Dear Referees,

We sincerely thank the referees for their careful evaluation of our manuscript and for the constructive comments. We have carefully considered all comments and revised the manuscript accordingly. Below, we provide a point-by-point response to each comment (referee comments in *italics*, our responses in regular type, and relevant changes made in the manuscript in blue). All line numbers refer to the revised manuscript.

Sincerely,
Mingyu Han
On behalf of all authors

Response to Referee #1

Comment 1:
1. Lack of Intercomparison with Existing Data Products
The manuscript currently lacks a robust comparison with other widely used ocean DO data products. To establish the reliability of this new product, it is essential to contextualize its performance against existing datasets. I strongly recommend adding a comprehensive data intercomparison section to validate the BLENDR outputs. The authors should refer to and compare their results within the context of recent multi-product coordinated intercomparisons, such as the one presented by Ito et al. (2025). This will significantly enhance the credibility of your product.
Reference: Ito T, Garcia H E, Wang Z, et al. Assessing the observational uncertainties of dissolved oxygen climatology and seasonal cycle through a coordinated intercomparison project[J]. Global Biogeochemical Cycles, 2025, 39(11): e2025GB008751.

Response 1:
We thank the reviewer for this suggestion. In response, we added an intercomparison between our DO reconstruction and two recent products, GOBAI (Sharp et al., 2023) and ITO (Ito et al., 2024), using the filtered GLODAPv2 dataset (Olsen et al., 2016) as an independent validation benchmark. The results show that, at the global scale, our reconstruction achieves the lowest MAE and RMSE and the highest R^2 among the three products, indicating the best overall agreement with independent observations (Table R1). Within the GOBAI coverage, our reconstruction has a lower RMSE and higher R^2 than GOBAI, while its MAE is slightly higher and its mean difference is closer to zero. Within the ITO coverage, our reconstruction has lower MAE and RMSE and higher R^2 than ITO, while its mean difference is farther from zero. These results support the reliability of our product.

Table R1. Performance comparison on the filtered GLODAPv2

Product	MAE	RMSE	R^2	ΔDO
Our reconstruction	10.204	18.139	0.968	-0.334
GOBAI on filtered GLODAPv2	**11.101**	19.875	0.956	-0.971
Our reconstruction in GOBAI coverage	11.115	**19.658**	**0.963**	**-0.470**
ITO on filtered GLODAPv2	13.415	22.958	0.951	**-0.123**
Our reconstruction in ITO coverage	**11.824**	**19.966**	**0.964**	-0.544

We also added a profile-based comparison. Figure R1a shows the global mean vertical profiles of dissolved oxygen from our reconstruction, ITO Oxygen (Ito et al., 2024), and WOA23 climatology (Garcia et al., 2024) over 1965 – 2022. Near the surface, the three profiles are very close. In the 800-1000 m depth range, our reconstruction is close to WOA23, while ITO Oxygen is lower over part of this depth range. Below 1000 m, ITO Oxygen does not provide data, so the comparison is limited to our reconstruction and WOA23. Their profiles remain close through the deep ocean, indicating that our product gives a reasonable extension of dissolved oxygen fields below the depth range covered by ITO Oxygen. Figure R1b shows the global mean vertical profiles of our reconstruction, ITO Oxygen, and GOBAI over 2004 – 2020 for the upper 2000 m. The three products show a similar overall vertical structure, with the largest differences appearing in the 500-1000 m depth range. In this depth range, our reconstruction is generally higher than both ITO Oxygen and GOBAI, while the three profiles are closer near the surface. This comparison shows that our product reproduces the large-scale vertical pattern seen in existing datasets, while also showing differences in intermediate waters.

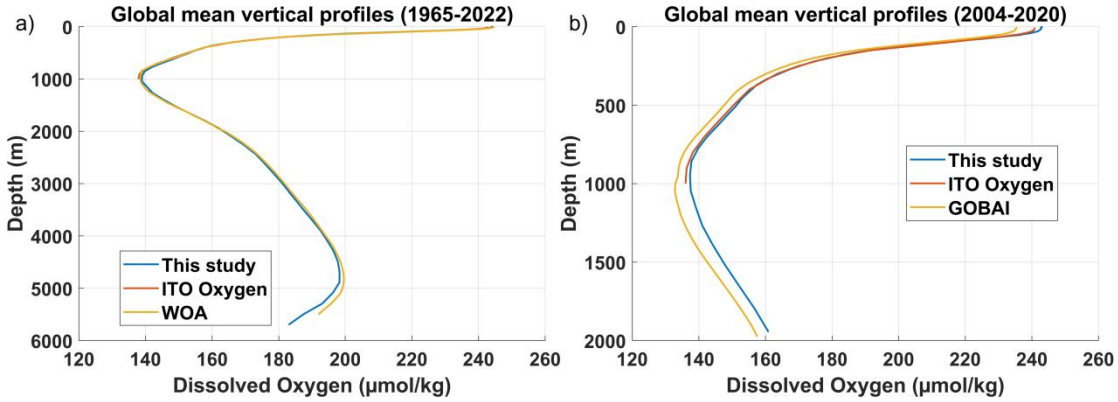

Figure R1. Global mean vertical profiles of dissolved oxygen from different products. (a) Profiles for this study, ITO Oxygen, and WOA23 over 1965 – 2022, shown from the surface to 5902 m. (b) Profiles for this study, ITO Oxygen, and GOBAI over 2004 – 2020, shown from the surface to 2000 m.

We further added a spatial comparison with the WOA23 climatology at several representative depths (Figure R2). Our reconstruction is closer to WOA23, particularly in the surface layer, and shows smaller differences in many low- and mid-latitude regions. At the surface layer around 10 m depth, our reconstruction shows small differences, generally within ±2 μmol kg^{-1}, except in some high-latitude regions. In comparison, ITO Oxygen exhibits broader regions of red, corresponding to negative differences of about 4–8 μmol kg^{-1} in the subtropical gyres, and more pronounced blue regions, corresponding to positive differences of about 6–10 μmol kg^{-1} under the Antarctic Circumpolar Current. At 30 m, the differences in our reconstruction remain small in the mid-latitude regions, with larger variability near boundary currents. In contrast, ITO Oxygen again shows larger negative differences in the subtropics and positive differences in the southern high latitudes. These results indicate that our reconstruction is generally closer to WOA23 in the surface ocean. At around 200 m, both our reconstruction and ITO Oxygen show larger departures from the WOA23 reference, reaching about ±10 μmol kg^{-1} in the tropical and subtropical regions. At around 700 m, our reconstruction and WOA23 remain within about ±8 μmol kg^{-1} over large parts of the Atlantic and Pacific basins, indicating good agreement at mid-depths. These spatial maps complement the statistical comparisons by showing that our product remains close to a widely used climatological reference across multiple depth levels.

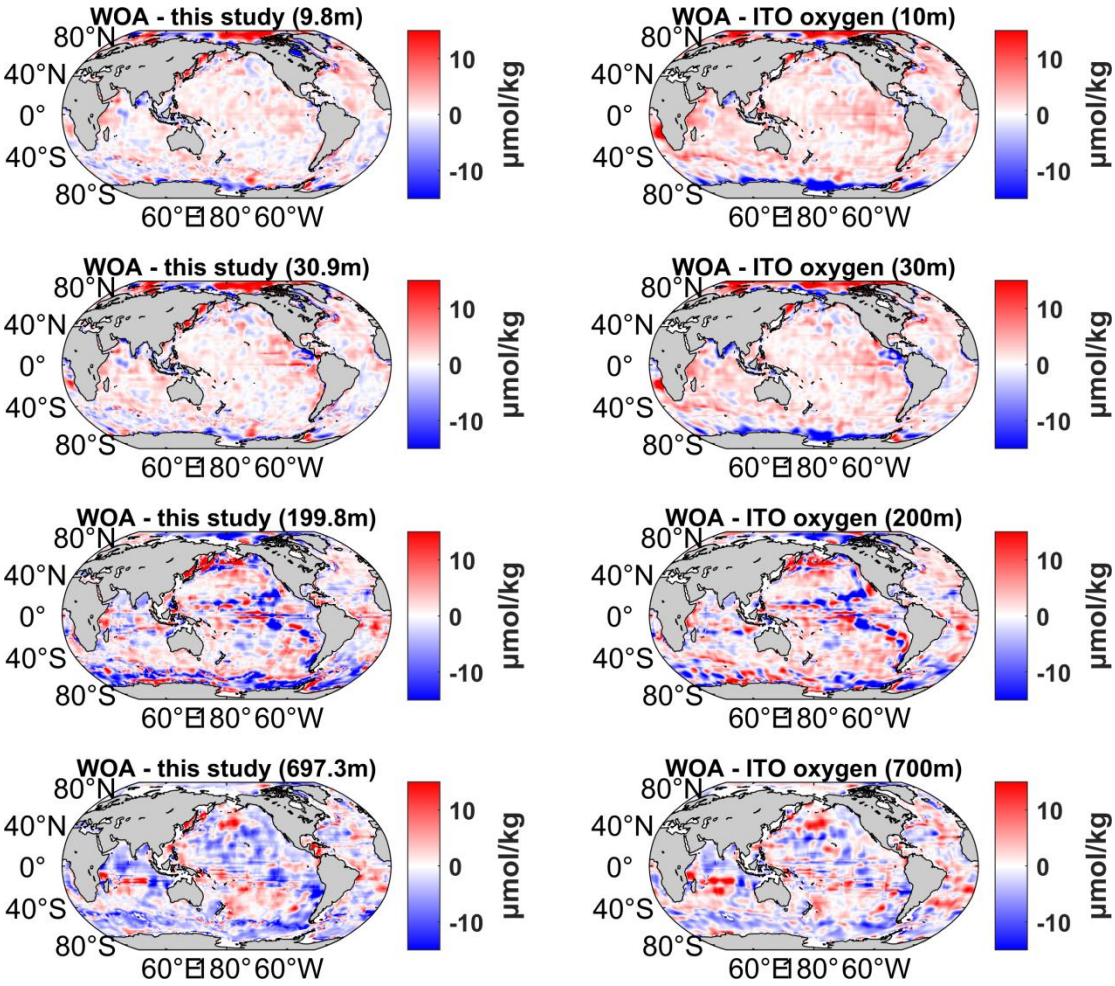

Figure R2. Spatial differences from WOA23 at four representative depths for this study and ITO Oxygen. Left panels show WOA23 minus this study at 9.8, 30.9, 199.8, and 697.3 m. Right panels show WOA23 minus ITO Oxygen at 10, 30, 200, and 700 m. Units are μmol kg^{-1}.

