Response to referee comment 1

Comment:

The variation of Chl-a concentration in the global ocean surface is influenced by various complex factors, which poses challenges to accurately retrieve Chl-a concentration. In this manuscript, the authors selected Chl-a data products retrieved from satellite data as a reference, supplemented by reanalysis data to provide environmental factor information. By combining the advantages of machine learning in big data analysis and simulation, they ultimately reconstructed a global-scale, long-term time series of Chl-a concentration dataset. The results show that the OCNET model performs very well in reconstructing Chl-a concentrations, accurately capturing the temporal variations of these features. This suggests that the model has strong potential for use in large-scale ocean color data reconstruction, and may even be able to predict Chl-a concentration trends in response to changes in the environment.

The dataset is extremely valuable for a wide range of researchers, policymakers, and managers involved in the monitoring and management of aquatic ecosystems. This study and its results are really novel and impressive to me.

Overall, I find this study and its results to be highly promising and valuable for the scientific community. With the suggested clarifications and improvements, this manuscript has the potential to make a significant contribution to the field of Chl-a concentration retrieval and ocean color data reconstruction.

Response:

We really appreciate these overall comments and recommendation by this reviewer. Our point-by-point responses to the reviewer's comments are given as follows.

Specific Comments:

1) Pg. 3, Lines 65-66: Please provide an explanation for the difference between OCNET and the CNNs used in previous studies (Cao et al., 2020; Jin et al., 2021; Cen et al., 2022; Yussof et al., 2021). Clarifying this distinction would enhance the reader's understanding of the novelty of the OCNET model.

Response:

Thanks for this comment. OCNET, a modified U-Net architecture, deviates from the general Convolutional Neural Network (CNN) in terms of network structure. Most data reconstruction methods based on machine learning, including CNNs and random forest, primarily rely on the spatiotemporal correlations within the data, using valuable spatiotemporal sequences to infer missing regions. However, such methods struggle to achieve satisfactory results when dealing with extensive, irregularly distributed missing data. The OCNET machine learning method proposed in this study draws inspiration from oceanic biogeochemical models and incorporates environmental variables that influence the distribution of chlorophyll-a concentration into data reconstruction. By learning the impact of environmental factors on the variation of chlorophyll-a concentration, it enables long-term and large-scale reconstruction of missing data. We will further elaborate on the distinctive features of OCNET in the revised manuscript.

Modifications: The difference between OCNET and other CNNs mentioned in previous studies were added in lines 65-75.

2) Pg. 6, Lines 125-126: The authors state that "We have selected three environmental variables, i.e., sea surface temperature (SST), salinity (SAL), and photosynthetically active radiation (PAR) as the input data for the OCNET model." However, it appears that the input data also contain SSP. Please address this discrepancy and provide clarity on whether SSP is included in the input data or not.

Response:

Thanks for this comment. In fact, we did select four environmental variables, including SSP, as inputs. The three variables mentioned here (SST, SAL, and PAR) are directly related to phytoplankton growth, as further explained in the manuscript. SSP primarily reflects the dynamic environmental factors at the ocean's surface, which influence the spatial distribution of phytoplankton. To eliminate any potential ambiguity, we have made corresponding modifications in the manuscript at the relevant sections.

Modifications: The description of the input data was clarified in lines 133-141.

Response to community comment 1

Comment:

I agree with reviewer #1 that this dataset has the potential to be extremely valuable (I found it while searching for a gapless daily chl dataset to test an idea on), and I thank the authors for making their data publicly available. However, I'm leaving this comment to suggest that the authors provide the data as netCDF files, perhaps one for each year, which would meet this journal's request that the data be provided in a non-proprietary community-established format that is findable, accessible, interoperable, and reusable. In the repository linked in the manuscript, the data are provided as 74 .rar files, each containing a number of ascii files. Every interested user (including myself, right now) will have to download and extract all 74 rar files, then write their own code to read the non-standard format and take a guess at some of the missing metadata (e.g., units).

Response:

Thank you very much for your suggestions. It is important for us to receive these feedbacks to further improve the data set. The new version of the dataset has been re-uploaded. Please refer to https://doi.org/10.5281/zenodo.10011908. This version of the dataset consists of one NetCDF file per year. Information regarding data format, latitude and longitude, handling of missing values, units, and other data specifications has been incorporated into the netCDF files. Subsequent updates to the dataset DOI will be provided in the revised manuscript. Thank you once again for your valuable suggestions!

Modifications: DOI of new version of OCNET dataset was updated in abstract and section 5.

Response to referee comment 2

Comment:

This manuscript describes a large dataset from 2001-2021 for global daily gap filled chlorophyll-a. The authors developed a convolutional neural network called OCNET to reconstruct global chlorophyll-a concentration in open oceans. This dataset is very useful and important for the scientific community. The manuscript in general is well written, but would benefit from some minor clarifications, and adjustments before publication.

Response: Thanks for thoroughly reviewing the manuscript and making such encouraging comments. It is important for us to receive these feedbacks to further improve the data set and the manuscript. Comments and issues mentioned in referee comment 2 have been addressed and are illustrated as follows.

General Comments:

In line 155, the authors stated that they excluded regions from seas with surface salinities below 25. On the other hand, the minimum value of salinity shown in table 2 is 0. How? Please clarify this in the text. According to ESSD, the DOI of the dataset and its in-text citation must be given in the abstract (https://www.earth-system-science-data.net/submission.html). Please add them. Please Provide more details in section 5 (data availability section) about the dataset which is very relevant for a data description paper. For example, you can explain how are the data organized in different files (per year or ..), and how are the files named (chl_OCNET_.. followed by Date (day, month, year). What are the type of files (".asc", or ".csv".. etc) , separator (if any)… etc

Response:

Thanks for this comment. When defining the open ocean areas, we used the multi-year mean of WOA2013 data as the reference and selected regions where salinity is greater than 25 PSU. However, it is important to note that the seasonal and interannual variations in salinity in these regions may not always exceed 25 PSU. It should be emphasized that, to establish the study boundaries, it is not a strict requirement for salinity to always exceed 25 PSU, but rather for the mean value to meet this criterion.

Furthermore, the study utilized salinity data from the Ocean Reanalysis System 5, and according to the documentation for this dataset, the minimum salinity value is 0. Therefore, in Table 2, the salinity is described as having a minimum value of 0.

Thanks for the comment about the dataset information. The DOI for the dataset will be added in the abstract, and any updates to the dataset will be provided through the same link. Please refer to the instructions on the dataset's publishing website for details (the data sets are available online with a DOI: https://doi.org/10.5281/zenodo.10011908).

Initially, the first version of the dataset was released in ASC format with daily data files. Due to the high volume of files in the first version, we have opted to release a new version and have converted the data into netCDF files. In the second version, each file corresponds to one year of data, and the respective year is indicated in the file name. Information regarding data format, latitude and longitude, handling of missing values, units, and other data specifications has been incorporated into the netCDF files.

Modifications: The source of the minimum value is added in title of Table 2. DOI of new version of OCNET dataset was updated in abstract and section 5.

Specific Comments:

1) Line 15 in the abstract: missing data not "data missing"

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

2) Line 24 in the abstract: phytoplankton biomass not “phytoplankton mass”

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

3) Line 24-25 in the abstract: The authors state that the “OCNET model achieves good performance in the reconstruction of global ocean Chl-a concentration data... etc.”. We don’t know how the model perform in polar regions or high latitudes (higher than 25) or coastal areas. It would be more precise to use the term “global open ocean”. The sentence should then become as follows: “the OCNET model achieves good performance in the reconstruction of global open ocean Chl-a … etc.”

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

4) Line 25 in the abstract: “captures temporal variations”. It is recommended to write spatiotemporal.

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

5) Line 125: “Four” not “three” environmental variables

Response:

Thanks for this comment. There is indeed an issue with clarity in the text. What we intended to convey is that the study selected three variables that affect the growth of marine phytoplankton, namely SST, SAL, and PAR, along with one variable that influences their distribution, namely SSP. SST, SAL, and SSP are derived from reanalysis data, while PAR is obtained from satellite data products. To address this, we have made appropriate revisions to the data description section from lines 125 to 135 to clarify this point.

Modifications: The description of the input data was clarified in lines 133-141.

6) Line 259: add degree (°) to “0.25”

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

7) Please ensure consistency in the terminology used throughout the text and in the figures. For example, in line 333 and 341, the terms 'training set,' 'validation set,' and 'test set' are used, while in Figure 6, they are referred to as 'training set,' 'validating set,' and 'testing set.

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

Modifications: The terminology used for the training set, validating set, and testing set has been standardized throughout the entire manuscript.

8) Line 346-347: I recommend that you use “compared to” instead of “while”. The statement would

become: “Based on the results of the OCNET model, regions 2, 3, and 5 show larger decreasing magnitudes, compared to other regions, which also exhibit a decreasing trend.”

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

9) Line 369-370: I recommend that you use “compared to” instead of “while”. The statement would then be: “From the results of bias, the training set shows a clear tendency of underestimation (Fig. 7d), compared to the validation and testing sets, which exhibit a less pronounced underestimation.

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

10) Line 372: global open ocean Chl-a concentration instead of “global Chl-a concentration”

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

11) I recommend that you change all statements starting with: “it can be seen” or “it can be observed” and ending with (figure #) to: “By referring to figure #”, or “Figure # indicates/shows …. Etc”. Below are some examples:

> *Line 387-388: “It can be seen that the output data of the OCNET model show a similar distribution to NOAA MSL12 data in the global tCC distribution (Fig.8(a–b))”*
>
> *Line 332: “it can be seen that the model performs well (Fig.6)*

Response:

Thanks for this comment. It will be corrected in the revised manuscript.

Modifications: The statements ending with (figure #) were changed into “referring to figure #”, or “Figure # indicates/shows …. Etc” everywhere.

12) Line 335-336, the authors stated that OCNET performed well but shows poor performance in individual regions, and stated that “region 9, being mostly near the American continent, is heavily influenced by human activities, and the satellite data quality in coastal areas is also poorer...”. Then, in their conclusion (Line 510-512), the authors concluded that “the model's performance was somewhat weaker in the eastern Pacific region compared to other areas. This may be due to specific climate characteristics that have a significant impact on phytoplankton growth and distribution or low quality of satellite-based dataset in this region”. Isn’t region 9 supposed to be the eastern Pacific? If yes, then please state the same reasons in both statements, and provide more details or examples on such climate characteristics that are specific to the eastern Pacific.

Response:

Thanks for this comment. In these two sentences, there is indeed a lack of clarity in our statements. In fact, the summaries of these two sections refer to slightly different regions. Region 9 encompasses parts of the Eastern Pacific and the western North Atlantic, all of which are close to the American continent.

However, when we mention that "the model performs poorly in the eastern tropical Pacific," it actually refers to the "eastern tropical Pacific," which is closer to the equator (as shown in Figure 7) and should be clarified in the revised manuscript.

Since the OCNET model proposed in the study does not consider the influence of human activities in coastal areas, the poor performance of the model in the coastal regions near the American continent could likely be attributed to human activities. On the other hand, the model's poor performance in the "eastern tropical Pacific" region may be more likely to be affected by specific climate characteristics. These unique climate variations might not be captured by the OCNET model within the relatively short training time span (2018-2021). In fact, the results of the OCNET model for Region 9 during the period of 2018-2021 closely align with the satellite-merged OCCCI dataset. This suggests that OCNET performs well only within the time frame covered by the target dataset NOAA MSL12, and it exhibits poorer performance in earlier periods (2001-2017).

There are several studies that focus on anomalies in the eastern tropical Pacific. For example, Geng et al. suggest that increased sea surface temperature variability due to global warming may manifest in the eastern Pacific earlier than central Pacific. Duteil et al. discuss the important impact of future changes in atmospheric synoptic variability (ASV) on ocean properties and primary productivity in the eastern tropical Pacific.

Regarding the issue of the model's poor performance in this particular region, we plan to conduct further analysis and exploration in future work. Once again, we appreciate your feedback!

Modifications: In the conclusion section, “eastern Pacific” was changed into “eastern tropical Pacific”. The citations for the two mentioned articles have also been included in the manuscript.

13) Figure 6 shows the evaluation indices of the training, testing and validating sets. Meanwhile, there is no indication to which dataset correspond the evaluation metrics shown in table 4. Although can be inferred by comparison, I recommend that you indicate in the text or table caption that they correspond to the training set. Readers shouldn’t guess.

Response:

Thanks for the comment. We indeed omitted an explanation of the datasets included in the evaluation metrics in Table 4. The median values in these metrics represent the median of all evaluation results, including those from the training set, validation set, and test set. Since Figure 6 already provides separate visualizations of the evaluation results for the training set, validation set, and test set using box plots, Table 4 only shows the overall evaluation summary. We have added an explanation to the title of Table 4 to clarify this.

Modifications: The explanation of evaluation metrics in Table 4 was added in lines 339-340.

14) Several References lack DOI. Please add the corresponding DOI. Below are some examples of references lacking DOI:

Behrenfeld, M. J., O’Malley, R. T., Siegel, D. A., McClain, C. R., Sarmiento, J. L., Feldman, G. C., Milligan,A. J., Falkowski, P. G., Letelier, R. M., and Boss, E. S.: Climate-driven trends in contemporary ocean productivity, Nature, 444, 752-755, 2006. Chen, S., Hu, C., Barnes, B. B., Xie, Y., Lin, G., and Qiu,

Z.: Improving ocean color data coverage through machine learning, Remote Sensing of Environment, 222, 286-302, 2019. Groom, S., Sathyendranath, S., Ban, Y., Bernard, S., Brewin, R., Brotas, V., Brockmann, C., Chauhan, P., Choi, J.-k., Chuprin, A., Ciavatta, S., Cipollini, P., Donlon, C., Franz, B., He, X., Hirata, T., Jackson, T., Kampel, M., Krasemann, H., Lavender, S., Pardo-Martinez, S., Mélin, F., Platt, T., Santoleri, R., Skakala, J., Schaeffer, B., Smith, M., Steinmetz, F., Valente, A., and Wang, M.: Satellite Ocean Colour: Current Status and Future Perspective, Frontiers in Marine Science, 6, 485 2019.

