Detailed Responses

Here, we provide detailed responses to the referee #1' comments. The comments from the referees are shown in black. Our responses to the critics of the referees are supplied in normal font and **blue**. The appropriate correction in the manuscript has been repeated in **red** font in the response letter.

Referee #1:

General comments:

The paper by Zhang et al. presents the first AI-driven product for Phytoplankton Function Types (PFT) for the global ocean (AIGD-PFT). The AIGD-PFT consists of a L4 gap-free product including 8 PFT at daily and 4-km resolution for the period 1998-2023. AIGD-PFT is generated using an extended ensemble modelling approach (STEE-DL), which is based on machine and deep learning technologies and includes 100 models. Each model is built on statistical relationships between the physical environment and phytoplankton community and incorporates in situ HPLC data, ocean colour satellite observations whose missing data have been reconstructed throughout a cost-efficient DCT-PLS method, physical data from reanalysis and biogeochemical inputs from hindcast simulations.

Overall, the study falls within the scope of ESDD, methods are robust, and the manuscript is well written and detailed. Moreover, I believe that the AIGD-PFT product will be a very useful tool for all scientists interested in detecting climate-induced changes in the phytoplankton community. Therefore, I recommend this paper for publication, although I feel that some clarifications should be addressed to strengthen the way it is presented.

Response:

We are very grateful for reviewing our manuscript and providing us with your recognition and valuable advices on our work. Your comments and suggestions will definitely help us improve the manuscript.

We have revised the manuscript according to your specific comments and improved the quality. Please check the flowing item-by-item response, as well as the revised manuscript. Note that the appropriate corrections in the manuscript have been repeated in red font in the response letter.

Specific comments:

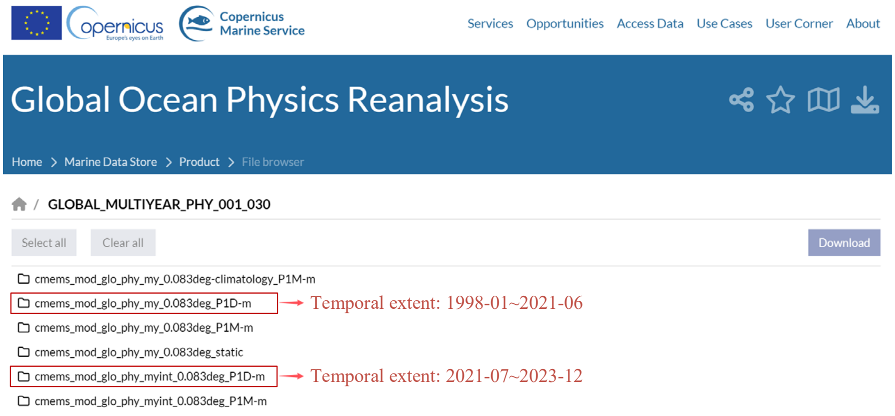
Authors present the AIGD-PFT as the product with the longest time span, covering 26 years (i.e., 1998-2023). However, I double checked the data sets used to create it and found some discrepancies that need to be clarified. In particular, except for the ESA-OC-CCI data set, which covers the whole period, I found that SST data from https://doi.org/10.48670/moi-00169 and biogeochemical variables from https://doi.org/10.48670/moi-00019 are available until October 2022 and December 2022, respectively, while SSS from https://doi.org/10.48670/moi-00016 is available from January 2022 to June 2024. So, I am not sure how authors create a 26-year product using some data sets that do not cover the same period.

Response:

Thank you for your detailed review. We apologize for the errors and confusion in our manuscript. We would like to clarify the specifics of the data used in our research as follows:

(1) **Sea Surface Temperature (SST) Data**: For SST, we utilized data from the ESA SST CCI and C3S reprocessed sea surface temperature analyses (DOI: https://doi.org/10.48670/moi-00169) which covers up to October 2022. For the period from November 2022 onwards, we employed the Global Ocean OSTIA Sea Surface Temperature and Sea Ice Analysis (DOI: https://doi.org/10.48670/moi-00165).

(2) **Sea Surface Salinity (SSS) Data**: We utilized the dataset Global Ocean Physics Reanalysis for SSS data (DOI: https://doi.org/10.48670/moi-00021, Fig. #1-1). This dataset includes the subset cmems_mod_glo_phy_my_0.083deg_P1D-m covering data before June 2021, and the subset cmems_mod_glo_phy_myint_0.083deg_P1D-mcovering from June 2021 onwards.