Relevant changes made in the manuscript: We added a new Section 4, “Intercomparison with existing oxygen products” (lines 411 – 483 in the revised manuscript), which includes the statistical comparison against the filtered GLODAPv2 dataset, the global mean vertical profile comparison, spatial comparisons with WOA23, and the oxygen content anomaly comparison.

Comment 2:

2. Spatial-Temporal Representativeness of the Validation Set

The authors state that for each profile in GLODAPv2, they searched the CTD and OSD records for matches within ±1° and the same month, excluding those that matched. This filtered the dataset down to 8,020 profiles. However, the manuscript lacks a spatial-temporal distribution map of this filtered validation set. It is critical to prove the coverage and representativeness of these remaining 8,020 profiles. Without a distribution map showing the cruise tracks or sampling locations, readers cannot determine whether the validation set represents a global oceanic assessment or if it is merely biased toward a few localized, data-rich sub-regions. Please provide maps and temporal histograms of the validation set.

Response 2:

We thank the reviewer for this suggestion. To clarify the coverage and representativeness of the filtered GLODAPv2 validation set, we have added a new figure showing both its spatial distribution and temporal histograms (Figure R3). The filtered dataset contains 8,020 unique profiles and spans all major ocean basins, indicating that the validation is not limited to a small number of localized regions. The yearly histogram shows that the profiles are distributed across multiple decades, although sampling becomes denser in the more recent period, while the monthly histogram shows coverage throughout the year with a clear peak in July.

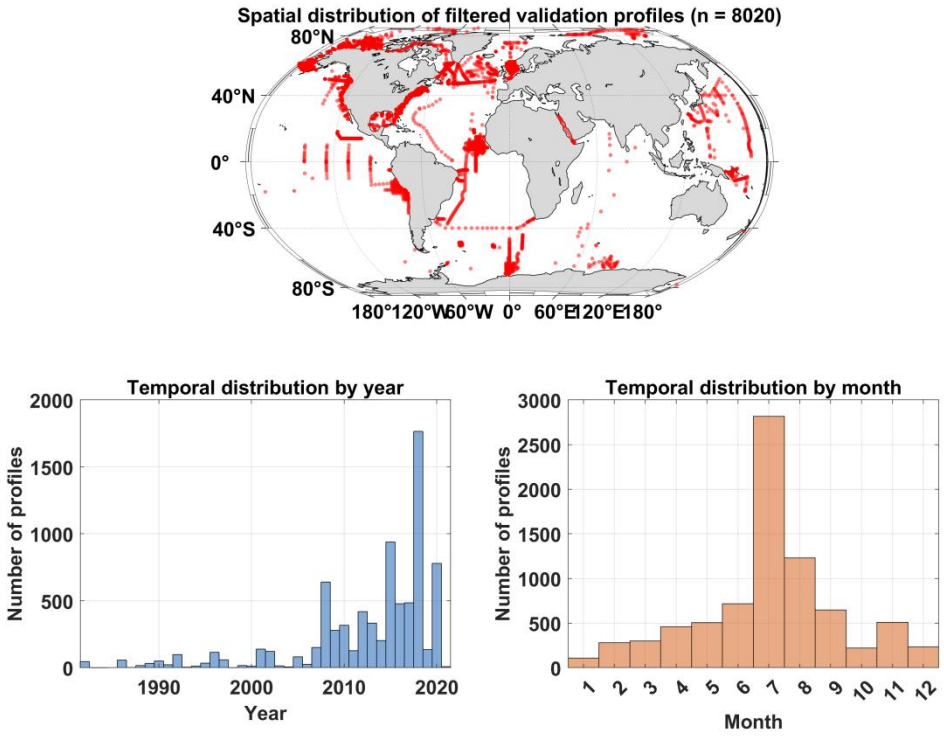

Figure R3. Spatial and temporal distribution of the filtered GLODAPv2 validation set. The upper panel shows the locations of the 8,020 unique validation profiles after excluding profiles matched with the CTD and OSD records used in model training. The lower panels show the temporal distribution of these profiles by year and by month.

Relevant changes made in the manuscript: In lines 128 – 129 of the revised manuscript, we added the sentence "The spatial distribution and temporal histograms of the filtered GLODAPv2 dataset are shown in Figure S2." We also added Figure S2 to the Supplementary Information, titled "Spatial and temporal distribution of the filtered GLODAPv2 validation set."

Comment 3:

3. Potential Spatial Discontinuity in the Weight Allocation Strategy

While the dynamic weighting strategy is conceptually interesting, its current mathematical formulation may benefit from further justification. The transition mechanism between dynamic and static weights could potentially lead to spatial discontinuities. For instance, suppose grid cell A contains an observation, and the adjacent grid cell B does not. In cell A, the dynamic weight might heavily favor a specific model that perfectly fits the local observation; however, in cell B, the weight instantaneously reverts to the global average static weights (w_i) of the 6 models. This abrupt transition ("hard switch") between observed and unobserved regions might produce artificial gradients or step-changes at the boundaries, which may not fully align with the continuous nature of oceanographic variables. I recommend the authors discuss this potential limitation to ensure physical continuity.

Response 3:

Thank you for this comment. We carefully re-examined our original dynamic weighting strategy in light of your concern. Although we did not observe obvious spatial discontinuities in the final reconstructed dissolved oxygen field, we agree that the reviewer's concern is theoretically well founded. In the original formulation, the model weight at an observed grid cell was determined by the collocated local error, whereas at a neighboring unobserved grid cell the weight reverted directly to the global prior weight. This observation-dependent switching could indeed introduce an abrupt transition in the weight field. To address this issue, we revised the dynamic weighting scheme by introducing a spatially smoothed weighting framework that preserves the original local error-based weighting at observation-supported grid cells while allowing the influence of these locally constrained weights to extend continuously into neighboring regions. The revised formulation remains conceptually consistent with ensemble weighting based on model skill and with locally calibrated weighting strategies that use nearby observations to inform local adjustment (Raftery et al., 2005; Kleiber et al., 2011; Brunsdon et al., 1996).

First, the global prior weight of model i is still defined from its time-cross-validation RMSE ε_i as

$$\omega_i = \frac{\exp(-\beta\varepsilon_i)}{\sum_{j=1}^{M}\exp(-\beta\varepsilon_j)},$$

where M=6 is the number of base models and β is the prior-weight sensitivity parameter. At grid

cells with valid observations, we retain the original local error-based weighting by first defining the local score of model i as

$$\mathrm{s}_i(x) = \exp\left[-\alpha\left|p_i(x) - O(x)\right|\right],$$

where $p_i(x)$ is the prediction of model i, O(x) is the observation, and α controls the sensitivity of the local weighting to model error. These scores are then normalized across all M models to obtain the effective local weight at observation-supported grid cells:

$$l_i^{obs}(x) = \frac{s_i(x)}{\sum_{j=1}^{M} s_j(x)}.$$

To introduce spatial continuity, we then smooth these effective local weights using neighboring observation-supported grid cells weighted by a Gaussian kernel, following the general idea of kernel-weighted local estimation for spatially varying relationships (Brunsdon et al., 1996):

$$K_{\mathrm{h}}(x, x_n) = \exp\left(-\frac{d_{xy}(x, x_n)^2}{2\sigma_{xy}^2} - \frac{d_z(x, x_n)^2}{2\sigma_z^2}\right),$$

$$\widetilde{l}_i(x) = \frac{\sum_{x_n \in N(x)} K_h(x, x_n) l_i^{obs}(x_n)}{\sum_{x_n \in N(x)} K_h(x, x_n)},$$

where x is the target grid cell, x_n denotes neighboring observation-supported grid cells and N(x) is the set of neighboring locations with valid effective local weights. In this way, the revised method retains the original locally constrained weighting at observed locations while producing a spatially continuous extension of these weights into neighboring areas.