Response:

Thanks for the comment. The DOI of all references will be modified in the revised manuscript.

A list of modifications in the manuscript

Modification position is referred to the marked-up manuscript

Comment	Modification	Modification position
Modifications based on RC1		
RC1 General comment	None	None
RC1 Specific(1)	The difference between OCNET and other CNNs mentioned in previous studies were added in section 1	P3, L68-L74
RC1 Specific(2)	The description of the input data was clarified in section 2.1	P7, L134-L142
Modifications based on CC1		
CC1 General comment	The new version of the dataset has been re-uploaded as netCDF files. Please refer to https://doi.org/10.5281/zenodo.10011908.	None
Modifications based on RC2		
RC2 General comment1	None	None
RC2 General comment2	The source of the minimum value is added in title of Table 2. The new version of the dataset has been re-uploaded as netCDF files and the DOI is updated in the revised manuscript. (https://doi.org/10.5281/zenodo.10011908)	P1, L27 P8, L189-L190 P25, L525 P27, L618-L619
RC2 Specific(1)	"data missing" was changed into "missing data"	P1, L15
RC2 Specific(2)	"phytoplankton mass" was changed into "phytoplankton biomass"	P1, L25

RC2 Specific(3)	"global ocean" was changed into "global open ocean"	P1, L25
RC2 Specific(4)	"tenporal" was changed into "spatiotemporal"	P1, L26
RC2 Specific(5)	The description of the input data was clarified in section 2.1	P7, L134-L142
RC2 Specific(6)	The unit degree (°) was added	P12, L271
RC2 Specific(7)	The terminology used for the training set, validating set, and testing set has been standardized throughout the entire manuscript.	Everywhere
RC2 Specific(8)	"while" was changed into "compared to"	P18, L364
RC2 Specific(9)	"while" was changed into "compared to"	P20, L389
RC2 Specific(10)	"global Chl-a" was changed into "global open ocean Chl-a"	P20, L391-L392
RC2 Specific(11)	The statements ending with (figure #) were changed into “referring to figure #”, or “Figure # indicates/shows …. Etc” everywhere.	Everywhere
RC2 Specific(12)	In the results section and conclusion section, “eastern Pacific” was changed into “eastern tropical Pacific”. The citations for the two mentioned articles have also been included in the manuscript.	P20, L381 P21, L411-L412 P25, L535-L538
RC2 Specific(13)	The explanation of evaluation metrics in Table 4 was added.	P17, L346-L347
RC2 Specific(14)	The DOI of all references was added in the revised manuscript.	References

A global daily gap-filled chlorophyll-a dataset in open oceans during 2001–2021 from multisource information using convolutional neural networks

Zhongkun Hong[1,2], Di Long[1,2*], Xingdong Li[1,2], Yiming Wang[1,2], Jianmin Zhang[1,2], Mohamed A.Hamouda[3,4] and Mohamed M.Mohamed[3,4]

1. State Key Laboratory of Hydroscience and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing, China
2. Department of Hydraulic Engineering, Institute of Ocean Engineering, Tsinghua University, Beijing 100084, China
3. Department of Civil and Environmental Engineering, United Arab Emirates University, Al Ain, United Arab Emirates
4. National Water and Energy Center, United Arab Emirates University, Al Ain, United Arab Emirates

(*Correspondence to Di Long at dlong@tsinghua.edu.cn)

Abstract. Ocean color data are essential for developing our understanding of biological and ecological phenomena and processes, and also important sources of input for physical and biogeochemical ocean models. Chlorophyll-a (Chl-a) is a critical variable of ocean color in the marine environment. Quantitative retrieval from satellite remote sensing is a main way to obtain large-scale oceanic Chl-a. However, missing data ~~missing~~ ~~is~~ are~~a~~ a major limitation in satellite remote sensing-based Chl-a products, due mostly to the influence of cloud, sun glint contamination, and high satellite viewing angles. The common methods to reconstruct (gap filling) missing data often consider spatiotemporal information of initial images alone, such as data interpolation empirical orthogonal function, optimal interpolation, Kriging interpolation, and extended Kalman filter. However, these methods do not perform well in the presence of large-scale missing values in the image and ~~ignore~~ overlook the ~~potential of other~~ valuable information ~~on~~ available from other datasets ~~missing pixels in the~~for data reconstruction. Here we developed a convolutional neural network (CNN) named OCNET for Chl-a concentration data reconstruction in open ocean areas, considering environmental variables that are associated with ocean phytoplankton growth and distribution. Sea surface temperature (SST), salinity (SAL), photosynthetically active radiation (PAR), and sea surface pressure (SSP) from reanalysis data and satellite observations were selected as the input of OCNET to correlate with the environment and phytoplankton biomass~~mass~~. The developed OCNET model achieves good performance in the reconstruction of global open ocean Chl-a concentration data, and captures spatiotemporal ~~temporal~~ variations of these features. The reconstructed Chl-a data are available online at https://doi.org/10.5281/zenodo.10011908 (Hong et al., 2023). This study also shows the potential of machine learning in large-scale ocean color data reconstruction and offers the possibility to predict Chl-a concentration trends under a changing environment.

Key words. Chlorophyll-a; U-Net; Satellite remote sensing; Data reconstruction

Introduction

Chlorophyll-a (Chl-a), the primary pigment responsible for photosynthesis in plants, plays a vital role in the global carbon cycle and serves as a key indicator of the health and productivity of aquatic ecosystems (Righetti et al., 2019; Sun et al., 2021; Mouw et al., 2016). Chl-a is a measure of the amount of phytoplankton present in water bodies, and changes in its concentration can indicate shifts in the balance of these ecosystems, including the onset of harmful algal blooms or declines in productivity (Ho et al., 2019). Accurate and timely measurement of chlorophyll-a concentrations is therefore of ~~great~~ paramount importance for understanding and predicting the carbon fluxes and other elemental cycles in the oceans (Salgado-Hernanz et al., 2019; Laufkotter et al., 2016).

In recent years, satellite remote sensing has become a widely used method for monitoring chlorophyll-a concentrations on a global scale (Hu et al., 2012; Hu et al., 2019a; Feng et al., 2021). Satellite sensors can provide synoptic coverage of large areas, with a temporal resolution that ranges from daily to monthly. However, there are a lot of missing data in satellite products caused by cloud, sun glint contamination, and high satellite viewing angles (Feng and Hu, 2016; Mikelsons and Wang, 2019). For example, there are over 70% missing data in global daily ocean color products from MODIS-Terra/Aqua and VIIRS-SNPP (~~(Fig.1)~~referring to Figure 1) ~~-~~(Feng and Hu, 2016; Liu and Wang, 2018). In addition, the spatial and temporal resolution of these measurements is often limited, and they are subject to various sources of error and uncertainty. These include atmospheric effects, such as scattering and absorption of light, which can distort the signal and introduce biases in the measurements (Hu et al., 2019a; Zheng and Digiacomo, 2017). To address these limitations, it is useful to combine satellite remote sensing data with other sources of information, such as in situ measurements, model output, and ancillary data (Nikolaidis et al., 2014). Conventional methods for reconstructing missing data, such as data interpolation, DINEOF (Data Interpolating Empirical Orthogonal Functions), optimal interpolation, Kriging interpolation, and extended Kalman filter, often rely on the spatiotemporal information of the initial images alone (Wang and Liu, 2014; Hilborn and Costa, 2018; Catipovic et al., 2023; Liu and Wang, 2018). However, these geostatistical methods are not always effective in the presence of large-scale missing values and do not take into account the potential contribution of other information to the reconstruction of missing pixels (Konik et al., 2019).

The development of robust and efficient methods for synthesizing and integrating multisource information is becoming increasingly important as the availability and diversity of data sources continue to grow (Li et al., 2020). The integration of multisource information is not a trivial task, as the data sources may have different spatial and temporal scales, resolutions, and uncertainties, and may be subject to different biases and errors. These differences can make it challenging to reconcile and combine the data in a meaningful and reliable way (Catipovic et al., 2023). With the proliferation of sensors and platforms, the volume of data being generated is increasing at an exponential rate, making it difficult to manage and analyze in a traditional way. Machine learning techniques, such as convolutional neural networks (CNNs), offer a promising approach for handling and extracting meaningful insights from this large and complex data stream (Zhang et al., 2018). CNNs are a class of deep learning algorithms that have proven to be highly effective for image recognition and analysis tasks. They are particularly well suited to this problem, as they can automatically learn features and patterns from data and can handle large amounts of data with high dimensionality and complexity. CNNs have been applied to a wide range of remote sensing applications, including the analysis of satellite imagery and the integration of multisource data. A number of studies have demonstrated the effectiveness of CNNs for analyzing global or regional daily chlorophyll-a products (Cao et al., 2020; Jin et al., 2021; Cen et al., 2022; Yussof et al., 2021). Most machine learning-based data reconstruction methods, such as Convolutional Neural Networks (CNNs) and Random Forests, predominantly leverage spatiotemporal correlations inherent in the data. They utilize valuable spatiotemporal sequences to predict missing regions. Nevertheless, these techniques face significant challenges in yielding satisfactory outcomes when confronted with extensive and irregularly distributed missing data. Here we propose a CNNs-based approach named OCNET for the reconstruction of global daily chlorophyll-a products from multisource information. By emphasizing the significance of incorporating spatiotemporal complete environmental variables for chlorophyll gap-filling, OCNET demonstrates remarkable data reconstruction performance.

The OCNET model developed here is an improved version based on the general U-Net. One advantage of U-Net is its ability to handle large images while maintaining high-resolution segmentation results (Li et al., 2020; Ronneberger et al., 2015; Andersson et al., 2021). This is achieved by using skip connections, which allow the network to "skip" certain layers and merge higher-resolution information from early layers into the final prediction (Ronneberger et al., 2015; Wagle et al., 2020). This helps preserve fine-grained details of the input image and generates more accurate segmentation results (Krug et al., 2017). Here we utilized this characteristic of OCNET for global-scale input of big data, and successfully accomplished the task of data reconstruction. Given that the input image contains multi-level information elements at the global scale, it places high demands on how the model extracts feature information and captures its inherent correlations (Moran et al., 2022; Chen et al., 2019). Another advantage of U-Net is its ability to utilize contextual information from the entire image. Compared to other machine learning methods such as multiple linear regression and random forest, U-Net excels in learning complex nonlinear relationships between input data and output predictions (Ronneberger et al., 2015; Li et al., 2020). This is due to the use of nonlinear activation functions and the ability to learn hierarchical features through convolutional layers. Because artificial neural networks (ANNs) often face limitations in processing large images and struggle to incorporate global backgrounds into their predictions (Catipovic et al., 2023), U-Net outperforms traditional ANNs in various image segmentation tasks. Unlike ANNs, U-Net can handle high-resolution images and effectively incorporate global context information into its predictions (Andersson et al., 2021; Li et al., 2020).

In the big-data era, the effective integration and utilization of multisource information on the ocean are of importance for studying ocean color. The primary objective of this study was to propose the OCNET model which could be trained with environmental variables that are associated with ocean phytoplankton growth and distribution, in order to reconstruct high-quality gap-filled Chl-a data in open oceans. The Chl-a dataset covers the period from 2001 to 2021, with a daily temporal resolution and a spatial resolution of 0.25 °. Compared to traditional interpolation methods, this approach takes full advantage of environmental information mainly provided by ERA5 data, and considers the key factors that influence the growth and distribution of surface phytoplankton in the oceans. Furthermore, this method is not limited by the size of the ocean region or the temporal span covered by satellite data. By providing reliable environmental information, OCNET enables the retrospective analysis of Chl-a concentration data from the pre-satellite era and the prediction of future changes in global marine phytoplankton.

2. Data and methodology

2.1 Training data considerations

The Ocean-Colour Climate Change Initiative (OCCCI) version 5 and National Oceanic and Atmospheric Administration multi-sensor DINEOF global gap-filled data (termed as NOAA MSL12 hereafter) are two Chl-a products used in training the OCNET model (Table 1). OCCCI's data sources include the Moderate Spectral Resolution Imaging Spectroradiometer (MERIS) sensor from the European Space Agency, the SeaWiFS (~~Ocean Observation Wide Field~~ Sea-viewing Wide Field-of-view Sensor) and MODIS-Aqua (Moderate Resolution Imaging Spectroradiometer-Aqua) sensors from NASA, and the National Oceanic and Atmospheric Administration's VIIRS sensor (Visible and Infrared Imaging Radiometer Suite) (Sathyendranath et al., 2019). Data can be obtained starting from 1997. The Multi-Sensor Level-1 to Level-2 (MSL12) is the NOAA official enterprise VIIRS ocean color data processing system (Liu and Wang, 2022). The NOAA MSL12 dataset provides near-real-time, gap-free global maps of chlorophyll-a concentration by merging data from VIIRS and OLCI-Sentinel-3A satellites and utilizing the DINEOF method to fill in missing pixels caused by clouds, sun glint, and other factors (Liu and Wang, 2022). The strength of this dataset lies in its broader spatial coverage, showcasing more marine features in coastal and inland waters and enhancing data accuracy. In addition, Chl-a data from OLCI-Sentinel-3B have not been applied in the production of OCCCI V5 or NOAA MSL12 datasets. Therefore, Sentinel-3B data were used for the evaluation and comparison of the final performance of the OCNET model as an independent product.

Table 1 Full names, spatiotemporal resolution, temporal coverage, sources, and other information of data used in this study.

