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

Comments from the Editor:

We are pleased to inform you that the open discussion of your following ESSD manuscript was closed:
Title: GEOXYGEN: a global long-term dissolved oxygen dataset based on biogeochemistry-aware machine learning framework and multi-source observations
Author(s): Zhenguo Wang et al.
MS No.: essd-2025-699
MS type: Data description article
No more referee comments and short comments will be accepted. Now the public discussion shall be completed as follows:
You - as the contact author - are requested to respond to all referee comments (RCs) by posting final author comments (ACs) on behalf of all co-authors no later than 26 Feb 2026 (final response phase). Please log in using your Copernicus Office user ID 838731 at: https://editor.copernicus.org/essd-2025-699/final-response
When replying to the referee comments (RCs) it is sufficient to post one author comment (AC) by starting a new discussion thread. Please also consider replying to community comments (CCs) from the scientific community.
After your AC posts, you have to explicitly finalize the final-response form through the button "Finalize". You will then receive a separate email asking you to prepare and submit your revised manuscript for peer-review completion and potential final publication in ESSD.
Preparation and submission of a revised manuscript is encouraged only if you can satisfactorily address all comments and if the revised manuscript meets the high quality standards of ESSD (https://www.earth-system-science-data.net/peer_review/review_criteria.html). In case of doubt, please ask the handling topic editor directly whether they would encourage submission of a revised manuscript or not.

Please note also that the submission of a revised manuscript does not ensure publication in ESSD. The topic editor will carefully assess your revised manuscript in view of the interactive public discussion and may forward it to the original or new referees for further commenting.
You are invited to monitor the processing of your manuscript via your MS overview at: https://editor.copernicus.org/ESSD/my_manuscript_overview
Thank you very much in advance for your cooperation. In case any questions arise, please do not hesitate to contact me.

Response: We thank all the editors and reviewers for their valuable comments and suggestions. We have carefully revised the manuscript to enhance clarity and improve readability for the community. We have fully addressed the comments and concerns of the editors and reviewers. Our point-to-point responses are presented in the following.

Reviewer #1:

The authors created a 4-D global ocean dissolved oxygen atlas at 0.5˚x0.5˚ resolution by using multiple data sources and machine learning approaches. The ensemble machine learning submodel framework used to derive this data product is interesting, and this new ocean dissolved oxygen product shows slight improvement over previous products. This dataset has the potential to be used by the oceanography community to assess oxygen evolution under a changing climate. My recommendation is minor revision.

General comments:

1. I would recommend the authors create a web visualization tool to allow users easy access to this data product. The currently archived dataset on Zenodo is not easy to explore or play with. However, this suggestion is optional and should not affect the publication of this work.

Response: We thank the reviewer for this valuable suggestion. As suggested, we uploaded EOXYGEN to the Oceanographic Data Center, Chinese Academy of Sciences (CASODC) and added basic usage examples to enable efficient exploration without specialized infrastructure. Accordingly, we revised the Code and data availability section in the revised paper.

In Lines 588-590 of the revised manuscript, we stated:

"The GEOXYGEN dataset produced in this study, together with the basin province mask used for regionalization, is publicly available at the Oceanographic Data Center, Chinese Academy of Sciences (CASODC) (Wang et al., 2026, https://doi.org/10.12157/IOCAS.20260223.002), where details of the data files and metadata are documented."

2. This dissolved oxygen (DO) data product spans from 1960 to 2024. However, the authors only performed a climatological comparison with other reconstructed datasets. It would be helpful to add an additional validation or trend analysis of ocean DO using this dataset.

Response: Following the reviewer's advice, we have added a dedicated trend- and variability-oriented analysis in the revised manuscript (new Sect. 5.2; Fig. 10), focusing on deseasonalized 1–100 m depth-averaged DO anomalies to highlight the multi-decadal signal and to facilitate inter-product benchmarking. In this new section, we further compare our reconstruction with ML4O2 to provide contextual validation of the reconstructed temporal evolution.

In Lines 493-508 of the revised manuscript, we stated:

The full and reduced reconstructions exhibit strong agreement over 1960–2022 (Fig. 10), yielding a consistent depiction of upper-ocean (1–100 m) low-frequency variability: a sustained positive-anomaly regime through the 1970s–1980s followed by a marked decline after ~1990 and a transition into a persistently negative-anomaly regime by ~2000 (relative to the monthly climatology).

As an external benchmark, ML4O2 reproduces comparable variability during ~1965–2010 but shows more negative anomalies in the most recent decade, suggesting a stronger upper-ocean deoxygenation signal in anomaly space (relative to its monthly climatology) than GEOXYGEN over the same period; however, part of this divergence may arise from inter-product differences in baseline climatology, sampling, and reconstruction methodology.

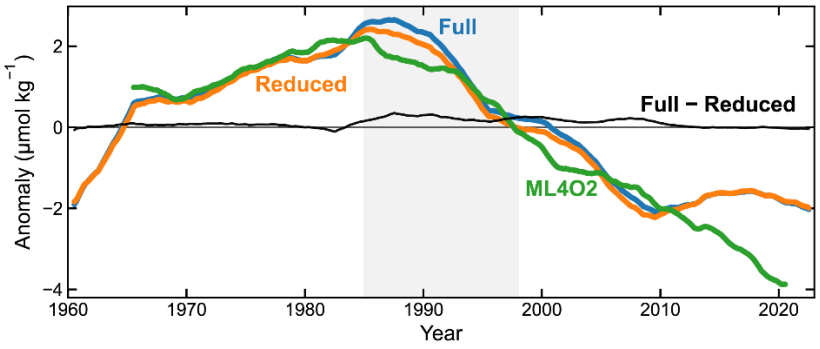


Figure 10: Depth-averaged (1–100 m) monthly dissolved-oxygen anomalies (1960–2022). The figure compares deseasonalized anomalies from the Full predictor reconstruction, the Reduced predictor reconstruction (excluding sea-surface predictors), and ML4O2, and overlays the difference between the Full and Reduced reconstructions (Full − Reduced). Gray shading indicates 1985–1997, corresponding to the rapid expansion of satellite-derived sea-surface observations.

Specific Comments:

1. Figure 1a: I was confused when looking at this figure. OSD, CTD and Argo are different approaches for DO concentration data and CCHDO, GLODAP, GEOTRACES and etc. are different sources of DO data. It seems they should not be mixed in this figure. It is unclear whether

the authors intend to show changes in sampling approaches over time or changes in data sources.

Response: We thank the reviewer for pointing out the conceptual inconsistency. We agree that sampling platforms (e.g., OSD/CTD, Argo) and data compilations/repositories (e.g., CCHDO, GLODAP, GEOTRACES) represent different classification levels, and mixing them without clarification can obscure whether the panel is intended to show platform evolution or source composition.

To address this, we have restructured Fig. 1a and the associated text, aiming to illustrate how contributions from major dissolved-oxygen data sources change over time. Importantly, the entries "OSD/CTD" and "Argo" in Fig. 1a are not the original sources of the observations; they denote the ship-based and Argo sub-components within the dissolved-oxygen dataset of Gouretski et al. (2024). We show these two sub-components separately to highlight the transition from ship-based measurements to autonomous profiling-float observations within the same data product, while keeping all other categories at the data-product level. The figure caption has been revised accordingly to make this hierarchy explicit.

In Lines 123-127 of the revised manuscript, we stated:

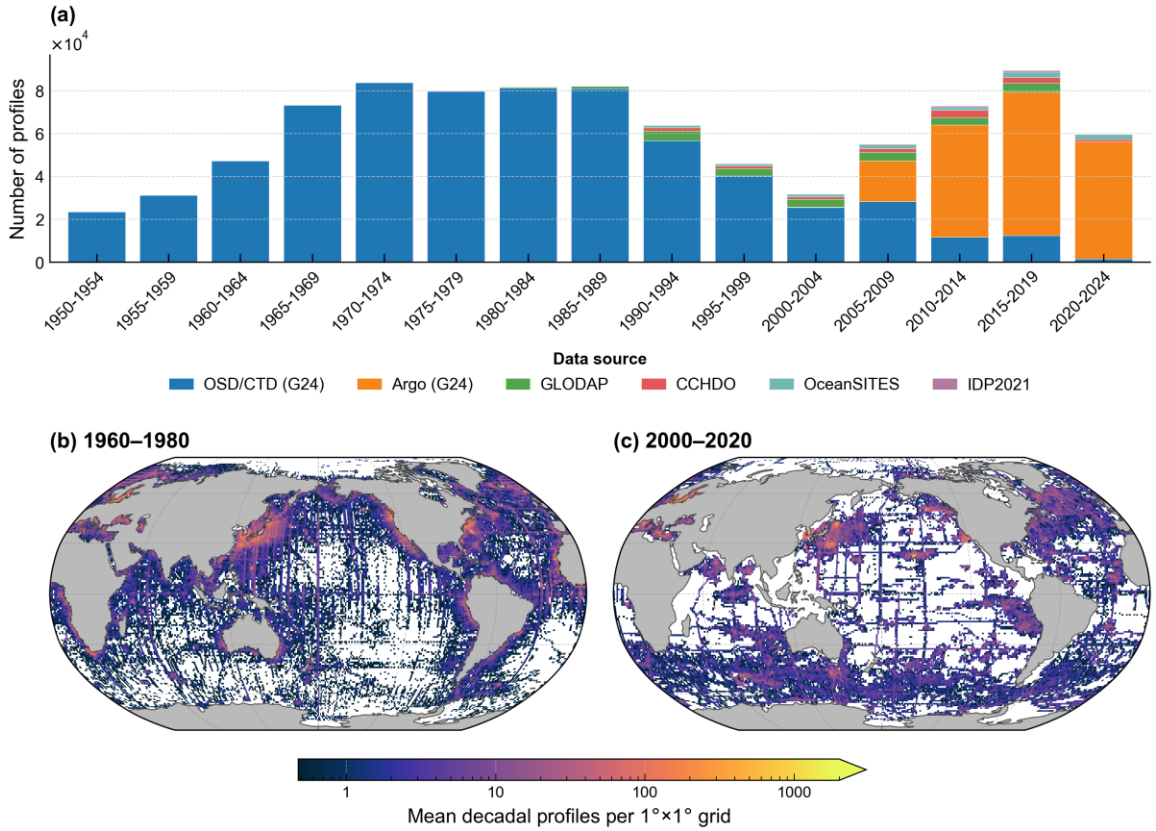


Figure 1: Global distribution and temporal coverage of DO profiles. (a) Changes in the number of profiles from major **data sources** during 1950–2024. (b–c) Spatial distribution of decadal-mean profile counts for 1960–1980 and 2000–2020, computed on a 1° × 1° grid. The color bar indicates the decadal-

mean number of profiles per grid cell (log scale). G24 denotes the oxygen compilation of Gouretski et al. (2024).

