Dear Referee,

We are very grateful for the referee's constructive comments. Many of the points raised are also issues that we have been considering since the submission of the manuscript. We have already examined several of these questions and prepared additional analyses. Below, we provide detailed responses to each comment. The major updates include: **(i) a quantitative summary of the number of available dissolved oxygen profiles below 2,000 m for each basin and for four deep-ocean layers (Table R1); (ii) an explanation of why ORAS5 was used rather than EN4, together with a discussion of how uncertainties in ORAS5 deep-ocean variables may affect the reconstruction; (iii) a sensitivity test using a stricter spatiotemporal exclusion criterion for the filtered GLODAPv2 validation set (Table R2); (iv) new comparisons of our reconstruction with GOBAI, ITO, WOA23, and the Roach and Bindoff product (Table R3; Figs. R1-R5); and (v) an explanation of the choice of tree-based models, together with the results from our initial model selection.** A point-by-point reply follows below (referee comments in *italics*, our responses in regular type).

On behalf of all authors, sincerely,

Mingyu Han
Shanghai Jiao Tong University

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. We will revise the manuscript accordingly to include this new table and to state this limitation more explicitly in the discussion.

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

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 will revise 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.

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	10.316	18.212	0.967	-0.276
±5°,same year	2982	9.719	18.545	0.968	-0.788

We will add this stricter sensitivity test to the revised manuscript and clarify that the filtered GLODAPv2 dataset should be regarded as an external validation benchmark with different levels of spatiotemporal exclusion, rather than as a perfectly independent dataset in a strict geostatistical sense.

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 will incorporate these results into the revised manuscript.

Table R3. Performance comparison on the filtered GLODAPv2

Product	MAE	RMSE	R^2	ΔDO
Our reconstruction	10.316	18.212	0.967	-0.276
GOBAI on filtered GLODAPv2	**11.101**	19.875	0.956	-0.971
Our reconstruction in GOBAI coverage	11.236	**19.731**	**0.963**	**-0.399**
ITO on filtered GLODAPv2	13.415	22.958	0.951	**-0.123**
Our reconstruction in ITO coverage	**11.937**	**20.045**	**0.964**	-0.485

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.

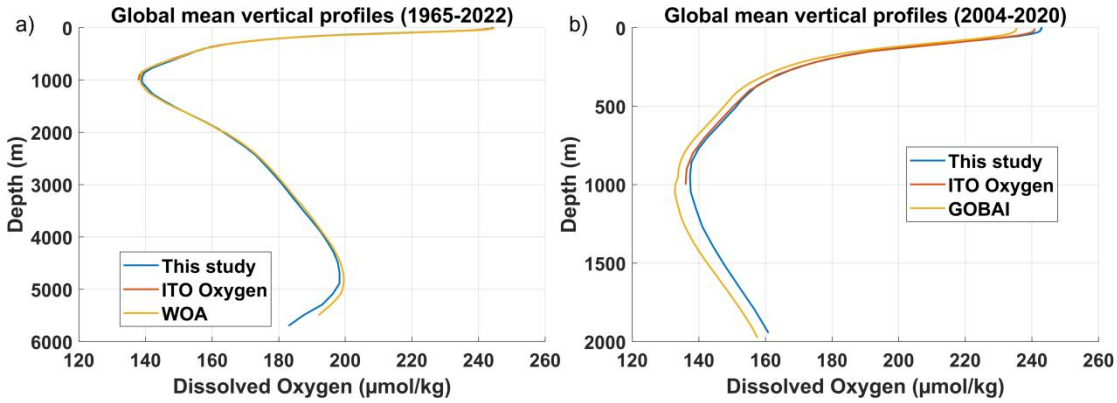

Figure R1. Global mean vertical profiles of dissolved oxygen from different products. (a) Profiles for this study, ITO Oxygen, and WOA23 over 1965 – 2022, shown from the surface to 5902 m. (b) Profiles for this study, ITO Oxygen, and GOBAI over 2004 – 2020, shown from the surface to 2000 m.

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.