Data	Variables	Abbreviation	Unit	Temporal resolution	Spatial resolution	Temporal coverage	References
OCCCI V5	Chlorophyll a	Chl-a	mg/m^3	daily	4km	1997.9.4–2021	(Sathyendranath et al., 2019)
MODIS-Aqua	Photosynthetically Available Radiation	PAR	einstein/(m^2 d)	daily	4km	2002.7.4–present	https://oceancolor.gsfc.nasa.gov/l3
MODIS-Terra	Photosynthetically Available Radiation	PAR	einstein/(m^2 d)	daily	4km	2000.2.24–present	https://oceancolor.gsfc.nasa.gov/l3
VIIRS-SNPP	Photosynthetically Available Radiation	PAR	einstein/(m^2 d)	daily	4km	2012.1.2–present	https://oceancolor.gsfc.nasa.gov/l3

OLCI S3B NRT	Chlorophyll a	Chl-a	mg/m^3	daily	4km	2018.5.14–present	www.star.nesdis.noaa.gov
NOAA MSL12	Chlorophyll a	Chl-a	mg/m^3	daily	9km	2018.2.9–present	(Liu and Wang, 2022)
ERA5	Surface Pressure	SSP	Pa	hourly	0.25 °	1940.1.1–present	(Hersbach et al., 2020)
ERA5	Sea Surface Temperature	SST	K	hourly	0.25 °	1940.1.1–present	(Hersbach et al., 2020)
ORAS5	Salinity	SAL	PSU	monthly	0.25 °	1958.1.1–present	(Zuo et al., 2019)
ETOPO1	Depth	Dep	m	–	1'	–	(Information and Doc/Noaa/Nesdis/Ncei National Centers for Environmental Information, 2009)
WOA2013	Salinity	SAL	PSU	–	0.25 °	–	(Levitus et al., 2014)

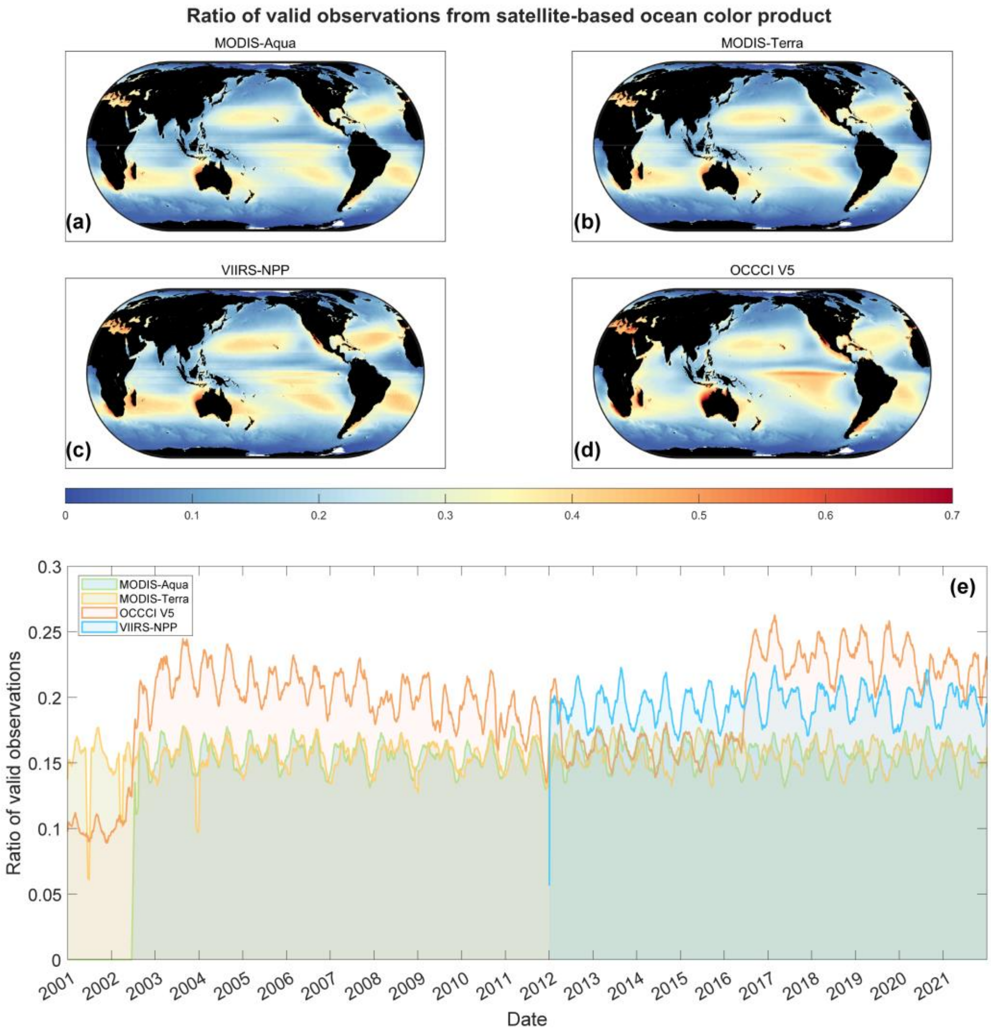


Figure 1 Valid data proportion of each satellite-based Chl-a product during 2001–2021. The global distribution of valid chlorophyll-a (Chl-a) observations ratio was examined using (a) MODIS-Aqua, (b) MODIS-Terra, (c) VIIRS-SNPP, and (d) OCCCI satellite datasets, along with an examination of the (e) temporal variation over their respective coverage periods.

The ocean Chl-a data of the OCCCI product cover more than 20 years. Compared with a single satellite product, OCCCI products that integrate multiple sources of data improve data availability by complementing different data sources~~ (Fig.1)~~ (~~-~~referring to Figure 1). Due to changes in satellite data sources used in different years, the valid data proportion of OCCCI varies greatly in different time periods. In addition, OCCCI has been significantly improved with the introduction of more satellite data. However, ~~Figure 1 shows that ~~valid observations from OCCCI are unevenly distributed globally (referring to Figure 1~~Figure 1~~)~~ (Fig.1d)~~. ~~And m~~Missing data on more than 70% of satellite-based products still pose a huge obstacle to the study of ocean color (Feng and Hu, 2016). The NOAA MSL12 achieved the spatiotemporal continuity of chlorophyll concentration products by the DINEOF method, but NOAA MSL12 are only available after February 9, 2018. Given the high coincidence of OCCCI and NOAA MSL12 datasets in the selection of satellite sources, these two datasets were selected as the main data sources. Other Chl-a data products from single-mission satellites, such as MODIS-Aqua/Terra and VIIRS-SNPP, which have more severe missing values (~~(Fig.1)~~referring to Figure 1), were only used for comparison in this study and were not directly applied.

We have selected ~~three~~ four environmental variables, i.e., sea surface temperature (SST), salinity (SAL), ~~and~~ photosynthetically active radiation (PAR), and sea surface pressure (SSP) as the input data for the OCNET model. These variables play a significant role in influencing the growth and distribution of marine phytoplankton (Flynn, 2001; Han and Zhou, 2022)~~. SST and SAL are two important environmental factors influencing marine phytoplankton growth~~. SST affects algal metabolic rates, enzymatic activity, cell division rates, and growth cycles, among other biological processes (Nelson et al., 2020). Variations in salinity can influence osmoregulation in marine phytoplankton and ion balance within cells (Nelson et al., 2020). Consequently, SST and SAL are considered pivotal input variables in the OCNET model. Furthermore, from a hydrodynamic perspective, changes in wind patterns and ocean currents can also affect the distribution of surface algae. To capture this impact, we have chosen to represent changes in ocean surface pressure with the parameter SSP. Therefore, we selected reanalysis data ERA5's SSP, SST, and Ocean Reanalysis System 5's SAL as input data for the OCNET model.

In addition to SST and SAL, PAR is a crucial energy source for plant photosynthesis, and its distribution is of great importance for studying plant growth and photosynthetic processes (Xing and Boss, 2021). Its spatiotemporal variations can impact the photosynthetic efficiency, biomass accumulation, and yield of plants (Righetti et al., 2019). Here we selected PAR data from satellite sources, specifically MODIS-Terra/Aqua and VIIRS-SNPP, as part of the model input. To address spatial gaps in satellite data and correct biases among different datasets, preprocessing and fusion techniques were applied to the PAR data from different satellite products (see Section 2.2).

Both ETOPO1 and WOA13 data were used as auxiliary data for determining the study area and were not input for the OCNET model. The ETOPO Global Relief Model is a global digital elevation model developed by the National Geophysical Data Center (NGDC), a NOAA department (Information and Doc/Noaa/Nesdis/Ncei National Centers for Environmental Information, 2009). It provides elevation data for the Earth's surface and finds applications in areas such as topographic maps, hydrological models, oceanography, and other related fields. Data of ETOPO1 were selected because of the 1-min resolution it offers. ETOPO1 is widely utilized in scientific and research communities due to its high accuracy, serving various purposes like mapping, visualization, resource management, and environmental modeling (Moran et al., 2022; Righetti et al., 2019). The World Ocean Atlas 2013 (WOA2013) is a comprehensive collection of objectively analyzed climatology data for various oceanic parameters, including temperature, salinity, oxygen, phosphate, silicate, and nitrate (Zweng et al., 2013). It was provided by NOAA's National Oceanographic Data Center - Ocean Climate Laboratory. Salinity data provided by WOA13 are often used as a reference to analyze abnormal variations in ocean salinity (Righetti et al., 2019; Li et al., 2017).

The study area considered here mainly focuses on the middle and low latitudes of the open ocean area, constrained primarily due to limitations in satellite data sources. In particular, ~~Figure 1 indicates that~~ satellite-based Chl-a products exhibit a substantial number of missing values in high latitudes and coastal regions (referring to Figure 1~~Figure 1~~) ~~(Fig.1)~~. Additionally, the accuracy of chlorophyll concentration retrievals is affected mostly by the presence of high concentrations of suspended matter resulting from sediment discharge from rivers in coastal areas. To mitigate the influences stemming from complex coastal environments on the analysis of ocean color, we excluded regions from seas shallower than 200 m and from seas with surface salinities below 25, as determined by ETOPO1 and WOA2013 datasets, respectively (Righetti et al., 2019).

2.2 Data preprocessing

For the OCCCI V5 data, we selected its climatology product as the background field. Because the OCCCI climatology data only provide valid observations for 12 months, temporal smoothing interpolation was performed to cover each ocean grid cell from January 1, 2001, to December 31, 2021. Due to the presence of missing values in both the daily and monthly data products of OCCCI V5, it is not suitable for direct use as model input. Therefore, the climatology product without missing values in the spatial domain was used to set the Chl-a baseline.

As PAR data from different satellite sources were used in this study, preprocessing and bias correction were applied. The overlapping period of MODIS and VIIRS data from 2012 to 2021 was chosen as the reference, using a ratio-based method with MODIS-Aqua as the baseline for bias correction. In cases of missing values in the spatial domain, the three different products were used for complementarity. If effective observational values were not available, linear spatial interpolation was performed. Finally, a spatiotemporally continuous PAR dataset was obtained for model input.

For the reanalysis datasets, as they are already spatiotemporally continuous with a spatial resolution of 0.25°, no additional preprocessing is required. The average of the first five levels of SAL data (approximately 5.14 m) from ORAS5 was taken as the input. It should be noted that ORAS5 has a spatial resolution of 9 km near the polar regions. However, this study does not consider the inversion of Chl-a data in high-latitude areas. Considering the different spatial resolutions of the data, apart from the reanalysis data, the other input data for the model in this study were resampled to 0.25° using the nearest interpolation method.

When using the data mentioned above as inputs for the OCNET model, normalization is necessary. For environmental variables (SST, SSP, SAL, and PAR), normalization was performed according to Eq.1 where the parameters used in the formula were pre-calculated (Table 2). Due to the presence of numerous low values in the Chl-a concentration data in open waters, it is first natural logarithm transformed and then normalized to achieve a uniform distribution of the input data (Eq.2).

Table 2 Maximum, minimum, and mean values of environmental variables obtained f~~or the~~rom satellite and reanalysis datasets ~~environmental variables~~.

Variables	max	min	mean	units
SST	310.06	269.17	286.821	K
SSP	106980	54834	96643	Pa
PAR	70.329	0	32.2007	einstein/(m^2 d)
SAL	43.467	0	34.169	PSU

$$X_{\mathrm{N}} = \frac{X - \overline{X}}{X_{\max} - X_{\min}} \tag{1}$$

$$C_{\mathrm{N}} = \frac{\ln(C) + 4.61}{4.61 \times 2} \tag{2}$$

where X represents different environmental variables, subscript N represents the normalized variables, subscripts "max" and "min" correspond to the maximum and minimum values in Table 2, and $\bar{X}$ represents the mean. C represents Chl-a data. Values of Chl-a concentration lower than 0.01 mg/m^3 were all set to 0.01 mg/m^3. Actually, the accuracy of satellite retrievals cannot reach such a small value.

2.3 Model architecture

Data-driven deep learning algorithms can extract high-level information from multisource input data using multiple non-linear processing layers (Li et al., 2020; Cen et al., 2022). In the research of large-scale, long-term, and multi-data scenarios, deep learning algorithms excel at discovering data patterns and inherent connections (Li et al., 2020; Andersson et al., 2021). Given the applicability of CNNs to satellite remote sensing imagery and climate model data, we constructed the global OCNET model consisting of 405 regional CNNs. Specifically, each CNN employed in the individual regions was based on the U-Net model (referring to ~~(Fig.2)as~~ Figure 2~~ shows~~). U-Net, initially designed for medical image segmentation, is a variant of the CNN (Ronneberger et al., 2015). Across various applications, U-Net has been consistently proven to be highly effective in terms of learning accuracy and pixel-wise mappings (Andersson et al., 2021; Urakubo et al., 2019; Wagner et al., 2019).
Here we applied the OCNET to reconstruct global Chl-a concentration data in open ocean areas, considering environmental variables that are associated with ocean phytoplankton growth and distribution. SST, SAL, and SSP from reanalysis data and PAR from satellite observations were selected as the input of OCNET to correlate with the environment and phytoplankton mass. The whole area considered in this study covers latitude 45 °N to 45 °S, and longitude 180 °W to 180 °E. The open ocean is divided into 45 horizontal and 9 vertical zones, 405 in total. Each area has a size of 16 ° × 16 °and a side length of 64 grid cells. There is an 8 °overlap in the latitude direction between each pair of adjacent regions at the same latitude. Additionally, there is a 6.25 °overlap in the longitude direction between each pair of adjacent regions at the same longitude. This is to reduce the boundary effect caused by dividing regions for network training separately.

