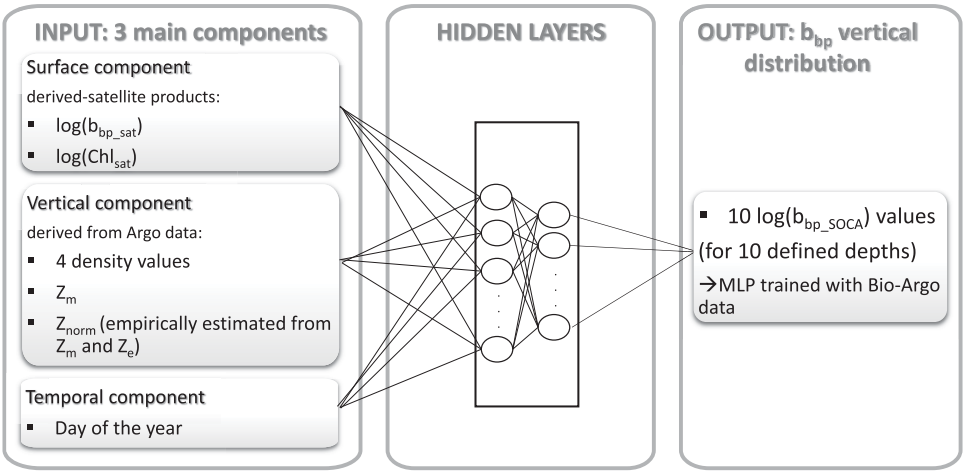

Figure 4. Schematic overview of the SOCA-BBP MLP-based algorithm that retrieves the vertical distribution of b_{bp} from merged ocean color satellite and Argo data associated with the day of the year of the considered satellite-to-Argo matchup.

determined using the satellite-derived chlorophyll *a* concentration [*Morel and Maritorena*, 2001]; (2) the total attenuation coefficient, K_{PAR}, is retrieved from K_{d490} [*Rochford et al*., 2001]; and (3) finally, Z_e is retrieved from K_{PAR} using the exponential decrease of light over depth. The most statistically significant relationship between Z_{norm} and both the Z_e and Z_m parameters was found when stratified conditions are discriminated from mixed conditions based on the ratio of Z_e to Z_m (i.e., $Ze > Zm$: stratified; $Ze < Zm$: mixed) [*Morel and Berthon*, 1989; *Uitz et al*., 2006]. We obtain the following optimal statistical relationships for a stratified water column:

$$\log(Z_{norm}) = 0.12 \cdot \log(Z_m) + 1.04 \cdot \log(Z_e). \quad (5)$$

and for a mixed water column:

$$\log(Z_{norm}) = 0.64 \cdot \log(Z_m) + 0.51 \cdot \log(Z_e). \quad (6)$$

The relationship between Z_{norm} computed from the fluorescence in situ profiles measured by the Bio-Argo floats and modeled with the statistical relationships presented above for the two hydrological regime of the water column (stratified or mixed) is satisfactory with a median absolute percent difference, MAPD, of 14% (for more details see supporting information Figure S2). Finally, Z_{norm} used to scale b_{bp} profiles ranges from 20 to 805 m in the Bio-Argo database (see supporting information Figure S3). Scaling the b_{bp} profiles with respect to ζ enables the merging all the profiles regardless of their vertical shape and range of variation while simultaneously accounting for their variability.

3. SOCA-BBP Algorithm Development

3.1. General Principles of Multi-Layer Perceptron (MLP)

The type of artificial neural network chosen in this study is a Multi-Layer Perceptron (MLP) [*Bishop*, 1995; *Rumelhart et al*., 1988]. A MLP is composed of several layers: one input layer, one output layer, and one or more intermediate levels (i.e., the so-called hidden layers). Each layer is composed of neurons, which are elementary transfer functions that provide outputs when inputs are applied. Each neuron is interconnected with the others by weights (Figure 4). The matrix of these weight values is iteratively adjusted during the training phase of the MLP and is computed by minimizing a cost function defined as the quadratic difference between the desired and computed outputs. The technique used for this minimization is the back-propagation conjugate-gradient, which is an iterative optimization method adapted to MLP development [*Bishop*, 1995; *Hornik et al*., 1989].

To determine the weights of the MLP, the training data set is randomly split into two subsets (50% of the data each), the so-called "learning" and "test" data sets. These two subsets are used during the training

process of the MLP to prevent overlearning [*Bishop*, 1995]. The validation data set used to evaluate independently the final performance of the MLP is composed of 20% of the entire initial database.

3.2. Developing a MLP to Retrieve the Vertical Distribution of b_{bp}

After multiple tests, the following set of three input components was selected as optimal (see Figure 4): (1) a temporal component, i.e., the day of year; (2) a surface component, defined by the satellite-derived log-transformed particulate backscattering coefficient (b_{bp_MODIS}) and chlorophyll *a* concentration (Chl_{MODIS}) (see section 2.2); and (3) a vertical component, i.e., the normalization depth Z_{norm}, the mixed layer depth Z_m and four potential density values along the vertical profile, of which three taken at shallow depths and one at depth. The dimensionless depths according to density inputs were chosen using a principal component analysis to minimize redundancy in the selected input data (not shown). The MLP returns simultaneously 10 normalized values of $\log(b_{bp})$ as output according to 10 depths taken at regular intervals within the 0–1.3 ζ layer.

The elementary transfer function (sigmoid nonlinear function) that provides outputs when inputs are applied to the MLP varies within the range $[-1;1]$. Therefore, to take advantage of the nonlinearity of this function, the inputs and outputs of the MLP ($x_{i,j}$) are centered and reduced to match the $[-1;1]$ domain as follows:

$$x_{i,j} = \frac{2}{3} \cdot \frac{x_{i,j} - \mathrm{mean}(x_i)}{\sigma(x_i)}. \tag{7}$$

with σ the standard deviation of the considered input or output variable x. Obviously, the outputs need to be "denormalized" using the above equation with appropriate mean and standard deviation for each dimensionless depth of restitution.

With respect to the temporal component, we applied a specific normalization procedure that accounts for the periodicity of the day of the year (i.e., day one of the year is very similar from a seasonal perspective to day 365). Thus, similar to the method developed by *Sauzède et al*. [2015], the temporal input is transformed in radians using the following equation:

$$\mathrm{Day}_{rad} = \frac{\mathrm{Day} \cdot \pi}{182.625}. \tag{8}$$

where Day_{rad} is the day of the year in radian units and the coefficient 182.625 accounts for half the number of days per year (365.25).

Once the optimal input and output variables were determined, additional tests were performed to establish the best architecture of the MLP: one or two hidden layers with a number of neurons in each layer varying between 1 and 50 and 1 and 20, respectively. The architecture with minimum error of validation and minimum number of neurons was selected as optimal. The resulting optimal MLP is here composed of two hidden layers with eight neurons in the first hidden layer and six neurons in the second one. In order to evaluate the MLP robustness, different subsets of the training data set have been tested and no significant difference in the predictive skills of the MLP was observed.

3.3. Evaluation of Method Performance

The SOCA-BBP method is validated using independent data sets of b_{bp} profiles acquired by Bio-Argo floats or as part of an AMT field cruise (see section 2.2). For each profile used for the validation of the method, the 10 b_{bp} values simultaneously retrieved by SOCA-BBP (b_{bp_SOCA}), associated with the 10 dimensionless depths taken at regular intervals within the 0–1.3 ζ layer, are compared to b_{bp} values measured by the Bio-Argo floats (b_{bp_Floats}) or measured during the AMT cruise (b_{bp_AMT}) at each corresponding depth. To evaluate the performance of SOCA-BBP in inferring the vertical distribution of b_{bp}, several statistical parameters are considered. First, the determination coefficient (R^2) and the slope of the linear regression between the log-transformed values of b_{bp_SOCA} and b_{bp_Floats} (or b_{bp_AMT}) are computed. Second, we estimate the model error using the Median Absolute Percent Difference (MAPD, %) calculated as follows:

$$\mathrm{MAPD} = \mathrm{median}\left(\frac{|b_{bp\_SOCA} - b_{bp\_Floats}|}{b_{bp\_Floats}}\right) \cdot 100. \tag{9}$$

