A Novel Global Gridded Ocean Oxygen Product Derived from a Neural Network Emulator and in-situ observations

ESSD

August 12, 2025

Said Ouala[1], Oussama Hidaoui [2], Zouhair Lachkar [3],
Contact: said.ouala@imt-atlantique.fr

[1] IMT Atlantique, CNRS UMR Lab-STICC, INRIA Team Odyssey, Brest, France.
[2] African Institute for Mathematical Sciences, South Africa.
[3] Mubadala Arabian Center for Climate and Environmental Sciences, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates.

Abstract

The authors would like to thank the anonymous reviewer for their valuable comments and suggestions. In this document, we address the issues raised to the best of our ability. The modifications made in response to the reviewer's comments are highlighted in blue in the tracked-changes version of the manuscript.

1 Reviewer's comments

Reviewer Comment 1

I would like to thank the authors for the interesting and timely work. The paper presents a novel approach to generating a gridded dissolved oxygen product by integrating direct observations with ML-based emulations derived from temperature and salinity profiles, followed by optimal interpolation. The methodology is simple to follow and technically sound, the results are compelling, and the product demonstrates clear improvements over existing datasets, especially in capturing long-term trends and reducing uncertainties. I recommend acceptance with minor revisions, but I would like to note that my review is primarily focused on the ML aspect.

Response

The authors appreciate the feedback on our work. Every comment is addressed carefully below, and the modifications can be found in blue in the tracked-changes version of the manuscript.

Reviewer Comment 2

The training/test split was done randomly, how the authors ensure there is no data leakage? It would have been more interesting if the trains was done in a temporal way.

Response

We thank the reviewer for raising this important point. We agree that data leakage and the nature of the training/test split are crucial to evaluating the robustness of machine learning models.

To clarify, we used three separate subsets, training, validation, and test datasets, to design and assess our neural network model. The training and validation sets were obtained via a random split: 80% of the data was used for training and 20% for validation. This random split applies only to the training/validation phase.

The test set is completely independent of both the training and validation sets. It consists of 23 spatially distinct regions across major ocean basins, as introduced in Figures 1 and 2. These regions were excluded from the training/validation data and were used exclusively for testing, ensuring no data leakage.

We acknowledge that our original manuscript may not have made this distinction clear. To clarify, we have revised the paragraph around line 85 as follows:

"The dataset used to train the ML model consists of collocated pairs of T, S and (DO) data from 1965 to 2022. The dataset was divided into training, validation, and test subsets. The test set comprises 23 independent 1×1 regions, distributed across all major ocean basins. The locations of these test regions, along with the performance of the machine learning-based emulator, are shown in Figure 2. The remaining data were allocated to training and validation, with 80% used for training and the remaining 20% for validation."

Regarding the reviewer's suggestion to adopt a temporal split, we chose a spatial test split instead of a temporal one for two reasons. First, we aim for the model to learn long-term deoxygenation trends related to climate change; excluding certain years from training could limit the model's ability to capture these patterns. Second, our goal was to evaluate how well the model generalizes across different oceanographic regimes, such as oxygen minimum zones or well-oxygenated regions, which motivated the use of geographically distinct test regions.

Reviewer Comment 3

It would have been also more robust to use a validation dataset, instead of only train/test.

Response

We apologize for any confusion caused by our wording. As explained in the previous response, we use separate training, validation, and test sets. The test set is fully independent from the training and validation sets, as it consists of data from locations that do not overlap with those used for training and validation.

Reviewer Comment 4

Any reason why the test locations do not include any points near Europe?

Response We thank the reviewer for this relevant observation. There is no specific reason why the test locations do not include points near Europe. Our test regions were designed to cover all major ocean basins, including the North Equatorial and South Pacific, North and South Atlantic, and the Indian Ocean. Within each basin, we selected test locations that capture a diversity of oxygen dynamics, including both oxygen minimum zones (e.g., Regions F, G and H in the Indian Ocean, Region D in the Atlantic, and Regions A, B, C in the Pacific) and highly oxygenated regions (e.g., Regions P, Q and R). The European seas were not explicitly included because the oxygen dynamics in those areas are not different from the ones already sampled in our selected test regions.

Reviewer Comment 5

Any reason why using Month of the year + Day of the month in the MLP inputs instead of just using Day of the year?

Response

We thank the reviewer for this insightful comment. Our initial choice to use the month of the year as an input feature follows common practice in the literature (e.g., [1]), where it is shown to help capture seasonal patterns in MLP-based models. We added the day of the month to allow for finer resolution of intra-month variability, which led to a slight improvement in performance. However, our explainability analysis (XAI) confirms that this variable has relatively low importance compared to other predictors. We acknowledge that using the day of the year is a valid alternative and appreciate the suggestion.