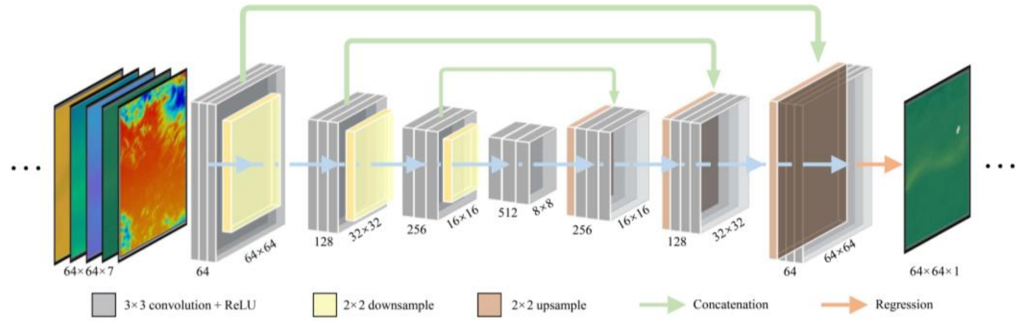


Figure 2 Flowchart of the developed OCNET model in each zone. The OCNET model, comprised of deep learning U-Net models, receives three monthly averaged variables (SST, SAL, and PAR) and two daily real-time variables (SST and SSP) as input. The climatology Chl-a of OCCCI and daily Chl-a data of NOAA MSL12 were treated as background and target set, respectively.

Inputs to the network include Chl-a_OCCCI, Chl-a_N, SST, SAL, SSP, PAR and SST_d. Chl-a_OCCCI is the climatology data from OCCCI, as the background field of the dataset with only one value per month. Considering the typical monthly growth cycle of phytoplankton, we calculated the environmental factors influencing marine algae growth by averaging the data from the preceding month as input variables. Therefore, SST, SAL, and PAR took the average of one month forward as input to OCNET. In addition, the values of SST_d and SSP were also taken as the input of the day, respectively.

There are totally 405 zones of size 64×64 globally. Each zone has its own independent U-Net. Each network undergoes a maximum of 100 training steps to ultimately output the network model for each region. First, the input data with a size of 64×64×7 are passed through the initial convolutional layer, which consists of 64 filters. Each filter has a grid size of 3×3 and a stride of 1. Subsequently, an activation function is applied to the data, and the dimension of the feature map is reduced to half of its original size, resulting in a size of 32×32×64, through a pooling layer operation of size 2×2. After completing this initial step, the subsequent operations follow a similar pattern. The feature map undergoes a halving of its spatial dimension through pooling, while the number of channels is doubled through convolution. The final feature map obtained from these operations has a size of 8×8×512, and it serves as input for the subsequent decoding process. The decoding process mirrors the encoding process described earlier. It is important to note that the encoding and decoding networks are connected through skip connections, enabling the preservation of information that may be lost during downscaling. This U-Net structure facilitates the preservation of detailed information from previous layers during the subsequent decoding stage. Finally, the last layer consists of a single filter that outputs a feature map with a size of 64×64×1, representing a single channel of data. Finally, by inputting environmental information from 2001 to 2021 into the OCNET model, a spatiotemporal continuous dataset of Chl-a concentration was reconstructed, covering the period from 2001 to 2021.

2.4 Statistical tests

2.4.1 Evaluation of OCNET output

In the simulation performed by OCNET, the data from the year 2021 was selected as the testing set. This portion of the data was excluded from model training and validation, and was solely used for evaluating the quality of the final data. The commonly used evaluation metrics, including CC, bias, and RMSE, were employed for this purpose. The specific formulae used for the calculations can be found in Table 3, while the evaluation results are presented in Section 3.2.

Table 3 Statistical metrics used in evaluating the reconstructed Chl-a (C) against the observed data (Cg) from the NOAA MSL12 during the testing period. An overbar donates the mean during evaluation periods. N denotes the number of data pairs. Cov denotes the covariance and σ is the standard deviation.

Performance Score	Score symbol	Equation	
Pearson's correlation coefficient	CC	$\mathrm{CC} = \dfrac{\mathrm{cov}(C, C_{\mathrm{g}})}{\sigma(C)\sigma(C_{\mathrm{g}})}$	(3)
Bias	Bias	$\mathrm{BIAS} = \dfrac{\sum (C - C_{\mathrm{g}})}{\overline{C_{\mathrm{g}}}}$	(4)
Root mean square error	RMSE	$\mathrm{RMSE} = \sqrt{\dfrac{\sum (C - C_{\mathrm{g}})^2}{N}}$	(5)

2.4.2 Evaluation using the ETC method

Due to the lack of enough and reliable in-situ measurements for the assessment of global ocean Chl-a, the extended triple collocation (ETC) method was used to indirectly evaluate the quality of OCNET model output data (Mccoll et al., 2014). The ETC method uses exactly the same assumptions as the triple collocation (TC) method. The TC method utilizes three mutually independent datasets to assess the relative errors of the data without requiring the knowledge of the true value. This method was initially developed by Stoffelen (1998) and has been widely used for soil moisture assessment (Dorigo et al., 2010; Miralles et al., 2010). The ETC method, improved by Mccoll et al. (2014) from the TC method, provides the correlation coefficient as another performance index. The ETC method has also been extensively applied, such as in the evaluation of sea surface temperature data (Gentemann, 2014).

Because the Sentinel-3B data are not used in the OCCCI and NOAA MSL12 datasets, it was selected as an independent dataset for evaluation. Chl-a data products from Sentinel-3B, NOAA MSL12 and OCNET were used in ETC method. Considering the available time period of Sentinel-3B data, the evaluation covers the period from June 7, 2019, to December 31, 2021. Due to the presence of numerous missing values in the Sentinel-3B data products, grid cells with severe missing values, i.e., grid cells with fewer than 30 valid days, were excluded, and the remaining grid cells were retained for evaluation. It should be noted that since OCNET was trained using NOAA MSL12 as the target set, they cannot be considered mutually independent datasets. This evaluation mainly utilizes Sentinel-3B data as a third-party source to validate the reliability of the OCNET model. It is possible that the results of the ETC in some grid cells may yield a negative square of the correlation coefficient or root mean square error. This can happen if the sample size is too small, or if one of the assumptions of ETC is violated. In the final presentation of results, these grid cells were excluded.

The calculation method is based on Eq.6-Eq.11, where C_{ij} represents the covariance between the i-th and j-th data points. The calculated correlation coefficient (tCC) and root mean square error (tRMSE) based on the TC method are denoted as ρ and σ, respectively. It should be noted that the magnitude of the tCC and tRMSE only reflects the relative performance as opposed to the absolute values.

$$\sigma_{t,1} = \sqrt{C_{11} - C_{12}C_{13} / C_{23}} \tag{6}$$

$$\sigma_{t,2} = \sqrt{C_{22} - C_{21}C_{23} / C_{13}} \tag{7}$$

$$\sigma_{t,3} = \sqrt{C_{33} - C_{31}C_{32} / C_{21}} \tag{8}$$

$$\rho_{t,1} = \pm\sqrt{C_{12}C_{13} / C_{11}C_{23}} \tag{9}$$

$$\rho_{t,2} = \pm\mathrm{sign}(C_{13}C_{23})\sqrt{C_{12}C_{23} / C_{22}C_{13}} \tag{10}$$

$$\rho_{t,3} = \pm\mathrm{sign}(C_{12}C_{23})\sqrt{C_{13}C_{23} / C_{33}C_{12}} \tag{11}$$

3. Results

3.1 Spatial variations and trends in global Chl-a estimates during 2001–2021

We have developed high-quality gap-filled Chl-a data in open oceans using the OCNET model. The dataset covers the time period from 2001 to 2021 and has a spatial resolution of 0.25°, with a daily temporal resolution. We applied the natural logarithm transformation to the Chl-a concentration values when generating global maps (referring to ~~(Fig.3)as~~ Figure 3~~ shows~~). This transformation was necessary due to the relatively low Chl-a concentrations in most sea areas but the relatively high concentrations in areas experiencing algal blooms. ~~It can be observed that~~~~By referring to Figure 3, t~~There are high chlorophyll concentrations in the sea areas near the west coast of Africa (~2.2 mg/m^3), the east coast of Asia (~1.1 mg/m^3), and the west coast of the Americas (~2.3 mg/m^3), which indicates a higher likelihood of algal blooms in these regions. Chl-a concentrations near the equator and in regions above 30° latitude are higher than in open ocean regions between 10° and 20° latitude. In addition, oceanic regions far from the continents, such as the Pacific Ocean, Indian Ocean, and Atlantic Ocean, exhibit low chlorophyll concentration distributions (less than 0.05 mg/m^3). This also suggests a higher possibility of algal blooms in coastal areas to some extent.

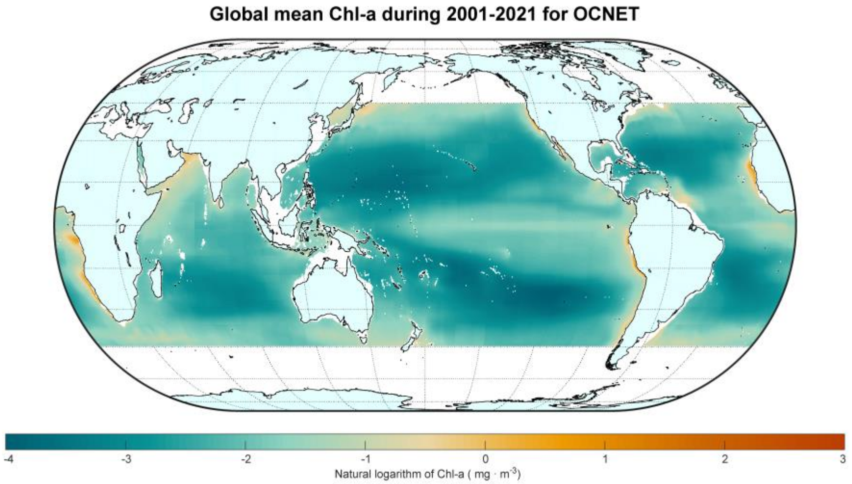


Figure 3 Natural logarithm of the OCNET model output Chl-a during 2001–2021. Light blue represents land areas. White denotes areas that are not considered in this study.

To ensure spatial continuity in the global Chl-a concentration product, the data underwent regional processing before being input into the OCNET model. Subsequently, overlapping region processing and image stitching were performed, resulting in a seamless global Chl-a concentration product without noticeable discontinuity or fragmentation. Although the OCNET model was trained separately for each region, the final results obtained after adequate data preprocessing and sufficient training steps were consistent and globally continuous. This outcome further highlights the effectiveness of the OCNET model in global data reconstruction.

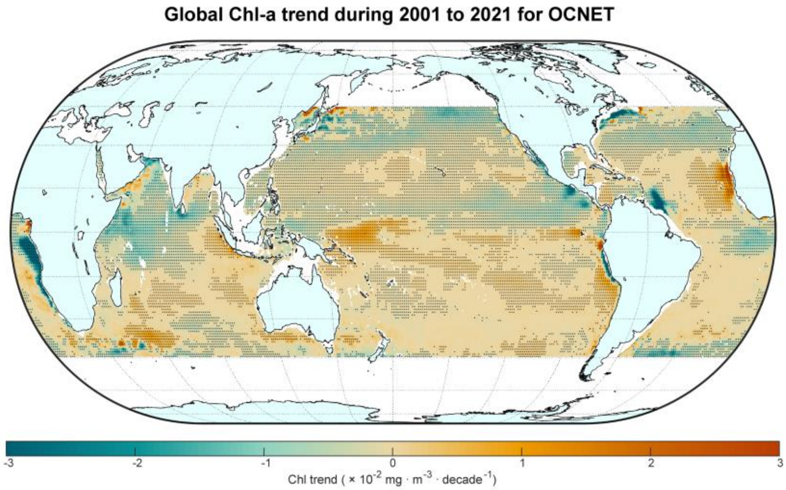


Figure 4 Global Chl-a trends from OCNET over the period Jan 2001–Dec 2021. Regions with significant trends (*p*<0.05) are marked with black dots.

Trends ~~ofin~~ in Chl-a concentration in the global ocean area from 2001 to 2021 according to the output of the OCNET model were derived (referring to Figure 4~~Fig.4~~). To emphasize regions exhibiting clear trends, data in this section were not subjected to natural logarithm transformation and was magnified instead (please note the unit is 10^{-2} mg m^{-3} decade^{-1}). ~~It can be observed that~~~~By referring to Figure 4, i~~In general, the sea areas closer to continental land exhibit more significant trends (referring to Figure 4). Although the sea areas near the west coast of Africa show high chlorophyll concentrations (~~(Fig.3)~~referring to Figure 3), the two hemispheres, northern and southern, exhibit different trend patterns. Specifically, the sea areas on the western side of the northern hemisphere of Africa show a clear upward trend in chlorophyll concentration (~4×10^{-2} mg m^{-3} decade^{-1}), while the sea areas on the western side of the southern hemisphere show a significant downward trend (~-8×10^{-2} mg m^{-3} decade^{-1}). The sea areas near North America predominantly exhibit a noticeable downward trend (~-5×10^{-2} mg m^{-3} decade^{-1}). The islands around the northern part of South America show a pronounced decrease in chlorophyll concentration (~-5×10^{-2} mg m^{-3} decade^{-1}), while the sea areas on the western side exhibit distinct increasing or decreasing trends at different latitudes. The chlorophyll concentration variation around Japan in eastern Asia shows the most significant trend. The sea areas near Japan demonstrate a decrease in chlorophyll concentration at lower latitudes and an increase at higher latitudes. In general, there are more areas in the open oceans worldwide where Chl-a concentration shows a decreasing trend than areas where it shows an increasing trend.