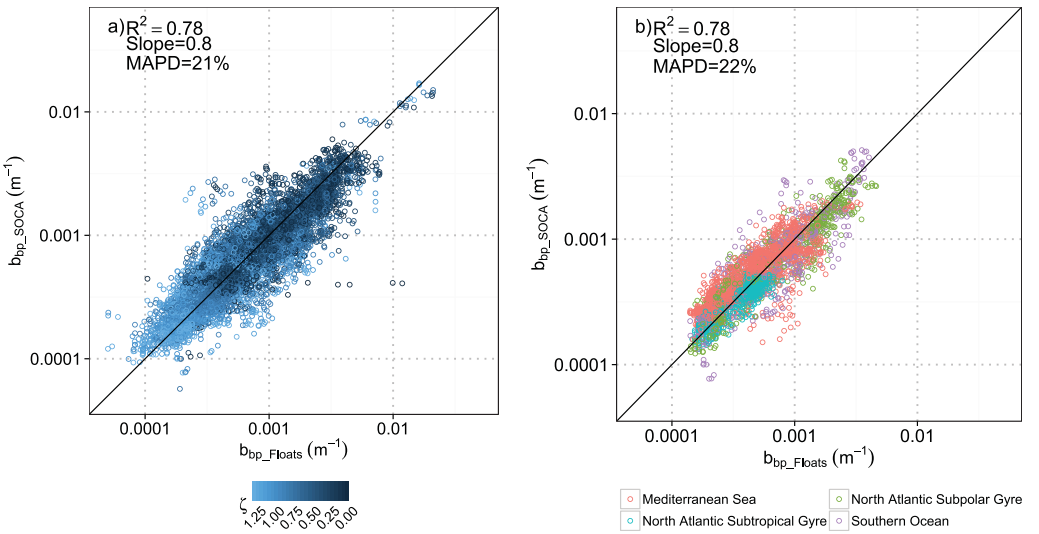

Figure 5. Comparison of the b_{bp} values retrieved by SOCA-BBP (b_{bp_SOCA}) to the reference b_{bp} measurements acquired by the Bio-Argo floats (b_{bp_Floats}) using two different data sets: (a) the validation database (i.e., 20% of the entire database chosen randomly) with data ordered according to the dimensionless depth ζ; (b) the independent data acquired by four Bio-Argo floats not integrated in the training and validation databases with the color code indicating the oceanic basins in which the Bio-Argo floats were deployed. The 1:1 line is shown in each plot. The calculation details of statistics are provided in section 3.3.

Note that b_{bp_Floats} is replaced by b_{bp_AMT} for the validation against bbp profiles from AMT cruise. We also evaluate the sensitivity of SOCA-BBP to uncertainties in the origin of the satellite data used as input to the MLP. For this purpose, a test is performed which consists in replacing the MODIS-Aqua-derived bio-optical products by VIIRS-derived products.

4. Results and Discussion

4.1. Retrieval of the Vertical Distribution of the Particulate Backscattering Coefficient

Using the validation database (i.e., 20% of our initial database), the ability of the method is evaluated through a comparison of the 10 values retrieved from SOCA-BBP (b_{bp_SOCA}) with corresponding values measured by the Bio-Argo floats (b_{bp_Floats}). The scatterplot of b_{bp_SOCA} versus b_{bp_Floats} reveals that SOCA-BBP predicts b_{bp} without systematic bias (i.e., global error of retrieval of 21%; see Figure 5a and Table 3). This figure shows that most of the b_{bp_SOCA} values are retrieved with substantial accuracy and that only a limited number of data points diverge significantly from the 1:1 line (Figure 5a). No bias according to the dimensionless depth of estimation seems to be identified from the Figure 5a. To be more precise, we tested statistically the performance of the method with respect to the vertical dimension (see Figure 6 and supporting information Figure S4 and Table 3). The Figure 6 presents the median of APD that is ~20% for each of the 10 dimensionless output depths. The APD appears somewhat lower for the 0–0.84 ζ layer suggesting that the method performs slightly better for the upper layers. For the deep layers, the b_{bp} values are very low, which may lead to large errors even when the difference between the predicted and reference values is minor. Nevertheless, the APD remains still low for these deep layers (~22% for the median). Supporting information Figure S4 presents the scatterplots of b_{bp_SOCA} versus b_{bp_Floats} for five layers of the water column (chosen from the dimensionless depth ζ). This figure reveals that b_{bp} is predicted without systematic bias according to the vertical dimension and that the slight deterioration of statistic results (see Table 3) for the deepest layer (1–1.3 ζ layer) might come from the lower range of b_{bp} values to predict.

4.2. Sensitivity of SOCA-BBP to Satellite Input Data

We evaluate the sensitivity of the method to satellite input data by replacing in the validation data set the MODIS-Aqua products by VIIRS-derived products. It is important to note that the purpose here is not to compare MOIS-Aqua and VIIRS products in term of accuracy but to evaluate the impact to use other satellite input data (e.g., VIIRS estimates rather than MODIS-Aqua) on the performance of the SOCA-BBP method. The VIIRS-to-Argo matchups are computed using standard level 3 VIIRS composites (reprocessing R2014.0)

Table 3. Statistics of the Comparison of the b_{bp} Values Predicted by SOCA-BBP to Bio-Argo Reference Measurements[a]

	n	R^2	Slope	MAPD (%)
Validation data set: total	8860	0.78	0.8	21
Validation data set: 0–0.2 ζ layer	1772	0.77	0.8	19
Validation data set: 0.2–0.5 ζ layer	1772	0.75	0.76	21
Validation data set: 0.5–0.7 ζ layer	1772	0.7	0.72	22
Validation data set: 0.7–1 ζ layer	1772	0.76	0.8	20
Validation data set: 1–1.3 ζ layer	1772	0.62	0.68	23
Independent data set: total	3140	0.78	0.8	22
Independent data set: Mediterranean Sea	1440	0.71	0.7	30
Independent data set: North Atlantic Subtropical Gyre	650	0.81	0.81	12
Independent data set: Austral Ocean	500	0.72	0.84	23
Independent data set: North Atlantic Subpolar Gyre	550	0.85	0.85	21

[a]The number of values (n) to compute the determination coefficient (R^2) and slope of the linear regression between the retrieved and reference values. The MAPD (Median Absolute Percent Difference) between the retrieved and reference values is also indicated (see section 3.3 for the calculation details).

with a 4 km resolution and 8 day binning period (9 km resolution as for MODIS-Aqua not available). Among the 886 Bio-Argo profiles of the validation data set, 649 profiles had concomitant VIIRS and MODIS-Aqua products available. The sensitivity of the method to both types of satellite input data is therefore evaluated using these 649 profiles (see statistics in Table 4). Obviously, as the neural network was trained using MODIS Aqua products as input data, the use of VIIRS data slightly reduces the skills of the method (i.e., decrease in the determination coefficient by 0.07). However, the scatterplot of b_{bp_SOCA} (using VIIRS data as input) versus b_{bp_Floats} shows that the data points are still fairly well scattered around the line 1:1 (Figure 7) and the accuracy of the method remains satisfactory when VIIRS data are used as input (Table 4). Finally, SOCA-BBP appears robust to reasonable noise in the input satellite data. Use of VIIRS-derived instead of MODIS-Aqua-derived products yield accurate results (global retrieval error of 21%; Table 4) despite a MAPD in the VIIRS products compared to the MODIS-Aqua products of 44% and 15% for b_{bp} and Chl, respectively.

4.3. Additional Validation of SOCA-BBP Using Independent Data Sets

Time series of the vertical profile of the particulate backscattering coefficient collected from four Bio-Argo profiling floats deployed in several oceanic basins (Southern Ocean, North Atlantic Subtropical Gyre, North Atlantic Subpolar Gyre, and North Western Mediterranean Sea) were removed from the initial database to obtain an independent data set for further validation of SOCA-BBP (see section 2.2). A comparison of b_{bp_SOCA} with the corresponding b_{bp_Floats} in each basin is presented in Figure 5b. This comparison suggests

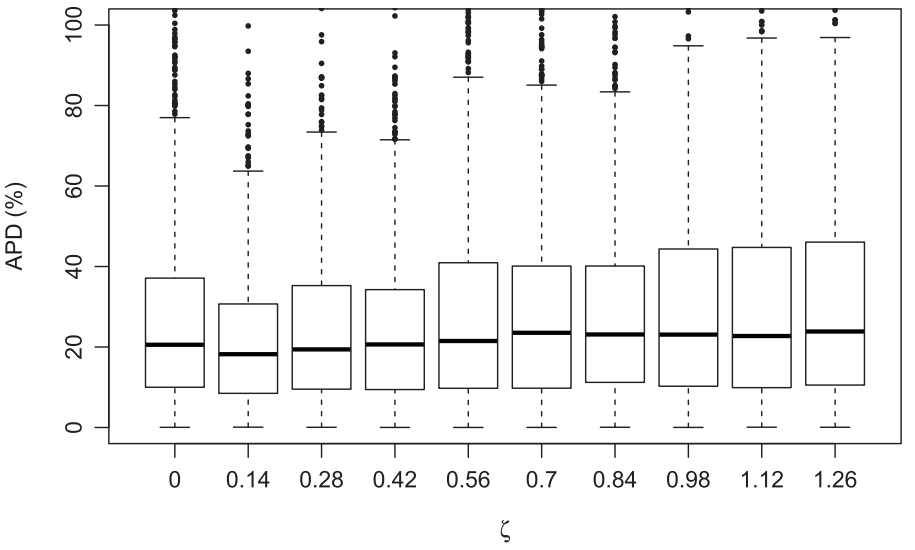

Figure 6. Boxplots of the Absolute Percent Difference, APD (%), between the retrieved b_{bp_SOCA} and the reference b_{bp_Floats}, according to the 10 dimensionless depths, ζ, that are the output of b_{bp_SOCA}. The box represents the upper quartile and the lower quartile with the middle line representing the median of the values. The points represent the outliers.