2. Sections 2 and 3: It will be helpful to add a flow chart to show the data cleaning and model training pipeline

Response: As the reviewer suggested, we have added a new flow chart (Fig. 2) to explicitly summarize the end-to-end pipeline, including data collection and preprocessing, quality control/cleaning, heterogeneity-based partitioning, feature engineering/selection, model training and evaluation, and the final 4-D gridded DO reconstruction. We also revised the relevant text in Sections 2–3 to refer readers to Fig. 2 for an integrated overview.

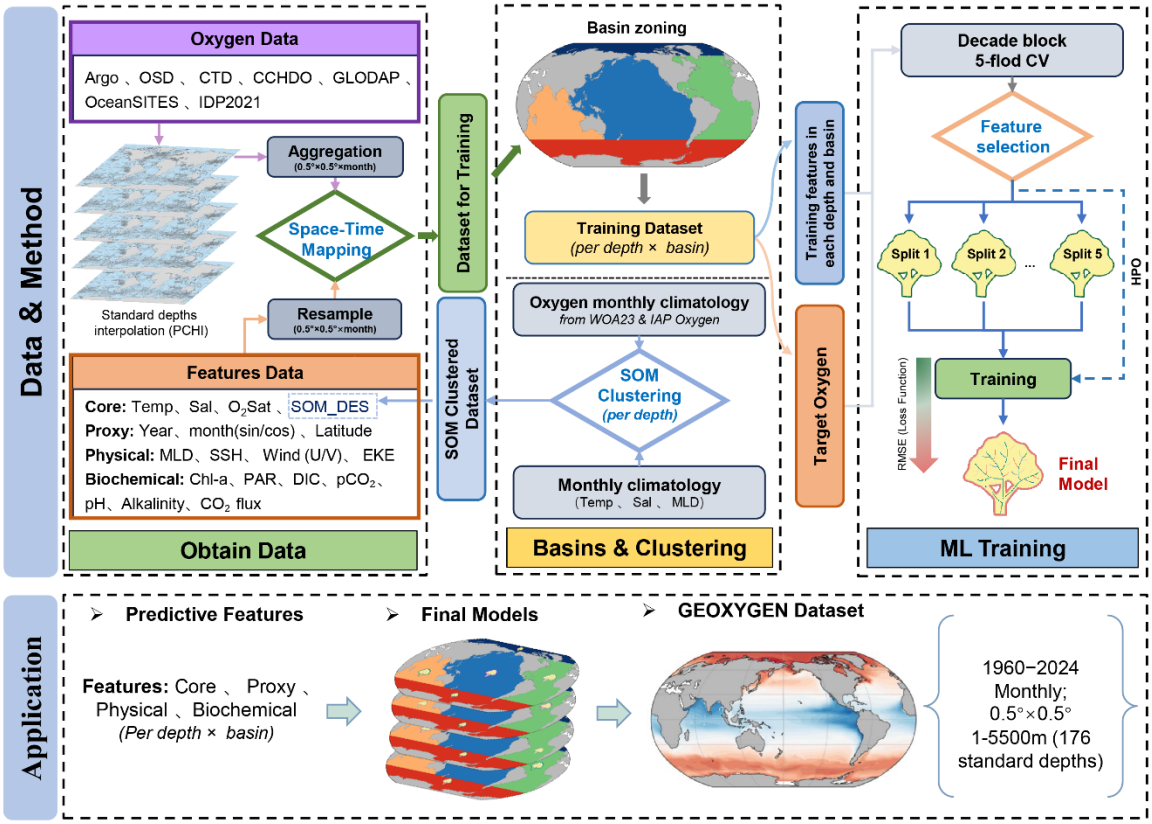


Figure 2: Overall workflow of the GEOXYGEN dataset construction.

3. Lines 108-109: I would be more cautious about this outlier detection approach especially in some dynamic regions like ENTP, where local DO concentration could change a lot within 10-day window.

Response: We fully agree that in highly dynamic regions such as the eastern North Tropical Pacific (ENTP), substantial sub-monthly variability can occur within a ~10-day window, and a short-window, threshold-based outlier filter (e.g., a 3σ rule) may inadvertently remove true physical–biogeochemical signals rather than measurement artifacts.

To address this comment, we adopted more robust QC procedures and performed monthly-grid aggregation processing, aiming to align with the characteristic distribution of the monthly-grid reconstruction target.

In Lines 95-99, we stated:
"A rigorous dual-stage QC protocol ensured observational reliability. The primary stage involved standardizing metadata formats and units across disparate sources, retaining only observations flagged as "good" or "probably good." Spurious terrestrial signals were omitted via land-masking, and duplicate profiles—defined by coincidence criteria of ≤ 1 km spatial distance and ≤ 24 h temporal difference—were identified across archives. In cases of redundancy, we prioritized profiles with the highest vertical sampling density."

In Lines 118-122, we stated:
"A second QC stage was applied to the standard levels. First, we excluded records where DO exceeded the 0–600 μmol kg^{-1} range. Next, to limit uncertainty from vertical interpolation, data with $\sigma_{\mathrm{interp}} > 3\ \mu$mol kg^{-1} were removed, based on the 95th percentile of σ_{interp} across all standard-level samples. Finally, following TEOS-10, oxygen saturation percentage (Sat%) was computed using in-situ temperature and salinity data. For standard levels deeper than 200 m, records with O_2 saturation ≥ 120% were considered erroneous and removed."

In Lines 178-182 we stated:
"To match the spatiotemporal scale of the supervised learning labels with the target reconstruction grid and to reduce sample-weight bias caused by uneven sampling in space and time, all observations are aggregated to a monthly 0.5° by 0.5° grid. Within each grid cell, observations are summarized into a single representative value, defined as the median to limit the influence of outliers and extreme events on the labels. The within-unit dispersion is also computed using the median absolute deviation (MAD), which serves as an empirical proxy for uncertainty."

4. Figure 3 and related texts: The partitioning of the global ocean into different provinces is central to the machine learning method used in the manuscript, as the submodels are trained. However, this partitioning is not clearly justified. Some questions:

Response: To address the reviewer's comments, we carefully revised the methodology and clarified the partitioning strategy in the revised manuscript.

1) the partitioning does not distinguish OMZ from other region. For example, ETSP OMZ is included within the whole South Pacific province,

Response: We fully agree that OMZs are a core scientific focus. Rather than defining OMZ boundaries a priori (which can be depth- and season-dependent and may introduce subjective, threshold-driven artifacts), we adopted a more general solution that allows the model to learn and

refine low-oxygen structures under varying backgrounds. In the revised manuscript, we introduce SOM_DES, a categorical environmental-state descriptor that encodes the large-scale climatological DO background and associated hydrographic structure, thereby enabling the model to condition its DO–environment mapping on regime-specific states (including low-oxygen regimes).

In Lines 206-222 we stated:
"Within each depth, the SOM_DES predictor accounts for localized variations in water-mass properties and frontal dynamics by encoding the large-scale, climatological background state. Specifically, for each month and grid cell, we form a four-dimensional vector [SST, SSS, MLD, DOclim] from the corresponding climatological monthly fields and assign the grid cell to a discrete SOM class, yielding a categorical environmental-state label. The global ocean is clustered into 25 biogeochemical regimes. 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. This feature-engineering step leverages the SOM's ability to map multivariate climatological structure onto a discrete set of regimes while preserving topological dependencies, thereby providing clear seasonal and spatial context for month-scale DO reconstruction. The background climatological fields are anchored in WOA23 (upper 1500 m) and the IAP Global Ocean Oxygen gridded product (IAP Oxygen; Cheng and Gouretski, 2024), ensuring vertically and horizontally comprehensive baseline states. Overall, SOM_DES captures the large-scale climatological structure of DO and supplies background-state information that supports refined month-scale DO modeling, improving diagnostic consistency across heterogeneous regimes and reinforcing the physical interpretability of the global DO product."

2) the authors state that the province partitioning is based on Fay and McKinley (2014), equatorial biomes identified in Fay and McKinley (2014) are not included,

Response: We thank the reviewer for pointing this out. In the revised manuscript, we changed the horizontal partitioning from the Fay and McKinley (2014) surface-biome scheme to five major ocean-basin domains following the basin definitions in WOA23 (Garcia et al., 2024). This basin partition provides a stable, depth-consistent training scaffold, while regime variability across seasons (and associated spatial structure) is captured through the SOM_DES environmental-state labels, which are defined for each month and depth based on climatological background fields. We train basin-specific submodels at each standard depth (Sect. 3.2).

In Lines 194-201 we stated:

"Our reconstruction framework utilizes a spatiotemporally stratified approach to address the shifting controlling mechanisms of ocean deoxygenation across basins and depths (Ma et al., 2025; Ito et al., 2024a). Following the basin definitions in the World Ocean Atlas 2023 (WOA23; Garcia et al., 2024), we additionally treat the Southern Ocean as a dedicated domain. Accordingly, the horizontal grid is divided into five primary modeling domains (Atlantic, Pacific, Indian, Southern, and Arctic), which are held constant across vertical layers to maintain training coherence (Fig. 3). By avoiding highly intricate province boundaries, this design reduces sensitivity of variability and trend estimates to boundary effects and makes cross-boundary continuity easier to maintain. We further mask the Mediterranean, Red Sea, and other semi-enclosed marginal seas to focus the reconstruction on open-ocean dynamics dominated by large-scale circulation and transport."

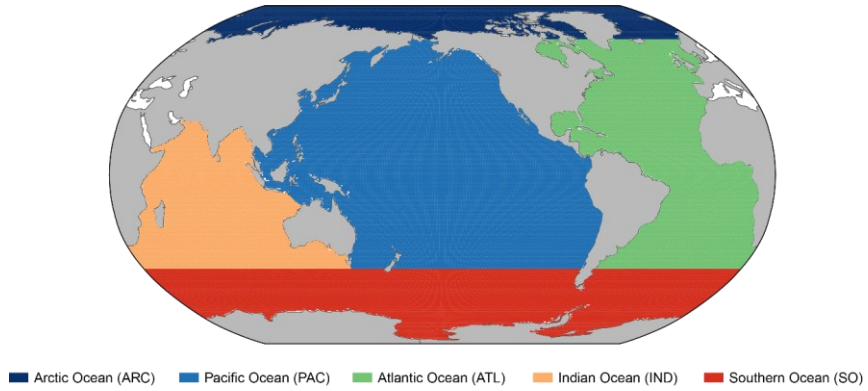


Figure 3: Partitioning of the global open ocean into five basins.

3) this global dissolved oxygen product is 4-D, and the province partitioning is same for every year and does not account for time shifts, which is also an important point in Fay and McKinley (2014)

Response: We agree with the reviewer on this point. As stated above, in the revised framework, we adopted the basin partition that provides a stable, depth-consistent training scaffold, while the regime variability across seasons (and associated spatial structure) is captured through the SOM_DES environmental-state labels, which are defined for each month and depth based on climatological background fields. This design allows the effective "regime partition" seen by the model to vary with seasonal background states, without requiring a manually prescribed, time-evolving province boundary map.