To avoid another abrupt transition between observation-rich and observation-sparse regions, we further define an observation-support factor

$$S(x) = \sum_{x_n \in N(x)} K_h(x, x_n),$$

$$\rho(x) = \frac{S(x)}{S(x) + c},$$

where c is a shrinkage parameter. The final model weight is then written as

$$w_i(x) = \rho(x)\widetilde{l}_i(x) + \left[1 - \rho(x)\right]\omega_i,$$

Finally, the ensemble reconstruction is calculated as

$$\hat{O}(x) = \sum_{i=1}^{M} w_i(x) p_i(x).$$

Under this revised formulation, the influence of observations no longer changes instantaneously between observed and unobserved grid cells. Instead, the effective local weights derived at observation-supported grid cells are propagated continuously through the kernel function, and these smoothed local weights are gradually shrunk toward the global prior weights as observational support decreases. Therefore, the revised method preserves the strengths of the original framework, namely local observational constraint and global prior stability, while improving the spatial smoothness of the weight field.

Because RF generally receives relatively higher weights than the other five models in observation-supported regions, its spatial weight pattern provides a clearer illustration of the transition from the discontinuous structure in the original method to the more continuous structure in the revised method. As shown in Figure R4, we compared the global mean surface weight distribution of the RF model at 0.5 m obtained from the original and revised weighting strategies. The original method exhibits uneven spatial patterns and more abrupt transitions in the RF weight field, particularly in regions where observational coverage changes rapidly, which is consistent with the reviewer's concern regarding potential discontinuities induced by the hard switch. In contrast, the revised method produces a much smoother and more continuous global weight distribution. The sharp transitions present in the original formulation are reduced, and the influence of observation-rich regions extends more gradually into neighboring areas. This comparison shows that the revised weighting framework reduces the potential spatial discontinuity in the original method. We have revised the manuscript accordingly by replacing the previous formulation with the new smoothed dynamic weighting scheme.

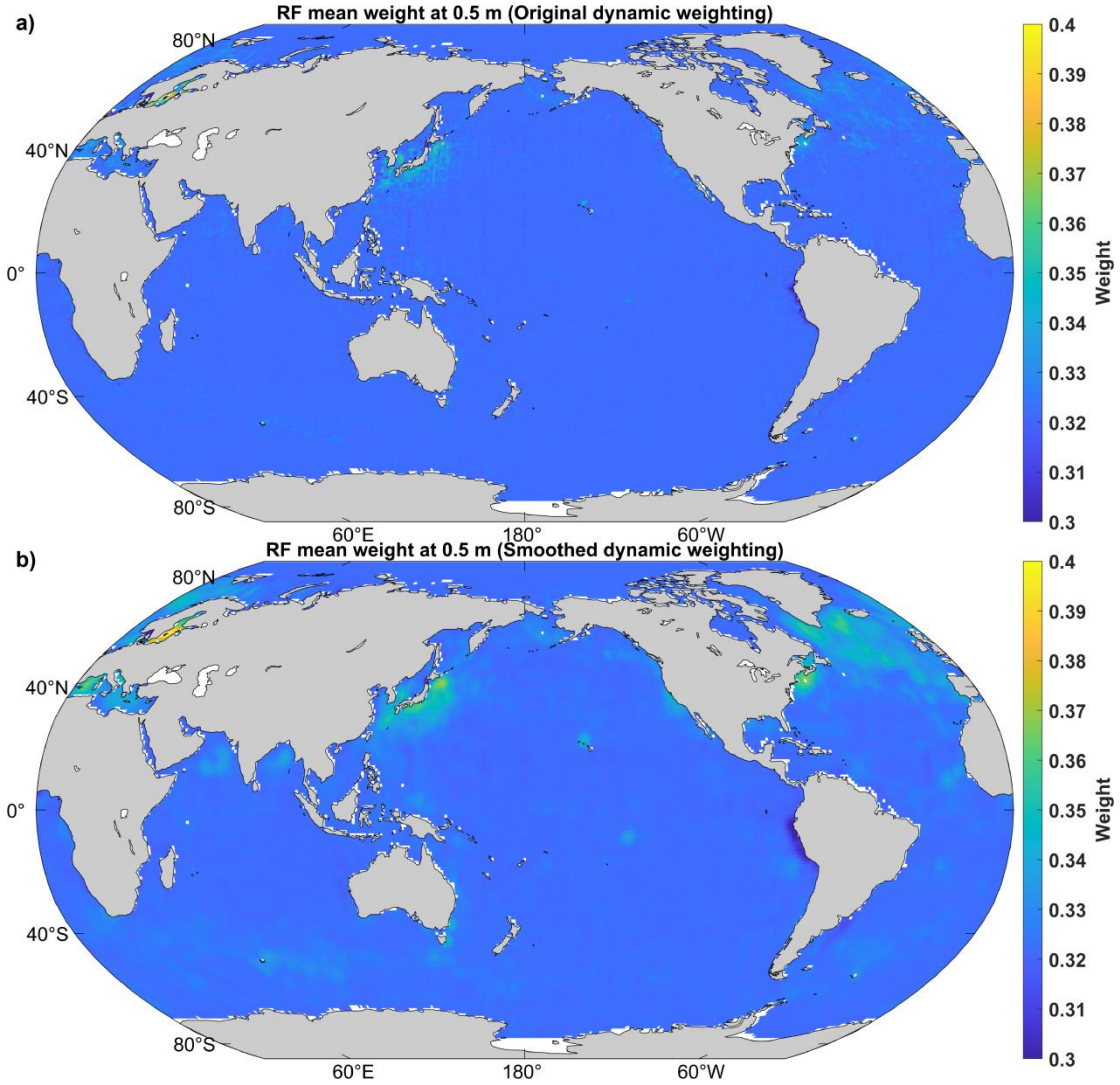

Figure R4. Global mean RF weight distribution at 0.5 m under the (a) original weighting scheme and (b) revised smoothed dynamic weighting scheme.

Although the weighting strategy was revised, the overall dissolved oxygen fields remain essentially unchanged. As shown in Figure R5, the global mean dissolved oxygen distributions

produced by the original and revised methods are nearly identical at representative depths from the surface to the deep ocean (0.5, 199.8, 856.7, and 4093.2 m). No obvious basin-scale shifts are introduced by the revised method. This indicates that the new smoothed dynamic weighting mainly improves the spatial continuity of the weighting field, while maintaining the original reconstruction results and their large-scale oceanographic structure.

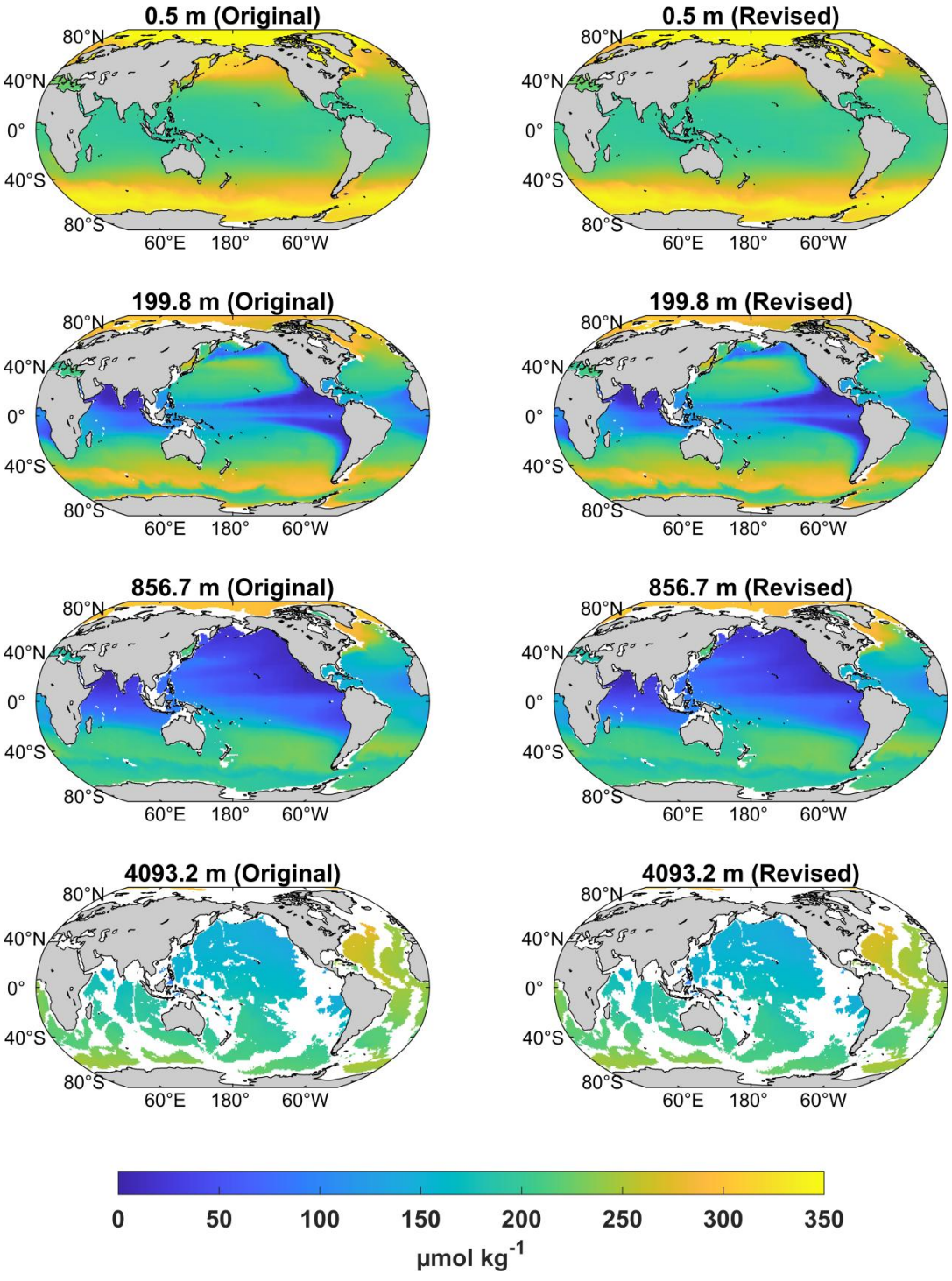

Figure R5. Global distributions of time-mean dissolved oxygen concentration at four representative depths, 0.5, 199.8, 856.7, and 4093.2 m, for the original reconstruction (left column) and the revised reconstruction (right column). Values were averaged over all months and years in the study period.

