Consistent long-term observations of surface phytoplankton functional types from space

Hongyan Xi[1*,] Marine Bretagnon[2], Ehsan Mehdipour[1,3], Julien Demaria[2], Antoine Mangin[2], Astrid Bracher[1,4]

[1]Alfred Wegener Institute, Helmholtz-Centre for Polar and Marine Research, Bremerhaven, 27570, Germany
[2]ACRI-ST, Sophia Antipolis Cedex, France
[3]School of Business, Social & Decision Sciences, Constructor University, Bremen, Germany
[4]Institute of Environmental Physics, University of Bremen, Bremen, 28359, Germany

Correspondence to: Hongyan Xi (hongyan.xi@awi.de)

Author Comments in response to Referee #2

This paper introduces a machine learning-based correction method to harmonize global phytoplankton functional type (PFT) data obtained from various ocean color sensors, addressing the discrepancies caused by differences in sensor characteristics. The authors propose the use of a random forest-based ensemble learning method (MLBE) for this task. By correcting the OLCI-derived PFT data to match the merged sensor-derived PFT data, this method ensures more consistent and reliable global PFT observations. The study demonstrates the utility of this correction for analyzing long-term trends in PFTs, revealing significant changes in the biomass of diatoms and dinoflagellates, while showing more stable trends for haptophytes and prokaryotes. Additionally, it examines anomalies in PFTs, noting significant increases in diatom and dinoflagellate Chla concentrations, particularly in higher latitudes and coastal regions.

Overall, the paper is clear in its objectives, and the methodologies employed are robust. The innovative use of machine learning to calibrate and harmonize PFT data from different sensors is a particularly valuable contribution. This approach significantly improves the accuracy and reliability of the resulting PFT time series, showcasing the potential of machine learning in enhancing ocean color application. Given the importance and novelty of this research, I recommend the publication of this paper. Before final acceptance, I have a few suggestions (listed below) that I hope the authors will consider to further refine and enhance the quality of the work.

We thank very much the reviewer for the positive feedback and constructive comments. We have carefully considered the suggestions during the revision. Below please find our response/clarificafication to each comment.

1. During model training, the authors randomly partitioned the dataset into a training set (70%) and 100 test sets (30%), achieving good validation results. However, the inclusion of latitude and longitude as input features may raise concerns about potential data leakage and shortcut learning, as spatial dependencies could result in overly optimistic estimates of model accuracy (https://doi.org/10.1038/s41559-023-02162-1). To mitigate this risk and improve the rigor of the validation, I suggest the authors incorporate temporal partitioning in addition to random partitioning. By dividing the dataset based on time and ensuring that the training and test sets are

strictly independent in terms of temporal coverage, the MLBE model's ability to generalize across both time and space can be more rigorously assessed.

We appreciate the reviewer’s suggestion and the reference provided. We understand the logic and reason why the inclusion of temporal partitioning is important. However, we don’t think it applies to our case here. Our MLBE model is basically a correction scheme, that is trained based on 12 months satellite data spaning only one year (the overlapping period of the two sensor sets), with the model we wanted to set up a regression model through random forest learning trying to identify better the spatial variation of the PFT data from the two sensor sets, so that it can fit one pattern to the other on the whole global scale. We considered all pixel data (over 4 million available data points) from the 12 monthly products, and we wanted to cover as complete as possible the whole global region to make sure the training learns the pattern globally. However, by applying the suggested temporal partitioning we would lose data, e.g., in high latitudes, if we exclude a certain month in the training. This may cause biases in the learning process. Then the trained model would very likely not be applicable to the test data set (because though they would be temporally independent from the training set, the spatial information in e.g. high latitudes which is not included in the training set could not be learnt from the training and thus the ML model might fail in the test data and also in the global products in later years).

Though we applied a straightforward random splitting in this study, the training and test data sets were nearly homogeneously divided over space and time by the random splitting due to the large amount of the data points (> 4 millions), as shown by (CDF, for each of the input variables) in Fig. R1 the cumulative distribution function. This ensures that the trained model takes the most of the knowledge of the available data sets within the limited time period that can be used in the correction model.