Table 4. Statistics of the Comparison of the b_{bp} Values Predicted by SOCA-BBP Using MODIS-Aqua or VIIRS-Derived Products as Input[a]

Type of Satellite Products	n	R^2	Slope	MAPD (%)
MODIS-Aqua	6490	0.79	0.79	21
VIIRS	6490	0.72	0.78	21

[a]Determination coefficient (R^2) and slope of the linear regression between the retrieved and reference values. The MAPD (Median Absolute Percent Difference) between the retrieved and reference values is also indicated (see section 3.3 for the calculation details).

that the method has similar accuracy when tested with this independent data set (Table 3) as with the validation data set comprising 20% of the initial database. The skill of the method is slightly reduced in the Mediterranean Sea compared to other areas (increase of MAPD by ~8%).

To obtain fully depth-resolved vertical profiles of b_{bp}, we applied a linear interpolation between each of the 10 b_{bp_SOCA} values provided by the MLP. The resulting predicted time series are compared to their float counterparts for the four basins (Figure 8). We note that the absence of b_{bp_SOCA} data (white bands in Figures 8b, 8d, 8f, and 8h) reflect missing satellite-to-Argo matchups caused by a lack of satellite image in cloudy areas/seasons (i.e., usually high-latitude environments in winter). The vertical patterns of b_{bp} predicted by SOCA-BBP are very consistent with those observed by the profiling floats in the 0–300 m layer. For the North Atlantic Subtropical Gyre, an area with extremely low b_{bp} values, SOCA-BBP reproduces the seasonal deepening of the b_{bp} maximum in spring and early summer and the shoaling in June–July (Figures 8a and 8b). For the Southern Ocean, the b_{bp_SOCA} values are consistent with the Bio-Argo float measurements with respect to an increase of b_{bp} from December to April in the 0–100 m layer (Figures 8c and 8d). In the North Atlantic Subpolar Gyre, the retrieved b_{bp} values agree with float measurements for both years of the time series (i.e., 2013 and 2014, Figures 8e and 8f). SOCA-BBP seems to underestimate b_{bp} in the Mediterranean Sea, especially from July to October when a deep b_{bp} maximum develops at ~50 m (Figures 8g and 8h). This underestimation is possibly caused by the relatively coarse vertical resolution of the SOCA-BBP outputs. The b_{bp} maximum may be missed by the 10 output depths from which the entire vertical profile is derived. Finally, a comparison of b_{bp} values integrated within the 0-Z_m layer estimated from SOCA-BBP and measured by the Bio-Argo floats show good agreement for the four examined areas (R^2 of 0.92 and MAPD of 20%; Figure 9a). Yet this result is not surprising because the four floats used for this additional validation exercise, although not used for the training process, were deployed in the same areas as the floats represented in the training and validation data sets (see Figure 1). The comparison of depth-integrated estimations of b_{bp} allows to smooth errors of estimation due to the noise in the in situ b_{bp} profiles.

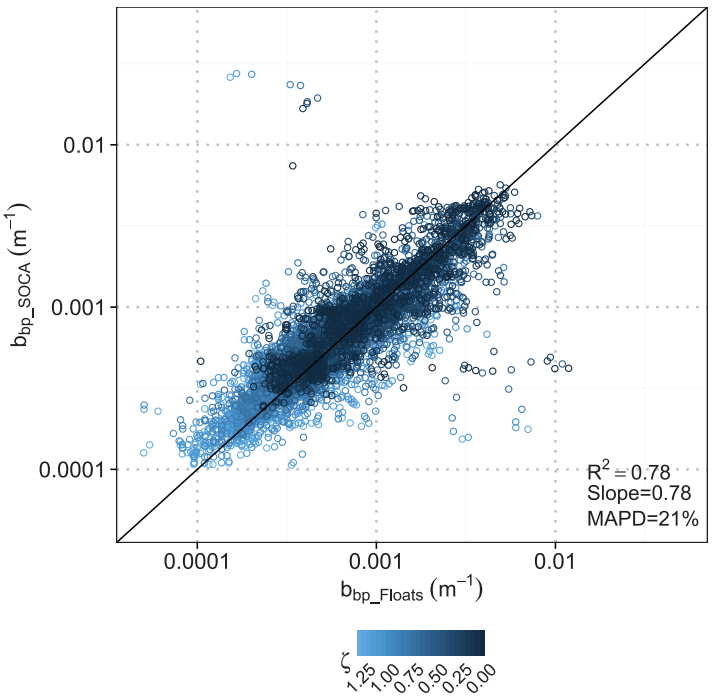

Figure 7. Predictive skills of SOCA-BBP when using as satellite input VIIRS-derived products instead of MODIS-Aqua products. Comparison of b_{bp} retrieved by SOCA-BBP (b_{bp_SOCA}) to reference b_{bp} measurements acquired by Bio-Argo floats (b_{bp_Floats}) with data ordered according to the dimensionless depth ζ. The 1:1 line is represented in black. The calculation details of statistics are provided in section 3.3.

The performance and the good generalization of the method was also evaluated using a totally independent set of data (i.e., different oceanic zone of sampling and different sampling sensors) from a Atlantic Meridional Transect (AMT) cruise conducted in 2009. The geolocation of the 16 profiles of temperature, salinity, and b_{bp} with concomitant MODIS-Aqua-derived products is shown in Figure 1 (dark blue crosses). The