Relevant changes made in the manuscript: We replaced the entire Section 2.2.4, “Multi-model fusion and dynamic weighting scheme” (lines 257 – 304 in the revised manuscript), with the new spatially coherent dynamic weighting scheme.

Comment 4:

4. Insufficient Assessment of Deep-Ocean Accuracy

A major selling point of this dataset is its extension to 5,902 m depth. However, the manuscript lacks a rigorous, depth-specific accuracy assessment for the deep ocean. Validating the entire water column collectively obscures potential biases in the bathypelagic zone. To substantiate the claims regarding deep-ocean reconstruction, I suggest conducting a direct comparison of your deep-ocean results with the DIVA-based dataset by Roach and Bindoff (2023). This comparison will help verify if your machine-learning ensemble correctly captures the subtle deep-water mass structures compared to variational analysis methods.

Reference: Roach, C. J., & Bindoff, N. L. (2023). Developing a New Oxygen Atlas of the World's Oceans Using Data Interpolating Variational Analysis. Journal of Atmospheric and Oceanic Technology.

Response 4:

We thank the reviewer for this suggestion. We agree that, because our product extends to 5902 m, a depth-specific assessment is needed to evaluate its performance in the deep ocean rather than relying only on whole-water-column statistics. To address this point, we added a direct comparison with the DIVA-based dataset of Roach and Bindoff (2023), which extends to 6800 m and provides an independent reference for deep-ocean structure. Figures R6 and R7 show the 1965 – 2022 mean vertical profiles from our reconstruction, ITO Oxygen, WOA23, and Roach and Bindoff (RB), with Figure R6 focusing on the upper 1000 m and Figure R7 extending the comparison to 6000 m. In the upper 50 m, our reconstruction and RB show very similar profiles, and both are close to WOA23. Between about 100 and 400 m, RB is slightly closer to WOA23, whereas from about 400 to 1000 m our profile approaches WOA23 and remains close to both reference products. From 1000 to 3500 m, the profiles of our reconstruction, RB, and WOA23 are nearly indistinguishable, indicating very similar large-scale deep-ocean structure in this depth range. Below 3500 m, the RB profile remains slightly closer to WOA23, but the differences among the profiles are still small.

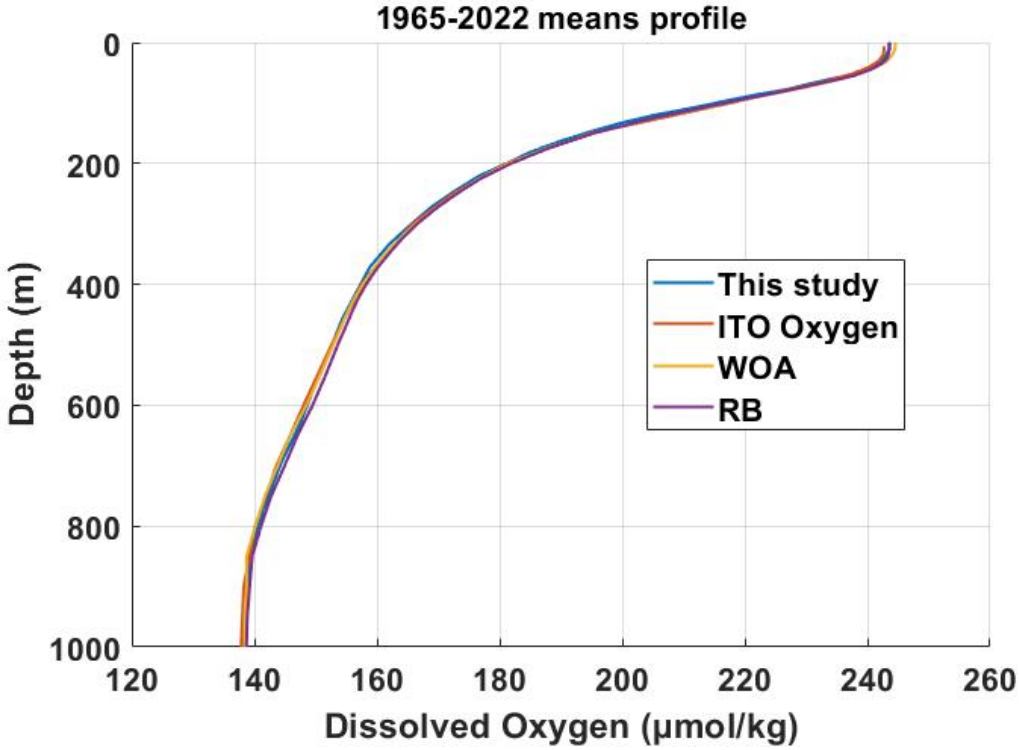

Figure R6. Global mean vertical profiles of different dissolved oxygen products (1965 – 2022). Solid lines show our reconstruction (blue), Roach & Bindoff's reconstruction (purple), ITO Oxygen (orange) and WOA23 climatology (yellow), plotted from the surface down to 1000 m.

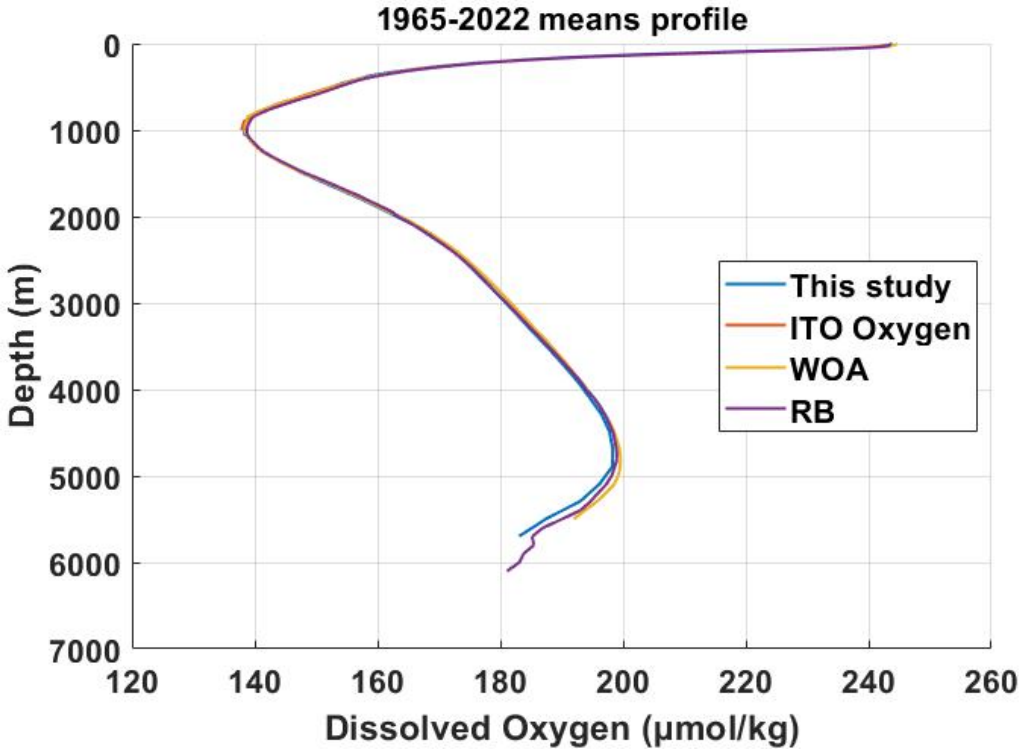

Figure R7. Global mean vertical profiles of different dissolved oxygen products (1965 – 2022). Solid lines show our reconstruction (blue), Roach & Bindoff's reconstruction (purple), ITO Oxygen (orange) and WOA23 climatology (yellow), plotted from the surface down to 6000 m.

We further added a spatial comparison with WOA23 at four representative depths to assess the deep-ocean performance of our reconstruction against the Roach and Bindoff product (Figure R8). At the surface layer around 10 m, our reconstruction is generally closer to WOA23 than RB, with smaller spatial differences over much of the open ocean. At around 700 m, both our reconstruction and RB show relatively large differences from WOA23. At around 1500 m, the differences decrease in both products, and at around 4000 m they decrease further, with both products showing generally smaller differences relative to WOA23. Overall, this comparison indicates that our product agrees more closely with WOA23 in the surface layer, while in the deep ocean both our reconstruction and RB show reduced differences below 1500 m.