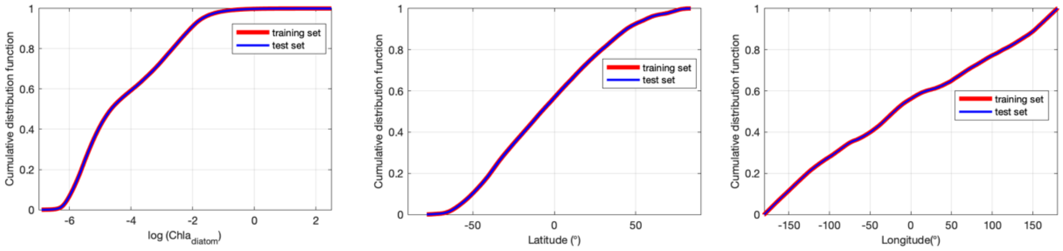

Figure R1. Cumulative distribution functions of input variables (PFT, lat, lon) involved in the MBLE training set, taking diatom as representative.

Reviewer 1 also posted a similar comment and we understand that there are limitations existing in the training and testing procedure and a discussion is necessary to clarify this point. Therefore, we add a paragraph of discussion about model caveats to cover this aspect:

“However, the MLBE model training was based on 12 months satellite data spanning only one year (the overlapping period of the two sensor sets), trying to identify the spatial variation of the PFT data from the two sensor sets, so that it could fit one pattern to the other on the whole global scale. It has been reported that random splitting between training and test sets may produce data leakages (Meyer et al., 2018; Stock et al., 2023) which result in overoptimistic performance in the test data but less good performance in actual applications to other data sets. To avoid data leakage data temporal partitioning has been suggested to ensure that the training and data sets are independent. However, random split was applied in the study as the temporal partitioning does not apply to our case. The MLBE model is basically a correction scheme trained based on all pixel data (over 50 million available data points) from 12 monthly PFT products. The purpose was to cover as complete as possible the global region to ensure that the training learns the pattern globally. By applying the suggested temporal partitioning we would lose data, e.g., in high latitudes, if we exclude a certain month in the training. This can cause biases in the learning process, then the trained model would very likely not be applicable to either the test set or other data sets that contain the missing periods. The straightforward random splitting in our study ensured the homogeneous splitting between the training and test data sets over space and time thanks to the large amount of data points, so that the trained model learned the most knowledge from the available data within the limited time period. Though such random partitioning has been widely used (e.g., Li. et al. 2023; Zoffoli, et al. 2025), one should keep in mind that having data for only a single year is challenging because the year may present conditions that are specific to that year only which may cause unrealistic predictions for other years. It is therefore noteworthy that target-oriented data splitting and cross-validation such as considering spatial and temporal blocks should be applied in machine learning based studies when data set allows (e.g., Zhang et al. 2024)."

2. When calculating the relative difference (RD in %) using PFT data, is the RD calculation performed on the log-transformed data or the raw PFT data? It is recommended to clarify this in the paper.

The RD calculation is based on non-log transformed (real) PFT Chla concentrations to describe the true relative difference between products from two different sensors, similar to the other commonly used statistical parameters in ocean color models such as mean percentage difference, which are also based on the Chla concentration (e.g. Xi et al. 2020; 2021). This has been clarified in the revised manuscript. We have added a subsection to clarify it better:

2.5 Statistical metrics

To evaluate the correction ensemble performance, relative difference (RD), median absolute difference (MAD) and median absolute relative difference (MARD) have been calculated based on the Chl*a* data of each PFT, which are defined as below.