3.2 Temporal variations in Chl-a estimates in different ocean regions

To facilitate the analysis and evaluation of regional data, we divided the study area into 10 regions based on latitude, longitude, and the ranges of oceans (referring to ~~(Fig.5)~~~~as~~ Figure 5~~ shows~~). The division of sea areas considered the characteristics of the regions and the influence of ocean currents, taking into account the division of biogeochemical provinces (Reygondeau et al., 2013). To avoid excessive complexity resulting from overly detailed regional divisions, a final selection of 10 regions was determined. This study calculated and presented the Chl-a concentration products for these 10 regions in a 20-year time series ~~(Fig.5)~~. Due to the OCNET model's target dataset being NOAA MSL12, the output results of the OCNET model are consistent with NOAA MSL12 after February 9, 2018. However, ~~it can be observed~~~~Figure 5 indicates that~~ the results of the OCNET model are noticeably lower than the results of OCCCI V5, particularly in regions 2, 4, 6, 7, 8, and 9 (referring to Figure 5). The primary reason for this systematic bias is the discrepancy between the NOAA MSL12 data and the OCCCI V5 data products.

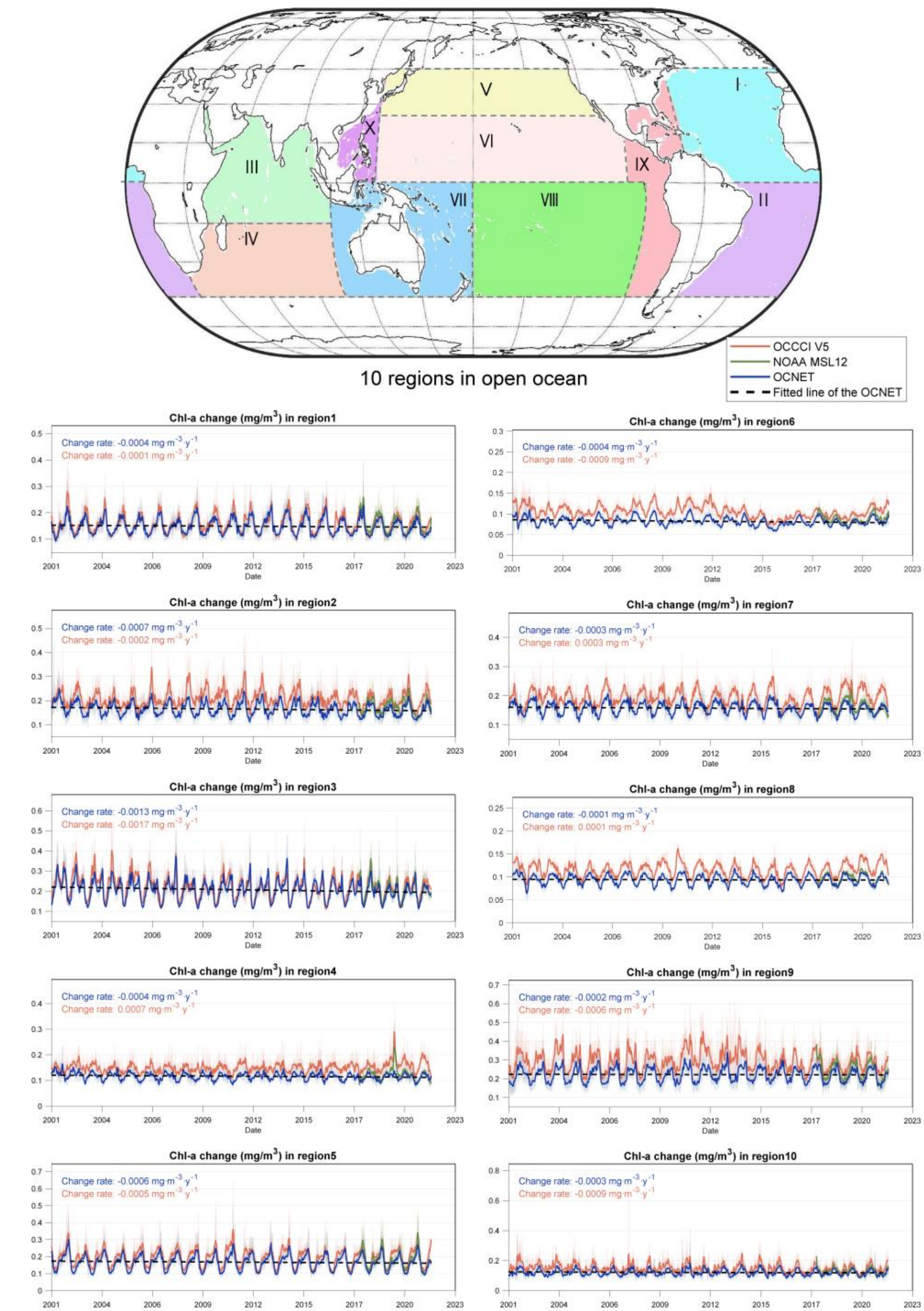


Figure 5 Global open ocean was divided into 10 regions in this study, and the temporal variations of Chl-a from 2001 to 2021 are shown for each region. The blue line represents the output results of the OCNET model, the red line represents the results from OCCCI V5, the green line represents the results from NOAA MSL12 data, and the dark dashed line represents the linear fit of OCNET. The trends of OCCCI V5 and the OCNET model outputs during 2001 to 2021 are indicated with their respective color

labels in the top left corner of the temporal variation plot. For comparison purposes, we only consider and display calculations based on grid cells with valid values from OCCCI V5.

~~It can be observed that~~~~By referring to Figure 5, t~~The long-term Chl-a concentration trends in most regions are relatively small, with changes within 0.001 mg/m^3 per year, except for Region 3 (referring to Figure 5). In terms of seasonal variations, regions 3, 5, and 9 exhibit larger intra-annual fluctuations. On the other hand, regions 4, 7, and 8, which encompass a wider range of low Chl-a concentrations (~~(Fig.3)~~referring to Figure 3), show smaller seasonal fluctuations. It is worth noting that OCCCI V5 and OCNET show significant deviations in Region 9, where there are higher Chl-a concentrations (particularly during the period from 2010 to 2015). Considering that Region 9 mainly covers the sea areas surrounding the Americas ~~(Fig.5)~~, it is likely influenced by human activities. Additionally, the satellite retrieval of Chl-a concentration data in this region is of poorer quality due to high sediment concentrations and turbidity near the coastline. This partially explains the significant interannual variability observed in OCCCI V5 products for Region 9. Furthermore, both OCCCI V5 and NOAA MSL12 products have instances of unusually high Chl-a values, such as in regions 7 and 10 for OCCCI V5, and Region 4 for NOAA MSL12. These abnormally high Chl-a concentrations, which surpass typical values for the respective years, could be due to algal blooms or satellite data quality issues. Overall, from the long-term trends, most regions show small magnitudes of change, with more regions exhibiting a decreasing trend.

3.3 Evaluation of OCNET's performance

The target dataset for the OCNET model is the NOAA MSL12 data product with a time span from February 9, 2018, to December 31, 2021. Details of model construction are explained in Section 2.3, where the data were divided into training, validation, and testing sets in a ratio of 7:1.5:1.5. Three statistical metrics, i.e., CC, bias, and RMSE, were selected to evaluate the training performance of the OCNET model (~~Fig.~~Figure 6 and Table 4) for different regions (referring to ~~Fig.~~Figure 5).

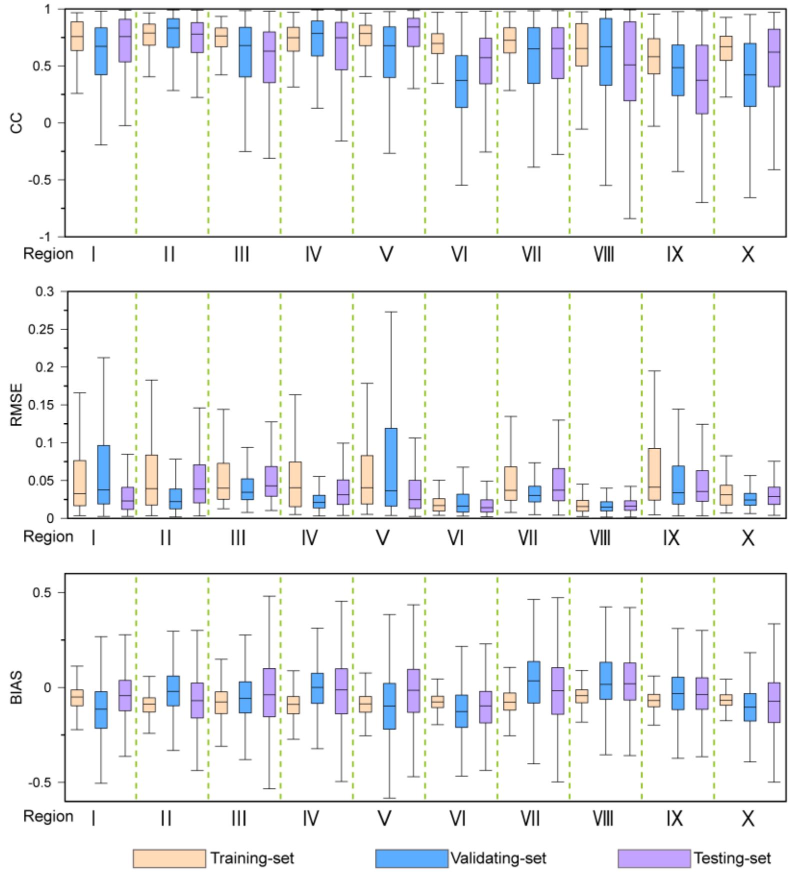


Figure 6 Boxplot of evaluation results of the OCNET model in each region.

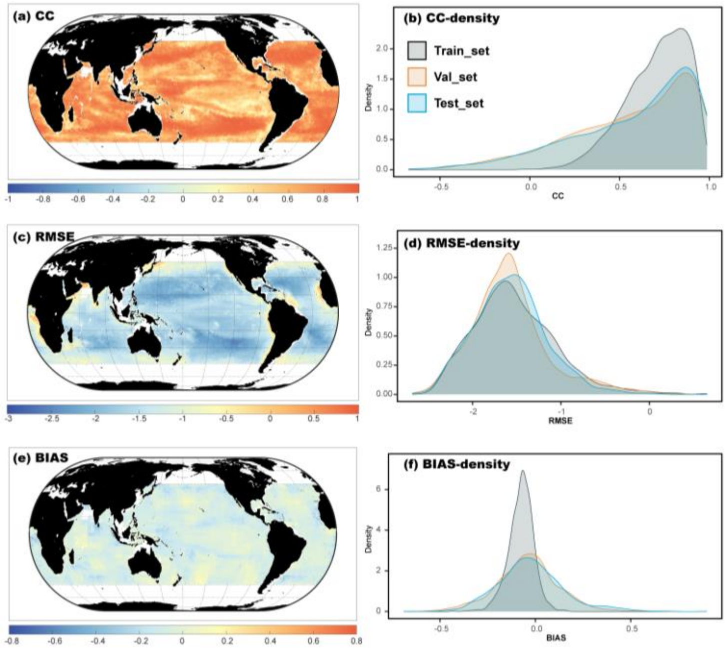
Table 4 Chl-a concentration change rates for each region from three datasets and the median values of evaluation metrics (including the training set, validating set, and testing set) for the OCNET model.

Region	Change rate (×10^{-4} mg m^{-3} year^{-1})			Evaluation Index		
	OCNET	OCCCI	NOAA MSL12	CC	Bias	RMSE
1	-3.5	-1.1	-103.3	0.73	-0.06	0.03
2	-7.1	-1.9	-1.5	0.78	-0.08	0.04
3	-13.3	-16.7	-91.2	0.75	-0.07	0.04
4	-4.0	7.1	-17.6	0.72	-0.08	0.04
5	-5.9	-4.9	-94.4	0.76	-0.09	0.04
6	-3.5	-8.6	3.5	0.66	-0.09	0.02
7	-3.1	3.2	-26.0	0.71	-0.06	0.04
8	-0.8	1.0	6.1	0.63	-0.03	0.02
9	-1.6	-6.1	-96.0	0.56	-0.07	0.04

10	-3.5	-9.1	-29.8	0.64	-0.08	0.03	

From the daily evaluation, it is shown that ~~Figure 6 showsit can be seen that~~ the model performs well (referring to Figure 6) ~~(Fig.6)~~. The median values of CC for the training set are mostly above 0.6. The performance of the ~~validation~~ validating set and the testing set is similar, but individual regions show poor performance. For example, in the validating~~validation~~ set, the median values of CC for regions 6, 9, and 10 are around 0.4 and 0.5, and for region 9 in the testing set, the median value of CC is around 0.4. This corroborates the findings in Section 3.1 that region 9, being mostly near the American continent, is heavily influenced by human activities, and the satellite data quality in coastal areas is also poorer. In terms of bias, the performance of the training set is excellent, with biases within a small range for each region. The boxplot ranges for the ~~validation~~validating set and the testing set also fluctuate within 0.2. It is worth noting that, due to the low Chl-a concentrations in most marine areas, the calculated biases are defined as relative biases (with the denominator being the mean of the target dataset). Therefore, it is possible to have higher biases in regions with low Chl-a concentrations. For the RMSE, both the training set, ~~validation~~ validating set, and ~~test~~testing set are below 0.2, with most of them below 0.1, indicating excellent performance. Regions 6 and 8 have the lowest RMSE values. This may be because regions 6 and 8 mostly cover low Chl-a concentration offshore areas with minimal seasonal fluctuations (~~(Fig.5)~~referring to Figure 5).

According to the results of the Chl-a concentration rate of change in each region, it can be observed that most regions show relatively small trends ~~(~~as Table 4 shows~~)~~. Most regions exhibit a decreasing trend, which is consistent with the conclusions of existing related studies (Le Grix et al., 2021; Beaulieu et al., 2013; Signorini et al., 2015). Based on the results of the OCNET model, regions 2, 3, and 5 show larger decreasing magnitudes, ~~while~~ compared to the other regions, which also exhibit a decreasing trend. According to the results of OCCCI, except for regions 4, 7, and 8, which show small increasing trends, the other regions demonstrate a decreasing trend, with regions 3, 6, 9, and 10 showing more pronounced declines. As for NOAA MSL12, except for regions 6 and 8, which show an upward trend, the other regions display a decreasing trend. Due to the relatively short time series of NOAA MSL12, it cannot reflect long-term trend characteristics. It can be seen that NOAA MSL12 shows a significant decrease in Chl-a concentration in regions 1, 3, 5, 7, 9, and 10. This overall decline exhibited by NOAA MSL12 directly influences the training results of the OCNET model. Therefore, OCNET and OCCCI share similarities in long-term trends but may have differences in individual regions.



