



**HAL**  
open science

# NEW GLOBAL VERTICAL DISTRIBUTION OF GRIDDED PARTICULATE ORGANIC CARBON AND CHLOROPHYLL-A CONCENTRATION USING MACHINE LEARNING FOR CMEMS

Raphaëlle Sauzède, Hervé Claustre, Remanan Pannimpullath, Julia Uitz,  
Stéphanie Guinehut

► **To cite this version:**

Raphaëlle Sauzède, Hervé Claustre, Remanan Pannimpullath, Julia Uitz, Stéphanie Guinehut. NEW GLOBAL VERTICAL DISTRIBUTION OF GRIDDED PARTICULATE ORGANIC CARBON AND CHLOROPHYLL-A CONCENTRATION USING MACHINE LEARNING FOR CMEMS. 9th EuroGOOS International conference, Shom; Ifremer; EuroGOOS AISBL, May 2021, Brest, France. hal-03335370v1

**HAL Id: hal-03335370**

**<https://hal.science/hal-03335370v1>**

Submitted on 6 Sep 2021 (v1), last revised 24 Sep 2021 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# NEW GLOBAL VERTICAL DISTRIBUTION OF GRIDDED PARTICULATE ORGANIC CARBON AND CHLOROPHYLL-A CONCENTRATION USING MACHINE LEARNING FOR CMEMS

Raphaëlle Sauzède <sup>(1)</sup>, Hervé Claustre <sup>(2)</sup>, Pannimpullath Remanan Renosh <sup>(2)</sup>, Julia Uitz <sup>(2)</sup> and Stéphanie Guinehut <sup>(2)</sup>

<sup>(1)</sup> Institut de la Mer de Villefranche, CNRS-INSU, Sorbonne Université, Villefranche-Sur-Mer, France

<sup>(2)</sup> Laboratoire d'Océanographie de Villefranche, CNRS-INSU, Sorbonne Université, Villefranche-Sur-Mer, France

<sup>(3)</sup> Collecte Localisation Satellites, Environment and Climate, Ramonville Saint-Agne, France

## Abstract

As part of Copernicus Marine Environmental Monitoring Service (CMEMS), the multi-observations thematic assembly center aims to provide products based on observations and data fusion techniques (Guinehut *et al.*, 2021). Sauzède *et al.*, (2016) have demonstrated the potential of using hydrological measurements and ocean color satellite observations to infer the vertical distribution of backscattering coefficient, a proxy for the stock of particulate organic carbon (POC). The 'Satellite Ocean-Color merged with Argo data to infer bio-optical properties to depth' (SOCA) method is a neural-network-based method trained using the Biogeochemical-Argo database. SOCA has been upgraded to improve the POC retrieval and additionally retrieve the chlorophyll-a concentration (Chl). Using this method with CMEMS hydrological and satellite products, weekly 3-dimensional fields of POC and associated uncertainty were retrieved for the 1998-2018 period and made available from the CMEMS online portal since July 2020. The 3-dimensional products of SOCA-retrieved Chl will be made available by the end of 2021. Both of these products will be updated yearly as new input data become available. These new CMEMS products represent a most valuable source of data useful not only for supporting the quality control of Biogeochemical-Argo float observations but also for data assimilation and initialization/validation of biogeochemical models.

**Keywords:** Particulate organic carbon, Chlorophyll-a concentration, Machine learning, Multi-observations, CMEMS

## 1. Introduction

Phytoplankton is a central component as well as a regulator of marine biogeochemical cycles and ecosystems. It indeed largely influences the magnitude of the so-called biological carbon pump, which results from the sinking and sequestration to the deep oceans of part of the photosynthetically produced Particulate Organic Carbon (POC) stock. Furthermore, phytoplankton is the first link of the marine food chains with the magnitude of photosynthesis being an important driver of fish standing stocks. A better mechanistic understanding of phytoplankton dynamics with respect to its fate (food chain, export at depth) is essential in the problematic of societal relevance like climate change or fisheries management.

Climate change (i.e. warming, CO<sub>2</sub> uptake), which impacts on surface physico-chemical forcing (i.e. enhanced stratification, reduced upward nutrient fluxes, acidification), deeply influences the magnitude of phytoplankton photosynthesis, the phytoplankton communities and hence marine ecosystems. The development of predictive modeling capabilities to quantitatively address future climate change effects on carbon sequestration or on fisheries management requires a better understanding of key biogeochemical processes in the upper oceanic layer.