RD_i = ($\mathrm{Chla}_i^{\mathrm{OLCI}}$ - $\mathrm{Chla}_i^{\mathrm{Merged}}$) / $\mathrm{Chla}_i^{\mathrm{Merged}}$, where *i* is the *i*th PFT

$$\mathrm{RD}_{\mathrm{PFT}} = \frac{(\mathrm{Chl}a\_\mathrm{PFT}_{\mathrm{OLCI}} - \mathrm{Chl}a\_\mathrm{PFT}_{\mathrm{merged}})}{\mathrm{Chl}a\_\mathrm{PFT}_{\mathrm{merged}}} * 100\% \tag{1}$$

$$\mathrm{MAD}_{\mathrm{PFT}} = \text{median of } (\mathrm{Chl}a\_\mathrm{PFT}_{\mathrm{OLCI}} - \mathrm{Chl}a\_\mathrm{PFT}_{\mathrm{merged}}) \tag{2}$$

$$\mathrm{MARD}_{\mathrm{PFT}} = \text{median of } \frac{|\mathrm{Chl}a\_\mathrm{PFT}_{\mathrm{OLCI}} - \mathrm{Chl}a\_\mathrm{PFT}_{\mathrm{merged}}|}{\mathrm{Chl}a\_\mathrm{PFT}_{\mathrm{merged}}} * 100\% \tag{3}$$

To validate the corrected PFT Chl*a* data with in situ data, statistical metrics including regression slope, determination coefficient (R^2), root mean square difference (RMSD, mg m^{-3}), and median percent difference (MDPD, %) have been used. For definition equations of these terms please refer to Xi et al. (2020). Note that the slope and R^2 are calculated in the base 10 logarithmic scale.

3. Similarly, when calculating the PFT anomaly, is the calculation performed on the log-transformed data or the raw PFT data?

   The PFT relative anomaly (%) was also calculated based on the original PFT Chl-a values. This has been clarified in the revised manuscript in section 2.5.

4. When calculating the average, it is important to clarify whether the authors computed a weighted average based on factors such as latitude, or if they simply calculated the unweighted mean of all pixels. Given that PFT variations are primarily observed in high-latitude regions, using a latitude-weighted average would be more reasonable.

   We agree with the two reviewers about the area distortion when calculating the mean spatially, and that a latitudinal weighted average can compensate the geographical distortion by taking into account proportional contribution. We therefore have modifed the calculations of the global and the four regional averages when generating PFT time series. The latitude-weighted averaging was applied to the logarithmic transformed Chla concentrations to get the log based mean which are then converted to their natural values. So for each monthly product over a certain region, the average was calculated based on the equation below:
   $Mean_{chla\_diatom} = \exp\left(\frac{\sum \cos(lat) \cdot \ln\left(Chla_{diatom}\right)}{\sum \cos(lat)}\right).$

   We have added the following text in Section 2.4 of the revised manuscript: “PFT time series of different spatial scales were calculated by applying the weighted average (taking cosine of the latitude as weights) to the monthly PFT data over the defined regions, to take into account the proportional contribution of each pixel to the global surface ocean due to area distortion in the gridded dataset. The latitude-weighted averaging was applied to the logarithmic transformed PFT Chla to get the log-based mean which are then converted to natural values.”

   The time series plots in Figure 4 have also been updated, showing overall slight changes in the trends, however the Chla magnitudes of the PFT time series at global scale are in general slightly lower for the weighted average, except for prokaryotes. This is mainly due to much lower prokaryotes Chla concentrations in high latitudes (compared to lower latitudes) contributed less with the latitude weighting applied, contrasting to other PFTs which have in general higher Chla in higher latitudes. The trend of the global diatom Chla was slightly decreased (from 0.0014 to 0.0011 mg m-3 per decade), while their increase at high latitudes is still very prominent, as the proportional weights are also considered in the ‘divider’ which is the weighted total number of observations. Accordingly, we have updated Figure 4 and the statistical description related to this figure in Section 3 of the revised manuscript.