In Lines 206-213 we stated:

"Specifically, for each month and grid cell, we form a four-dimensional vector [SST, SSS, MLD, DOclim] from the corresponding climatological monthly fields and assign the grid cell to a discrete SOM class, yielding a categorical environmental-state label. The global ocean is clustered into 25 biogeochemical regimes. 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

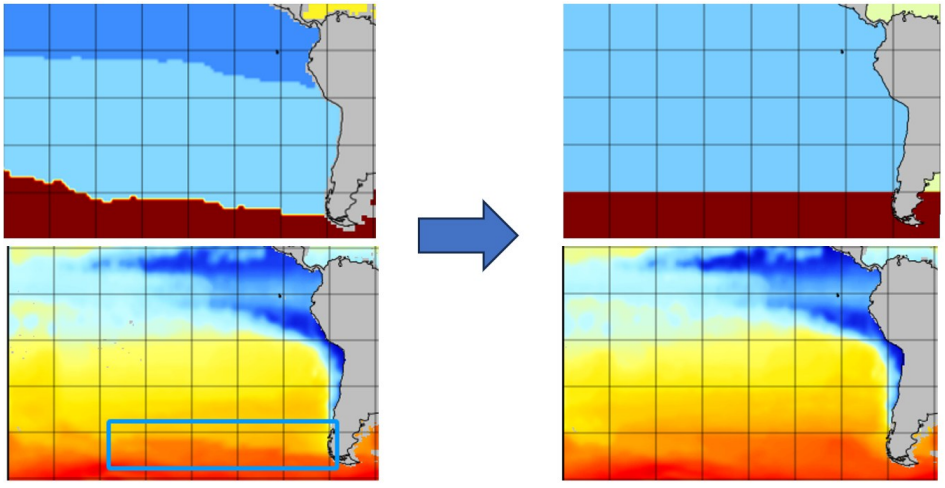
seasonally varying, dynamic partitioning. The SOM is trained directly on these monthly climatological fields, such that the DO climatology implicitly exerts a stronger influence on the resulting state partition and ensures that the partitioning evolves coherently with the seasonal DO cycle. “

4) How sensitive is this machine learning approach to the province selection?

Response: We thank the reviewer for this thoughtful and constructive comment. We were aware that training a ML model and applying it to different regions is generally suboptimal, due to covariate shifts, changes in the underlying physical drivers and differences in sampling bias across regions. As such, we opt to train region-specific models that reflect the distinct environmental regimes of individual provinces.

That said, training separate models across complex provinces also presents its own challenges. Province boundaries often coincide with regions of sharp physical gradients—such as oceanic fronts or mesoscale eddy fields—where extrapolation across these boundaries can introduce additional uncertainties and lead to discontinuities in the reconstructed fields.

Our sensitivity analyses indicate that this change primarily affects the reconstructed fields in the vicinity of regional boundaries, with larger discrepancies occurring in areas where the boundaries are geometrically complex. As illustrated in the figure below, which compares results obtained using a complex-boundary partitioning versus the current basin-based partitioning with simplified horizontal boundaries, the differences are largely confined to boundary zones. This supports that our current partitioning choice is robust with respect to the basin-scale and large-scale signals of the reconstruction.



Complex boundary Simplification boundary

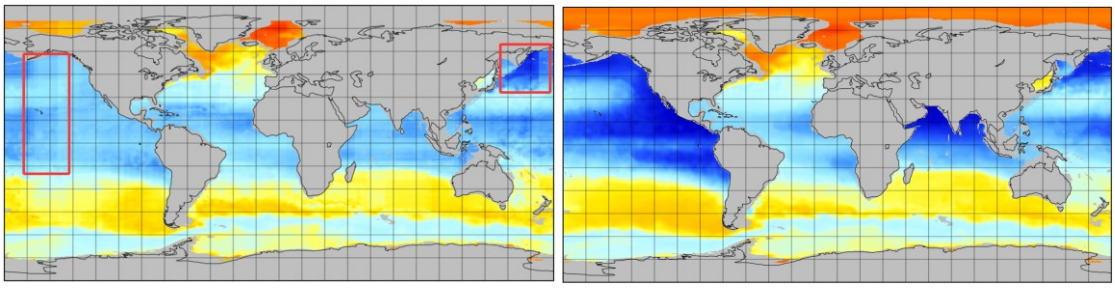
We believe these refinements improve the physical consistency of the reconstruction while minimizing artifacts at domain boundaries. We again thank the reviewer for helping us strengthen this aspect of the study.

In Lines 198-200 we stated:

"By avoiding highly intricate province boundaries, this design reduces sensitivity of variability and trend estimates to boundary effects and makes cross-boundary continuity easier to maintain. "

5. Line 218: why longitude is not included as a predictor while latitude is included? Please note that the coordinate information (longitude and latitude) might also need to be transformed like time to represent true geographical distances.

Response: We thank the reviewer for this thoughtful comment. In our early experiments, we found that adding longitude—even in transformed form—can induce spurious longitude-aligned stripe artifacts in data-sparse regions (notably in the North Pacific), where the reconstruction exhibited unrealistically sharp differences on the two sides of certain longitudes. This behavior strongly suggests that the model was partially using longitude as a shortcut proxy rather than learning physically consistent DO–environment relationships. To clearly document this issue, we present an early example of the artifact pattern caused by longitude for your reference.



Early example (including longitude) Current version

The current product (without longitude predictor) is substantially improved relative to the experimental version. The two panels shown here correspond to the same year, month, and depth (January 2000; 600 m).

By contrast, latitude is retained because it provides a physically interpretable large-scale constraint that aligns with meridional gradients in radiative forcing, stratification, and water-mass structure, and we did not observe analogous artifact patterns from latitude inclusion. Therefore, in the revised manuscript we do not include longitude as a predictor, and instead rely on a richer set of physically and biogeochemically meaningful covariates—together with our hierarchical training design—to represent spatial structure more robustly.

We have revised the predictor description to clarify this design choice and to explain that longitude predictors were tested but ultimately excluded due to the above artifact risk. We sincerely thank you for your guidance.

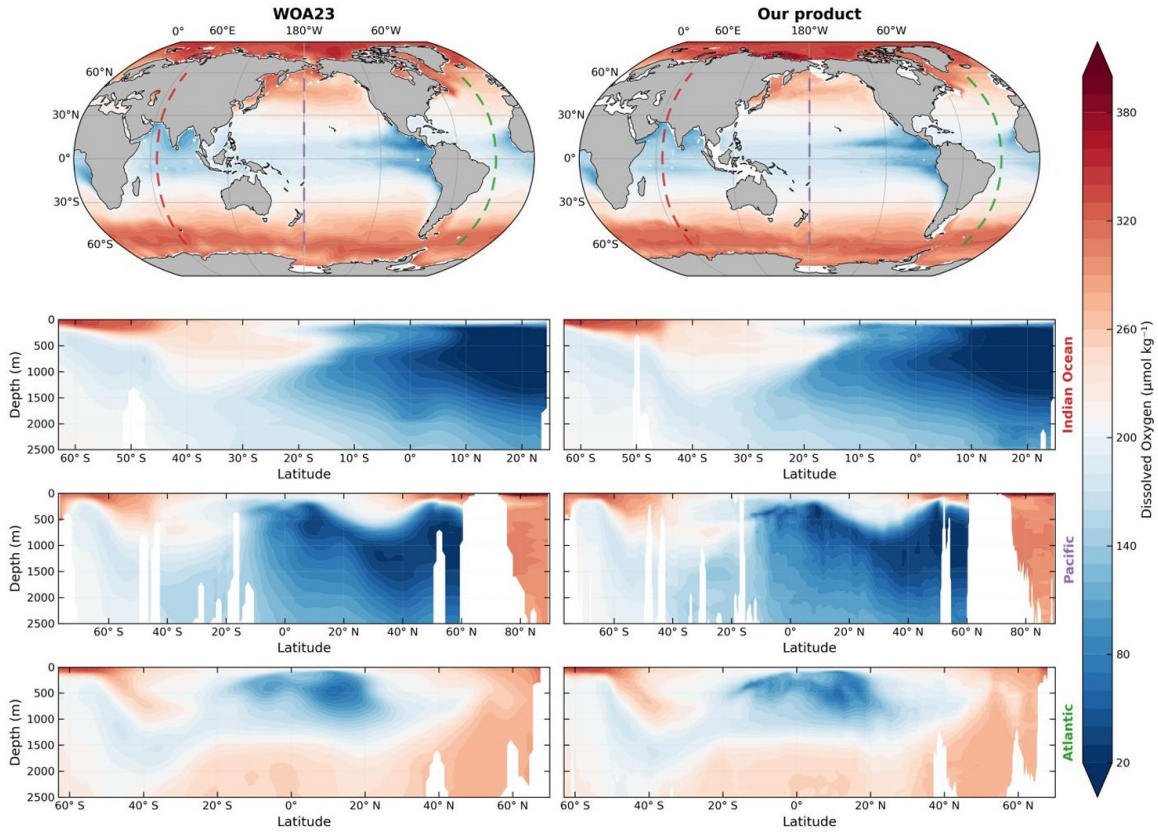
In Lines 234-235, we stated:

"We tested longitude as a candidate predictor; however, it induced spurious banded (stripe-like) artifacts in data-sparse regions and was therefore excluded from the final predictor set. "

6. Figure 8: the red dashed lines on the figures are misleading and they should not cross land.

Response: Thank you very much for this careful comment. To address the reviewer's concern, we revised Fig. 8 so that the dashed transect lines no longer intersect land and instead remain strictly within the ocean domain. The updated figure improves the clarity and interpretability of the transects. For transparency, we provide the original and revised versions of Fig. 8 below for comparison. We also note that, in the course of this revision round, the updated reconstruction and presentation are overall more consistent with the large-scale climatological structure in WOA23, with spatial patterns that appear more physically coherent. We sincerely appreciate your guidance—this adjustment makes the presentation more rigorous and user-friendly.