For a long time, most *in situ* measurements used for the characterization of upper ocean biogeochemical processes were acquired through ship-based observations with the obvious consequence of critical under-sampling and associated observational gaps. Some of these measurements can now be addressed remotely/autonomously either through *in situ* robots or ocean color satellites. In particular, the recently launched Biogeochemical-Argo (BGC-Argo) program (Claustre *et al.*, 2020) opens the new perspective of rapidly acquiring bio-optical proxies of key biogeochemical variables at pertinent observation scales.

POC can be addressed through the estimation of the particulate backscattering coefficient (bbp) (e.g. Cetinić *et al.*, 2012), which is an ocean color product as well as a reference measurement of the BGC-Argo program. Chlorophyll-a concentration (Chl), which is a key proxy for phytoplankton biomass can be also measured by both ocean color remote sensing and autonomous measurements of fluorescence, a proxy for Chl. Combining Chl and POC both available from satellite and float observations offers great potential for deriving 3D/4D fields of biogeochemical products. A machine learning-based method has already been developed with this aim, the so-called SOCA2016-BBP method (for Satellite Ocean Color merged with Argo data to infer the vertical distribution of bbp; Sauzède *et al.*, 2016).

The success of SOCA2016-BBP motivated the effort to create depth-resolved global proxy of POC with higher space-time resolution, a prerequisite for improving the characterization and quantification of export carbon fluxes. Thus, in the context of the European Copernicus Marine Environment Monitoring Service (CMEMS), one of the challenges of the MULTIOBS Thematic Assembly Center (TAC) was dedicated to improve SOCA2016-BBP to obtain high-level 4D gridded global products of POC (with associated estimation errors), and to develop SOCA for Chl in support of biogeochemical model data requirements. In particular, the resulting Chl products benefit the biogeochemical modeling community for the initialization and validation of biogeochemical models.

## 2. Machine learning-based methods to infer the vertical distribution of biogeochemical variables

### 2.1 Particulate Organic Carbon

SOCA2016-BBP proposed a novel method that merges satellite ocean color bio-optical products with Argo temperature-salinity profiles to infer the vertical distribution of  $b_{\text{pp}}$  (Sauzède *et al.*, 2016). This neural-network-based method used three main input components: 1) ocean color satellite-based surface estimates of bbp and Chl matched up in space and time with 2) depth-resolved physical properties derived from temperature-salinity profiles measured by Argo profiling floats (i.e. 4 density values over the euphotic layer and the Mixed Layer Depth, MLD) and 3) the day of the year of the considered satellite-Argo matchup. The developed neural network, more specifically a multi-layer perceptron (MLP), was trained and validated using a database including ~5,000 simultaneous profiles of temperature-salinity and bio-optical properties collected by BGC-Argo floats, with concomitant satellite-derived products.

Recently, the SOCA2016 method has been upgraded to SOCA2020 to obtain the CMEMS high level 3D gridded products. Specifically, SOCA2020 includes several improvements: (1) SOCA2020 relies on a much larger BGC-Argo database than SOCA2016, owing to the new data acquired since 2016 (i.e. 6 times more data and a dataset with less geographical bias of sampling compared to SOCA2016); (2) it includes additional input information such as the satellite-based sea level anomaly (SLA) which gives information about mesoscale processes possibly impacting the vertical distribution of phytoplankton biomass (e.g. McGillicuddy, 2016); (3) it uses as input ocean color remote sensing reflectance (Rrs) at five wavelengths instead of the satellite-based Chl and bbp used in SOCA2016, in order to avoid additional errors due to ocean color algorithms; (4) the representation of the temperature-salinity inputs has been improved by replacing the 4 values of density in the euphotic layer by the first most significant components of the Principal Component Analysis (PCA) of the whole temperature and salinity profiles; (5) the vertical resolution of the  $b_{\text{pp}}$  output products

