Ref: essd-2025-699

GEOXYGEN: a global long-term dissolved oxygen dataset based on biogeochemistry-aware machine learning framework and multi-source observations

Zhenguo Wang[1], Weiwei Fu[1,2], Cunjin Xue[3,4], Guihua Wang[1]

[1]Department of Atmospheric and Oceanic Sciences, Fudan University, Shanghai, 200438, China

[2]Institute of Eco-Chongming (IEC), 1050 Baozhen, Lühua Town, Chongming District, Shanghai 202151, China

[3]International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China

[4]Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

Response Letter to Reviewers' Comments

Reviewer #1:

The authors have improved and clarified their machine learning architecture in the revised manuscript. Using SOM_DES as an additional feature for model training is creative, and the reconstructed ocean oxygen fields look more robust. The authors have also made efforts to discuss the uncertainties and limitations on their reconstructed oxygen fields. Overall, the manuscript has been improved and could be considered for publication after some technical corrections.

1. The authors have changed the data repository to the Oceanographic Data Center, Chinese Academy of Sciences (CASODC) (Wang et al., 2026; https://doi.org/10.12157/IOCAS.20260223.002). However, I noticed that users are required to register and log in before downloading the dataset. I am not sure whether this data repository is compatible with the ESSD data availability policy.

Response: We thank the reviewer for the suggestion. An open-access link through **Zenodo was added** in addition to the CASODC repository.

In **Code and data availability:**

The GEOXYGEN dataset produced in this study can be found at https://doi.org/10.12157/IOCAS.20260223.002 and https://doi.org/10.5281/zenodo.19703198 (last access: 24 April 2026).

2. There is no full name of SOM_DES in the manuscript and there should be some more texts to describe how SOM_DES is generated in the manuscript.

Response: The full name of SOM_DES was defined in Lines 207-213 of the revised manuscript, as: "Within each depth, the self-organizing-map-derived descriptor (SOM_DES) is used as a categorical predictor to encode the large-scale climatological background state. Specifically, we construct a four-dimensional climatological vector [SST, SSS, MLD, DO] from monthly background fields. These vectors are then used to train a self-organizing map (SOM), and each grid cell is assigned to one of 25 discrete classes. The clustering is based on the joint patterns of multiple environmental fields, and, at this step, the O_2 climatology is implicitly assigned a higher weight to obtain a seasonally varying, dynamic partitioning. The SOM is trained directly on these monthly climatological fields, such that the DO climatology implicitly exerts a stronger influence on the resulting state partition and ensures that the partitioning evolves coherently with the seasonal DO cycle."

Reviewer #2:

The authors have done a very good job addressing the previous comments, and I support publication after minor revision. I only have two remaining suggestions.

1. The manuscript notes that uncertainty in nearshore and shelf regions is more than double that in the open ocean. This is an important and commendably transparent result. However, once the dataset is publicly released, many users may apply it directly without fully appreciating these limitations. I therefore recommend that the authors provide a more practical usage note for coastal applications, for example by suggesting an appropriate level of spatial aggregation or cautionary guidance for nearshore analyses. This could be added in the data availability section or in a short recommended use paragraph.

Response: We agree with the reviewer that a usage note would be helpful for coastal applications.

In Sect. 5.1, Lines 483-486, we added:
"It should be noted that GEOXYGEN has higher uncertainty in coastal and shelf regions. Therefore, for coastal applications, we recommend using multi-grid-cell averages and focusing on monthly to seasonal or longer timescales. Use of the native 0.5° grid near the coast should therefore be approached with caution and, wherever feasible, cross-validated against local observations or higher-resolution regional products."

In "Code and data availability", we added:
"A recommended-use note: because uncertainty in nearshore regions is substantially higher than in the open ocean, coastal and shelf applications should preferentially rely on spatially aggregated and

temporally averaged fields rather than on individual native-grid cells."

2. The paper presents a well-integrated workflow, from multi-source quality control to interpolation, and uncertainty decomposition. This is useful and carefully executed. However, parts of the manuscript still read as if the study introduces a fundamentally new methodological framework. In my view, the main contribution is better described as a thoughtful and comprehensive implementation of existing machine-learning reconstruction strategies for a challenging dissolved oxygen problem, rather than as a conceptual breakthrough in method development. I therefore encourage the authors to moderate some of the stronger wording regarding methodological novelty.

Response: We thank the reviewer for this thoughtful and constructive comment. We agree with the reviewer and have toned down the wording in several places in the revised manuscript.

In Introduction,

"……through a regionally structured, depth-aware, and adaptively constrained machine-learning framework." was revised as:

"……through an integrated machine-learning reconstruction workflow, which combines regional partitioning, depth-wise modeling, adaptive feature selection, and physically informed predictors."

In Method

"…we develop a biogeochemistry-aware machine-learning ..." was revised as: "…we implement a biogeochemistry-aware machine-learning ..."

In Conclusion, Lines 559-563,

"By integrating multi-source physical and biogeochemical predictors with an adaptive feature-selection strategy, we develop a biogeochemistry-aware hierarchical modeling framework. Building on this framework, we introduce GEOXYGEN—a monthly, four-dimensional global ocean dissolved oxygen (DO) product spanning 1960–2024 at 0.5° × 0.5° resolution—designed to mitigate long-standing limitations arising from observation sparsity and strong spatiotemporal heterogeneity in historical DO records."

we now stated: "By combining multi-source physical and biogeochemical predictors with an adaptive feature-selection strategy, we constructed a hierarchical modeling framework that accounts for underlying biogeochemical controls. Using this framework, we produced GEOXYGEN—a monthly, four-dimensional global ocean dissolved oxygen (DO) product at 0.5° × 0.5° resolution spanning 1960–2024. This product is intended to help address some of the long-standing challenges posed by sparse observations and strong spatiotemporal heterogeneity in historical DO records."