Original manuscript figure 8:



Revised manuscript figure 8:

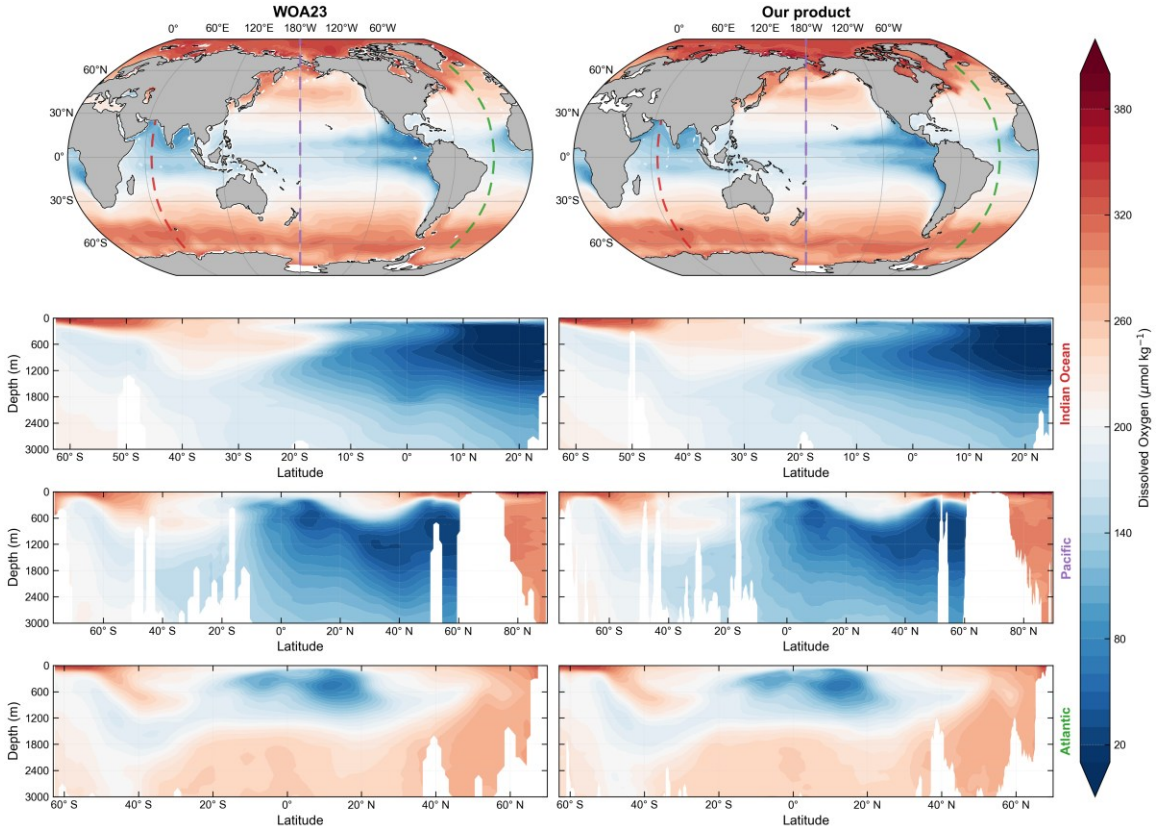


Figure 8: Climatological comparison between WOA23 and our product (GEOXYGEN). Row 1 shows the global-mean DO distribution averaged over 0–300 m. Colored dashed lines mark the locations of three sections: 65°E (red), 180° (purple), and 30°W (green). Rows 2–4 show DO cross-section distributions along these three transects.

Reviewer #2:

The manuscript presents a gridded dissolved oxygen (DO) dataset, GEOXYGEN, characterized by a high spatial resolution of 0.5° and an exceptionally dense vertical discretization of 187 layers from 1960 to 2024. However, several fundamental issues regarding data integrity, methodological rigor, and physical consistency must be addressed. The paper does not provide sufficient evidence that the reconstructed long-term oxygen changes, especially in data-sparse regions such as the Southern Ocean in early years, are not artifacts of the machine-learning methodology and predictor availability. Major revisions are therefore required before the dataset can be considered reliable for climate-scale analyses.

Response: We thank the reviewer for the thoughtful comments. In the revised manuscript, we fully addressed these concerns and substantially strengthened both the manuscript and the validation framework. We believe the revised version is more robust and transparent.
In the revision, we specifically addressed the potential "artifact" pathways that the reviewer highlighted:

1. New out-of-time evaluation design. We used decade-block cross-validation for model selection and calibration by reserving a strictly independent decade-stratified withheld-year test set exclusively for final performance assessment and inter-product benchmarking. This design reduces the risk of temporal leakage and provides an explicit out-of-time check on generalization.
2. Sensitivity to predictor availability (satellite-era covariates). To directly test whether satellite-era sea-surface information can imprint artificial long-term changes, we added a Full vs. Reduced predictor sensitivity experiment (new Sect. 5.2; Fig. 10), where the Reduced configuration excludes the entire sea-surface predictor group while keeping the remainder of the pipeline unchanged. The comparison focuses on deseasonalized upper-ocean anomalies and shows that the inferred low-frequency evolution remains stable without satellite-era predictors, supporting the robustness of the long-term signal.
3. Robustness in the Southern Ocean during early, data-sparse years. We added a ship-only (Argo-excluded) Southern Ocean analysis to evaluate whether early reconstructions are influenced by the introduction of Argo observations and platform sampling shifts. The ship-only results indicate that early reconstructions are broadly consistent and are not driven by Argo inclusion, indicating the early-year signal is not a methodological artifact.
4. Major refinement of the horizontal partitioning strategy for depth-consistent reconstruction. To address the concerns about physical consistency and regime stability across depth and time, we implemented a revised partitioning scheme that uses five major ocean basins as stable training domains and complements this scaffold with a regime descriptor derived from climatological background states to represent large-scale structural variability. This change reduces sensitivity to complex province boundaries, improves depth-wise consistency, and helps avoid boundary-induced artifacts, thereby strengthening the physical plausibility and robustness of the global 4-D reconstruction.

5. Separation of uncertainties. We expanded the uncertainty framework by explicitly separating observation-related uncertainty from mapping uncertainty, enabling users to identify periods and regions where uncertainty is elevated (including early, data-sparse regimes) and to interpret long-term changes with appropriate caution.

Major comments:

1. Observational quality control and cross-source consistency

A central weakness of the manuscript is the limited and insufficiently documented quality control applied to the observational data. The authors combine multiple observational products and implicitly assume that systematic biases among different sources are negligible. This assumption is unlikely to hold. Historical dissolved oxygen measurements are well known to exhibit source-dependent biases related to measurement technique, calibration practice, and processing methodology, even when data are flagged as "good" in the public datasets.

The treatment of Argo oxygen data is particularly concerning. The manuscript does not clearly state whether any sensor bias correction is applied, nor does it assess the potential impact of known Argo oxygen sensor issues on the reconstruction. Given the dominant role of Argo in the modern observing system, this omission undermines confidence in the training target used by the machine-learning model.

Response: We thank the reviewer for pointing this out. To address the reviewer's comment, we substantially strengthened the observational QC description and, most importantly, revised the source of Argo oxygen. In the revised manuscript, we explicitly adopt the internally consistent OSD/CTD and Argo DO dataset of Gouretski et al. (2024). In this dataset, OSD/CTD and Argo profiles are merged under a single automated QC framework, and **Argo DO biases are evaluated and corrected** using contemporaneous reference measurements, thereby reducing platform-dependent systematic differences before any machine-learning training is performed.

In Lines 89-94, we stated:

"To reconstruct a long-term, global ocean DO data set, we compiled several complementary in situ data products. These include CLIVAR and the Carbon Hydrographic Data Office (CCHDO), GLODAP, the GEOTRACES Intermediate Data Product (IDP2021), the OceanSITES mooring network, and the internally consistent OSD/CTD and Argo DO data set of Gouretski et al. (2024). The OSD/CTD and Argo profiles are merged under a single automated QC framework, and Argo DO biases are evaluated and corrected using contemporaneous reference measurements. This makes the quality more consistent across platforms and helps reduce platform-related systematic differences that can affect later modeling."

2. Validation strategy and independence at decadal time scales

The validation strategy relies primarily on holding out a limited number of years for testing. While this approach may be adequate for assessing short-term predictive skill, it is insufficient for evaluating decadal to multi-decadal variability, which is the central motivation for the dataset.

Training and validation subsets drawn from adjacent years inevitably share the same observing system, spatial sampling patterns, and predictor relationships, leading to optimistic skill estimates. This is particularly problematic when the dataset is intended for long-term trend analysis. The manuscript does not demonstrate that reconstructed trends in the 1960s–1980s, especially in poorly observed regions, are robust and not dominated by relationships learned from the dense Argo-era data.

Response: To address the reviewer's comment, we substantially revised our validation strategy to strengthen independence at climate-relevant time scales:

1) Decade-block cross-validation for model selection and calibration.
We replaced the previous year-wise random cross-validation with a decade-block (multi-year group) cross-validation strategy, in which entire decade-scale blocks are held out during each fold. This design explicitly reduces temporal dependence between training and validation subsets and mitigates the risk that performance metrics are inflated by shared observing-system structure.

A strictly independent, decade-stratified withheld-year test set for final evaluation and benchmarking. Because the modeling workflow relies on decade-block CV, we additionally constructed an independent test set by stratified sampling of one withheld year per decade. This withheld-year set is used only for final performance assessment and for comparisons with other products, and it is kept fully separate from the cross-validation used for model development.

In Lines 265-281 we stated:

3.4 Hyperparameter optimization and validation

Independent hyperparameter optimization for each basin-depth unit is performed using Bayesian inference (Optuna), targeting the objective of validation RMSE minimization (Table 2). This automated search is integrated with a decadal-block five-fold cross-validation scheme to address the challenges of non-stationary ocean signals. By grouping observations into multi-year blocks, we decouple validation results from the short-term temporal dependencies that often inflate predictive skill in traditional random-split CV (Salazar et al., 2022). This structural separation ensures that the model learns large-scale climatic drivers rather than localized temporal artifacts. In each cross-validation fold, early stopping is activated if validation RMSE fails to improve for 50 consecutive iterations, after which the optimal iteration count is recorded. To secure a robust final model, we set the terminal iteration count to the median of these recorded values across all folds. This iteration-locked retraining on the complete calibration set—with early stopping disabled—prevents overfitting and ensures a stable convergence state.
For an unbiased final assessment, a strictly independent global test set is constructed via decade-stratified random sampling (1960–2024), where one full calendar year per decade is entirely excluded from feature selection and hyperparameter tuning. The resultant test years (1961, 1970, 1984, 1993, 2003, 2012, and 2020) provide a temporally representative benchmark. This withheld-

year test set is reserved exclusively for final performance assessment and inter-product benchmarking; all model selection and calibration are conducted via decade-block cross-validation within the remaining training data. Integrated predictions from regional sub-models are then evaluated against this set using RMSE, mean absolute error (MAE), and the coefficient of determination (R^2).

2) Robustness of early-period trends against Argo-era influence (new control experiment; Fig. 11).

We fully acknowledge your concern that reconstructed trends in the 1960s–1980s, especially in poorly observed regions, could potentially be dominated by relationships learned from the dense Argo era. To directly test this, we designed an additional control experiment (new Fig. 11) that evaluates the early-period reconstruction under an observation-availability constraint, thereby quantifying whether the inferred low-frequency evolution persists without reliance on Argo-era information. The results support that the early-period signal is not an artifact of Argo-era learned relationships, and we now explicitly document this evidence in the revised manuscript.