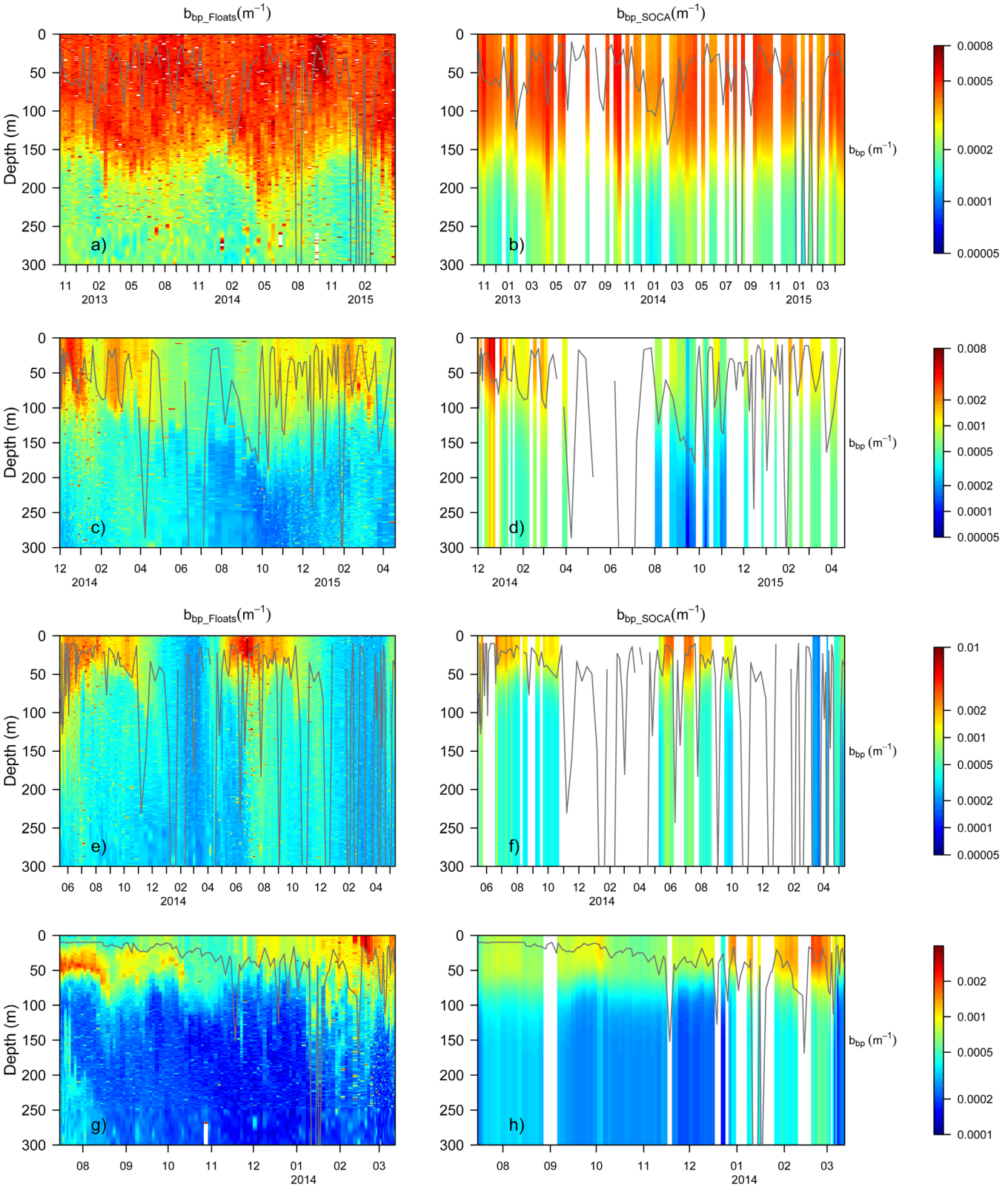

Figure 8. Comparison of the reference b_{bp} measurements acquired by Bio-Argo floats, b_{bp_Floats} (a, c, e, and g), with the values predicted by SOCA-BBP), b_{bp_SOCA} (b, d, f, and h). Time series for the Bio-Argo floats deployed (a and b) in the North Atlantic Subtropical Gyre (WMO = 6901472), (c and d) in the Southern Ocean (WMO = 6901493), (e and f) in the North Atlantic (WMO = 6901523), and (g and h) in the Mediterranean Sea (WMO = 6901496). The WMO numbers are official numbers of the World Meteorological Organization. The location of time series for each float is represented in orange in Figure 1. The grey line in each plot indicates the depth of the mixed layer.

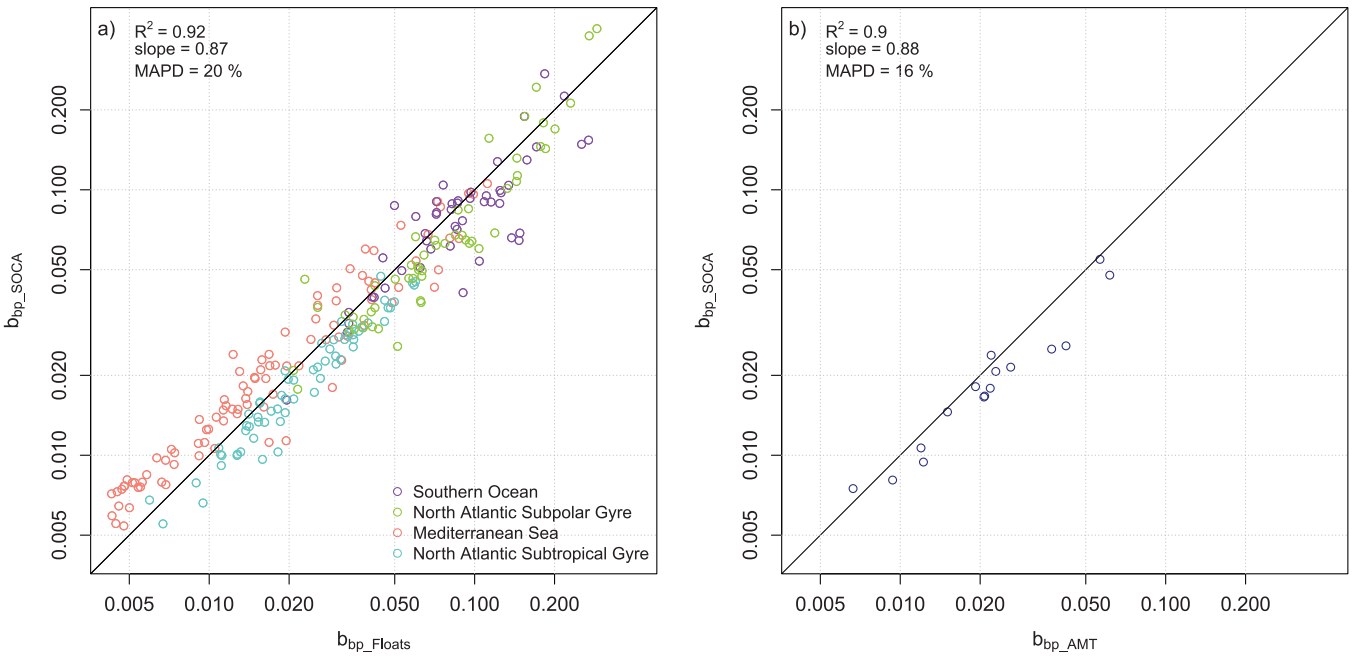

Figure 9. Comparison of b_{bp} integrated within the 0-Z_m layer (scale given in optical thickness) predicted by the SOCA-BBP method (b_{bp_SOCA} dimensionless) and calculated from the reference measurements collected by the Bio-Argo floats (b_{bp_Floats} dimensionless) or during an AMT cruise (b_{bp_AMT} dimensionless). This comparison makes use of reference measurements from two independent data sets: (a) collected by the four Bio-Argo floats not represented in the training and validation databases of the MLP with the color code indicating the float deployment basins; and (b) acquired during an AMT oceanographic cruise in 2009. The identity line is shown in black in each plot. The calculation details of statistics are provided in section 3.3.

SOCA-BBP retrieved b_{bp} values, b_{bp_SOCA}, were compared to the reference b_{bp} ship-based measurements integrated within the 0-Z_m layer, b_{bp_AMT} (Figure 9b). Similar to previous validation tests, the results appear highly satisfactory (i.e., median absolute percent difference of 18%). In addition, this validation exercise based on totally independent set of data demonstrates that SOCA-BBP may be applicable to conditions/regions not included in the training database (e.g., Atlantic Equatorial Zone, see Figure 1).

4.4. Potential Application of the SOCA-BBP Method: Development of Global 3-D Climatologies of b_{bp}

Before the emergence and use of profiling floats equipped with backscattering sensors, vertical profiles of b_{bp} in the ocean were quite scarce and highly heterogeneous. SOCA-BBP provides a way to estimate vertical profiles of b_{bp} using basic ocean color products merged with Argo data. A natural application of this method is the development of depth-resolved b_{bp} climatologies. As an example, we develop a 3-D climatology of b_{bp} for the global ocean for the months of June and December. We use as input satellite-based monthly composites of b_{bp} and Chl. These data are merged with monthly temperature and salinity data from the Argo global climatologies [*Roemmich and Gilson*, 2009]. As the Mediterranean Sea is not represented in the Argo climatology, we use climatological data from the World Ocean Database (WOD) [*Levitus et al.*, 2013] for this basin.

First, we compare the b_{bp_SOCA} surface values with the corresponding satellite estimates, with the b_{bp_SOCA} surface values defined as b_{bp} averaged within the layer comprised between the surface and the penetration depth, $Z_{pd} = Z_e/4.6$ [*Morel and Berthon*, 1989]. This comparison is conducted at a 1°-resolution (i.e., Argo climatology resolution). Overall, the geographical patterns of the SOCA-BBP retrieved surface b_{bp} values for the months of June and December (Figures 10a and 10b) are consistent with those observed by the satellite (Figures 10c and 10d). In June, high values of b_{bp} are consistently recorded in the high-latitude regions of the Northern Hemisphere. Reciprocally, high values of b_{bp} are recorded in the high latitudes of the Southern Hemisphere in December. Unsurprisingly, the equatorial band, the upwelling systems associated with Eastern Boundary Currents and other near-coastal areas (Figure 10a) show high b_{bp} values with weak seasonal variability. SOCA-BBP yields low b_{bp} values compared to satellite-based estimates in the subtropical gyres (Figure 10e). A general bias between the SOCA-retrieved surface b_{bp} values and the corresponding satellite