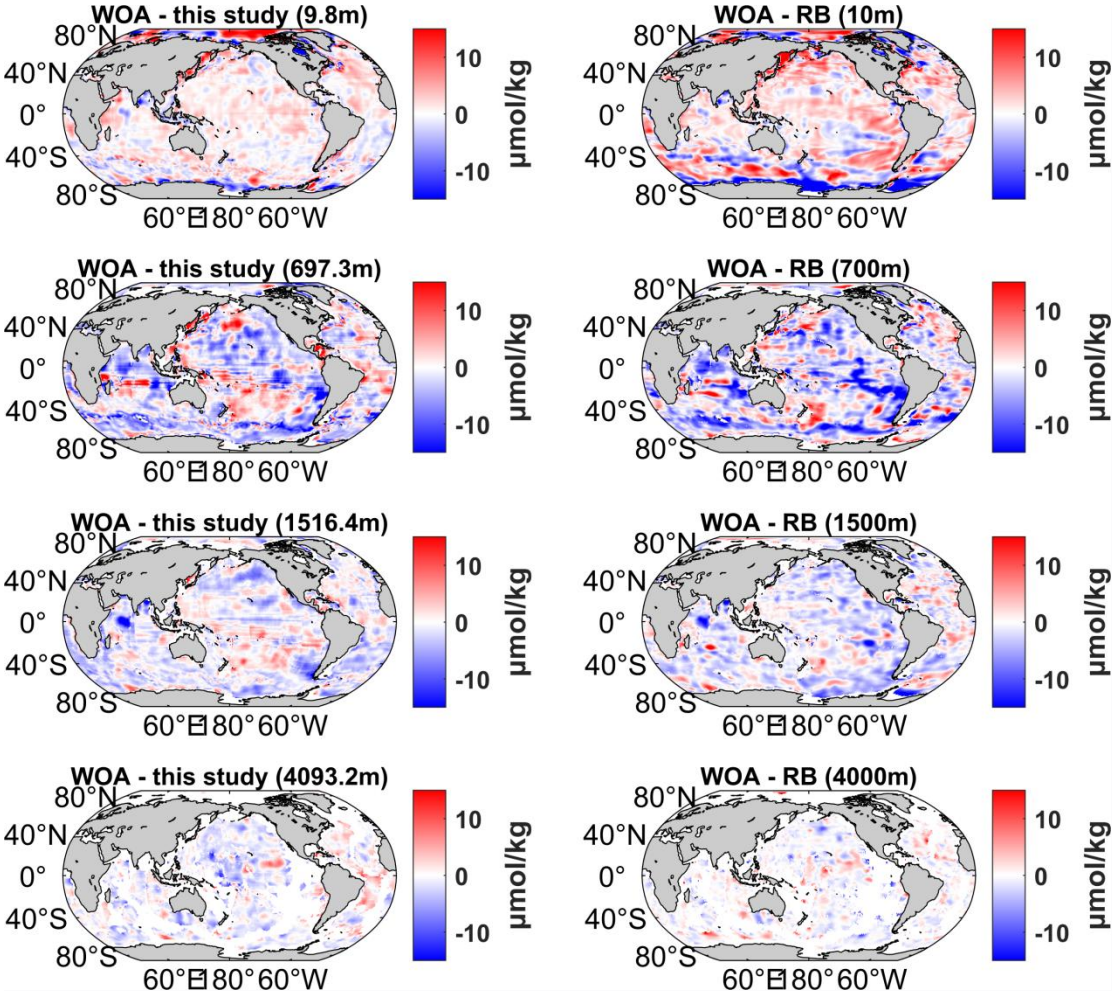

Figure R8. Spatial differences from WOA23 at four representative depths for this study and the Roach and Bindoff (RB) product. Left panels show WOA23 minus this study at 9.8, 697.3, 1516.4, and 4093.2 m. Right panels show WOA23 minus RB at 10, 700, 1500, and 4000 m. Units are µmol kg^{-1}.

Relevant changes made in the manuscript: Together with the revisions described in Response 1, we added these analyses to the new Section 4, "Intercomparison with existing oxygen products," in the revised manuscript. Specifically, the comparisons with the Roach and Bindoff (RB) product were added in Sections 4.2 and 4.3 (lines 424 – 470 in the revised manuscript).

Other comments:

Lines 34-35: The word "biogeochemical" is used multiple times within a single or adjacent sentence. Please rephrase to avoid redundancy and improve the flow of the text.

Line 87: Typographical error. There appears to be an extraneous space after the degree symbol (°). Please remove the space for proper formatting.

Lines 125-126: Typo in the dataset name. The text reads "...use this filtered GLODA v2...". Please correct this misspelling to "GLODAPv2".

Line 306 (Equation 11): The equation for is written as . Without parentheses, this mathematically implies that is subtracted after the summation of is complete. Please add parentheses to ensure mathematical correctness: .

Lines 456-457: The sentence states, "However, the expansion rates increased again beyond 1,600 m because of expansion in the NP". Using the word "expansion" twice in such close proximity makes the sentence clunky.

Lines 624-625 (Data Availability): While the authors provided URL links for the Argo and WOD databases, the reference link or DOI for the GLODAPv2 dataset is missing. Please add the official link or DOI for GLODAPv2 to maintain consistency and adhere to ESSD's data accessibility standards.

Response 5:

Thank you for the technical corrections. We have implemented every suggested change and double-checked the manuscript. Your detailed remarks have improved the clarity and accuracy of the manuscript.

Response to Referee #2

Comment 1:

Novelty and Data Sparsity in the Deep Ocean: The authors claim that while previous studies have focused on specific regions, temporal/spatial resolutions, or time spans, it remains challenging to simultaneously address all aspects. However, the core methodological approach appears highly derivative of recent works (e.g., Ito et al., 2024), with the primary novelty being the extension down to 5,902 m (To my knowledge, Ito et al., 2024 is a monthly product that spans a similar time period than what the paper is proposing). The authors explicitly state that historical DO measurements below 2,000 m remain sparse. Given this sparsity, extending the reconstruction to ~6,000 m requires rigorous justification. How are the authors confident that the reconstruction beyond 2000m is good ? Given that there are only few training data used to calibrate the model bellow 2000m. The manuscript must include a quantitative analysis (e.g., density plots or a table) of the number of available profiles per region below 2,000 m to prove that the machine learning algorithms are actually learning from sufficient data rather than blindly extrapolating based on upper-ocean trends.

Response 1:

We thank the reviewer for this suggestion. We agree that the observational support below 2,000 m should be quantified explicitly rather than described only qualitatively. We therefore added a new quantitative summary of the number of available dissolved oxygen profiles below 2,000 m for each basin (Table R1). To better resolve the depth dependence of data availability, we further divided the deep ocean into four layers: 2000 – 3000 m, 3000 – 4000 m, 4000 – 5000 m, and 5000 – 5902 m. The results show that deep-ocean sampling is uneven among basins, but it is not absent. The number of available profiles below 2,000 m is highest in the North Atlantic and North Pacific, followed by the Southern Ocean and Arctic Ocean, while the other basins still contain several thousand profiles each. In the deepest layer (5000 – 5902 m), the number of available profiles decreases substantially, as expected, but remains non-zero in most basins. These additions provide the quantitative basin-by-basin assessment requested by the reviewer and clarify that the observational constraint below 2,000 m is region dependent and becomes weaker toward the deepest layers.

Table R1. Number of available dissolved oxygen profiles below 2,000 m for each basin

Region	2000-3000m	3000-4000m	4000-5000m	5000-5902m
North Pacific	22867	10469	7230	3352
Equatorial Pacific	5988	3443	2307	809
South Pacific	5349	3357	2117	669
Arctic Ocean	8673	2555	58	3
North Atlantic	34629	12012	5678	1722
Equatorial Atlantic	4179	2953	2000	533
South Atlantic	4192	3380	2054	674
Southern Ocean	10098	6172	3262	545
North Indian Ocean	3664	2053	1351	340
South Indian Ocean	5148	2981	2043	648

Relevant changes made in the manuscript: In lines 120 – 121 of the revised manuscript, we added the sentence "Although DO observations below 2,000 m become sparser with depth, they remain available across most basins (Table S1)." We also added Table S1 to the Supplementary Information, titled "Number of available dissolved oxygen profiles below 2,000 m for each basin."

Comment 2:

Choice of Reanalysis Data: The reconstruction relies on temperature, salinity, and velocity fields from the Ocean Reanalysis System 5 (ORAS5). The authors need to justify the selection of ORAS5 over other standard products like EN4 (used by Ito et al. 2024). Furthermore, because reanalysis products also suffer from high uncertainty in the deep ocean, the authors should discuss how the inherent uncertainties in ORAS5 deep-ocean variables propagate into the BLENDR DO reconstruction.

Response 2:

We thank the reviewer for this comment. We agree that the choice of environmental predictors should be justified. We selected ORAS5 rather than EN4 because our BLENDR framework requires a dynamically consistent set of predictors including not only temperature and salinity, but also zonal and meridional velocity fields. ORAS5 is a global ocean reanalysis produced within ECMWF's OCEAN5 system and provides the full set of physical predictors used in this study. By contrast, EN4 is an observation-based monthly subsurface temperature and salinity analysis product and does not provide the current fields required by our framework (Good et al., 2013). In addition, although EN4 has been widely used, previous studies have noted limitations of EN4 that are relevant for some long-term analyses. Chen and Tung (2024) showed that, in EN4-based AMOC proxy analyses, increasing observational coverage in sparsely observed early decades can produce artificial rises in variance. Cheng et al. (2020) pointed out that EN4 shows a large spurious upward shift in upper-ocean salinity during 2000 – 05 relative to their improved estimate. For this reason, besides the absence of current fields, we consider ORAS5 to be better suited to the present framework.