Reviewer Comment 6

Can the authors describe the hyper parameter search procedure to tune the MLP?

Response We thank the reviewer for this important question. Our approach to tuning the MLP architecture was as follows. We began by testing the model on simulated data (a coupled ROMS and BGC model in the Indian Ocean) to assess whether an ML model could effectively predict oxygen concentrations from temperature, salinity, spatiotemporal coordinates, and surface chlorophyll-a. Initial experiments indicated that temperature and salinity alone were sufficient to achieve strong predictive performance.

We then transitioned to real data, using MODIS satellite-derived CHLA-II as an additional input. Similarly to the experiment on simulated data, we observed that including satellite CHLA-II did not improve the model's performance, so we opted to keep the architecture simple and relied only on in-situ observations to design the model.

Regarding the architecture, we performed a stepwise increase in complexity: starting with 2 hidden layers, we incrementally added more layers (up to 4) and increased the number of

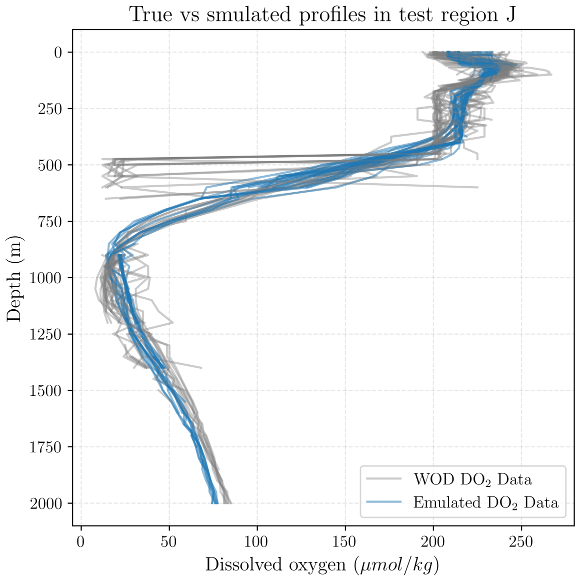

Figure 1: Measured profiles and emulator prediction in test region J.

neurons per layer. We selected the final architecture based on the point at which additional complexity no longer yielded performance improvements on the validation set.

Reviewer Comment 7

Figure 1 would have been more informative if the plots where done per test region.

Response We thank the reviewer for raising this point. We agree with the reviewer that a per-region breakdown provides more insight. We have updated Figure 1 accordingly to show the scatter plots for each individual test region. This revised figure shows that the model performs consistently across different test regions.

Reviewer Comment 8

Any explanation of what's happening at depth 500 in test region J (Figure 2)?

Response This is a very interesting point raised by the reviewer. We analyzed the profiles at depth 500 in test region J and found that several **measured** profiles in this region show abrupt variations around 500 m depth, leading to an inflated standard deviation at that depth level. These anomalies are likely due to sensor errors rather than physical processes. Figure 1 illustrates these outliers by comparing measured and predicted profiles. Notably, the emulator outputs remain smooth and do not reproduce these irregular patterns.

Reviewer Comment 9

It would be interesting to use any XAI method to study feature importance for the MLP.

Response

We appreciate the comment from the reviewer. We have added a new section in the appendix presenting an analysis of feature importance using Integrated Gradients. The results confirm that the most influential features for predicting dissolved oxygen are the geographical location of the profiles (latitude and longitude) along with physical variables, namely temperature and salinity. These features reflect the importance of the regional context through dominant physical ventilation regimes, biogeochemical dynamics, and oxygen solubility in explaining oxygen variability in the ocean.

Reviewer Comment 10

Any plans to share the code used and not only the dataset?

Response

We thank the reviewer for raising this point. We do plan to share the code. Currently, the code consists of several modules developed and hosted by different contributors, covering data extraction and preprocessing, model training, emulation, quality control of emulated profiles, and interpolation. Due to this "distributed" development, publicly releasing the code in its current form is not feasible. However, we are happy to provide it upon request.

However, we are in the process of cleaning and organizing the code to make it publicly available as a single package. We appreciate the reviewer's interest and are committed to ensuring the code is shared as soon as it reaches an appropriate level of clarity and documentation.

Reviewer Comment 11

Typos:
* Line 35: "weather forecasting" instead of "forecasting"
* Many citations are badly formatted, /citet vs /citep

Response We thank the reviewer for spotting these typos. They have been corrected in the new version of the manuscript. Regarding the citations, all references are cited in the text using the /citet.

References

[1] Takamitsu Ito, Ahron Cervania, Kaylin Cross, Sanika Ainchwar, and Sara Delawalla, "Mapping dissolved oxygen concentrations by combining shipboard and argo observations using machine learning algorithms," *Journal of Geophysical Research: Machine Learning and Computation*, vol. 1, no. 3, pp. e2024JH000272, 2024.