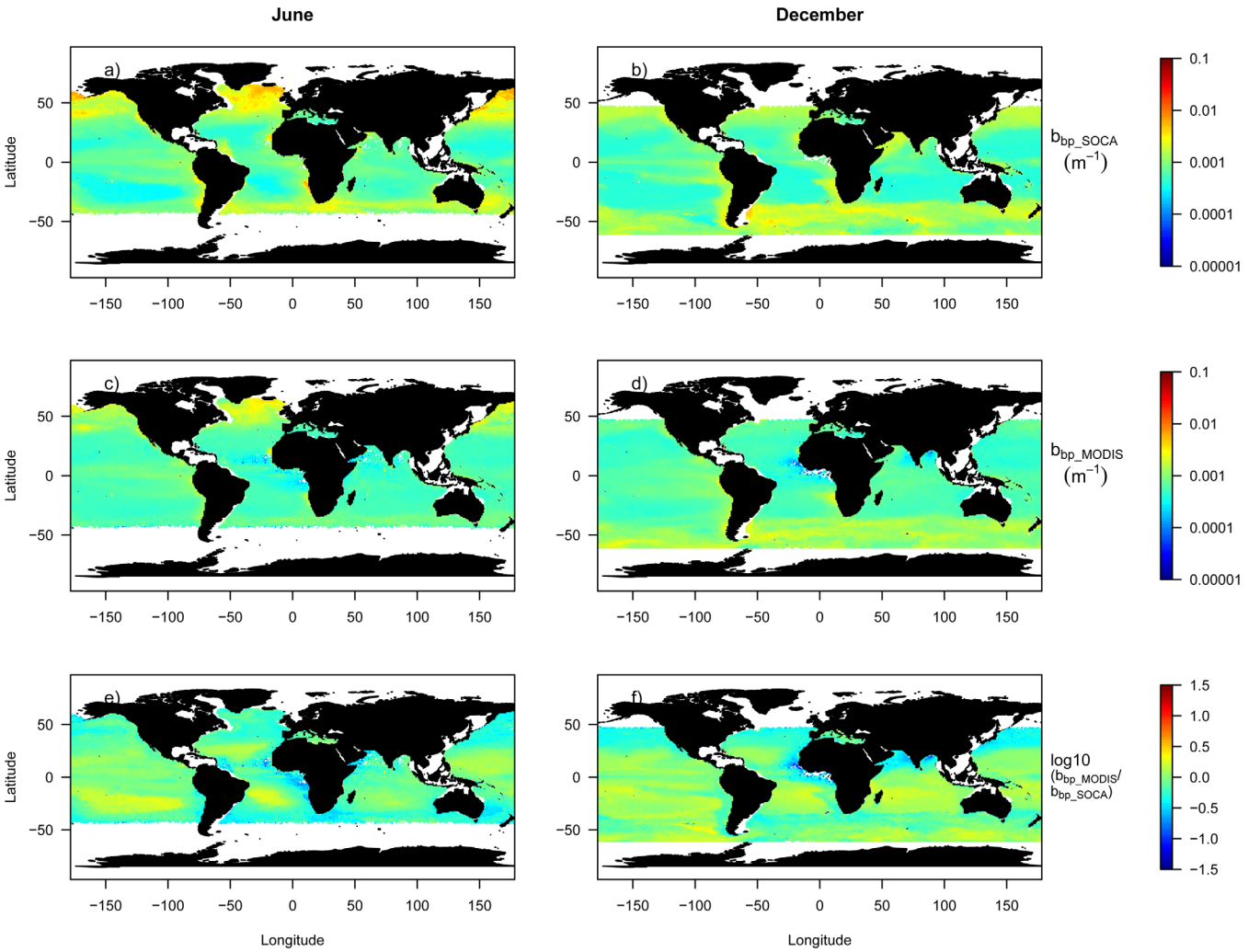

Figure 10. Surface climatology of the particulate backscattering coefficient with a 1° resolution for the month of (left) June and (right) December. (a and b) Surface b_{bp} (i.e., averaged over the 0-Z_{pd} layer) obtained from the SOCA-BBP algorithm; (c and d) MODIS-Aqua-derived estimates of b_{bp}; (e and f) $\log_{10}$ ratio of the satellite-based b_{bp} to the SOCA-BBP-retrieved b_{bp}.

values can be identified in supporting information Figure S5 that shows a comparison of the surface b_{bp} values retrieved by SOCA-BBP and derived from MODIS-Aqua climatological products for the month of June. This observation might result either from a global overestimation of the predicted b_{bp}, or from a global underestimation of b_{bp_MODIS} derived from the QAA model. The SOCA-retrieved b_{bp} versus VIIRS-estimated climatological b_{bp} for the month of June is displayed in supporting information Figure S6 for a comparison with supporting information Figure S5. There is no general bias between the SOCA-BBP retrieved and satellite-derived values using the VIIRS matchup data. This suggests that the bias observed in supporting information Figure S5 is probably caused by an underestimation of b_{bp} using the QAA algorithm with satellite MODIS-Aqua products.

Besides the bias between the two estimates of b_{bp}, b_{bp} levels yield by the model are lower than the satellite-derived values especially in the subtropical gyres (see Figure 10). Using in situ data from the BIOSOPE cruise (Biogeochemistry and Optics South Pacific Experiment) [*Claustre et al.*, 2008], several studies have shown an overestimation of backscattering satellite estimates in the most oligotrophic conditions of the South Pacific Subtropical Gyre [*Brown et al.*, 2008; *Huot et al.*, 2008]. In fact, inherent optical properties are very difficult to estimate from satellite-based measurements in these extremely clear waters and it is now acknowledged that semianalytical algorithms of b_{bp} retrieval from satellite data lead to a systematic overestimation [e.g., *Brown et al.*, 2008; *Lee and Huot*, 2014]. Interestingly, the model improves the retrieval

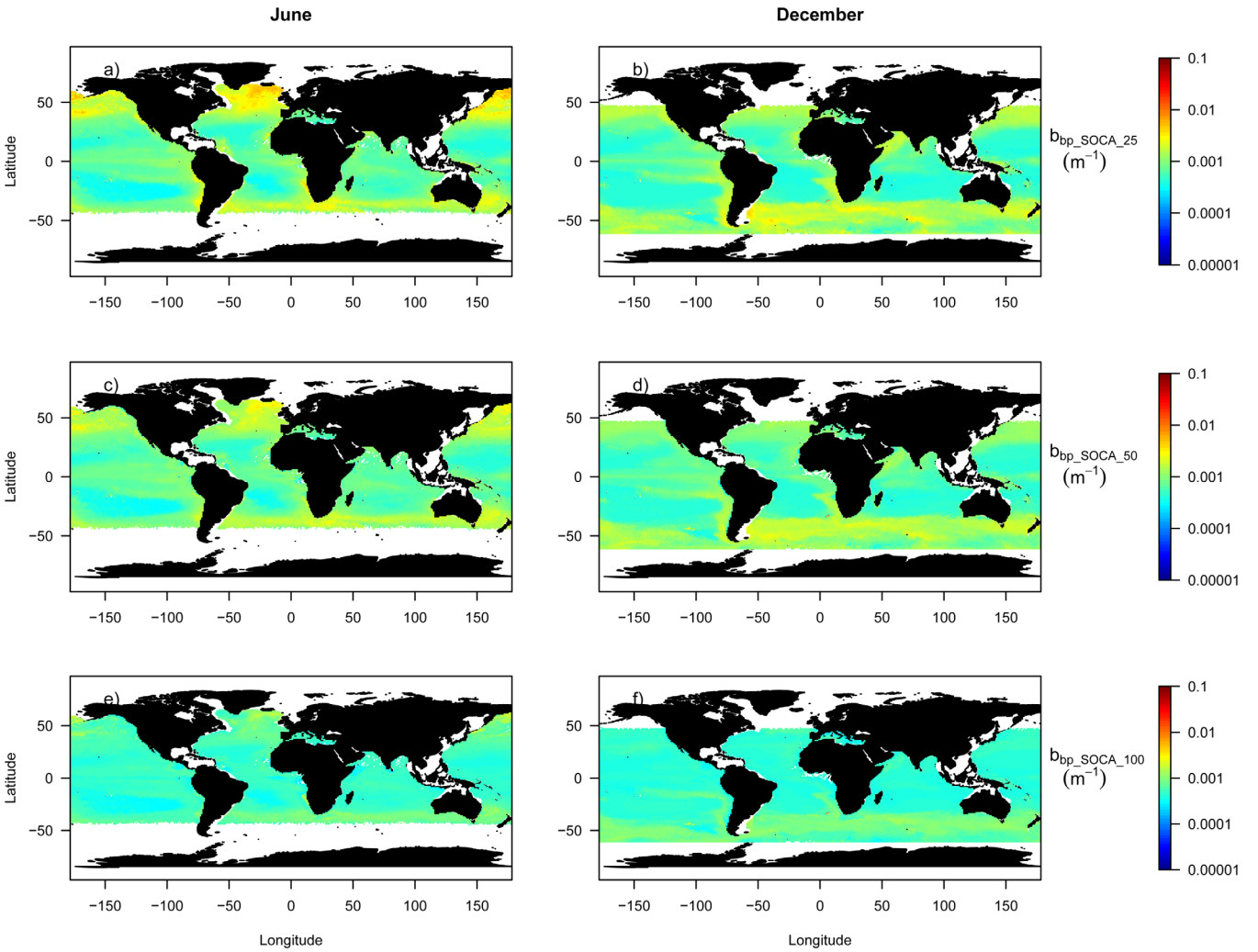

Figure 11. Depth-resolved climatology of the particulate backscattering coefficient with a 1° resolution for the month of (left) June and (right) December. (a and b) $b_{bp\_SOCA\_25}$ at 25 m depth (averaged ±5 m) obtained from the SOCA-BBP algorithm; (c and d) $b_{bp\_SOCA\_50}$ at 50 m depth; (e and f) $b_{bp\_SOCA\_100}$ at 100 m depth.