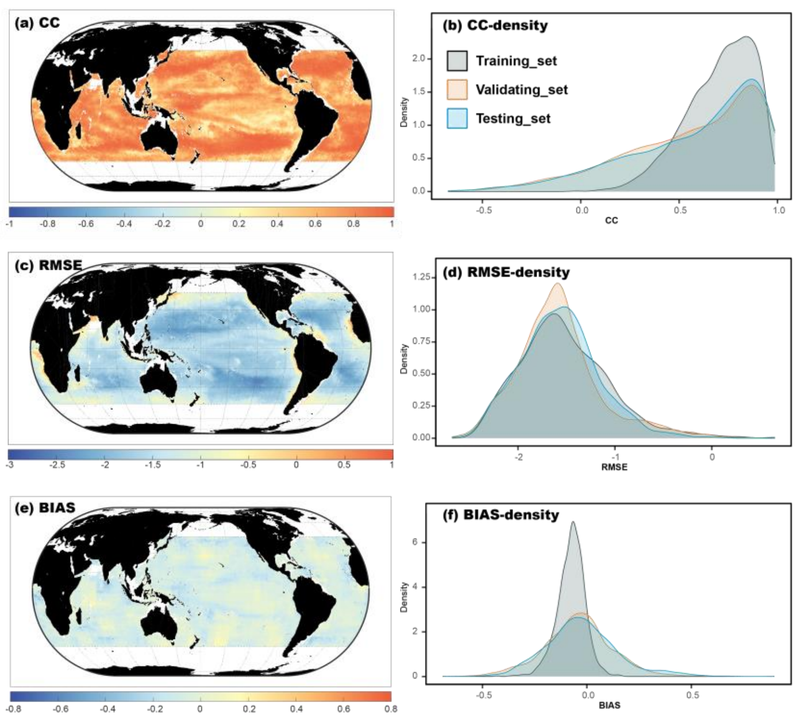


Figure 7 Evaluation results of OCNET based on the NOAA MSL12 dataset. a), c), and e) represent the global distribution maps of CC, RMSE, and bias from February 9, 2018, to December 31, 2021, respectively. b), d), and f) display the density distributions of

the evaluation results for CC, RMSE, and bias across the training, ~~validation~~validating, and testing sets, respectively. Note that RMSE took logarithm base 10.

From the comparison results with the target data NOAA MSL12~~ (Fig.7)~~, the OCNET model has effectively learned the relationship between environmental data and Chl-a concentration variations (referring to Figure 7). At the global scale, the overall performance of CC is good, with most regions above 0.7. ~~Regions~~ Regions with lower CC are mainly concentrated in the eastern tropical~~central~~ Pacific, where the OCNET model output shows apparent systematic biases compared to OCCCI (referring to Figure 5)~~ (Fig.7a)~~. Due to the low~~er~~ mean Chl-a concentration in Region 8 ~~((Fig.3)~~referring to Figure 3), the~~its~~ RMSE and ~~bias~~ BIAS of Region 8~~performance~~ are ~~also~~ better than other regions~~ (Fig.7 (c-e))~~. The preliminary evaluation results in Region 8 suggest that OCNET's performance is not as good as in other areas. This may be related to the specific climate characteristics or low satellite data quality in that region. The complex factors ultimately result in OCNET's less optimal learning effect in Region 8. For the density distribution maps of the evaluation results for the training, validation, and testing sets, the performance of the training set is generally excellent. The performance of the validation and testing sets is comparable. From the results of bias, the training set shows a clear tendency of underestimation (referring to Figure 7d) ~~(Fig.7d)~~, ~~while~~ compared to the ~~validation~~ validating and testing sets, which~~ also~~ exhibit ~~some~~ a less underestimation but less pronounced. This may be due to the smoothing effect of OCNET on some abnormally high values in the satellite data (Section 3.2). In summary, the evaluation results indicate that OCNET performs exceptionally well in the reconstruction of global open ocean Chl-a concentration data.

3.4 Extended triple collocation evaluation

Output of the OCNET model, NOAA MSL12, and Sentinel-3B's Chl-a concentration data were selected for the ETC evaluation method~~ (Fig.8)~~. It should be noted that the Sentinel-3B dataset was considered independent of the other two datasets, while output of OCNET is not independent of the NOAA MSL12 dataset. Therefore, the evaluation results are biased towards OCNET and NOAA MSL12 data and may underestimate Sentinel-3B data. The purpose of the ETC method evaluation here was to demonstrate the quality of OCNET output data compared to NOAA MSL12 data. The low absolute values of the evaluation results do not necessarily imply that Sentinel-3B dataset is unreliable. Additionally, due to algorithmic reasons, grid cells with outlier data were excluded. To highlight relevant information, ~~Fig.8~~Figure 8 only includes the results of Sentinel-3B data in the interval distributions (e) and (f) while omitting the global distribution of the metrics (which mostly perform worse). It should be noted that tCC and tRMSE mentioned in Section 3.4 are different from those in Section 3.3. The metrics in Section 3.4 can only reflect the relative ranking.

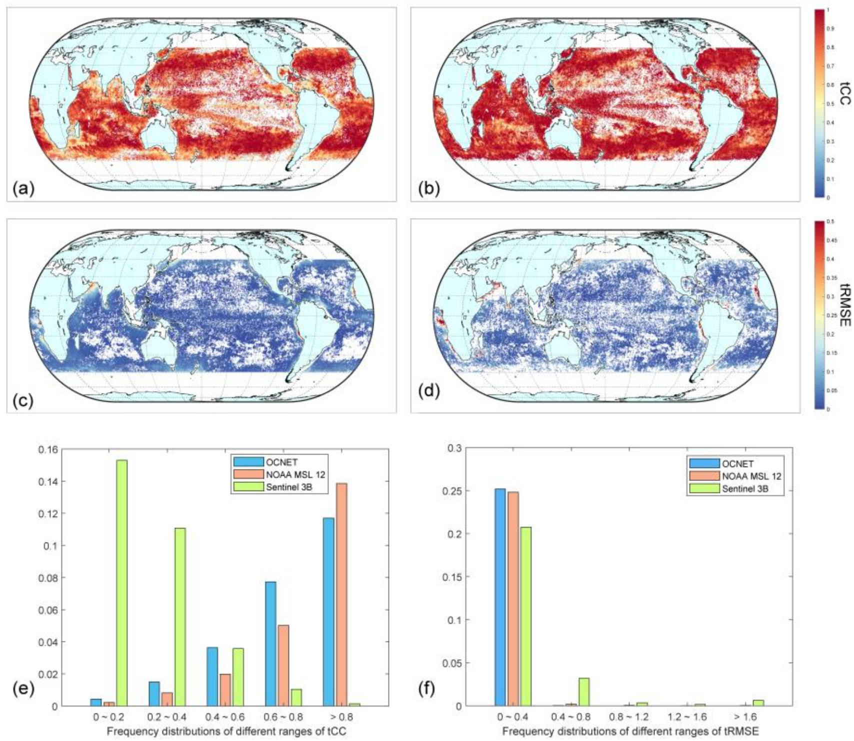


Figure 8 Evaluation results based on the ETC evaluation method. Global distribution map of the tCC of Chl-a for a) OCNET and b) NOAA MSL12. Global distribution map of the tRMSE of Chl-a for c) OCNET and d) NOAA MSL12. Interval distribution of e) tCCs and f) tRMSEs for the three products was calculated using the ETC method.

By referring to Figure 8(a–b), ~~It can be seen that~~ the output data of the OCNET model show a similar distribution to NOAA MSL12 data in the global tCC distribution ~~(Fig.8(a–b))~~, with most regions above 0.7. In the interval distribution ~~(Fig.8e)~~(referring to Figure 8e), the proportion of OCNET model output data exceeding 0.8 is approximately 12%, slightly lower than NOAA MSL12's 14%. Regions with poorer tCC evaluation results are mainly distributed in the ~~central~~ eastern tropical Pacific ~~Ocean~~ where there are significant missing values in tCC, and the performance of OCNET is slightly lower than other regions. It should be noted that in the ocean areas near the American continent, there is a prevalent occurrence of tCC values below 0.5. This is similar to the evaluation results in Section 3.3, where the training performance for Region 9 is also slightly lower than other regions ~~(Fig.7)~~(referring to Figure 7). Additionally, in the southern hemisphere region of the Atlantic Ocean, OCNET seems to exhibit higher tCC values in the middle compared to the northern and southern sides, which is a different characteristic from the NOAA MSL12 dataset.

By referring to Figure 8f, ~~From~~ from the results of tRMSE, the model output of OCNET is slightly better than NOAA MSL12 data. Specifically, NOAA MSL12 exhibits poorer tRMSE performance in the sea area near the west coast of Africa. This area is also characterized by high Chl-a concentration and significant interannual variations ~~(~~(referring to Fig~~.3~~ure 3 and Figure 4). ~~Fig.4).~~ While OCNET exhibits similar high tRMSE values in the ocean areas near the western side of Africa as NOAA MSL12, the distribution range is smaller compared to NOAA MSL12. Additionally, in the ocean areas near South America, both OCNET and NOAA MSL12 show small-scale high tRMSE values. It is worth mentioning that even for Sentinel-3B, the majority of tRMSE values are concentrated below 0.4. This may be related to the fact that most of the ocean Chl-a concentrations are relatively low.

4. Discussion

4.1 Factors affecting the distribution of marine phytoplankton

We focused on the surface chlorophyll-a (Chl-a) concentration as an indicator of the distribution of phytoplankton in the ocean surface layer. The distribution of marine phytoplankton is influenced by various factors, including light, temperature, nutrients, salinity, hydrodynamic conditions, and biological interactions (Behrenfeld et al., 2006; Ducklow et al., 2022; Feng et al., 2021). Among them, light, temperature, salinity, and hydrodynamics are directly reflected in the input data of the OCNET model. However, the influences of nutrients and biological interactions are more complex. Different phytoplankton communities require different major nutrients such as nitrogen, phosphorus, and silicon (Powell et al., 2015; Takeda, 1998). The biological interactions also include predation by zooplankton and the impact of human activities in coastal areas. Due to the lack of publicly available reliable quantitative data on these two aspects, they are not considered in this study.

Considering the correlation between SST, SAL, and PAR with the growth cycle of phytoplankton, when creating input data samples for the OCNET model, the mean values from one-month prior were selected as variables. However, hydrodynamic conditions have real-time effects on the distribution of planktonic algae, so SSP and SST were taken as daily values for input. It is worth mentioning that surface wind speed variations also have a direct impact on the movement of surface phytoplankton in the ocean. However, wind speed not only includes direction and magnitude, but it also fluctuates significantly in both direction and magnitude within a day. Therefore, simply taking daily averages as model inputs would not suffice. Additionally, selecting too many variables can lead to overfitting or poor training performance due to limitations in the quantity of sample data.

According to the result of OCNET model (referring to Figure 3), ~~it can be observed that~~ regions with higher Chl-a concentration are generally located near continents ~~(Fig.3)~~. From the perspective of hydrodynamic conditions, hydrological factors such as water currents, ocean currents, and tides have a significant influence on the distribution and aggregation of phytoplankton. They affect the horizontal migration, vertical mixing, and nutrient transport of phytoplankton. The nearshore waters have relatively low seawater velocity, coupled with features such as coastlines, underwater ridges, and archipelagos, which to some extent contribute to the retention and aggregation of phytoplankton. In addition, river inflows into the ocean often bring abundant nutrients (Slomp, 2011; Wang et al., 2016), creating favorable conditions for the growth of phytoplankton (Liu et al., 2022).

Global variations in ocean temperature also have an important impact on the growth of phytoplankton. With the continued increase in global sea temperatures, temperature anomalies can also lead to anomalies in Chl-a concentration (Liu et al., 2022;

Gruber et al., 2021; Le Grix et al., 2021). Global ocean warming results in more pronounced stratification of the ocean, altering the depth of the mixed layer and reducing vertical mixing between the surface layer and the cold, nutrient-rich layer below (Liu et al., 2022; Le Grix et al., 2021). The reduction in nutrients ultimately leads to a decrease in Chl-a concentration in the ocean surface layer. However, the declining trend in Chl-a concentration over the 20 years does not necessarily indicate a reduction in algal blooms. On the contrary, the frequency of extreme events associated with algal blooms may be continuously increasing due to the influence of climate change (Feng et al., 2021; Dai et al., 2023).

In conclusion, understanding and studying these influencing factors are crucial for comprehending the ecological and biogeochemical processes of marine phytoplankton. A thorough investigation of the interactions among these factors can lead to better predictions and explanations of the growth and distribution patterns of phytoplankton. Subsequent research can further focus on the impact of human activities in coastal areas on the growth of marine phytoplankton.

4.2 Uncertainty in ocean color data from satellite remote sensing

Satellite remote sensing is one of the important technologies to obtain long-term and large-scale ocean color data (Groom et al., 2019). However, there is uncertainty in the satellite data inversion process, degrading the accuracy (Groom et al., 2019; Hu et al., 2019a; Jiang and Wang, 2013). The first factor is the influence of atmospheric correction algorithms. The selection of models and parameters can affect the satellite data during the atmospheric correction process. In addition, coastal areas and inland lakes closer to land often have more turbid waters, and the presence of high concentrations of suspended particles in complex water environments makes it more difficult for satellites to accurately retrieve water color information from the water surface (Lian et al., 2021; Wang et al., 2021; Zheng and Digiacomo, 2017). Furthermore, weather and environmental factors such as clouds and fog can partially or completely obscure the target water areas, posing important challenges to satellite data acquisition (Zheng and Digiacomo, 2017; Wang et al., 2021).