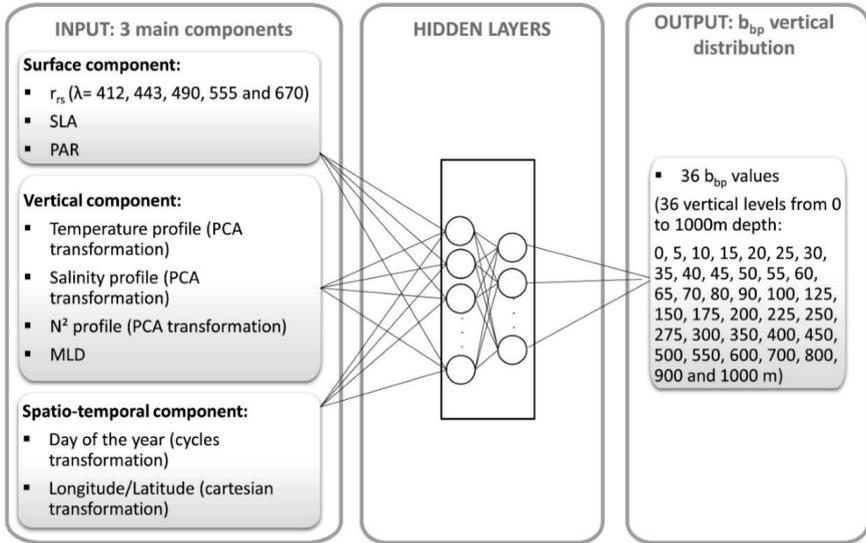


Fig. 1. Schematic overview of the SOCA2020 multi-layer perceptron that retrieves the vertical distribution of  $b_{bp}$  from merged satellite and Argo data associated with the longitude/latitude/day of the year of the considered satellite-to-Argo matchup.

is refined to 36 depth levels from surface down to 1000 m (determined by the 36 depth levels of the CMEMS ARMOR3D physical product used as input) instead of the 10 depth levels of SOCA2016-BBP solely focusing on the euphotic layer; and (6) the performance of the neural network has been improved by using an ensemble of the ten best MLPs (with different topologies, i.e. number of neurons in each hidden layer) which also allows to obtain uncertainties associated to the neural network estimations.

After multiple tests, the optimal architecture selected for the SOCA2020 MLP is composed of 27 inputs (7 surface inputs, 15 vertical inputs and 5 spatio-temporal inputs) and 36 outputs (see the schematic overview of SOCA2020 MLP in Figure 1). The day of the year is converted into radians, with the corresponding sinus and cosinus used as inputs so as to capture the cyclic behavior of this variable. The longitude and latitude were converted into Cartesian (x,y,z) inputs to better accommodate the Euclidean-centric distance calculation.

The performance of the neural network-based method is evaluated using an independent validation dataset corresponding to 20% of the initial SOCA2020 database chosen randomly. The comparison between SOCA2020  $b_{bp}$  retrieval against the BGC-Argo  $b_{bp}$  reference measurements is shown in Figure 2. SOCA2020 infers the vertical distribution of  $b_{bp}$  with a global Median Absolute Percent Difference (MAPD) error of 11% and without bias according to depth.

A dedicated POC vs.  $b_{bp}$  measured at 700 nm (the reference wavelength of  $b_{bp}$  measurements from BGC-Argo floats) relationship for the global ocean has been specifically developed for the second release of our CMEMS product (in May 2021). For this purpose, we used an *in situ* POC database matched up with satellite  $b_{bp}$  data from Evers-King *et al.*, (2017) which gathers 8,318 samples for the global ocean. For each station and depth for which POC was available in this database, corresponding  $b_{bp}$  values have been derived from SOCA2020. Then, different types of POC vs  $b_{bp}$  relationship have been tested (i.e. simple linear, multi-linear relationship, and exponential relationships). The best relationship that was found is an exponential relationship:  $POC = 38687 * b_{bp}^{0.95}$ . Finally, our relationship retrieves POC for the global ocean and for all seasons with a MAPD of 21%.