of the surface b_{bp} in the subtropical gyres (Figure 10 and supporting information Figure S5), even though satellite-based surface b_{bp} estimates used as input show overestimation. This is principally because the learning of the neural network is based on accurate Bio-Argo in situ b_{bp} profiles hence constraining the retrieved b_{bp} surface values. In addition, as the subtropical gyres are characterized by low Chl, the results shown in this study are consistent with the formulations of *Morel and Maritorena* [2001] and *Huot et al.* [2008] that account for a continuous decrease in b_{bp} with decreasing chlorophyll *a* concentrations for low chlorophyll *a* concentrations ($<$0.1 mg m^{-3}) instead of the constant b_{bp} with decreasing chlorophyll *a* concentrations as reported by *Behrenfeld et al.* [2005].

Finally, the vertical distribution of the depth-resolved b_{bp} climatologies is presented in Figure 11. This figure presents sections of b_{bp} for the global ocean at 25, 50, and 100 m depth for the months of June and December. Overall, the decrease of the SOCA-BBP retrieved b_{bp} values with depth is apparent for most of the oceanographic zones expect in most oligotrophic waters (see Figure 11). Indeed, for instance in the pacific south subtropical gyre, the decrease in b_{bp} levels is barely visible (from ∼0.0003 m^{-1} at the surface to ∼0.0002 m^{-1} at 100 m depth in June and from ∼0.0004 m^{-1} at the surface to ∼0.0003 m^{-1} at 100 m depth in December) because b_{bp} values decrease only at ∼150–200 m depth. Rapid decrease of b_{bp} are recorded in the high-latitude regions of the Northern Hemisphere in June (from ∼0.004 m^{-1} at the surface to ∼0.001 m^{-1} at 100 m depth). Reciprocally, rapid decrease of b_{bp} is recorded in the high-latitude regions of the Southern Hemisphere in December (from ∼0.003 m^{-1} at the surface to ∼0.001 m^{-1} at 100 m depth).

As for b_{bp} surface estimates, the vertical distribution of b_{bp} shows weak seasonal variability in the equatorial band, the upwelling systems associated with Eastern Boundary Currents and other near-coastal areas. In Figure 11, b_{bp} levels are shown for only three depths but it is important to note that SOCA-BBP method provides b_{bp} vertical distribution for the whole productive zone (0–1.3 ζ layer). Finally, these depth-resolved b_{bp} climatologies are an invaluable source of information on the vertical distribution of a key bio-optical property at a global scale with the potential to support investigations dedicated to carbon cycle, including carbon production and export.

5. Conclusion and Perspectives

We have demonstrated that, using a Multi-Layered Perceptron method, we can merge ocean color-based products with temperature-salinity Argo data to infer the vertical distribution of a bio-optical property estimated from both satellite and robotic platform measurements. The proposed method, SOCA-BBP, infers the vertical distribution of the particulate backscattering coefficient using three main input components: (1) a surface component, i.e., satellite-derived products; (2) a vertical component derived from temperature and salinity profiles measured by Argo floats; and (3) a temporal component, i.e., the day of the year of the considered satellite-to-Argo matchup. Because the training of the MLP-based method was conducted using a data set representative of the hydrologic and biogeochemical conditions prevailing in the global open ocean, the method is expected to be applicable to most open-ocean environments. Nevertheless, we note that SOCA-BBP has not been developed for applications on a profile-per-profile basis, where a single satellite-to-Argo matchup associated with a specific day would be used to retrieve an "accurate" vertical profile of b_{bp}. Instead SOCC-BBP should be considered as a method dedicated to relatively large-scale applications, such as the development of climatological products (see, e.g., section 4.4).

The natural variability of the vertical distribution of b_{bp} makes the prediction of this bio-optical parameter challenging. Compared to the reference measurements acquired by the Bio-Argo floats from the training data set (i.e., used to establish the SOCA-BBP underlying relationship), the retrieved b_{bp_SOCA} values from the training data set (i.e., used to establish the underlying relationship of the MLP) are retrieved with a median absolute percent difference of 19% (i.e., intrinsic error of the model). Therefore, the error of SOCA-BBP in retrieving the vertical distribution of b_{bp} (i.e., 21%) seems to be essentially induced by the natural variability of b_{bp}. The uncertainties associated with ocean color-based bio-optical products may also generate additional uncertainties in the retrieval of b_{bp}. Our analysis of the sensitivity of SOCA-BBP to the satellite data used as input indicates limited changes in the prediction of b_{bp} as one uses VIIRS products instead of MODIS-Aqua products (Figure 7 and Table 4). This suggests that the MLP is relatively insensitive to reasonable noise levels in the input satellite data because noise is accounted for in the training of the MLP. Based on this sensitivity analysis, we expect that the method can be safely used with satellite products other than those derived from MODIS-Aqua (e.g., SeaWiFS, MERIS, VIIRS, and OLCI) or with merged products (e.g., GlobColour and CCI-OC).

The present study provides an invaluable source of information on the vertical distribution of the b_{bp} allowing this key bio-optical property to be comprehensively described at a global scale. A major application of the method is obviously linked to the creation of a depth-resolved global proxy of POC and, possibly, phytoplankton carbon with high space-time resolution. This is a prerequisite for improving the characterization and quantification of key carbon fluxes such as net primary production or export fluxes. In particular, the data resulting from SOCA-BBP are valuable for the initialization or validation of biogeochemical models. The climatological data retrieved from SOCA-BBP also have the potential to serve as benchmarks against which temporal or regional trends could be evidenced.

Several published relationships link POC to b_{bp} either regionally or at the global scale [*Balch et al*., 2001; *Cetinić et al*., 2012; *Loisel et al*., 2001, 2002; *Stramski et al*., 1999, 2008]. A systematic and routine acquisition of b_{bp} vertical profile has started only recently (a decade ago) so that the number of concurrent b_{bp} and POC measurements for establishing robust regional or global b_{bp}-to-POC relationships is still limited. Obviously, the converted POC or phytoplankton carbon from b_{bp} estimated by SOCA-BBP method will integrate combined errors from SOCA-BBP method of b_{bp} and from the relationships b_{bp}-to-POC or b_{bp}-to-$C_{phyto.}$ Therefore, collecting systematic measurements of POC and b_{bp} is of critical need for refining the previously

published relationships and, ultimately, exploit in an optimal manner the growing b_{bp} data set acquired by Bio-Argo floats and the SOCA-BBP climatological products.

Apart from deriving POC, recent studies have highlighted the potential of b_{bp} as a phytoplankton carbon proxy [*Graff et al*., 2015]. Actually b_{bp} might be a more reliable proxy of phytoplankton carbon than Chl or POC. Hence, using the b_{bp}-to-phytoplankton carbon relationships presently available in the literature in combination with SOCA-BBP, it appears possible to propose global estimates of the vertical distribution of phytoplankton carbon with high space-time resolution. A potential consequence of obtaining improved estimates of the phytoplankton biomass is a possible reassessment of the sources of variability in the Chl. Using phytoplankton carbon estimates derived from satellite-based data of $b_{bp,}$ some studies have indeed shown that temporal changes in Chl over large oceanic regions may be predominantly caused by physiologically driven modifications in the cellular Chl-to-carbon ratio rather than by actual changes in phytoplankton biomass [*Behrenfeld et al*., 2005, 2009; *Mignot et al*., 2014; *Siegel et al*., 2013]. The combination of SOCA-BBP with other methods, which infer the vertical distribution of Chl from space [e.g., *Uitz et al*., 2006], could permit the variability in the phytoplankton carbon-to-Chl relationship to be examined over the vertical dimension. This would represent a significant step toward a better understanding of light and nutrient control of phytoplankton biomass and physiological status, a prerequisite for improving the characterization of the distribution and variability in primary production and carbon export.