In practical applications, in-situ measurements are typically used to calibrate and fit parameters for satellite data (Hu et al., 2012). However, obtaining a large amount of continuous shipborne measurement data is challenging, and publicly available in-situ data often suffer from problems such as inconsistent formats, varying measurement standards, complex composition of research institutions, and unclear data quality. Therefore, the application of in-situ data is limited for studying large-scale, long-term time series of Chl-a concentration variations. In this study, to further demonstrate the data quality of OCNET outputs comparable to NOAA MSL12, an indirect evaluation method using ETC was employed. The evaluation results not only confirmed the excellent training performance of the OCNET model but also indicated significant differences among different satellite products, as evidenced by the low evaluation indicators for Sentinel-3B (referring to Figure 7) ~~(Fig.7)~~. Therefore, when applying satellite products of Chl-a concentration data, it is important to carefully select and correct for biases (Krug et al., 2017). It should be noted that the global distribution map of ETC evaluation results shows a significant number of missing values ~~(Fig.7)~~ (referring to Figure 7). These missing values primarily stem from data gaps in Sentinel-3B and issues within the algorithm itself, resulting in negative squared evaluation metrics. Short data sequences or data that do not conform to the algorithm's underlying assumptions can lead to unusable results from the ETC algorithm. The final result analysis is based solely on grid cells with valid values.

The main purpose of our study was to address the serious issue of spatial missing values in existing satellite datasets. It should be noted that the satellite-derived water color data itself still have errors that are difficult to correct (Wang et al., 2021). To improve accuracy, algorithms can be applied to differentiate different concentrations, such as OCx and CI algorithms (Hu et al., 2012), or specific parameter fitting can be performed for different regions (Li et al., 2019). However, the accuracy of satellite sensors, resolution, and other factors still influence the inversion accuracy. The accuracy of satellite data is not the focus of this study. Nevertheless, satellite data can still provide important references for algal blooms on a global scale (Wang et al., 2021; Feng et al., 2021). An anomaly algorithm can also be used to reduce the impact of systematic biases (Wang et al., 2021; Stumpf, 2001). It may be beneficial to employ machine learning techniques for anomaly of Chl-a concentration, enabling better prediction of extreme events.

4.3 Applications of OCNET in the future

The variation of Chl-a concentration in the global ocean surface is influenced by various complex factors, which poses challenges to accurately retrieve Chl-a concentration. We selected Chl-a data products retrieved from satellite data as a reference, supplemented by reanalysis data to provide environmental factor information. By combining the advantages of machine learning in big data analysis and simulation, we ultimately reconstructed a global-scale, long-term time series of Chl-a concentration dataset.

It is worth noting that this study intentionally excluded coastal regions in the selection of the study region, due mostly to the poor performance of satellite data in coastal regions. Currently, most satellite data algorithms for Chl-a retrieval are based on the absorption peak of Chl-a in the blue spectral band (Hu et al., 2019b; Hu et al., 2012). This approach is highly applicable in open waters but can be significantly affected by interference in coastal regions, particularly in cases of high suspended matter concentration or colored dissolved organic matter (CDOM) (Blondeau-Patissier et al., 2014). Although adjustments can be made to the retrieval algorithms based on localized measurements, there is significant variability in water composition across different coastal regions. This has resulted in poor performance of current satellite retrieval algorithms for estimating global Chl-a concentrations in coastal areas (Dai et al., 2023).

The performance of OCNET in coastal areas is primarily limited by the quality of the input satellite data. The construction of the OCNET model can be affected if the training-set quality is poor or severely lacking. However, OCNET has demonstrated its potential application in open waters. In regional calculations, OCNET can effectively capture the interrelationships among various environmental factors in different zones and apply them to the reconstruction of Chl-a concentrations. There have also been successful studies applying machine learning to analyze Chl-a concentration variations at the regional scale (Chen et al., 2019; Roussillon et al., 2023), further demonstrating the potential application of machine learning methods in coastal areas. In the future, if reliable water color data from coastal areas can be obtained with a certain time span and spatiotemporal continuity for training OCNET, the reconstruction of Chl-a concentrations in coastal regions may also yield favorable results.

Overall, OCNET is capable of surpassing traditional machine learning methods such as multiple linear regression and random forest, as well as traditional artificial neural networks, because it can learn complex nonlinear relationships and incorporate global context into its predictions. This is of great significance for in-depth understanding and analysis of variable changes under the complex environmental influences in the context of big data.

5. Data availability

The reconstructed Chl-a data are archived and available at https://doi.org/10.5281/zenodo.10011908~~https://doi.org/10.5281/zenodo.8105194~~ (Hong et al., 2023).

6. Conclusion

We developed the OCNET model for the purpose of reconstructing global ocean Chlorophyll-a (Chl-a) concentration data. Chl-a is an important indicator of the health and productivity of marine ecosystems, and accurate measurements of Chl-a concentrations are essential for understanding the dynamics of these systems. The OCNET model is based on a convolutional neural network and considers a variety of environmental variables that are known to influence the growth and distribution of ocean phytoplankton, which are the primary producers of Chl-a.

Our results show that the OCNET model performs very well in reconstructing Chl-a concentrations, accurately capturing the temporal variations of these features. This suggests that the model has strong potential for use in large-scale ocean color data reconstruction, and may even be able to predict Chl-a concentration trends in response to changes in the environment. However, we did observe that the model's performance was somewhat weaker in the eastern tropical Pacific region compared to other areas. This may be due to specific climate characteristics that have a significant impact on phytoplankton growth and distribution (Geng et al., 2022; Duteil and Park, 2023) or low quality of satellite-based dataset in this region. The model's performance in the eastern tropical Pacific region requires further improvement in future work.

Overall, the OCNET model represents an important step forward in the use of machine learning techniques for predicting and reconstructing Chl-a concentrations. The model's strong performance in all regions of the globe suggests that it could be a valuable tool for understanding and predicting the dynamics of marine ecosystems on a global scale. OCNET and other machine learning tools will help us better understanding and predicting the change of marine phytoplankton under climate change. It is hoped that the results of this study will be of interest and relevance to a wide range of researchers, policymakers, and managers involved in the monitoring and management of aquatic ecosystems.

Author contributions

LD and HZ designed the research. HZ, LD, LX, and WY developed the approaches and data sets. HZ, LD, LX, WY, ZJ, MA and MM contributed to the analysis of results and writing of the paper.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

This study was supported by the National Water and Energy Center, United Arab Emirates University, through the Asian University Alliance (AUA) program (Grants 31R281-AUA-NWEC-4-2020 and 12R023-AUA-NWEC-4-2020). Reviewers and editors' comments on improving this study and manuscript will be highly appreciated. The authors thank the researchers and their teams for providing all the datasets used in this study.

References

Andersson, T. R., Hosking, J. S., Perez-Ortiz, M., Paige, B., Elliott, A., Russell, C., Law, S., Jones, D. C., Wilkinson, J., Phillips, T., Byrne, J., Tietsche, S., Sarojini, B. B., Blanchard-Wrigglesworth, E., Aksenov, Y., Downie, R., and Shuckburgh, E.: Seasonal Arctic sea ice forecasting with probabilistic deep learning, Nature Communications, 12, 10.1038/s41467-021-25257-4, 2021.

Beaulieu, C., Henson, S. A., Sarmiento, J. L., Dunne, J. P., Doney, S. C., Rykaczewski, R. R., and Bopp, L.: Factors challenging our ability to detect long-term trends in ocean chlorophyll, Biogeosciences, 10, 2711-2724, 10.5194/bg-10-2711-2013, 2013.

Behrenfeld, M. J., O'Malley, R. T., Siegel, D. A., McClain, C. R., Sarmiento, J. L., Feldman, G. C., Milligan, A. J., Falkowski, P. G., Letelier, R. M., and Boss, E. S.: Climate-driven trends in contemporary ocean productivity, Nature, 444, 752-755, 10.1038/nature05317, 2006.

Blondeau-Patissier, D., Gower, J. F. R., Dekker, A. G., Phinn, S. R., and Brando, V. E.: A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans, Progress in Oceanography, 123, 123-144, 10.1016/j.pocean.2013.12.008, 2014.

Cao, Z. G., Ma, R. H., Duan, H. T., Pahlevan, N., Melack, J., Shen, M., and Xue, K.: A machine learning approach to estimate chlorophyll-a from Landsat-8 measurements in inland lakes, Remote Sensing of Environment, 248, 10.1016/j.rse.2020.111974, 2020.

Catipovic, L., Matic, F., and Kalinic, H.: Reconstruction Methods in Oceanographic Satellite Data Observation-A Survey, Journal of Marine Science and Engineering, 11, 10.3390/jmse11020340, 2023.

Cen, H. B., Jiang, J. H., Han, G. Q., Lin, X. Y., Liu, Y., Jia, X. Y., Ji, Q. Y., and Li, B.: Applying Deep Learning in the Prediction of Chlorophyll-a in the East China Sea, Remote Sensing, 14, 10.3390/rs14215461, 2022.

Chen, S., Hu, C., Barnes, B. B., Xie, Y., Lin, G., and Qiu, Z.: Improving ocean color data coverage through machine learning, Remote Sensing of Environment, 222, 286-302, 10.1016/j.rse.2018.12.023, 2019.

Dai, Y., Yang, S., Zhao, D., Hu, C., Xu, W., Anderson, D. M., Li, Y., Song, X.-P., Boyce, D. G., Gibson, L., Zheng, C., and Feng, L.: Coastal phytoplankton blooms expand and intensify in the 21st century, Nature, 615, 280-284, 10.1038/s41586-023-05760-y, 2023.

Dorigo, W. A., Scipal, K., Parinussa, R. M., Liu, Y. Y., Wagner, W., de Jeu, R. A. M., and Naeimi, V.: Error characterisation of global active and passive microwave soil moisture datasets, Hydrology and Earth System Sciences, 14, 2605-2616, 10.5194/hess-14-2605-2010, 2010.

Ducklow, H., Cimino, M., Dunton, K. H., Fraser, W. R., Hopcroft, R. R., Ji, R., Miller, A. J., Ohman, M. D., and Sosik, H. M.: Marine Pelagic Ecosystem Responses to Climate Variability and Change, BioScience, 72, 827-850 10.1093/biosci/biac050, 2022.

Duteil, O. and Park, W.: Future changes in atmospheric synoptic variability slow down ocean circulation and decrease primary productivity in the tropical Pacific Ocean, Npj Climate and Atmospheric Science, 6, 10.1038/s41612-023-00459-3, 2023.

Feng, L. and Hu, C. M.: Comparison of Valid Ocean Observations Between MODIS Terra and Aqua Over the Global Oceans, Ieee Transactions on Geoscience and Remote Sensing, 54, 1575-1585, 10.1109/tgrs.2015.2483500, 2016.

Feng, L., Dai, Y., Hou, X., Xu, Y., Liu, J., and Zheng, C.: Concerns about phytoplankton bloom trends in global lakes, Nature, 590, E35-E47, 10.1038/s41586-021-03254-3, 2021.

Flynn, K. J.: A mechanistic model for describing dynamic multi-nutrient, light, temperature interactions in phytoplankton, Journal of Plankton Research, 23, 977-997, 10.1093/plankt/23.9.977, 2001.

Geng, T., Cai, W. J., Wu, L. X., Santoso, A., Wang, G. J., Jing, Z., Gan, B. L., Yang, Y., Li, S. J., Wang, S. P., Chen, Z. H., and McPhaden, M. J.: Emergence of changing Central-Pacific and Eastern-Pacific El Nino-Southern Oscillation in a warming climate, Nature Communications, 13, 10.1038/s41467-022-33930-5, 2022.

Gentemann, C. L.: Three way validation of MODIS and AMSR-E sea surface temperatures, Journal of Geophysical Research-Oceans, 119, 2583-2598, 10.1002/2013jc009716, 2014.

Groom, S., Sathyendranath, S., Ban, Y., Bernard, S., Brewin, R., Brotas, V., Brockmann, C., Chauhan, P., Choi, J.-k., Chuprin, A., Ciavatta, S., Cipollini, P., Donlon, C., Franz, B., He, X., Hirata, T., Jackson, T., Kampel, M., Krasemann, H., Lavender, S., Pardo-Martinez, S., Mélin, F., Platt, T., Santoleri, R., Skakala, J., Schaeffer, B., Smith, M., Steinmetz, F., Valente, A., and Wang, M.: Satellite Ocean Colour: Current Status and Future Perspective, Frontiers in Marine Science, 6, 485 10.3389/fmars.2019.00485, 2019.

Gruber, N., Boyd, P. W., Frölicher, T. L., and Vogt, M.: Biogeochemical extremes and compound events in the ocean, Nature, 600, 395-407, 10.1038/s41586-021-03981-7, 2021.

Han, Y. and Zhou, Y. T.: Investigating biophysical control of marine phytoplankton dynamics via Bayesian mechanistic modeling, Ecological Modelling, 474, 10.1016/j.ecolmodel.2022.110168, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Munoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Holm, E., Janiskova, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thepaut, J. N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146, 1999-2049, 10.1002/qj.3803, 2020.

Hilborn, A. and Costa, M.: Applications of DINEOF to Satellite-Derived Chlorophyll-a from a Productive Coastal Region, Remote Sensing, 10, 10.3390/rs10091449, 2018.

Ho, J. C., Michalak, A. M., and Pahlevan, N.: Widespread global increase in intense lake phytoplankton blooms since the 1980s, Nature, 574, 667-670, 10.1038/s41586-019-1648-7, 2019.

Hong, Z. K., Long, D., Li, X. D., Wang, Y. M., Zhang, J. M., Mohamed, A. H., and Mohamed, M. M.: OCNET global daily Chlorophyll-a products [dataset], https://doi.org/10.5281/zenodo.10011908, 2023.

Hu, C., Lee, Z., and Franz, B.: Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference: A NOVEL OCEAN CHLOROPHYLL a ALGORITHM, Journal of Geophysical Research: Oceans, 117, 10.1029/2011JC007395, 2012.

Hu, C., Feng, L., Lee, Z., Franz, B. A., Bailey, S. W., Werdell, P. J., and Proctor, C. W.: Improving Satellite Global Chlorophyll-a Data Products Through Algorithm Refinement and Data Recovery, Journal of Geophysical Research: Oceans, 124, 1524-1543 10.1029/2019JC014941, 2019a.