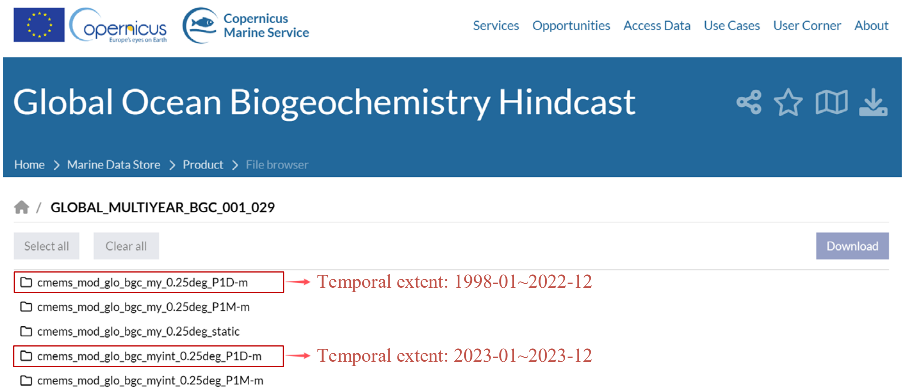
Global Ocean Physics Reanalysis

GLOBAL_MULTIYEAR_PHY_001_030

□ cmems_mod_glo_phy_my_0.083deg-climatology_P1M-m
□ cmems_mod_glo_phy_my_0.083deg_P1D-m → Temporal extent: 1998-01~2021-06
□ cmems_mod_glo_phy_my_0.083deg_P1M-m
□ cmems_mod_glo_phy_my_0.083deg_static
□ cmems_mod_glo_phy_myint_0.083deg_P1D-m → Temporal extent: 2021-07~2023-12
□ cmems_mod_glo_phy_myint_0.083deg_P1M-m

Fig. #1-1 Global Ocean Physics Reanalysis for SSS data. (DOI: https://doi.org/10.48670/moi-00021)

(3) **Biogeochemical Variables**: Regarding the biogeochemical variables, we used the Global Ocean Biogeochemistry Hindcast dataset (DOI: https://doi.org/10.48670/moi-00019, Fig. #1-2), which consists of two subsets. Until December 2022, we used the subset cmems_mod_glo_bgc_my_0.25deg_P1D-m, and from January 2023 onwards, we employed the subset cmems_mod_glo_bgc_myint_0.25deg_P1D-m.



/ GLOBAL_MULTIYEAR_BGC_001_029

- cmems_mod_glo_bgc_my_0.25deg_P1D-m → Temporal extent: 1998-01~2022-12
- cmems_mod_glo_bgc_my_0.25deg_P1M-m
- cmems_mod_glo_bgc_my_0.25deg_static
- cmems_mod_glo_bgc_myint_0.25deg_P1D-m → Temporal extent: 2023-01~2023-12
- cmems_mod_glo_bgc_myint_0.25deg_P1M-m

Fig. #1-2 Global Ocean Biogeochemistry Hindcast dataset. (DOI: https://doi.org/10.48670/moi-00019)

We have added a clear statement (see line 198-206 on page 11 of revised manuscript), as follows:

"The SST data are sourced from the ESA SST CCI (Climate Change Initiative) and C3S (Copernicus Climate Change Service) global Sea Surface Temperature Reprocessed product (https://doi.org/10.48670/moi-00169, covering the period from January 1998 to October 2022) and Global Ocean OSTIA Sea Surface Temperature and Sea Ice Analysis (https://doi.org/10.48670/moi-00165, covering the period from November 2022 to December 2023). The SSS data are obtained from Global Ocean Physics Reanalysis (https://doi.org/10.48670/moi-00021). Biogeochemical data include nitrate concentration (NC), phosphate concentration (PC), silicate concentration (SC), and dissolved oxygen (DO). These variables are critical for understanding the nutrient dynamics in marine ecosystems, which are fundamental factors influencing phytoplankton growth and distribution. The data for these biogeochemical variables are sourced from the global biogeochemical multi-year hindcast products (https://doi.org/10.48670/moi-00019)."

2) As reported in Sect. 2.2.3, all physical and biogeochemical data have been resampled to a 4 km resolution, and I believe that this was done to match the high spatial resolution

of the ESA-OC-CCI product. However, any time data are resampled to a higher resolution, a greater but false accuracy is introduced due to the assumption that all new pixels have the same value when it may only be true for one pixel. This is why, as far as I know, the remapping direction is typically from high to low resolution. I would therefore ask authors to discuss this choice and, if possible, include a reference to previous works applying the same strategy. An interesting paper that may help the discussion can be found at https://journals.ametsoc.org/view/journals/apme/60/11/JAMC-D-20-0259.1.xml.

Response:

Thank you for pointing out this important concern.

We agree with you. As demonstrated in the study by Rajulapati et al. (2021) that you recommended, resampling from a lower to a higher resolution indeed can alter the statistical properties of the data, thereby introducing potential inaccuracies. In our study, we opted to resample all physical and biogeochemical data to the same high 4 km resolution as the ESA-OC-CCI product primarily for consistency across datasets. We acknowledge that transforming data from a lower to a higher resolution often assumes that the newly generated pixel values are similar to the original ones, potentially introducing a so-called "false precision" that could lead to systematic biases.