Along with the progressive development of a global Bio-Argo program and associated float deployments, additional measurements of concurrent density and b_{bp} profiles will help to improve the relationship established in the MLP. This is especially expected for the regions currently undersampled in the Bio-Argo database used in this study (e.g., Indian Ocean Gyre and Arctic Ocean). It is indeed important to stress out the evolving aspect of this database and of the quality of the products that can be retrieved from it. This study has shown that neural network-based methods can link the vertical distribution of a given bio-optical property (i.e., particulate backscattering coefficient) to the corresponding near-surface value merged with vertically resolved physical properties. The development of analogous methods for other bio-optical properties, measured from both Bio-Argo floats and ocean color satellites (e.g., chlorophyll *a* concentration, CDOM), appears as a natural extension of the present study.

Acknowledgments
This paper is a contribution to the Remotely Sensed Biogeochemical Cycles in the Ocean (remOcean) project, funded by the European Research Council (grant agreement 246777), to the ATLANTOS EU project (grant agreement 2014-633211) funded by H2020 program, to the French Bio-Argo project funded by CNES-TOSCA, and to the French "Equipement d'avenir" NAOS project (ANR J11R107-F). This study is a contribution to the international IMBER project and was supported by the UK Natural Environment Research Council National Capability funding to Plymouth Marine Laboratory and the National Oceanography Centre, Southampton. This is contribution number 276 of the AMT program. G.D.O. acknowledges funding from the UK National Centre for Earth Observation. Density and particulate backscattering coefficient profiles acquired by Bio-Argo floats used in this study are freely available (http://www.coriolis.eu.org). Temperature and salinity data were collected and made freely available by the International Argo Program and the national programs that contribute to it (http://www.argo.ucsd.edu, http://argo.jcommops.org). The Argo Program is part of the Global Ocean Observing System. The authors deeply acknowledge the NASA for MODIS-Aqua and VIIRS imagery. The International Argo Program and the World Ocean Atlas are acknowledged for making freely available data that have allowed computation of the temperature and salinity climatologies used for our application. We are grateful to an anonymous reviewer for his valuable comments and suggestions.

References

Balch, W. M., D. T. Drapeau, J. J. Fritz, B. C. Bowler, and J. Nolan (2001), Optical backscattering in the Arabian Sea—Continuous underway measurements of particulate inorganic and organic carbon, *Deep Sea Res., Part I*, *48*(11), 2423–2452, doi:10.1016/S0967-0637(01)00025-5.

Behrenfeld, M. J., E. Boss, D. A. Siegel, and D. M. Shea (2005), Carbon-based ocean productivity and phytoplankton physiology from space, *Global Biogeochem. Cycles*, *19*, GB1006, doi:10.1029/2004GB002299.

Behrenfeld, M. J., et al. (2009), Satellite-detected fluorescence reveals global physiology of ocean phytoplankton, *Biogeosciences*, *6*(5), 779–794.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford Univ. Press, Cambridge, U. K.

Bishop, J. K. B. (2009), *Autonomous Observations of the Ocean Biological Carbon Pump*, Lawrence Berkeley Natl. Lab. [Available at http://escholarship.org/uc/item/1q5530cp.]

Bishop, J. K. B., and T. J. Wood (2009), Year-round observations of carbon biomass and flux variability in the Southern Ocean, *Global Biogeochem. Cycles*, *23*, GB2019, doi:10.1029/2008GB003206.

Boss, E., and M. Behrenfeld (2010), In situ evaluation of the initiation of the North Atlantic phytoplankton bloom, *Geophys. Res. Lett.*, *37*, L18603, doi:10.1029/2010GL044174.

Boss, E., and W. S. Pegau (2001), Relationship of light scattering at an angle in the backward direction to the backscattering coefficient, *Appl. Opt.*, *40*(30), 5503–5507, doi:10.1364/AO.40.005503.

Boss, E., D. Swift, L. Taylor, P. Brickley, R. Zaneveld, S. Riser, M. Perry, and P. Strutton (2008), Observations of pigment and particle distributions in the western North Atlantic from an autonomous float and ocean color satellite, *Limnol. Oceanogr.*, *53(5)*, 2112–2122.

Bricaud, A., C. Mejia, D. Blondeau-Patissier, H. Claustre, M. Crepon, and S. Thiria (2007), Retrieval of pigment concentrations and size structure of algal populations from their absorption spectra using multilayered perceptrons., *Appl. Opt.*, *46*(8), 1251–1260.

Brown, C., Y. Huot, P. WerdellL, B. Gentili, and H. Claustre (2008), The origin and global distribution of second order variability in satellite ocean color and its potential applications to algorithm development, *Remote Sens. Environ.*, *112*(12), 4186–4203, doi:10.1016/j.rse.2008.06.008.

Cetinić, I., M. J. Perry, N. T. Briggs, E. Kallin, E. A. D'Asaro, and C. M. Lee (2012), Particulate organic carbon and inherent optical properties during 2008 North Atlantic Bloom Experiment, *J. Geophys. Res.*, *117*, C06028, doi:10.1029/2011JC007771.

Claustre, H., A. Sciandra, and D. Vaulot (2008), Introduction to the special section Bio-optical and biogeochemical conditions in the South East Pacific in late 2004: The BIOSOPE program, *Biogeosci. Discuss.*, *5*(1), 605–640.

Claustre, H., et al. (2010a), Bio-optical profiling floats as new observational tools for biogeochemical and ecosystem studies: Potential synergies with ocean color remote sensing, in *Proceedings of the OceanObs 09: Sustained Ocean Observations and Information for Society Conference*, vol. 2, edited by J. Hall, D. E. Harrison, and D. Stammer, ESA Publ., Venice, Italy. [Available at http://www.obs-vlfr.fr/LOV/OMT/fichiers_PDF/Claustre_et_al_OceanObs_cwp_10.pdf.]

Claustre, H., et al. (2010b), Guidelines towards an integrated ocean observation system for ecosystems and biogeochemical cycles, in *Proceedings of the OceanObs 09: Sustained Ocean Observations and Information for Society Conference*, vol. 1, edited by J. Hall, D. E. Harrison, and D. Stammer, ESA Publ., Venice, Italy. [Available at http://www.obs-vlfr.fr/LOV/OMT/fichiers_PDF/Claustre_et_al_OceanObs_10.pdf.]

Dall'Olmo, G., and K. A. Mork (2014), Carbon export by small particles in the Norwegian Sea, *Geophys. Res. Lett.*, *41*, 2921–2927, doi:10.1002/2014GL059244.

De Boyer Montégut, C., G. Madec, A. S. Fischer, A. Lazar, and D. Iudicone (2004), Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology, *J. Geophys. Res.*, *109*, C12003, doi:10.1029/2004JC002378.

Duforêt-Gaurier, L., H. Loisel, D. Dessailly, K. Nordkvist, and S. Alvain (2010), Estimates of particulate organic carbon over the euphotic depth from in situ measurements. Application to satellite data over the global ocean, *Deep Sea Res., Part I*, *57*(3), 351–367, doi:10.1016/j.dsr.2009.12.007.

Falkowski, P. G., T. R. Barber, and V. Smetacek (1998), Biogeochemical controls and feedbacks on ocean primary production, *Science*, *281*(5374), 200–206, doi:10.1126/science.281.5374.200.

Friedrich, T., and A. Oschlies (2009), Neural network-based estimates of North Atlantic surface pCO_2 from satellite data: A methodological study, *J. Geophys. Res.*, *114*, C03020, doi:10.1029/2007JC004646.

Gardner, W. D., A. V. Mishonov, and M. J. Richardson (2006), Global POC concentrations from in-situ and satellite data, *Deep Sea Res., Part II*, *53*(5–7), 718–740, doi:10.1016/j.dsr2.2006.01.029.

Graff, J. R., T. K. Westberry, A. J. Milligan, M. B. Brown, G. Dall'Olmo, V. van Dongen-Vogels, K. M. Reifel, and M. J. Behrenfeld (2015), Analytical phytoplankton carbon measurements spanning diverse ecosystems, *Deep Sea Res., Part I*, *102*, 16–25, doi:10.1016/j.dsr.2015.04.006.

Gross, L., S. Thiria, R. Frouin, and B. G. Mitchell (2000), Artificial neural networks for modeling the transfer function between marine reflectance and phytoplankton pigment concentration, *J. Geophys. Res.*, *105*(C2), 3483, doi:10.1029/1999JC900278.

Hornik, K., M. Stinchcombe, and H. White (1989), Multilayer feedforward networks are universal approximators, *Neural Networks*, *2*, 359–366.

Huot, Y., A. Morel, M. S. Twardowski, D. Stramski, and R. A. Reynolds (2008), Particle optical backscattering along a chlorophyll gradient in the upper layer of the eastern South Pacific Ocean, *Biogeosciences*, *5*(2), 495–507.