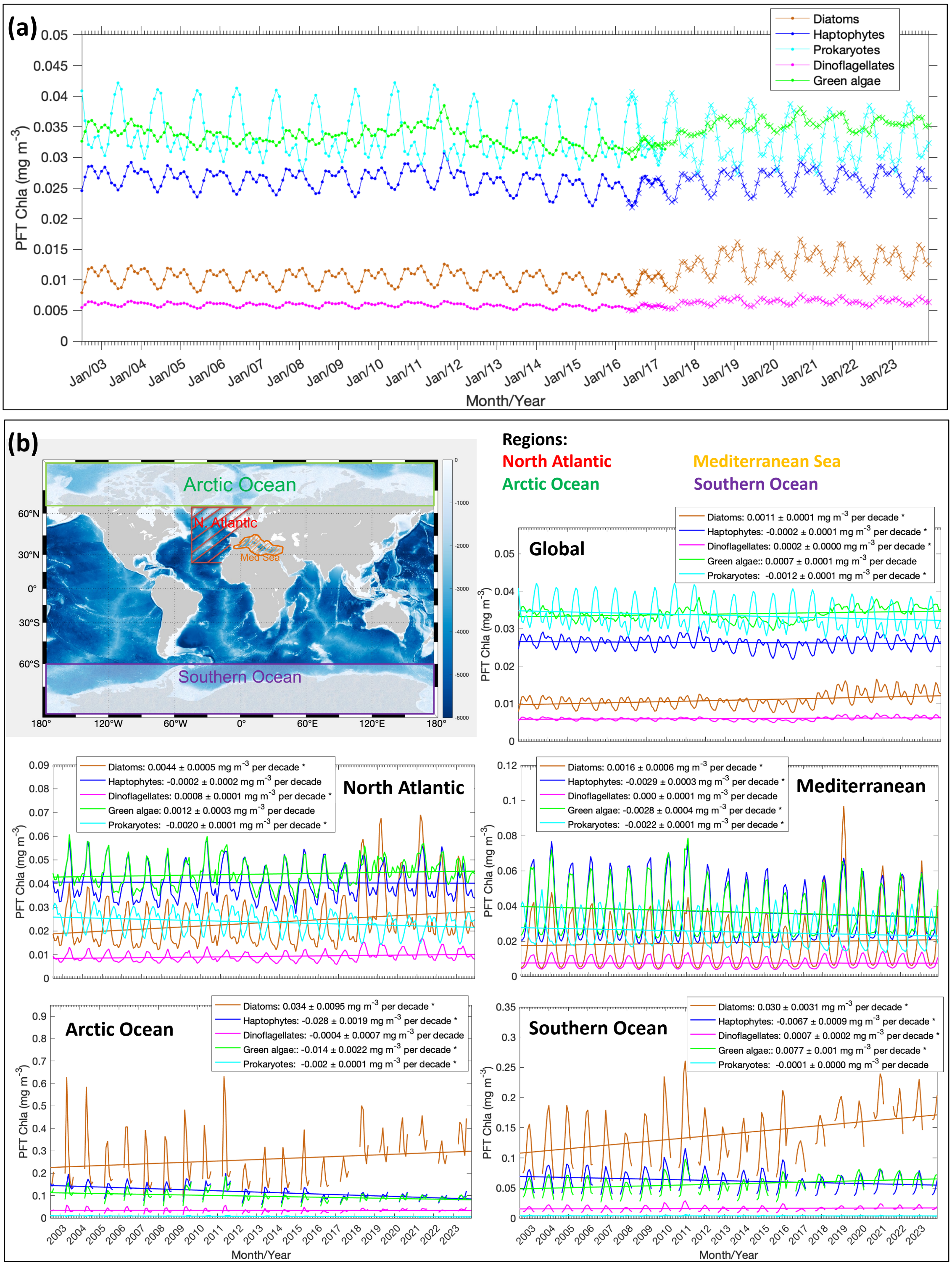

Figure 4. Panel (a): Updated (corrected) time series of the five PFT Chla based on the global mean from 2002 to 2023. Merged products cover the period of July 2002-April 2017 (indicated with dots), and OLCI products are for May 2016-Dec 2023 (indicated with crosses). Note that the OLCI products have been corrected to merged products based on MLBE. Panel (b): Trends of diatoms, haptophytes, dinoflagellates, green algae and prokaryotes Chla on the global scale and four regional scales (the North Atlantic Ocean, the Mediterranean Sea, the Arctic Ocean and the Southern Ocean), respectively. Trend slopes per decade with uncertainties have been indicated with significant trends marked with an asterisk (*).