To minimize the impact of false precision, the Inverse Distance Weighting (IDW) method was employed for spatial interpolation. The IDW identifies all available pixels around a target pixel based on a search radius of 8 pixels, and the weights of the identified available pixels are then calculated by the reciprocal of the square of the distance between the target pixel and the available pixels. This method is more likely to provide balanced estimates and reduce the risk of introducing false precision.

With advancements in technology, the availability of high-resolution ocean data is increasing, such as *Multi-Scale Ultra High Resolution (MUR) Sea Surface Temperature data* (1km resolution, DOI: https://doi.org/10.5067/GHGMR-4FJ04), which provides hope for fundamentally addressing these issues. However, at present, offering datasets with varying spatial and temporal resolutions seems impractical. The resampling approach we have taken is a compromise intended to maximize the use of existing data resources while minimizing the computational and data processing burden. How to reduce information loss during data processing will be an important focus for our future work.

Rajulapati, C. R., Papalexiou, S. M., Clark, M. P., and Pomeroy, J. W.: The Perils of Regridding: Examples Using a

Global Precipitation Dataset, J Appl Meteorol Clim, 60, 1561-1573, https://doi.org/10.1175/Jamc-D-20-0259.1, 2021.

Follow your concerns, we have added a clear explanation about resampling (see line 206-212 on page 11 of revised manuscript):

"All data undergo the following preprocessing steps: (1) resampling, where all data is resampled to a 4km resolution using the pysample library (https://doi.org/10.5281/zenodo.3372769). The Inverse Distance Weighting (IDW) method was employed for spatial interpolation. The IDW identifies all available pixels around a target pixel based on a search radius of 8 pixels, and the weights of the identified available pixels are then calculated by the reciprocal of the square of the distance between the target pixel and the available pixels. This resampling process may lead to missing pixels, which are then filled using the nearest neighbor method;"

Additionally, the Discussion section has been expanded to include the following content (see line 523-528 on page 30-31 of revised manuscript):

"Firstly, in this study, all physical and biogeochemical data were resampled to match the high resolution of 4 km, consistent with the OC-CCI product, primarily to ensure uniformity across datasets, and to maximize the use of existing data resources. However, resampling from a lower to a higher resolution can indeed alter the statistical properties of the data, potentially introducing inaccuracies. In future research, it is planned to incorporate more high-resolution data and to minimize the loss of information during the data processing stage."

3) Page 8, line 151: The sentence needs to be reworded because, as reported in the Product Guide (https://docs.pml.space/share/s/fzNSPb4aQaSDvO7xBNOCIw), the latest ESA-OC-CCI product (v6.0) also merges observations from OLCI-3A and OLCI-3B.

Response:

Thank you for your reminder. We rephrased the relevant text (see line 150 on page 8 of revised manuscript) as follows:

"This dataset is generated by band-shifting and bias-correcting SeaWiFS, MODIS, VIIRS, and Sentinel 3A and 3B OLCI data to match MERIS data, achieving a spatial resolution of 4 km”

4) I found the method used by authors to fill OC data gaps well described in Sect. 2.2.2. However, I think that specifying the number of available data before and after the filling procedure would be interesting and emphasize the effort authors have made. This information could also be presented by replacing Figure 3 with two Hovmöller diagrams showing the number of observations before and after the filling as function of time and latitude.

Response:

Thank you for your suggestion.

We have revised Figure 3 to include specific information on the changes in the quantity of available data before and after the filling process. Additionally, we have introduced two Hovmöller diagrams to visually represent these changes over time and latitude.