Hu, C. M., Feng, L., Lee, Z. P., Franz, B. A., Bailey, S. W., Werdell, P. J., and Proctor, C. W.: Improving Satellite Global Chlorophyll a Data Products Through Algorithm Refinement and Data Recovery, Journal of Geophysical Research-Oceans, 124, 1524-1543, 10.1029/2019jc014941, 2019b.

Information, N. N. C. f. E. and Doc/Noaa/Nesdis/Ncei National Centers for Environmental Information, N. N. U. S. D. o. C.: ETOPO1 1 Arc-Minute Global Relief Model, NOAA National Centers for Environmental Information [dataset], http://dx.doi.org/10.7289/V5C8276M, 2009.

Jiang, L. D. and Wang, M. H.: Identification of pixels with stray light and cloud shadow contaminations in the satellite ocean color data processing, Applied Optics, 52, 6757-6770, 10.1364/ao.52.006757, 2013.

Jin, D., Lee, E., Kwon, K., and Kim, T.: A Deep Learning Model Using Satellite Ocean Color and Hydrodynamic Model to Estimate Chlorophyll-a Concentration, Remote Sensing, 13, 10.3390/rs13102003, 2021.

Konik, M., Kowalewski, M., Bradtke, K., and Darecki, M.: The operational method of filling information gaps in satellite imagery using numerical models, International Journal of Applied Earth Observation and Geoinformation, 75, 68-82, 10.1016/j.jag.2018.09.002, 2019.

Krug, L. A., Platt, T., Sathyendranath, S., and Barbosa, A. B.: Ocean surface partitioning strategies using ocean colour remote Sensing: A review, Progress in Oceanography, 155, 41-53, 10.1016/j.pocean.2017.05.013, 2017.

Laufkotter, C., Vogt, M., Gruber, N., Aumont, O., Bopp, L., Doney, S. C., Dunne, J. P., Hauck, J., John, J. G., Lima, I. D., Seferian, R., and Volker, C.: Projected decreases in future marine export production: the role of the carbon flux through the upper ocean ecosystem, Biogeosciences, 13, 4023-4047, 10.5194/bg-13-4023-2016, 2016.

Le Grix, N., Zscheischler, J., Laufkoetter, C., Rousseaux, C. S., and Froelicher, T. L.: Compound high-temperature and low-chlorophyll extremes in the ocean over the satellite period, Biogeosciences, 18, 2119-2137, 10.5194/bg-18-2119-2021, 2021.

Levitus, S., Boyer, T. P., García, H. E., Locarnini, R. A., Zweng, M. M., Mishonov, A. V., Reagan, J. R., Antonov, J. I., Baranova, O. K., Biddle, M., Hamilton, M., Johnson, D. R., Paver, C. R., and Seidov, D.: World Ocean Atlas 2013 (NCEI Accession 0114815) [dataset], doi.org/10.7289/v5f769gt, 2014.

Li, H., Xu, F. H., Zhou, W., Wang, D. X., Wright, J. S., Liu, Z. H., and Lin, Y. L.: Development of a global gridded Argo data set with Barnes successive corrections, Journal of Geophysical Research-Oceans, 122, 866-889, 10.1002/2016jc012285, 2017.

Li, J., Gao, M., Feng, L., Zhao, H., Shen, Q., Zhang, F., Wang, S., and Zhang, B.: Estimation of Chlorophyll-a Concentrations in a Highly Turbid Eutrophic Lake Using a Classification-Based MODIS Land-Band Algorithm, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12, 3769-3783 10.1109/JSTARS.2019.2936403, 2019.

Li, X., Liu, B., Zheng, G., Ren, Y., Zhang, S., Liu, Y., Gao, L., Liu, Y., Zhang, B., and Wang, F.: Deep-learning-based information mining from ocean remote-sensing imagery, National Science Review, 7, 1584-1605, 10.1093/nsr/nwaa047, 2020.

Lian, F., School of Environmental, S., Engineering, S. U. o. S., and Technology, S. P. R. C.: Key issues in detecting lacustrine cyanobacterial bloom using satellite remote sensing, Journal of Lake Sciences, 33, 647-652 10.18307/2021.0301, 2021.

Liu, D., Zhou, C., Keesing, J. K., Serrano, O., Werner, A., Fang, Y., Chen, Y., Masque, P., Kinloch, J., Sadekov, A., and Du, Y.: Wildfires enhance phytoplankton production in tropical oceans, Nature Communications, 13, 1348 10.1038/s41467-022-29013-0, 2022.

Liu, X. and Wang, M.: Gap Filling of Missing Data for VIIRS Global Ocean Color Products Using the DINEOF Method, IEEE

Transactions on Geoscience and Remote Sensing, 56, 4464-4476, 10.1109/TGRS.2018.2820423, 2018.

Liu, X. and Wang, M.: Global daily gap-free ocean color products from multi-satellite measurements, International Journal of Applied Earth Observation and Geoinformation, 108, 102714 10.1016/j.jag.2022.102714, 2022.

McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., and Stoffelen, A.: Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target, Geophysical Research Letters, 41, 6229-6236, 10.1002/2014gl061322, 2014.

Mikelsons, K. and Wang, M.: Optimal satellite orbit configuration for global ocean color product coverage, Optics Express, 27, A445 10.1364/OE.27.00A445, 2019.

Miralles, D. G., Crow, W. T., and Cosh, M. H.: Estimating Spatial Sampling Errors in Coarse-Scale Soil Moisture Estimates Derived from Point-Scale Observations, Journal of Hydrometeorology, 11, 1423-1429, 10.1175/2010jhm1285.1, 2010.

Moran, N., Stringer, B., Lin, B. C., and Hoque, M. T.: Machine learning model selection for predicting bathymetry, Deep-Sea Research Part I-Oceanographic Research Papers, 185, 10.1016/j.dsr.2022.103788, 2022.

Mouw, C. B., Barnett, A., McKinley, G. A., Gloege, L., and Pilcher, D.: Phytoplankton size impact on export flux in the global ocean, Global Biogeochemical Cycles, 30, 1542-1562, 10.1002/2015gb005355, 2016.

Nelson, N. G., Munoz-Carpena, R., and Phlips, E.: Parameter uncertainty drives important incongruities between simulated chlorophyll-a and phytoplankton functional group dynamics in a mechanistic management model, Environmental Modelling & Software, 129, 10.1016/j.envsoft.2020.104708, 2020.

Nikolaidis, A., Georgiou, G. C., Hadjimitsis, D., and Akylas, E.: Filling in missing sea-surface temperature satellite data over the Eastern Mediterranean Sea using the DINEOF algorithm, Central European Journal of Geosciences, 6, 27-41, 10.2478/s13533-012-0148-1, 2014.

Powell, C. F., Baker, A. R., Jickells, T. D., Bange, H. W., Chance, R. J., and Yodle, C.: Estimation of the Atmospheric Flux of Nutrients and Trace Metals to the Eastern Tropical North Atlantic Ocean, Journal of the Atmospheric Sciences, 72, 4029-4045, 10.1175/jas-d-15-0011.1, 2015.

Reygondeau, G., Longhurst, A., Martinez, E., Beaugrand, G., Antoine, D., and Maury, O.: Dynamic biogeochemical provinces in the global ocean, Global Biogeochemical Cycles, 27, 1046-1058, 10.1002/gbc.20089, 2013.

Righetti, D., Vogt, M., Gruber, N., Psomas, A., and Zimmermann, N. E.: Global pattern of phytoplankton diversity driven by temperature and environmental variability, Science Advances, 5, eaau6253 10.1126/sciadv.aau6253, 2019.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, 10.1007/978-3-319-24574-4_28, 2015.

Roussillon, J., Fablet, R., Gorgues, T., Drumetz, L., Littaye, J., and Martinez, E.: A Multi-Mode Convolutional Neural Network to reconstruct satellite-derived chlorophyll-a time series in the global ocean from physical drivers, Frontiers in Marine Science, 10, 10.3389/fmars.2023.1077623, 2023.

Salgado-Hernanz, P. M., Racault, M. F., Font-Muñoz, J. S., and Basterretxea, G.: Trends in phytoplankton phenology in the Mediterranean Sea based on ocean-colour remote sensing, Remote Sensing of Environment, 221, 50-64 10.1016/j.rse.2018.10.036, 2019.

Sathyendranath, S., Brewin, R., Brockmann, C., Brotas, V., Calton, B., Chuprin, A., Cipollini, P., Couto, A., Dingle, J., Doerffer, R., Donlon, C., Dowell, M., Farman, A., Grant, M., Groom, S., Horseman, A., Jackson, T., Krasemann, H., Lavender, S., Martinez-Vicente, V., Mazeran, C., Mélin, F., Moore, T., Müller, D., Regner, P., Roy, S., Steele, C., Steinmetz, F., Swinton, J., Taberner, M., Thompson, A., Valente, A., Zühlke, M., Brando, V., Feng, H., Feldman, G., Franz, B., Frouin, R., Gould, R., Hooker, S., Kahru, M., Kratzer, S., Mitchell, B., Muller-Karger, F., Sosik, H., Voss, K., Werdell, J., and Platt, T.: An Ocean-Colour Time Series for Use in Climate Studies: The Experience of the Ocean-Colour Climate Change Initiative (OC-CCI), Sensors, 19, 4285, 10.3390/s19194285, 2019.

Signorini, S. R., Franz, B. A., and McClain, C. R.: Chlorophyll variability in the oligotrophic gyres: mechanisms, seasonality and trends, Frontiers in Marine Science, 2, 10.3389/fmars.2015.00001, 2015.

Slomp, C. P.: Phosphorus Cycling in the Estuarine and Coastal Zones: Sources, Sinks, and Transformations, Treatise on Estuarine and Coastal Science, Vol 5: Biogeochemistry, 201-229 pp., 10.1016/B978-0-12-374711-2.00506-4, 2011.

Stoffelen, A.: Toward the true near-surface wind speed: Error modeling and calibration using triple collocation, Journal of Geophysical Research-Oceans, 103, 7755-7766, 10.1029/97jc03180, 1998.

Stumpf, R. P.: Applications of satellite ocean color sensors for monitoring and predicting harmful algal blooms, Human and Ecological Risk Assessment, 7, 1363-U1315, 10.1080/20018091095050, 2001.

Sun, D., Pan, T., Wang, S., and Hu, C.: Linking phytoplankton absorption to community composition in Chinese marginal seas, Progress in Oceanography, 192, 102517 10.1016/j.pocean.2021.102517, 2021.

Takeda, S.: Influence of iron availability on nutrient consumption ratio of diatoms in oceanic waters, Nature, 393, 774-777, 10.1038/31674, 1998.

Urakubo, H., Bullmann, T., Kubota, Y., Oba, S., and Ishii, S.: UNI-EM: An Environment for Deep Neural Network-Based Automated Segmentation of Neuronal Electron Microscopic Images, Scientific Reports, 9, 10.1038/s41598-019-55431-0, 2019.

Wagle, N., Acharya, T. D., and Lee, D. H.: Comprehensive Review on Application of Machine Learning Algorithms for Water Quality Parameter Estimation Using Remote Sensing Data, Sensors and Materials, 32, 3879 10.18494/SAM.2020.2953, 2020.

Wagner, F. H., Sanchez, A., Tarabalka, Y., Lotte, R. G., Ferreira, M. P., Aidar, M. P. M., Gloor, E., Phillips, O. L., and Aragao, L.: Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images, Remote Sensing in Ecology and Conservation, 5, 360-375, 10.1002/rse2.111, 2019.

Wang, B. L., Liu, C. Q., Maberly, S. C., Wang, F. S., and Hartmann, J.: Coupling of carbon and silicon geochemical cycles in rivers and lakes, Scientific Reports, 6, 10.1038/srep35832, 2016.

Wang, M., Jiang, L., Mikelsons, K., and Liu, X.: Satellite-derived global chlorophyll-a anomaly products, International Journal of Applied Earth Observation and Geoinformation, 97, 102288 10.1016/j.jag.2020.102288, 2021.

Wang, Y. Q. and Liu, D. Y.: Reconstruction of satellite chlorophyll-a data using a modified DINEOF method: a case study in the Bohai and Yellow seas, China, International Journal of Remote Sensing, 35, 204-217, 10.1080/01431161.2013.866290, 2014.

Xing, X. and Boss, E.: Chlorophyll‐Based Model to Estimate Underwater Photosynthetically Available Radiation for Modeling, In‐Situ, and Remote‐Sensing Applications, Geophysical Research Letters, 48, 10.1029/2020GL092189, 2021.

Yussof, F. N., Maan, N., and Reba, M. N. M.: LSTM Networks to Improve the Prediction of Harmful Algal Blooms in the West Coast of Sabah, International Journal of Environmental Research and Public Health, 18, 10.3390/ijerph18147650, 2021.

Zhang, Q., Yuan, Q., Zeng, C., Li, X., and Wei, Y.: Missing Data Reconstruction in Remote Sensing Image With a Unified Spatial-Temporal-Spectral Deep Convolutional Neural Network, Ieee Transactions on Geoscience and Remote Sensing, 56, 4274-4288, 10.1109/tgrs.2018.2810208, 2018.

Zheng, G. and DiGiacomo, P. M.: Uncertainties and applications of satellite-derived coastal water quality products, Progress in Oceanography, 159, 45-72 10.1016/j.pocean.2017.08.007, 2017.

Zuo, H., Balmaseda, M. A., Tietsche, S., Mogensen, K., and Mayer, M.: The ECMWF operational ensemble reanalysis-analysis system for ocean and sea ice: a description of the system and assessment, Ocean Science, 15, 779-808, 10.5194/os-15-779-2019, 2019.

Zweng, M. M., Reagan, J. R., Antonov, J. I., Locarnini, R. A., Mishonov, A. V., Boyer, T. P., Garcia, H. E., Baranova, O. K., Johnson, D. R., and D. Seidov, M. M. B.: World Ocean Atlas 2013: Volume 2, Salinity, S. Levitus, Ed.; A. Mishonov, Technical

Ed.; NOAA Atlas NESDIS 74, 39 pp., 2013.