See Lines 510-532 copied below:

5.3 Ship-Only Analysis of Long-Term Trends

To quantify the dependency of the early reconstruction on modern autonomous sensing, a ship-only sensitivity experiment was conducted to assess potential retrospective signal contamination from the data-dense Argo era. Within the Southern Ocean (south of 45°S), we benchmarked an "Argo-excluded" configuration—utilizing non-Argo historical records—against an "Argo-included" configuration across the **1960–2000 period.** This region serves as a critical diagnostic domain due to its radical transition from sparse historical sampling to comprehensive Argo coverage over the last two decades. By maintaining identical external physical analysis constraints, the experiment isolates the impact of Argo observations on the reconstructed climatological mean and vertical structure (Fig. 11).

The results manifest pronounced morphological invariance in the upper-ocean (1–100 m) DO climatology across both configurations. The annular high-oxygen band and its relative position within the circumpolar system remain effectively stationary, implying that Argo inclusion does not induce systematic structural rearrangement. Notably, the difference map reveals localized positive ΔDO hotspots concentrated around the Antarctic margin, with enhanced amplitudes in several coastal/shelf sectors, whereas departures in the open-ocean circumpolar interior are minimal and barely discernible. Consistent with this pattern, the area-weighted mean vertical profiles are nearly overlapping across most depths (panel d), with only a subtle upper-ocean positive offset in the Argo-included configuration that remains below ~0.5 μmol kg^{-1}, indicating modest amplitude refinement rather than depth-dependent reorganization. Collectively, these features suggest that the

reconstructed vertical architecture is primarily constrained by consistent physical structure, while the denser modern sampling acts mainly to fine-tune regional magnitudes, thereby supporting the structural integrity of the historical reconstruction.

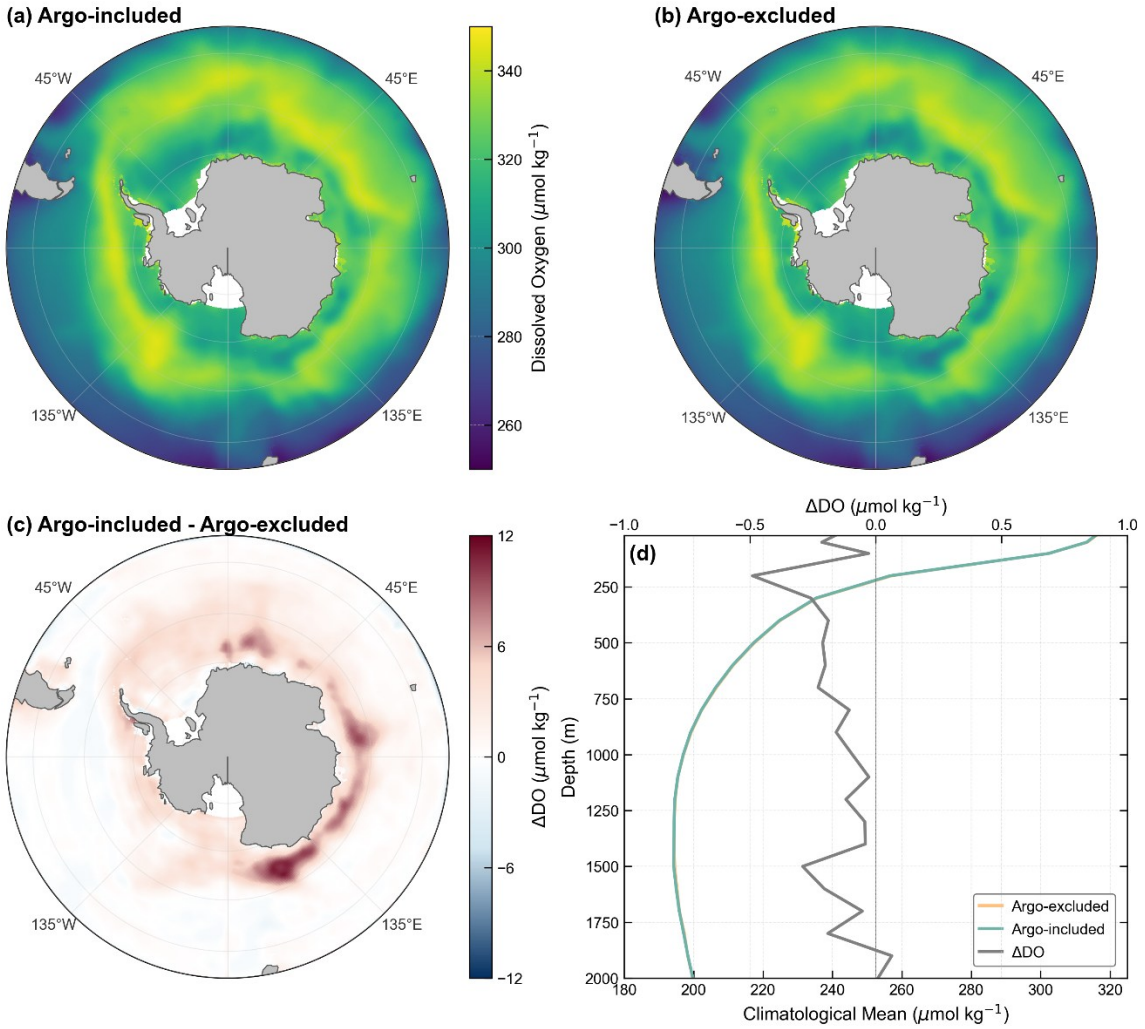


Figure 11: Southern Ocean dissolved oxygen (DO) climatology for 1960–2000. (a) Argo-included climatological mean DO averaged over 1–100 m. (b) Same as (a), but for the Argo-excluded configuration. (c) Difference map (Argo-included minus Argo-excluded) for the 1–100 m climatological mean DO. (d) Area-weighted mean vertical DO profiles south of 45°S from the Argo-included and Argo-excluded reconstructions; the upper x-axis shows the corresponding profile difference (Argo-included minus Argo-excluded).

3. Lack of uncertainty quantification

Another major limitation is the absence of a rigorous uncertainty framework. The manuscript reports standard skill metrics such as RMSE and R^2, but provides no quantitative estimate of uncertainty at the grid-cell level, nor for regionally integrated quantities such as basin means or OMZ volumes. For a dataset intended to support climate diagnostics and trend analysis, uncertainty estimates are essential. Users need to know how uncertainty varies spatially, temporally, and with

depth, and how it grows backward in time as observations become sparse. The lack of uncertainty propagation into derived metrics, such as OMZ volume, severely limits the scientific reliability and usability of the product.

Response: We have directly addressed the reviewer's comment by adding a dedicated uncertainty framework. Specifically, we added a new section, Sect. 5.1 "Uncertainty Analysis", where we provide a quantitative, interpretable uncertainty decomposition at the grid-cell level. We decompose the total uncertainty into two complementary components: (i) observation-related uncertainty and (ii) mapping uncertainty.

To make these uncertainty estimates practically useful to users, we further report how uncertainty varies spatially, with depth, and through time. In the revision we provide: (1) multi-depth maps of U_{obs}, σ_{map}, and U_{total}; (2) basin-scale summaries for the five major ocean basins; (3) a depth-dependent uncertainty profile; and (4) the decadal evolution of uncertainty over withheld test years, which serves as a diagnostic of how uncertainty changes as observational constraints become denser in the modern era.

We agree that the propagation of uncertainties is an important issue for the machine learning approaches. In the current revision, we quantify uncertainty primarily at the data level and for basin-integrated means. Due to scope and length constraints, we will discuss the uncertainty propagation into derived metrics (e.g., OMZ volume) in a separate paper.

See Lines 417-482 copied below:

5.1 Uncertainty Analysis

To quantify the credibility of the GEOXYGEN, we decompose total uncertainty into two components: observation-related uncertainty and mapping uncertainty. Observation-related uncertainty summarizes measurement error and the representativeness error introduced when observations are aggregated to the cell–month–depth scale. Mapping uncertainty describes prediction error from the machine-learning mapping between environmental predictors and DO. This decomposition separates label-side and model-side contributions. It also supports diagnosing elevated uncertainty in coastal regions and potential bias in earlier, data-sparse periods, and it improves traceability and clarity for product use.

Measurement error depends on the measurement technique and instrument. For Winkler titration, bottle measurements can include random errors on the order of 1 μmol kg^{-1} or smaller (Carpenter, 1965). Since observing platforms differ in sensor type, calibration strategy, and QC level, we assign a platform-specific measurement error scale σ_{meas} to each data source to represent differences in precision among platforms. Bottle-based sources including CCHDO Bottle, GLODAP, and GEOTRACES IDP are assigned 1 μmol kg^{-1}. Since Argo data are bias corrected, Argo, OSD and CTD, and CCHDO CTD are assigned 1.5 μmol kg^{-1}. OceanSITES is assigned 2.0 μmol kg^{-1}.

When forming supervised-learning labels at the cell–month–depth scale, observation-related uncertainty is defined as the combined contribution of measurement error, vertical mapping uncertainty, and representativeness error, as defined in Eq. (10).

$$U_{\mathrm{obs}} = \sqrt{\sigma_{\mathrm{meas}}^2 + \sigma_{\mathrm{interp}}^2 + \sigma_{\mathrm{rep}}^2} \quad (10)$$

Here, $\sigma_{\mathrm{interp}}$is the vertical mapping uncertainty introduced when a profile is mapped to standard levels, and σ_{rep} is the representativeness error estimated from within unit dispersion. Their definitions and computation have been described earlier.

Mapping uncertainty characterizes error generated during the mapping from environmental fields to DO. It reflects the combined effects of model structure, representativeness of training samples, and nonstationarity across regions and depth levels. On the independent test set, we define the residual following Eq. (11):

$$r = \hat{y} - y, \quad (11)$$

where $\hat{y}$ is the model prediction and y is the observed label for the corresponding test sample. Following Ito et al. (2024a), we estimate the mapping uncertainty at each grid cell from the test residuals by taking their second central moment, as defined in Eq. (12).

$$\sigma_{\mathrm{map}} = \sqrt{\overline{r^2} - (\overline{r})^2}, \quad (12)$$

Here, the overbar denotes the mean over all test samples within the same grid cell. This definition estimates the residual variance at the grid cell scale. It separates the systematic component $\bar{r}$from the random error component captured by σ_{map}.