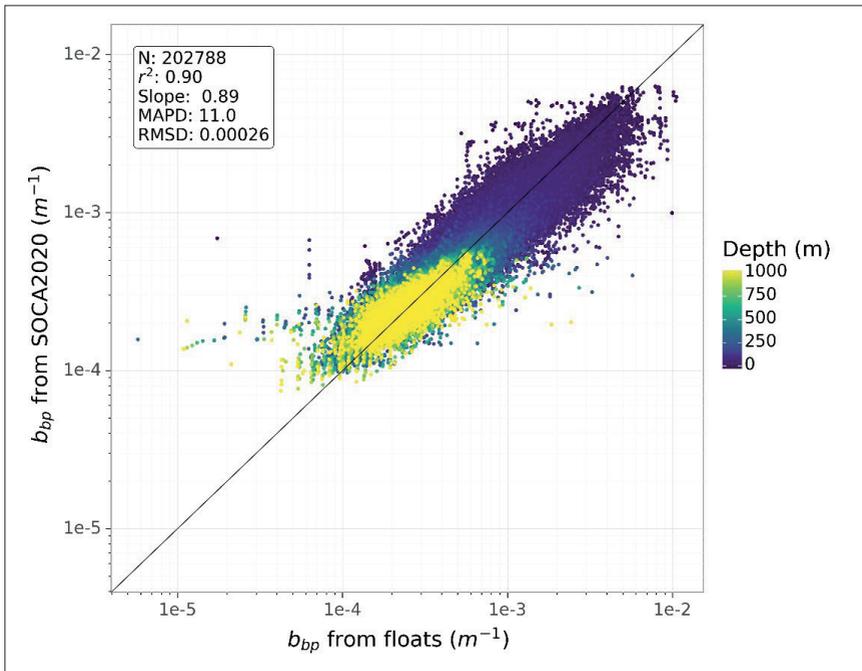


Fig. 2. Comparison of the  $b_{bp}$  values retrieved by SOCA2020 to the reference measurements acquired by BGC-Argo profiling floats from the independent validation dataset.

## 2.2 Chlorophyll-a concentration

In the same way as we developed SOCA2020 for  $b_{pp}$ , we are currently developing SOCA2020-CHL to retrieve the vertical distribution of Chl from satellite data merged with Argo temperature-salinity data. The main difference between SOCA2020-BBP and SOCA2020-CHL is the normalization of the Chl profiles with respect to depth. A dimensionless depth is introduced and calculated as the actual (physical) depth divided by the depth of the productive layer (i.e. the maximum between euphotic and mixed layer depths). Scaling the Chl profiles with respect to the productive layer depth enables to merge all profiles regardless of their vertical shape and range of Chl variation, simultaneously accounting for their variability. Indeed a typical profile of deep winter-mixing conditions in subpolar latitudes is highly different from a profile with a Deep Chlorophyll Maximum (DCM) characteristic of low-latitude permanently stratified systems.

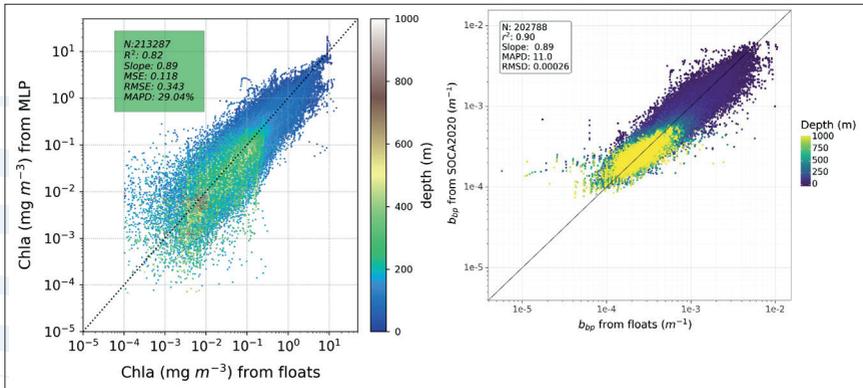


Fig. 3. Comparison of the Chl values retrieved by SOCA2020 to the reference measurements acquired by: a) BGC-Argo profiling floats from the independent validation dataset (left panel) and b) HPLC from MAREDAT database and LOV-IMEV database (right panel)

As for SOCA2020-BBP, the performance of the method is evaluated using an independent validation dataset (i.e. 20% of the initial SOCA2020 database chosen randomly), through the comparison between SOCA2020 Chl retrieval against the BGC-Argo Chl measurements (Figure 3 left panel). SOCA2020-CHL infers the vertical distribution of Chl with a global MAPD error of 29% and without systematic bias according to the depth of retrieval.