5. The MLBE model demonstrates excellent performance, but it would be valuable to explore whether it can be applied to data from other sensors. Given the potential for

future expansion, I recommend that the authors include a brief discussion in the paper about potential future work, particularly how the correction method could be further improved or extended to incorporate data from other satellite sensors. This discussion would not only highlight the adaptability and scalability of the method to other satellite datasets but also significantly enhance the broader impact and relevance of this research. I believe that including this discussion would add considerable value to the paper.

We agree with the reviewer. Indeed, such discussion is very necessary. We have added the following text in the discussion. We would also like to point out that this manuscript as a contribution to the Ocean State Report needs comply with the length limit, hence we tried to include the discussion as concise as possible.

“The correction scheme proposed in this study is specifically designed to address inter-sensor data inconsistencies in the current Copernicus Marine Service PFT products. The present trained model can be only used to correct the OLCI-derived PFT product to match the merged sensor-derived product. However, the underlying technical framework is adaptable to other common ocean color products, such as optical properties derived from multiple sensors, thereby enhancing the overall continuity and consistency of ocean color data. As a rapidly emerging and powerful technique, machine learning can be further leveraged in ocean color data services, supporting agencies and data platforms in delivering high-quality, consistent operational products.”

References

Li, Z., Sun, D., Wang, S., Huan, Y., Zhang, H., Liu, J., He, Y., 2023. A global satellite observation of phytoplankton taxonomic groups over the past two decades. Global Change Biology 29, 4511–4529. https://doi.org/10.1111/gcb.16766

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental Modelling & Software 101, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Stock, A., Gregr, E.J., Chan, K.M.A., 2023. Data leakage jeopardizes ecological applications of machine learning. Nat Ecol Evol 7, 1743–1745. https://doi.org/10.1038/s41559-023-02162-1

Xi, H., Losa, S. N., Mangin, A., Soppa, M. A., Garnesson, P., Demaria, J., Liu, Y., d’Andon, O. H. F., and Bracher, A., 2020. A global retrieval algorithm of phytoplankton functional types: Towards the applications to CMEMS GlobColour merged products and OLCI data. Remote Sensing of Environment, 240, 111704, https://doi.org/10.1016/j.rse.2020.111704

Xi, H., Losa, S. N., Mangin, A., Garnesson, P., Bretagnon, M., Demaria, J., Soppa, M. A., d’Andon, O. H. F., and Bracher, A., 2021. Global chlorophyll a concentrations of phytoplankton functional types with detailed uncertainty assessment using multi-sensor ocean color and sea surface temperature satellite products. Journal of Geophysical Research-Oceans, 126(5), https://doi.org/10.1029/2020JC017127

Zhang, Y., Shen, F., Li, R., Li, M., Li, Z., Chen, S., Sun, X., 2024. AIGD-PFT: the first AI-driven global daily gap-free 4 km phytoplankton functional type data product from 1998 to 2023. Earth Syst. Sci. Data 16, 4793–4816. https://doi.org/10.5194/essd-16-4793-2024

Zoffoli, M.L., Brando, V., Volpe, G., González Vilas, L., Davies, B.F.R., Frouin, R., Pitarch, J., Oiry, S., Tan, J., Colella, S., Marchese, C., 2025. CIAO: A machine-learning algorithm for mapping Arctic Ocean

Chlorophyll-a from space. Science of Remote Sensing 11, 100212. https://doi.org/10.1016/j.srs.2025.100212