Based on the two components above, total uncertainty (σ_{tot}) is defined in Eq. (13).

$$U_{\mathrm{total}} = \sqrt{U_{\mathrm{obs}}^2 + \sigma_{\mathrm{map}}^2}, \quad (13)$$

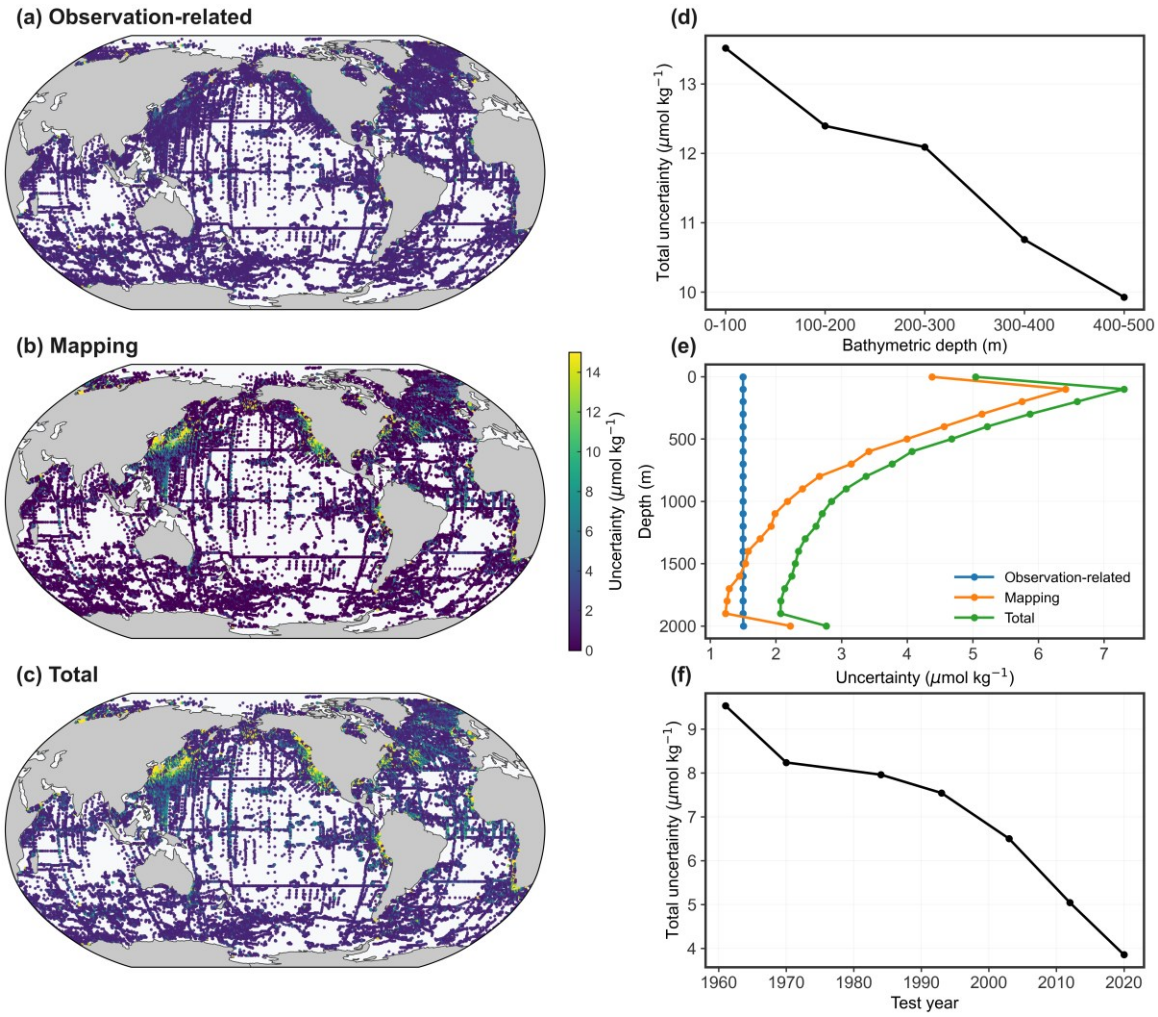


Figure 9: The left panels show the mean uncertainty distributions across multiple standard depth layers (0–2000 m), including observational uncertainty (a), mapping uncertainty (b), and total uncertainty (c). The right panels depict the nearshore spatial variability of total uncertainty (d), its vertical (depth-dependent) structure (e), and the decadal evolution of total uncertainty over the test years (f).

As shown in Fig. 9, the spatial distribution of uncertainty in GEOXYGEN reveals a distinct bifurcation between observational noise and mapping residuals. Observational uncertainty manifests pronounced spatial stationarity across pelagic domains. Conversely, mapping uncertainty exhibits a structured regional geometry, with elevated error bands aligning with high-gradient kinetic regimes such as western boundary currents and upwelling systems. The highest uncertainty is observed in the Pacific (7.385 μmol kg^{-1}), followed by the Atlantic (6.148 μmol kg^{-1}), Arctic (4.439 μmol kg^{-1}), Indian (4.084 μmol kg^{-1}), and Southern Ocean (3.652 μmol kg^{-1}). The higher uncertainty in the Pacific and Atlantic Oceans is primarily due to the structural complexity and dynamic intensity of their oceanographic systems, as well as the coastal distribution of early

observational data in these basins. The global-mean total uncertainty of 6.054 μmol kg^{-1} conceals a pronounced shallow-water divergence: in nearshore and shelf regions where bathymetric depth is shallower than ~200 m, total uncertainty rises to 12.917 μmol kg^{-1}—more than double the open-ocean baseline (5.970 μmol kg^{-1}). Consistent with the bathymetry-binned diagnostic, uncertainty increases monotonically with decreasing bathymetric depth, indicating progressively reduced predictability toward the shallow, dynamically heterogeneous coastal ocean (Fig. 9(d)). This intensification is predominantly driven by localized, high-frequency processes—including phytoplankton pulses, riverine influx, and tidal oscillation—which generate non-linear spatial gradients that challenge the transferability of pelagic-trained feature relationships (Gilbert et al., 2010; Regier et al., 2023; Giomi et al., 2023; Liu et al., 2024). These localized dynamics induce steep spatial gradients and temporal non-stationarity, which subsequently reduce the regional transferability of learned DO-environment associations (Valera et al., 2020). For nearshore and semi-enclosed bay environments, we recommend using GEOXYGEN with caution and interpreting results at larger spatial aggregation to reduce sensitivity to local high-frequency variability.

The vertical stratification of uncertainty reflects the underlying hydrographic stability and process coupling within the water column. As depicted in Fig. 9(e), the total uncertainty profile reaches its maximum in the epipelagic layer before undergoing monotonic attenuation toward the abyssal depths. This vertical variation is consistent with the change in model accuracy across depth layers. In contrast, the intermediate and deep layers provide stronger water-mass constraints and a more coherent oxygen field, facilitating a convergence of mapping uncertainty as the predictive relationship stabilizes.

Temporal trends in uncertainty serve as a diagnostic of the evolving global observing system. The progressive reduction in total uncertainty across the withheld test years (Fig. 9(f)) coincides with the expansion of the Argo float network, which transitioned ocean sensing from route-based ship surveys to a spatially distributed autonomous paradigm. This transition significantly improved the observational constraints in the Southern Hemisphere and remote ocean basins, effectively lowering the residual variance in model validation. While this decline highlights the structural evolution of sampling coverage, the resulting time series reflects the collective stability of the multi-decadal reconstruction rather than a year-specific local error estimate.

4. Time inconsistent predictors and regime consistency

The model uses a large number of predictors, many of which are derived from satellite products or reanalyses that are only available after the 1990s. The manuscript suggests that in earlier decades these predictors are effectively masked, leaving the model to rely primarily on temperature, salinity, and oxygen solubility. This raises a serious concern that the reconstruction may effectively be

governed by different rules in different time periods, potentially introducing artificial regime shifts or spurious trends. The manuscript does not demonstrate that reconstructions using the reduced predictor set are consistent with those obtained using the full predictor suite.

Response: We thank the reviewer for the thoughtful comment. We added a direct, controlled consistency test in the revision. Specifically, we implemented a two-configuration sensitivity experiment (new Sect. 5.2; Fig. 10) to explicitly assess whether satellite-era sea-surface information alters the reconstructed DO signal. The Full configuration uses the complete predictor suite, whereas the Reduced configuration excludes the entire sea-surface predictor group (U, V, SSH, EKE, MLD, PAR, Chl-a, DIC, pCO2, pH, alkalinity, and CO2 flux). All other components of the pipeline are held fixed, so the difference between the two reconstructions isolates the incremental effect of this predictor group on the same grid and over the same period.

To target the time-scale and depth range most susceptible to surface information, we compare deseasonalized 1–100 m depth-averaged DO anomalies, and we quantify the predictor impact using the configuration-induced component Full − Reduced. The results show high congruence between the Full and Reduced reconstructions across 1960–2022 (Fig. 10). Importantly, the Full − Reduced residual is stochastically negligible prior to the mid-1980s, indicating that the pre-satellite signal is not sensitive to the excluded surface predictors. Minor configuration-induced differences are confined to a narrow temporal window after the rapid expansion of satellite-derived observations and do not alter the inferred multi-decadal evolution or the long-term deoxygenation-related conclusions.

See Lines 483-509 copied below:

5.2 Impact of Removing Surface Predictors on Trends

A two-configuration sensitivity experiment was designed to test whether satellite-era sea-surface information alters the reconstructed dissolved oxygen (DO) signal. The full predictor configuration used the complete predictor suite, while the reduced predictor configuration excluded the sea-surface predictor group (U, V, SSH, EKE, MLD, PAR, Chl-a, DIC, pCO_2, pH, alkalinity, and CO_2 flux). All other parts of the reconstruction pipeline were kept the same, so differences between the two products isolate the effect of including this predictor group on the same grid and over the same period.

The comparison focuses on deseasonalized DO anomalies averaged over the 1–100 m depth range and further smoothed using a centered 12-month moving average. This metric targets the upper ocean, where sea-surface information is most likely to influence variability, while suppressing grid-scale noise that can obscure low-frequency signals. Predictor impact is quantified by the configuration-induced component, defined as Full − Reduced, which isolates the incremental effect of the excluded predictor group in anomaly space.

The full and reduced reconstructions exhibit strong agreement over 1960–2022 (Fig. 10), yielding a consistent depiction of upper-ocean (1–100 m) low-frequency variability: a sustained positive-anomaly regime through the 1970s–1980s followed by a transition toward persistently negative anomalies after ~1990 (relative to the monthly climatology). The Full–Reduced difference remains close to zero prior to the mid-1980s and stays small compared with the total anomaly amplitude thereafter, indicating that the basin-scale, decadal signal is only weakly sensitive to the inclusion of satellite-era surface predictors. As an external benchmark, ML4O2 reproduces comparable variability during ~1965–2010 but shows more negative anomalies in the most recent decade, indicating a stronger inferred upper-ocean deoxygenation signal (relative to its monthly climatology) than GEOXYGEN over the same period; however, part of this divergence may arise from inter-product differences in baseline climatology, sampling, and reconstruction methodology. Overall, the close concordance between configurations supports the robustness of GEOXYGEN for decadal-scale assessments, with configuration-dependent deviations being minor relative to the dominant multi-decadal evolution.