We also agree that uncertainties in ORAS5 deep-ocean variables can propagate into the BLENDR reconstruction and that this issue should be discussed. In our framework, ORAS5 temperature, salinity, and velocity fields are used as predictors, so biases or errors in these fields can influence the reconstructed dissolved oxygen through the learned nonlinear relationships between physical predictors and oxygen. This effect is expected to be more important in the deep ocean because direct oxygen observations become much sparser below 2000 m, meaning that the reconstruction relies more heavily on the large-scale structure provided by the physical predictors. Because all six component models in BLENDR use the same ORAS5 predictors, this source of uncertainty acts partly as a shared structural uncertainty and therefore cannot be fully removed by the ensemble weighting strategy. We revised the manuscript to acknowledge that our current uncertainty calculation does not explicitly propagate predictor uncertainty from ORAS5 and may therefore underestimate total uncertainty in poorly observed deep-ocean regions. We also note that ORAS5 itself includes an ensemble framework designed to sample uncertainty in forcing, observation locations, and initial ocean state, which provides a potential route for quantification of predictor-driven uncertainty propagation through the reconstruction.

Relevant changes made in the manuscript: We added a discussion of ORAS5-related predictor uncertainty in the "Conclusion and discussion" section of the revised manuscript, covering lines 742 – 756.

Comment 3:

Validation Robustness and Spatial Autocorrelation: To construct an independent validation subset from GLODAPv2, the authors applied a spatiotemporal filter to remove overlapping profiles within ±1° in latitude/longitude and the same calendar month. While this removes exact duplicates, it is an insufficient safeguard against data leakage. Oceanographic variables exhibit strong spatial autocorrelation; a ±1° radius is too narrow to ensure true independence. A more rigorous approach (e.g., removing profiles based on correlation scores or employing spatial block cross-validation) should be implemented.

Response 3:

We acknowledge that the original filtering criterion of ±1° and the same calendar month may not fully remove the effect of spatial autocorrelation. We performed an additional, much stricter sensitivity test. Specifically, we removed from the GLODAPv2 (Olsen et al., 2016) validation set all profiles located within ±5° in longitude and latitude of the training CTD/OSD records in the same year. Under this stricter criterion, the number of remaining validation profiles was reduced from 8,020 to 2,982.

We then re-evaluated the reconstruction using this more conservative validation set. The results remain very similar to those obtained with the original filtering criterion (Table R2). For the original ±1° and same-month filtering, the reconstruction achieved an MAE of 10.316 μmol kg^{-1}, an RMSE of 18.212 μmol kg^{-1}, an R^2 of 0.967, and a mean bias of -0.276 μmol kg^{-1}. Under the stricter ±5° and same-year filtering, the corresponding values were 9.719 μmol kg^{-1}, 18.545 μmol kg^{-1}, 0.968, and -0.788 μmol kg^{-1}, respectively. In other words, the skill does not show a substantial decline when the validation set is made much more conservative. This indicates that the main validation result is not simply an artifact of the original, weaker overlap criterion.

Table R2. Sensitivity of validation metrics to different exclusion criteria

	Remaining profiles	MAE	RMSE	R^2	ΔDO
±1°,same month	8020	10.204	18.139	0.968	-0.334
±5°,same year	2982	9.719	18.545	0.968	-0.788

Relevant changes made in the manuscript: In lines 359 – 361 of the revised manuscript, we added the sentence "A stricter sensitivity test using a wider spatiotemporal exclusion criterion yielded very similar validation metrics (Table S7), suggesting that the main evaluation results are not sensitive to the original filtering choice." We also added Table S7, "Sensitivity of BLENDR performance on the filtered GLODAPv2 validation set to different spatiotemporal exclusion criteria," to the Supplementary Information.

Comment 4:

Lack of Independent Baseline Comparisons: Validating solely against isolated GLODAPv2 profiles is inadequate for a global reconstruction product. To demonstrate true efficacy, the reconstructed fields climatologies, seasonality as well as deoxigination patterns and rates must be compared against established gridded climatologies (e.g., World Ocean Atlas, Ito et al.).

Response 4:

We agree that validation against filtered GLODAPv2 profiles alone is not sufficient for a global gridded reconstruction product. We have already carried out additional comparisons with existing gridded oxygen products and climatological references, and we incorporated these results into the revised manuscript.

Table R3. Performance comparison on the filtered GLODAPv2

Product	MAE	RMSE	R^2	ΔDO
Our reconstruction	10.204	18.139	0.968	-0.334
GOBAI on filtered GLODAPv2	**11.101**	19.875	0.956	-0.971
Our reconstruction in GOBAI coverage	11.115	**19.658**	**0.963**	**-0.470**
ITO on filtered GLODAPv2	13.415	22.958	0.951	**-0.123**
Our reconstruction in ITO coverage	**11.824**	**19.966**	**0.964**	-0.544

First, we compared our reconstruction with two recent machine-learning oxygen products, GOBAI (Sharp et al., 2023) and ITO (Ito et al., 2024), using the filtered GLODAPv2 dataset as an external benchmark (Table R3). The results show that, at the global scale, our reconstruction achieves the lowest MAE and RMSE and the highest R^2 among the three products, indicating the best overall agreement with external observations. Within the GOBAI coverage, our reconstruction has a lower RMSE and higher R^2 than GOBAI, while its MAE is slightly higher and its mean difference is closer to zero. Within the ITO coverage, our reconstruction has lower MAE and RMSE and higher R^2 than ITO, while its mean difference is farther from zero. These results support the reliability of our product.

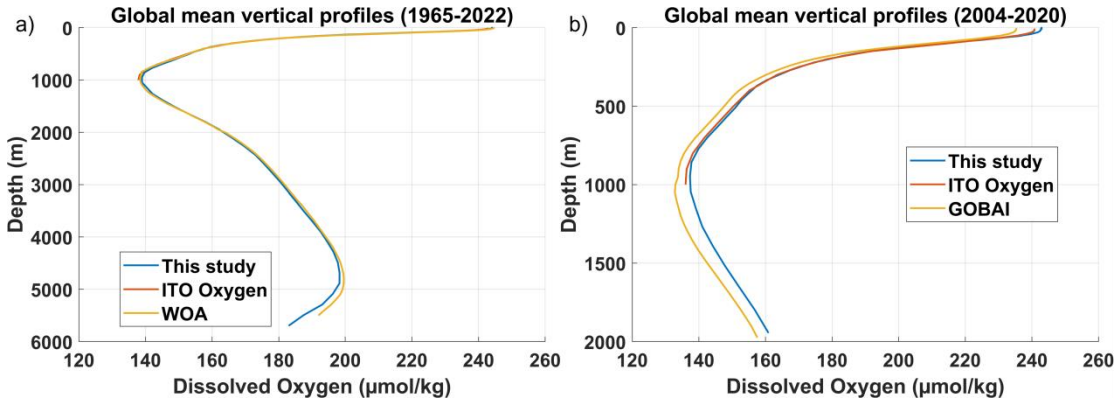

Figure R1. Global mean vertical profiles of dissolved oxygen from different products. (a) Profiles for this study, ITO Oxygen, and WOA23 over 1965 – 2022, shown from the surface to 5902 m. (b) Profiles for this study, ITO Oxygen, and GOBAI over 2004 – 2020, shown from the surface to 2000 m.

We also added profile-based comparisons with existing gridded products (Figure R1). In the global mean vertical profiles over 1965 – 2022, our reconstruction, ITO Oxygen, and WOA23 (Garcia et al., 2024) are very close near the surface. In the 800-1000 m depth range, our reconstruction is close to WOA23, while ITO Oxygen is lower over part of this depth range. Below 1000 m, ITO Oxygen does not provide data, so the comparison is limited to our reconstruction and WOA23. Their profiles remain close through the deep ocean, indicating that our product gives a reasonable extension of dissolved oxygen fields below the depth range covered by ITO Oxygen. Figure R1b shows the global mean vertical profiles of our reconstruction, ITO Oxygen, and GOBAI over 2004 – 2020 for the upper 2000 m. The three products show a similar overall vertical structure, with the largest differences appearing in the 500-1000 m depth range. In this depth range, our reconstruction is generally higher than both ITO Oxygen and GOBAI, while the three profiles are closer near the surface. This comparison shows that our product reproduces the large-scale vertical pattern seen in existing datasets, while also showing differences in intermediate waters.