Jamet, C., H. Loisel, and D. Dessailly (2012), Retrieval of the spectral diffuse attenuation coefficient K d (λ) in open and coastal ocean waters using a neural network inversion, *J. Geophys. Res.*, *117*, C10023, doi:10.1029/2012JC008076.

Kokhanovsky, A. A. (2012), *Light Scattering Reviews 7: Radiative Transfer and Optical Properties of Atmosphere and Underlying Surface*. Springer Science & Business Media, Chichester, UK.

Kostadinov, T. S., D. A. Siegel, and S. Maritorena (2010), Global variability of phytoplankton functional types from space: Assessment via the particle size distribution, *Biogeosciences*, *7*(10), 3239–3257, doi:10.5194/bg-7-3239-2010.

Krasnopolsky, V. M. (2009), Neural network applications to solve forward and inverse problems in atmospheric and oceanic satellite remote sensing, in *Artificial Intelligence Methods in the Environmental Sciences*, pp. 191–205, Springer, Amsterdam.

Lee, Z., and Y. Huot (2014), On the non-closure of particle backscattering coefficient in oligotrophic oceans., *Opt. Express*, *22*(23), 29,223–29,233, doi:10.1364/OE.22.029223.

Lee, Z., K. L. Carder, and R. A. Arnone (2002), Deriving inherent optical properties from water color: A multiband quasi-analytical algorithm for optically deep waters, *Appl. Opt.*, *41*(27), 5755–5772, doi:10.1364/AO.41.005755.

Lek, S., and J. F. Guégan (1999), Artificial neural networks as a tool in ecological modelling, an introduction, *Ecol. Modell.*, *120*(2–3), 65–73, doi:10.1016/S0304-3800(99)00092-7.

Levitus, S., et al. (2013), The World Ocean Database, *Data Sci. J.*, *12*, WDS229–WDS234.

Loisel, H., E. Bosc, D. Stramski, K. Oubelkheir, and P.-Y. Deschamps (2001), Seasonal variability of the backscattering coefficient in the Mediterranean Sea based on satellite SeaWiFS imagery, *Geophys. Res. Lett.*, *28*(22), 4203–4206, doi:10.1029/2001GL013863.

Loisel, H., J.-M. Nicolas, P.-Y. Deschamps, and R. Frouin (2002), Seasonal and inter-annual variability of particulate organic matter in the global ocean, *Geophys. Res. Lett.*, *29*(24), 2196, doi:10.1029/2002GL015948.

Loisel, H., J.-M. Nicolas, A. Sciandra, D. Stramski, and A. Poteau (2006), Spectral dependency of optical backscattering by marine particles from satellite remote sensing of the global ocean, *J. Geophys. Res.*, *111*, C09024, doi:10.1029/2005JC003367.

Martinez-Vicente, V., G. Dall'Olmo, G. Tarran, E. Boss, and S. Sathyendranath (2013), Optical backscattering is correlated with phytoplankton carbon across the Atlantic Ocean, *Geophys. Res. Lett.*, *40*, 1154–1158, doi:10.1002/grl.50252.

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

Morel, A., and J.-F. Berthon (1989), Surface pigments, algal biomass profiles, and potential production of the euphotic layer: Relationships reinvestigated in view of remote-sensing applications, *Limnol. Oceanogr.*, *34*(8), 1545–1562.

Morel, A., and S. Maritorena (2001), Bio-optical properties of oceanic waters: A reappraisal, *J. Geophys. Res.*, *106*(C4), 7163, doi:10.1029/2000JC000319.

Niang, A., F. Badran, C. Moulin, M. Crépon, and S. Thiria (2006), Retrieval of aerosol type and optical thickness over the Mediterranean from SeaWiFS images using an automatic neural classification method, *Remote Sens. Environ.*, *100*(1), 82–94, doi:10.1016/j.rse.2005.10.005.

Palacz, A. P., M. A. St. John, R. J. W. Brewin, T. Hirata, and W. W. Gregg (2013), Distribution of phytoplankton functional types in high-nitrate, low-chlorophyll waters in a new diagnostic ecological indicator model, *Biogeosciences*, *10*(11), 7553–7574, doi:10.5194/bg-10-7553-2013.

Raitsos, D. E., S. J. Lavender, C. D. Maravelias, J. Haralabous, A. J. Richardson, and P. C. Reid (2008), Identifying four phytoplankton functional types from space: An ecological approach, *Limnol. Oceanogr.*, *53*(2), 605–613.

Rochford, P. A., A. B. Kara, A. J. Wallcraft, and R. A. Arnone (2001), Importance of solar subsurface heating in ocean general circulation models, *J. Geophys. Res.*, *106*(C12), 30,923–30,938, doi:10.1029/2000JC000355.

Roemmich, D., and J. Gilson (2009), The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program, *Prog. Oceanogr.*, *82*(2), 81–100, doi:10.1016/j.pocean.2009.03.004.

Roemmich, D., G. Johnson, S. Riser, R. Davis, J. Gilson, W. B. Owens, S. Garzoli, C. Schmid, and M. Ignaszewski (2009), The Argo program: Observing the global oceans with profiling floats, *Oceanography*, *22*(2), 34–43, doi:10.5670/oceanog.2009.36.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1988), Learning representations by back-propagating errors, *Nature*, *323*, 533–536.

Sauzède, R., H. Claustre, C. Jamet, J. Uitz, J. Ras, A. Mignot, and F. D'Ortenzio (2015), Retrieving the vertical distribution of chlorophyll a concentration and phytoplankton community composition from in situ fluorescence profiles: A method based on a neural network with potential for global-scale applications, *J. Geophys. Res. Oceans*, *120*(1), 451–470, doi:10.1002/2014JC010355.

Siegel, D. A., S. Maritorena, N. B. Nelson, and M. J. Behrenfeld (2005), Independence and interdependencies among global ocean color properties: Reassessing the bio-optical assumption, *J. Geophys. Res.*, *110*, C07011, doi:10.1029/2004JC002527.

Siegel, D. A., et al. (2013), Regional to global assessments of phytoplankton dynamics from the SeaWiFS mission, *Remote Sens. Environ.*, *135*, 77–91, doi:10.1016/j.rse.2013.03.025.

Stramska, M. (2009), Particulate organic carbon in the global ocean derived from SeaWiFS ocean color, *Deep Sea Res., Part I*, *56*(9), 1459–1470, doi:10.1016/j.dsr.2009.04.009.

Stramski, D., R. A. Reynolds, M. Kahru, and B. G. Mitchell (1999), Estimation of particulate organic carbon in the ocean from satellite remote sensing, *Science*, *285*(5425), 239–242, doi:10.1126/science.285.5425.239.

Stramski, D., et al. (2008), Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern South Pacific and eastern Atlantic Oceans, *Biogeosciences*, *5*(1), 171–201.

Sullivan, J. M., and M. S. Twardowski (2009), Angular shape of the oceanic particulate volume scattering function in the backward direction., *Appl. Opt.*, *48*(35), 6811–6819, doi:10.1364/AO.48.006811.

Telszewski, M., et al. (2009), Estimating the monthly pCO_2 distribution in the north Atlantic using a self-organizing neural network, *Biogeosciences*, *6*, 1405–1421.

Uitz, J., H. Claustre, A. Morel, and S. B. Hooker (2006), Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll, *J. Geophys. Res.*, *111*, C08005, doi:10.1029/2005JC003207.

Volk, T., and M. I. Hoffert (1985), Ocean carbon pumps: Analysis of relative strengths and efficiencies in ocean-driven atmospheric CO_2 changes, in *The Carbon Cycle and Atmospheric CO2: Natural Variations Archean to Present, Geophys. Monogr. Ser.*, edited by E. T. Sundquist and W. S. Broecker, pp. 99–110, AGU, Washington, D. C. [Available at http://adsabs.harvard.edu/abs/1985GMS. . ..32. . .99V.]

Westberry, T., M. J. Behrenfeld, D. A. Siegel, and E. Boss (2008), Carbon-based primary productivity modeling with vertically resolved photoacclimation, *Global Biogeochem. Cycles*, *22*, GB2024, doi:10.1029/2007GB003078.

Wong, A., R. Keeley, T. Carval, and the Argo Data Management Team (2015), Argo Quality Control Manual for CTD and Trajectory Data, doi:10.13155/33951.

Xing, X., H. Claustre, J. Uitz, A. Mignot, A. Poteau, and H. Wang (2014), Seasonal variations of bio-optical properties and their interrelationships observed by Bio-Argo floats in the subpolar North Atlantic, *J. Geophys. Res. Oceans*, *119*, 7372–7388, doi:10.1002/2014JC010189.

Zhang, X., L. Hu, and M.-X. He (2009), Scattering by pure seawater: Effect of salinity, *Opt. Express*, *17*(7), 5698, doi:10.1364/OE.17.005698.