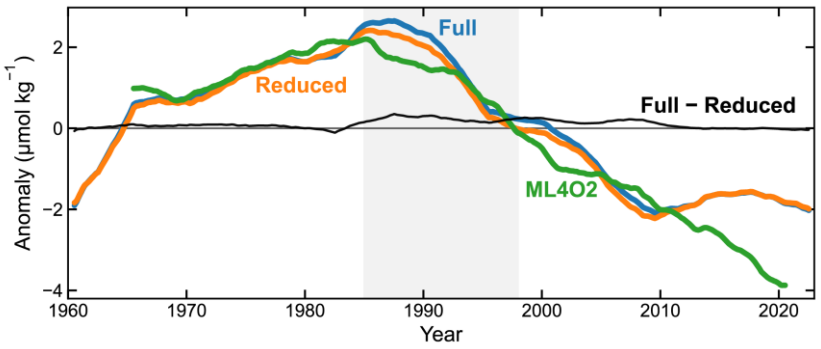


Figure 10: Depth-averaged (1–100 m) monthly dissolved-oxygen anomalies (1960–2022). The figure compares deseasonalized anomalies from the Full predictor reconstruction, the Reduced predictor reconstruction (excluding sea-surface predictors), and ML4O2, and overlays the difference between the Full and Reduced reconstructions (Full − Reduced). Gray shading indicates 1985–1997, corresponding to the rapid expansion of satellite-derived sea-surface observations.

5. Interpretation of predictors and mechanistic claims

Although the manuscript describes the framework as “biogeochemistry-aware,” many predictors originate from model or reanalysis products that themselves contain biases and uncertainties, and several predictors function primarily as proxies or coordinate variables rather than physical drivers. The discussion of predictor importance risks being interpreted as mechanistic attribution, despite the fact that machine-learning importance metrics do not imply causality.

Response: In the revised text, we avoided causal or mechanistic language when discussing predictors. Specifically, we now describe predictor effects as diagnostic, conditional relationships

learned by the model (i.e., statistical associations within the training distribution), rather than physical "drivers" or causal controls. We also added explicit caveats stating that feature-importance analyses are intended to improve interpretability and transparency of the model behavior, not to establish causality.

Consistent with this caution, we emphasize that GEOXYGEN's "biogeochemistry-aware" design refers to integrating physically and biogeochemically informed constraints and covariates—such as oxygen-solubility, depth-/basin-adaptive learning, and a SOM-derived environmental-state descriptor (SOM_DES) that encodes the large-scale climatological background structure—to improve robustness and interpretability, rather than claiming mechanistic attribution.

See Lines 302-317 copied below:

Feature-importance diagnostics are reported to describe model dependence and do not imply causality. The adaptive feature-selection results show that the model's reliance on predictors varies strongly with depth and region. In the upper 10 m, temperature and O_2 saturation are consistently among the most informative predictors, whereas at intermediate depths (1000–2000 m) salinity tends to contribute more strongly in certain high-latitude regimes, including the Arctic and Southern Oceans (Fig. 5). This depth-dependent pattern motivates the use of depth-specific predictor sets to better represent distinct hydrographic contexts. In addition, several regionally relevant covariates (e.g., SSH in the Indian Ocean and DIC/alkalinity in low-oxygen environments) are retained more frequently by the selection procedure, indicating that they provide useful contextual information for prediction under specific regimes (Franco et al., 2014).

Although latitude serves as a dominant proxy for broad-scale gradients (Milà et al., 2024), specialized SSEVs are vital for refining local accuracy. Our regionalized architecture prioritizes these idiosyncratic dynamics, utilizing variables like SOM_DES to resolve high-frequency seasonal variations in the Southern Ocean. This adaptive approach mitigates the biases inherent in spatially stationary parameterizations, ensuring the reconstruction respects the intrinsic heterogeneity of the global oxygen cycle.

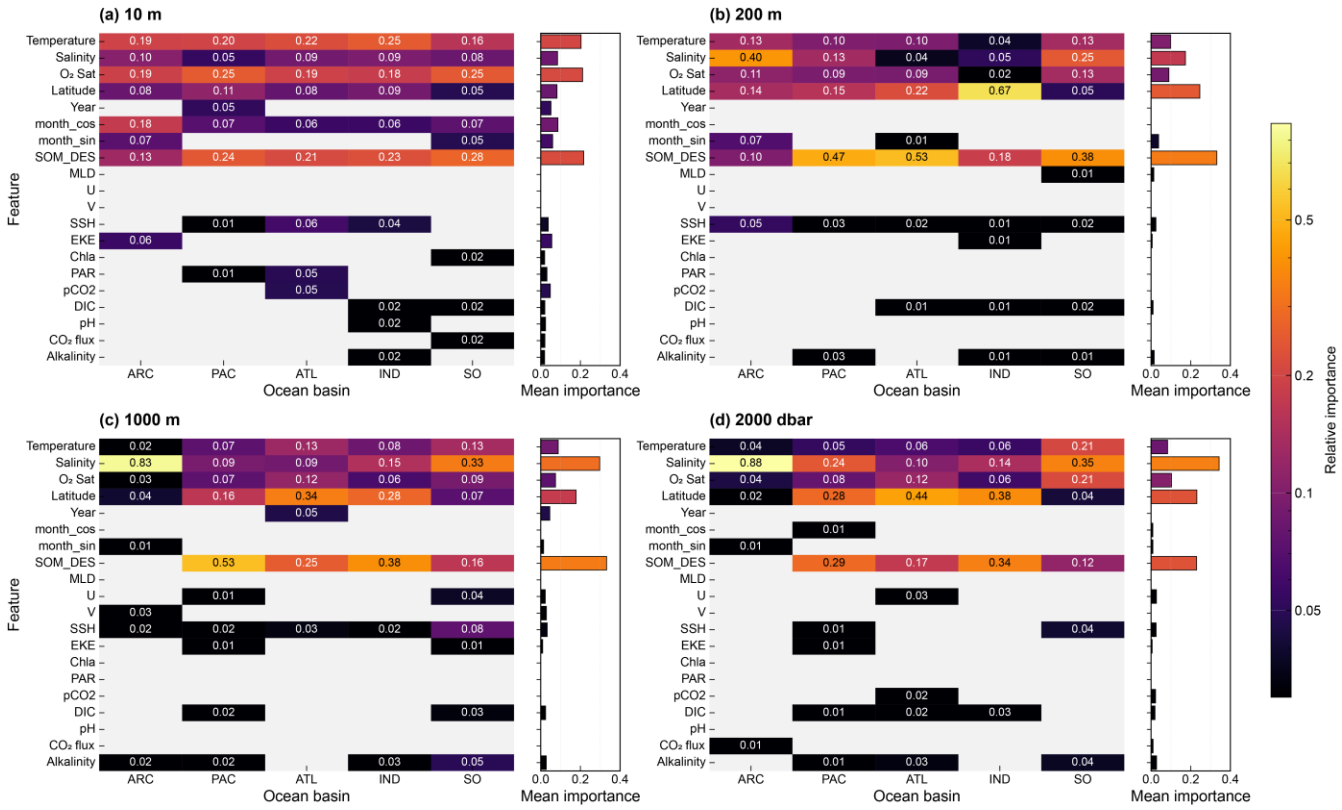


Figure 5: Heatmap of relative feature importance across depths and basins. Colors are on a logarithmic scale. The bar chart on the right shows each feature's mean importance computed over the basin provinces in which that feature is available.

6. Regional partitioning and boundary continuity

The authors acknowledge that a single global model may be inadequate and therefore adopt a regional partitioning strategy. While pragmatic, this approach introduces the risk of discontinuities at region boundaries. The manuscript does not provide sufficient quantitative evidence that boundary fusion fully resolves these issues, particularly for variability and trends. An explicit evaluation of continuity across regional boundaries is needed to demonstrate that the partitioning does not introduce artificial spatial artifacts.

Response: We are grateful for this incisive and constructive feedback. The reviewer raises an important point: although regional partitioning is a practical strategy for addressing heterogeneous DO–environment relationships, it carries the risk of introducing boundary discontinuities in the form of "step-effect" seams. We have addressed this issue at two complementary levels:

1) **Reducing the boundary-artifact risk at the source by simplifying the partitioning geometry.**

In the revised framework, we reduced the number of horizontal partitions and replaced the previous complex, biome-shaped boundaries with a simpler five-basin partitioning (Atlantic, Pacific, Indian, Southern, Arctic). These basin boundaries are geometrically straightforward (largely meridional–

zonal separations), which substantially lowers the likelihood of sharp seam artifacts compared with highly intricate province boundaries. This design choice was made explicitly to improve boundary continuity and robustness across depths.

See Lines 194-201 copied below:

Our reconstruction framework utilizes a spatiotemporally stratified approach to address the shifting controlling mechanisms of ocean deoxygenation across basins and depths (Ma et al., 2025; Ito et al., 2024a). **Following the basin definitions in the World Ocean Atlas 2023 (WOA23; Garcia et al., 2024)**, we additionally treat the Southern Ocean as a dedicated domain. Accordingly, the horizontal grid is divided into five primary modeling domains (Atlantic, Pacific, Indian, Southern, and Arctic), which are held constant across vertical layers to maintain training coherence (Fig. 3). **By avoiding highly intricate province boundaries, this design reduces sensitivity of variability and trend estimates to boundary effects and makes cross-boundary continuity easier to maintain.** We further mask the Mediterranean, Red Sea, and other semi-enclosed marginal seas to focus the reconstruction on open-ocean dynamics dominated by large-scale circulation and transport.

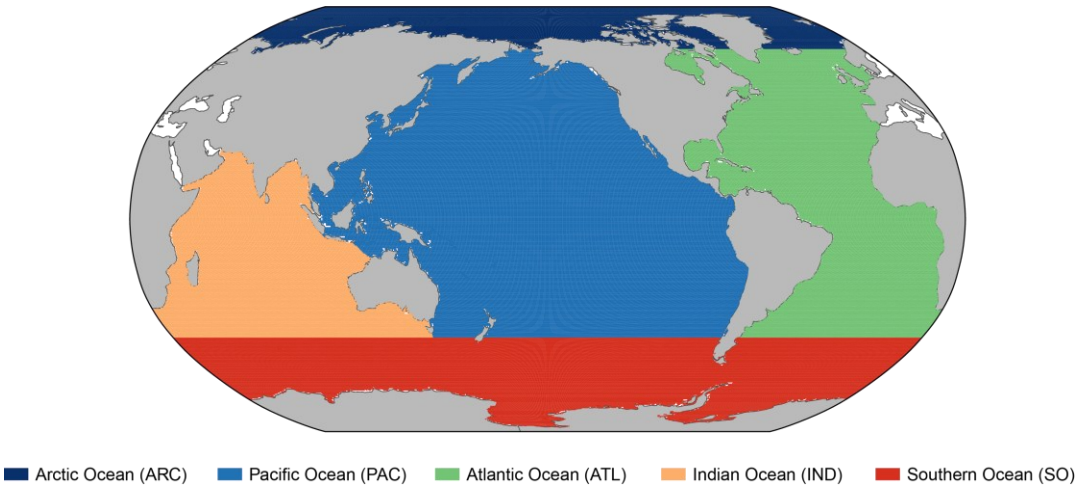


Figure 3: Partitioning of the global open ocean into five basins.