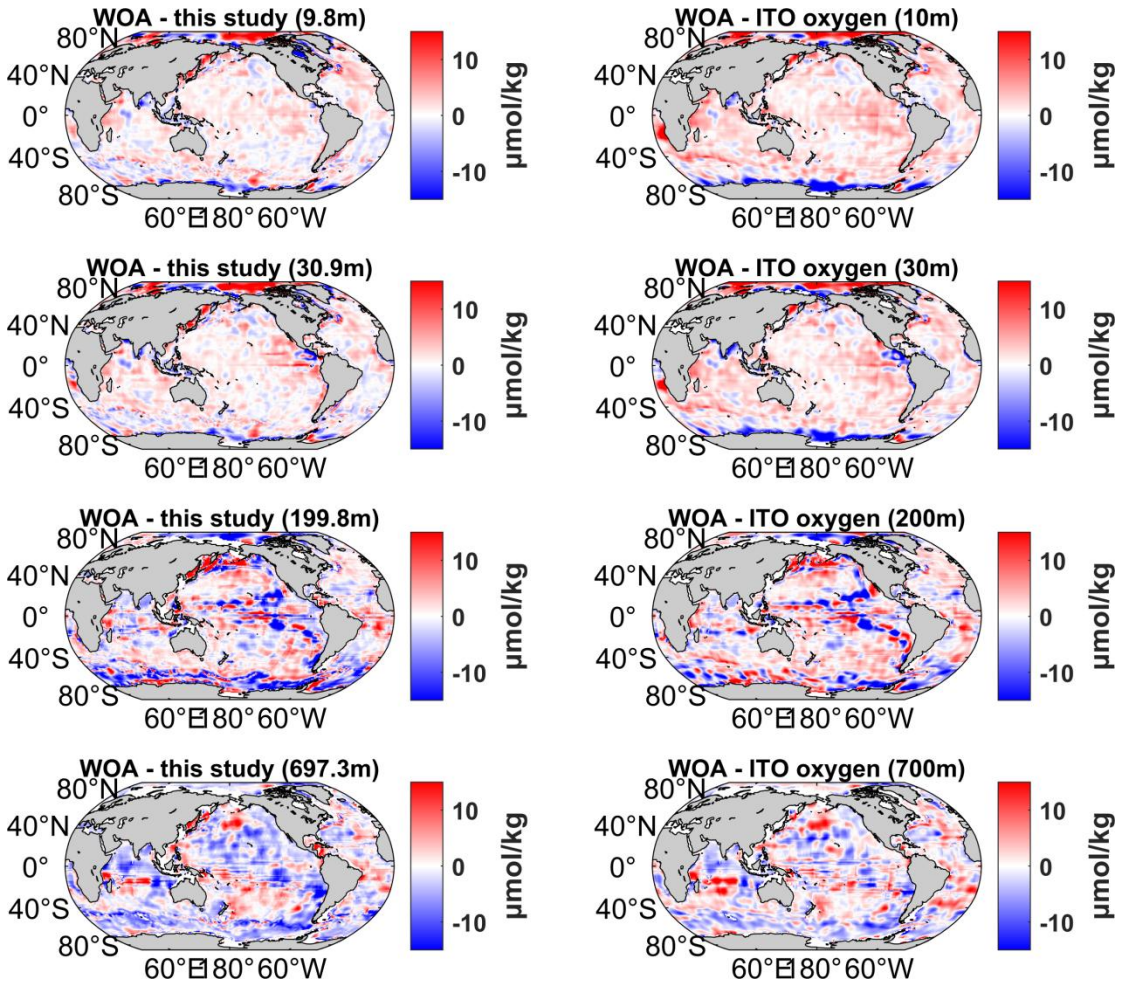

Figure R2. Spatial differences from WOA23 at four representative depths for this study and ITO Oxygen. Left panels show WOA23 minus this study at 9.8, 30.9, 199.8, and 697.3 m. Right panels show WOA23 minus ITO Oxygen at 10, 30, 200, and 700 m. Units are μmol kg^{-1}.

To provide a gridded comparison rather than only profile statistics, we further compared the spatial fields with WOA23 at several representative depths (Figure R2). Our reconstruction is closer to WOA23, particularly in the surface layer, and shows smaller differences in many low- and mid-latitude regions. At the surface layer around 10 m depth, our reconstruction shows small differences, generally within ±2 μmol kg^{-1}, except in some high-latitude regions. In comparison, ITO Oxygen exhibits broader regions of red, corresponding to negative differences of about 4 – 8 μmol kg^{-1} in the subtropical gyres, and more pronounced blue regions, corresponding to positive differences of about 6 – 10 μmol kg^{-1} under the Antarctic Circumpolar Current. At 30 m, the differences in our reconstruction remain small in the mid-latitude regions, with larger variability near boundary currents. In contrast, ITO Oxygen again shows larger negative differences in the subtropics and positive differences in the southern high latitudes. These results indicate that our reconstruction is generally closer to WOA23 in the surface ocean. At around 200 m, both our reconstruction and ITO Oxygen show larger departures from the WOA23 reference, reaching about ±10 μmol kg^{-1} in the tropical and subtropical regions. At around 700 m, our reconstruction and WOA23 remain within about ±8 μmol kg^{-1} over large parts of the Atlantic and Pacific basins, indicating good agreement at mid-depths. These

spatial maps complement the statistical comparisons by showing that our product remains close to a widely used climatological reference across multiple depth levels.

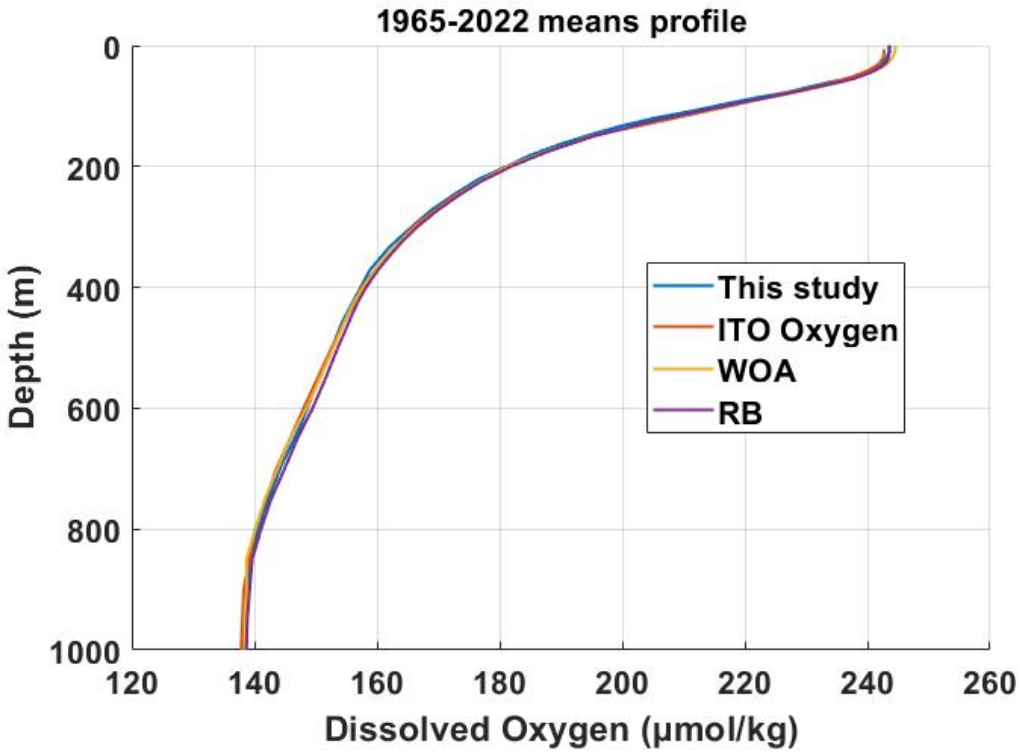

Figure R3. Global mean vertical profiles of different dissolved oxygen products (1965 – 2022). Solid lines show our reconstruction (blue), Roach & Bindoff's reconstruction (purple), ITO Oxygen (orange) and WOA23 climatology (yellow), plotted from the surface down to 1000 m.

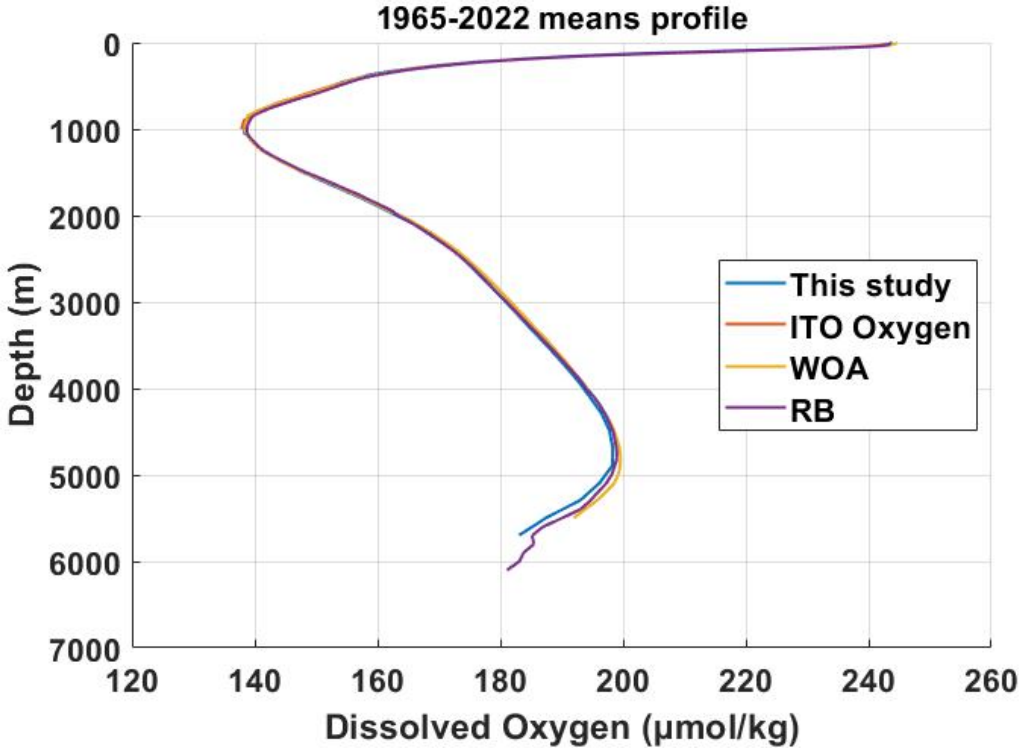

Figure R4. Global mean vertical profiles of different dissolved oxygen products (1965 – 2022). Solid lines show our reconstruction (blue), Roach & Bindoff's reconstruction (purple), ITO Oxygen (orange) and WOA23 climatology (yellow), plotted from the surface down to 6000 m.