The SOCA2020-CHL method is further evaluated against Chl determined by High Performance Liquid Chromatography (HPLC), used as reference measurements for Chl. The comparison between SOCA2020 Chl retrieval against the HPLC Chl reference measurements from a global database gathering MAREDAT data (Peloquin *et al.*, 2013) and additional data from the Laboratoire d'Océanographie de Villefranche (LOV) / Institut de la Mer de Villefranche (IMEV) database (total of ~1,700 stations at the global scale) is shown in Figure 3 (right panel). SOCA-CHL retrieves the Chl with an error of 40% compared to reference HPLC Chl estimations.

### 3. Global 3D gridded product of POC and Chl

The global 3D gridded product of POC and Chl is available from CMEMS portal (MULTIOBS\_GLO\_BIO\_BGC\_3D\_REP\_015\_010 product). This product provides data on a regular spatial grid (0.25°x0.25°) with two different temporal resolutions, i.e. weekly fields from January 1998 to December 2018 (2019 for the next CMEMS release of May 2021) and multi-year monthly climatological fields. Note that the Chl in the CMEMS product retrieved from SOCA-CHL will be available at the end of the year 2021.

### 4. Conclusion

The conjoint use of the POC and Chl fields offers a new path to examine the variability in the phytoplankton carbon-to-Chl relationship within vertical dimension. This represents a great opportunity for a better understanding of light and nutrient control of phytoplankton biomass and physiological status at a global scale. This is a crucial step for improving the characterization of the distribution and variability in ocean primary production and carbon export. Moreover, the new proposed gridded CMEMS products represent a most valuable source of synthetic data useful not only for supporting the quality control of BGC-Argo float observations but also for data assimilation and initialization/validation of biogeochemical models.

### Acknowledgements

The BGC-Argo 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 satellite data used in this study are freely distributed by CMEMS. This work has been funded by the European 'Copernicus Marine Environment Monitoring Service', the European Research Council (ERC) under the ERCAdG-2010 remOcean project (grant agreement 246577) and ERCAdG-2019 REFINE project (grant agreement 834177) and by the BlueCloud project (European Union's Horizon Programme call BG-07-2019-2020, topic: [A] 2019 – Blue Cloud services; grant agreement 862409).

## References

- Cetinić, I., Perry, M. J., Briggs, N. T., Kallin, E., D'Asaro, E. A., and Lee, C. M. (2012). Particulate organic carbon and inherent optical properties during 2008 North Atlantic Bloom Experiment. *Journal of Geophysical Research: Oceans*, 117(C6).
- Claustre, H., Johnson, K. S., and Takeshita, Y. (2020). Observing the global ocean with biogeochemical-Argo. *Annual review of marine science*, 12, 23-48.
- Evers-King, H., Martinez-Vicente, V., Brewin, R. J., Dall'Olmo, G., Hickman, A. E., Jackson, T., Kostadinov, T. S., Krasemann, H., Loisel, H., Röttgers, R., Roy, S., Stramski, D., Thomalla, S., Platt, T. and Sathyendranath, S. (2017). Validation and intercomparison of ocean color algorithms for estimating particulate organic carbon in the oceans. *Frontiers in Marine Science*, 4, 251.
- McGillicuddy Jr, D. J. (2016). Mechanisms of physical-biological-biogeochemical interaction at the oceanic mesoscale. *Annual Review of Marine Science*, 8, 125-159.
- Guinehut, S., Buongiorno Nardelli, B., Chau, T., Chevallier, F., Ciani, D., Claustre, H., Etienne, H., Gehlen, M., Greiner, E., Jousset, S., Mulet, S., Sauzède, R., and Verbrugge, N. (2021): The MULTI OBSERVATIONS Thematic Assembly Center of the Copernicus Marine Environment Monitoring Service. *Proceedings of 9<sup>th</sup> EuroGOOS conference*.
- Peloquin, J., Swan, C., Gruber, et al., and Wright, S. (2013). The MAREDAT global database of high performance liquid chromatography marine pigment measurements. *Earth System Science Data*, 5(1), 109-123.
- Sauzède, R., Claustre, H., Uitz, J., Jamet, C., Dall'Olmo, G., d'Ortenzio, F., Gentili, B., Poteau, A., and Schmechtig, C. (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. *Journal of Geophysical Research: Oceans*, 121(4), 2552-2571.