2) Explicit evaluation of boundary fusion and continuity (new Sect. 5.4 "Boundary Fusion Effects").

To directly respond to your request for quantitative evidence, we added a new section, Sect. 5.4 Boundary Fusion Effects, where we evaluate continuity across regional boundaries before and after fusion. We agree that fusion may not fully eliminate all seams. The revised manuscript now provides an explicit assessment of how boundary fusion alters cross-boundary consistency in key statistics relevant to climate diagnostics, including variability- and trend-related measures. This

analysis demonstrates that boundary fusion substantially reduces seam-like discontinuities and improves cross-boundary continuity, thereby mitigating partition-induced artifacts.

See Lines 533-553 copied below:

5.4 Boundary Fusion Effects

While basin-based modeling represents regional environmental heterogeneity, it often induces "step-effect" discontinuities at basin boundaries, resulting in unrealistic shifts in long-term trends and monthly variability at adjacent grid points. To address these boundary artifacts, we implemented a fusion method to smooth inter-basin transitions. We validated this approach by constructing boundary diagnostic samples on a 0.5° by 0.5° grid, specifically selecting adjacent points across basin interfaces.

The diagnostic experiment targeted the 100m depth layer, a region characterized by high spatial complexity and hydrographic sensitivity. Using a global non-partitioned model as a continuous baseline, we quantified discontinuity magnitudes via the differential trend (Δtrend) and standard deviation (Δstd) at adjacent cells. Contrasting fused and unfused basin-specific outputs against this reference isolated the consistency gains.

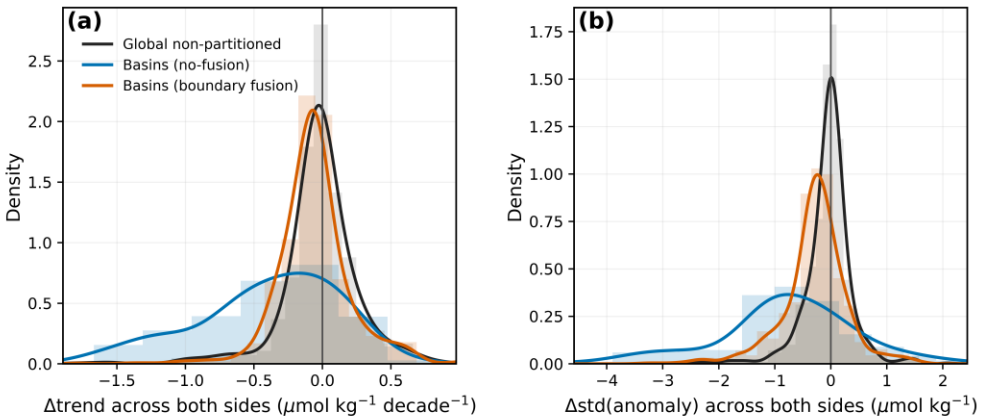


Figure 12: Spatial distribution of the "step-effect" before and after boundary fusion at the 100 m depth layer. (a) Compares the trend differences (Δtrend), and (b) compares the standard deviation differences (Δstd) across adjacent grid points at the basin boundaries.

Results reveal that the boundary fusion protocol mitigates abrupt statistical shifts, decreasing local biases in the partitioned map (Fig. 12). Improvements in Δtrend and Δstd demonstrate that the smoothing operator suppresses interface artifacts. Nevertheless, small discrepancies relative to the global non-partitioned baseline persist, indicating that independently trained regional submodels may retain slightly different statistical mappings that are only partially reduced by simple fusion.

Boundary fusion thus serves as an effective strategy for mitigating partition-induced "step-effect" artifacts, yielding a more coherent global spatial structure. While localized gradients remain in high-contrast hydrographic regions, their reduced magnitude indicates a meaningful alleviation of boundary inconsistencies. Future enhancements could involve adaptive fusion schemes or multi-task learning to explicitly share data across regional submodels while preserving specialization.

We sincerely appreciate your guidance—it prompted a major strengthening of both the methodology (simplified partitioning) and the evidence base (explicit continuity evaluation), making the reconstruction more physically consistent and more reliable for variability and trend analyses.

Specific comments:

1. Lines 100-105: The manuscript does not clearly describe how duplicates are defined and removed. In practice, the same ship-based profile often appears in multiple archives with small differences in time, position, or depth sampling, and the criteria used to identify such duplicates must be explicitly defined. A vague reference to duplicate removal is not adequate. Furthermore, the proposed fallback strategy, local gridded outlier detection, cannot substitute for a systematic assessment of inter-source biases.

Response: We thank the reviewer for this thoughtful comment. In the revised manuscript, we defined and documented an explicit duplicate-identification and removal protocol. In the updated QC description, duplicate profiles are identified across archives using coincidence criteria of ≤ 1 km spatial separation and ≤ 24 h temporal difference, and redundant records are consolidated accordingly. When multiple duplicates exist, we prioritize the profile with the highest vertical sampling density to retain the maximum information content. We also agree with the reviewer's second point that local gridded outlier detection cannot substitute for a systematic treatment of inter-source biases. In the revision, we clarify that robust aggregation is used only to stabilize grid–month labels in high-variability regimes, not as a cross-source bias-correction strategy. To address cross-source consistency more directly—especially for Argo oxygen—we now explicitly adopt the internally consistent OSD/CTD and Argo dataset of Gouretski et al. (2024), in which profiles are processed under a unified automated QC framework and Argo DO biases are evaluated and corrected using contemporaneous reference measurements.

In Lines 96-98, we stated:

"Spurious terrestrial signals were omitted via land-masking, and duplicate profiles—defined by coincidence criteria of ≤ 1 km spatial distance and ≤ 24 h temporal difference—were identified across archives."

2. The decision to exclude regions shallower than 200 m is supported by RMSE-based sensitivity

analysis and is reasonable from a modeling perspective. However, this exclusion removes many of the regions where oxygen variability is most societally relevant, including upwelling shelves and seasonally hypoxic coastal systems.

Response: To address the reviewer's comment, we removed the previous nearshore (<200 m) exclusion and now retain these regions in the reconstruction. Rather than discarding shallow-water data to improve RMSE, we adopt a more transparent and user-oriented approach: we explicitly quantify and discuss the elevated uncertainty in nearshore and shelf environments within our uncertainty framework (Sect. 5.1). The revised manuscript now highlights that total uncertainty increases markedly toward shallow bathymetry, reflecting stronger sub-grid variability, nonlinear coastal processes, and reduced predictability relative to the open ocean. This treatment preserves coverage in societally relevant regions while providing users with clear guidance on reliability and appropriate interpretation.

In Lines 460-470, we stated:

"The global-mean total uncertainty of 6.054 μmol kg^{-1} conceals a pronounced shallow-water divergence: in nearshore and shelf regions where bathymetric depth is shallower than ~200 m, total uncertainty rises to 12.917 μmol kg^{-1}—more than double the open-ocean baseline (5.970 μmol kg^{-1}). Consistent with the bathymetry-binned diagnostic, uncertainty increases monotonically with decreasing bathymetric depth, indicating progressively reduced predictability toward the shallow, dynamically heterogeneous coastal ocean (Fig. 9(d)). This intensification is predominantly driven by localized, high-frequency processes—including phytoplankton pulses, riverine influx, and tidal oscillation—which generate non-linear spatial gradients that challenge the transferability of pelagic-trained feature relationships (Gilbert et al., 2010; Regier et al., 2023; Giomi et al., 2023; Liu et al., 2024). These localized dynamics induce steep spatial gradients and temporal non-stationarity, which subsequently reduce the regional transferability of learned DO-environment associations (Valera et al., 2020). For nearshore and semi-enclosed bay environments, we recommend using GEOXYGEN with caution and interpreting results at larger spatial aggregation to reduce sensitivity to local high-frequency variability."

3. Section 4.1, in particular, does not deliver substantive new insight into oxygen dynamics, but instead reiterates known associations between oxygen, temperature, and stratification. This section should either be clearly reframed as a diagnostic assessment of model behavior or substantially strengthened with process-based analyses. As written, it overreaches relative to what the method can support.

Response: We have reframed Sect. 4.1 explicitly as a diagnostic assessment of model behavior,

rather than a mechanistic interpretation of oxygen dynamics. In the revision, we revised the texts to emphasize that feature-importance patterns describe statistical dependence and conditioning learned by the model within the training distribution, and do not imply causality or mechanistic attribution. Correspondingly, we removed causal wording (e.g., "drivers" or "controls") and added clear caveats to prevent over-interpretation of importance metrics.

In Lines 302-309, we stated:

"Feature-importance diagnostics are reported to describe model dependence and do not imply causality. The adaptive feature-selection results show that the model's reliance on predictors varies strongly with depth and region. In the upper 10 m, temperature and O_2 saturation are consistently among the most informative predictors, whereas at intermediate depths (1000–2000 m) salinity tends to contribute more strongly in certain high-latitude regimes, including the Arctic and Southern Oceans (Fig. 5). This depth-dependent pattern motivates the use of depth-specific predictor sets to better represent distinct hydrographic contexts. In addition, several regionally relevant covariates (e.g., SSH in the Indian Ocean and DIC/alkalinity in low-oxygen environments) are retained more frequently by the selection procedure, indicating that they provide useful contextual information for prediction under specific regimes (Franco et al., 2014)."

We sincerely thank the Editor for overseeing the review process and for giving us the opportunity to revise our manuscript. We also thank the reviewers for their time, care, and helpful guidance. The revision was shaped by your comments, and we believe the manuscript is now clearer and stronger in its data processing, validation design, uncertainty reporting, and interpretation. Our updated comparison results indicate that GEOXYGEN now exhibits substantially closer agreement with WOA23 in the large-scale climatological structure of dissolved oxygen than in previous versions. Compared with existing reconstructed DO products, GEOXYGEN provides higher spatial resolution, and evaluation on an independent test set suggests slightly higher accuracy than other products. We believe these improvements significantly increase the scientific value and long-term usability of GEOXYGEN, and we are grateful for the chance to improve the work under your guidance.