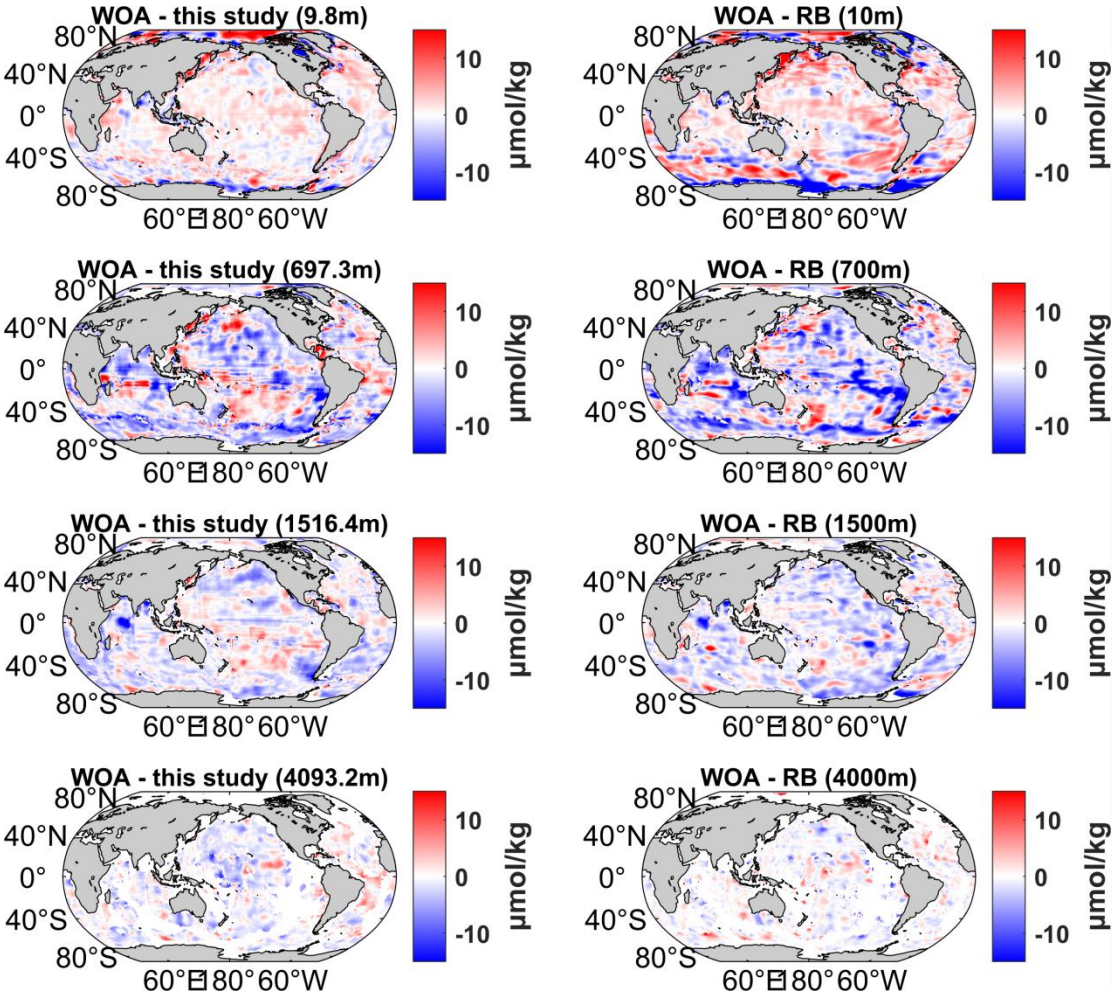

Figure R5. Spatial differences from WOA23 at four representative depths for this study and the Roach and Bindoff (RB) product. Left panels show WOA23 minus this study at 9.8, 697.3, 1516.4, and 4093.2 m. Right panels show WOA23 minus RB at 10, 700, 1500, and 4000 m. Units are μmol kg^{-1}.

Because the extension to 5902 m is a key feature of this dataset, we also carried out a depth-specific comparison with the DIVA-based product of Roach and Bindoff (2023). In the global mean vertical profiles, our reconstruction, Roach and Bindoff, and WOA23 are nearly indistinguishable from about 1000 to 3500 m, indicating very similar large-scale deep-ocean structure in this depth range. Below 3500 m, the Roach and Bindoff profile remains slightly closer to WOA23, but the differences are still small (Figure R3; Figure R4). We further compared the spatial differences from WOA23 at representative depths. At around 1500 m and 4000 m, the differences decrease in both products, indicating that both reconstructions remain close to WOA23 in the deep ocean (Figure R5).

Relevant changes made in the manuscript: We added a new Section 4, "Intercomparison with existing oxygen products" (lines 411 – 483 in the revised manuscript), which includes the

statistical comparison against the filtered GLODAPv2 dataset, the global mean vertical profile comparison, spatial comparisons with WOA23, and the oxygen content anomaly comparison.

Comment 5:

Benchmarking the BLENDR Framework: The authors justify their use of tree-based ensembles over neural networks by stating that preliminary NN trials did not yield consistent skill increases and were harder to tune. This is quite broad, particularly since lots of works (e.g. from Ito and Sharp, also from Ouala et al. 2026) used successfully neural networks for the exact same problem. The community can benefit from different methodological developments such as the ones used here, but they need clear justification. For instance, the reconstructed maps based on the BLENDER framework should be compared to the ones from a neural network model (with a similar architecture to the ones of the cited papers).

Response 5:

We agree that our original explanation for choosing tree-based models was too brief. In the manuscript, we only stated that preliminary neural-network tests did not lead to a clear improvement, but we did not show the result. We therefore provide the results from our initial model selection. At that stage, we tested Random Forest, a CNN, and a BP neural network. Random Forest performed much better than the other two models, with an R^2 of 0.97 and an RMSE of 12.57, compared with 0.81 and 36.01 for the CNN and 0.87 and 29.87 for the BP neural network. This is the main reason why we did not continue developing the reconstruction with these two neural-network models.

We also compared the final BLENDR product with recent machine-learning oxygen products that include neural-network-based approaches. As already shown in Response 4, our reconstruction gives lower RMSE and higher R^2 than GOBAI and ITO on the filtered GLODAPv2 benchmark, and it remains competitive in the corresponding coverage domains (Table R3). The added profile and map comparisons against GOBAI, ITO, WOA23, and the Roach and Bindoff product also show that our reconstruction reproduces the large-scale vertical and spatial structure reasonably well from the surface to the deep ocean (Figure R1-R5). At the same time, these analyses are not equivalent to a fully controlled comparison with a neural-network model trained on exactly the same inputs and train-test splits as BLENDR. We therefore limit our claim to the present model screening and the product-level comparisons shown above.

Relevant changes made in the manuscript: In lines 186 – 191 of the revised manuscript, we expanded the explanation of why neural-network models were not selected for the final reconstruction. The added text is: “In our initial model screening under the same input features and validation framework, the tested neural-network models, including a convolutional neural network and a back-propagation neural network, did not show an accuracy advantage over Random Forest, while requiring more tuning effort and providing less interpretability. Therefore, we focused the final ensemble on tree-based learners.”

Reference

Brunsdon C, Fotheringham A S, Charlton M E. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, **1996**, 28(4): 281-298.

Chen, X., & Tung, K. K. Evidence lacking for a pending collapse of the Atlantic Meridional Overturning Circulation. *Nature Climate Change*, **2024**, 14(1), 40-42.

Cheng, L., Trenberth, K. E., Gruber, N., et al. Improved estimates of changes in upper ocean salinity and the hydrological cycle. *Journal of Climate*, **2020**, 33(23), 10357-10381.

Garcia, Hernan E., et al. *World Ocean Atlas 2023, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, Dissolved Oxygen Saturation and 30-year Climate Normal.* **2024**.

Good, S. A., Martin, M. J., & Rayner, N. A. EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, **2013**, 118(12), 6704-6716.

Ito T, Cervania A, Cross K, et al. Mapping dissolved oxygen concentrations by combining shipboard and Argo observations using machine learning algorithms. *Journal of Geophysical Research: Machine Learning and Computation*, **2024**, 1(3): e2024JH000272.

Kleiber W, Raftery A E, Baars J, et al. Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Monthly Weather Review*, **2011**, 139(8): 2630-2649.

Olsen A, Key R M, Van Heuven S, et al. The Global Ocean Data Analysis Project version 2 (GLODAPv2)–an internally consistent data product for the world ocean. *Earth System Science Data*, **2016**, 8(2): 297-323.

Raftery A E, Gneiting T, Balabdaoui F, et al. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, **2005**, 133(5): 1155-1174.

Roach, C. J., Bindoff, N. L. Developing a new oxygen atlas of the world's oceans using data interpolating variational analysis. *Journal of Atmospheric and Oceanic Technology*, **2023**, 40(11): 1475-1491.

Sharp J D, Fassbender A J, Carter B R, et al. GOBAI-O2: temporally and spatially resolved fields of ocean interior dissolved oxygen over nearly 2 decades. *Earth System Science Data*, **2023**, 15, 4481–4518, https://doi.org/10.5194/essd-15-4481-2023.