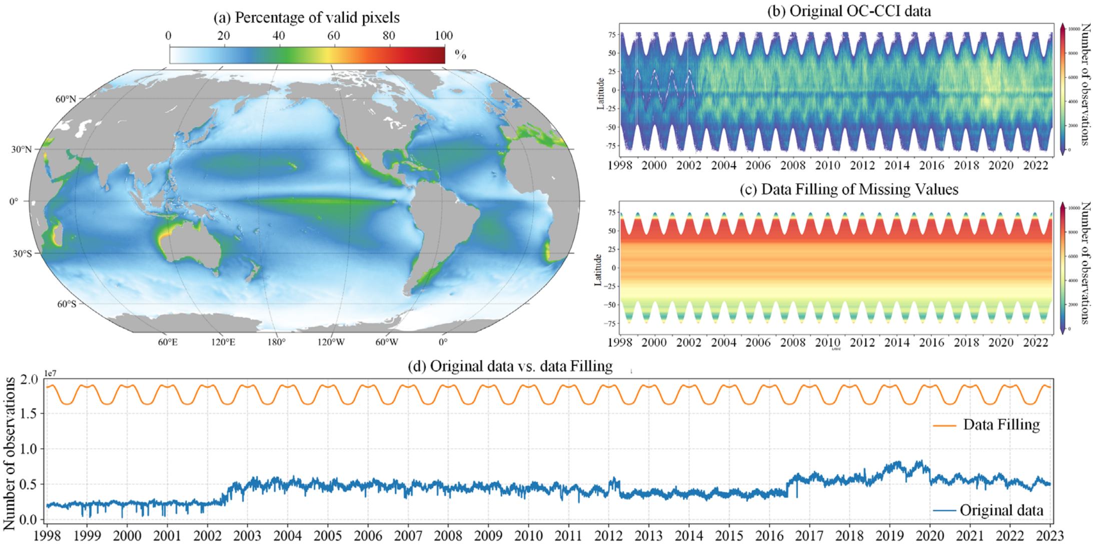


Figure 3 (a) Percentage of valid pixels in the OC-CCI v6.0 daily dataset; Hovmöller diagrams of (b) original OC-CCI data and (c) data after gap filling using the DCT-PLS method; (d) Comparison of the number of valid pixels between reconstructed and original data.

5) The choice to include the 8 PFTs as listed in the manuscript should be justified. I think that adding reference(s) should be enough to do that.

Response:

Thank you for your suggestion. We have added the relevant references (see line 137 on page 7 of revised manuscript):

"By utilizing an updated Diagnostic Pigment Analysis (DPA) methodology, along with newly adjusted weighting coefficients, we conducted DPA to ascertain in-situ PFT Chl-a concentrations. This analysis includes eight major PFTs: Diatoms, Dinoflagellates, Haptophytes, Pelagophytes, Cryptophytes, Green Algae, Prokaryotes, and Prochlorococcus, following conventional practices in the field (Xi et al., 2020; Xi et al., 2021)."

Xi, H. Y., Losa, S. N., Mangin, A., Garnesson, P., Bretagnon, M., Demaria, J., Soppa, M. A., D'Andon, O. H. F., and Bracher, A.: Global Chlorophyll a Concentrations of Phytoplankton Functional Types With Detailed Uncertainty Assessment Using Multisensor Ocean Color and Sea Surface Temperature Satellite Products, J Geophys Res-Oceans, 126, e2020JC017127, https://doi.org/10.1029/2020JC017127, 2021.

Xi, H. Y., Losa, S. N., Mangin, A., Soppa, M. A., Garnesson, P., Demaria, J., Liu, Y. Y., D'Andon, O. H. F., and Bracher, A.: Global retrieval of phytoplankton functional types based on empirical orthogonal functions using CMEMS GlobColour merged products and further extension to OLCI data, Remote Sens Environ, 240, 111704, https://doi.org/10.1016/j.rse.2020.111704, 2020.

6) The definition of ResNet models (i.e., residual neural networks) is given in Sect. 2.3.1, but I think it should be provided earlier as they are mentioned before Sect. 2.3.1.

Response:

Thank you for your suggestion. We have adjusted the definition of ResNet models, moving it to the first instance where the concept appears (see line 98-101 on page 4 of revised manuscript):

"Here, we propose a novel Spatial–Temporal–Ecological Ensemble model based on deep learning (STEE-DL), designed to produce a long time series PFT product. STEE-DL leverages an ensemble of 100 ResNet (residual neural networks) models, incorporating inputs from reconstructed missing ocean color data, physical reanalysis, biogeochemical, and spatiotemporal information."

7) I suggest authors to go through the manuscript and split some long sentences to make the text more readable. For example, the second sentence in the abstract, which starts on line 2 and ends on line 14, can be split into at least three sentences.

Response:

Thank you for your suggestion. We rephrased the relevant text (see line 8 on page 7 of revised manuscript) as follows:

"In this study, we integrated artificial intelligence (AI) technology with multi-source marine big data to develop a Spatial–Temporal–Ecological Ensemble model based on Deep Learning (STEE-DL). This model generated the first AI-driven Global Daily gap-free 4 km PFTs product from 1998 to 2023 (AIGD-PFT). The AIGD-PFT significantly enhances the accuracy and spatiotemporal coverage of quantifying eight major PFTs: Diatoms, Dinoflagellates, Haptophytes, Pelagophytes, Cryptophytes, Green Algae, Prokaryotes, and Prochlorococcus."

8) I found some errors in the reference list (e.g., Zhang and Shen, 2024a,b,c). Please, check them carefully against the references as cited in the abstract and main text.

Response:

Thank you for the reminder. We have corrected it.

To conclude, I would like to mention that, as stated by authors, model interpretability is beyond the scope of this manuscript and will be a focus of a future work. I look forward to that. So, keep up the good progress!

Response:

Thank you for your encouragement comments. We appreciate your support and are committed to making model interpretability a key focus of our future research.