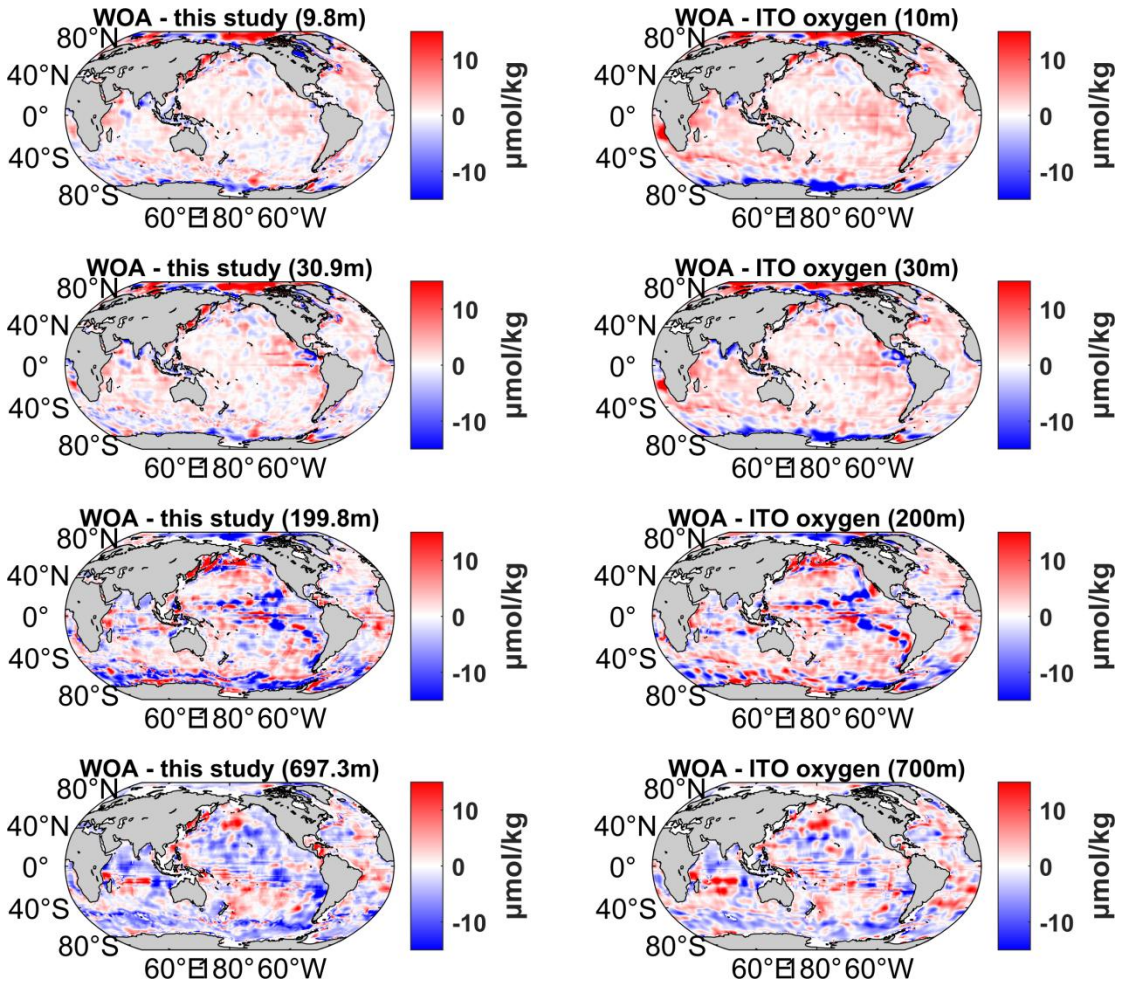

Figure R2. Spatial differences from WOA23 at four representative depths for this study and ITO Oxygen. Left panels show WOA23 minus this study at 9.8, 30.9, 199.8, and 697.3 m. Right panels show WOA23 minus ITO Oxygen at 10, 30, 200, and 700 m. Units are μmol kg^{-1}.

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.

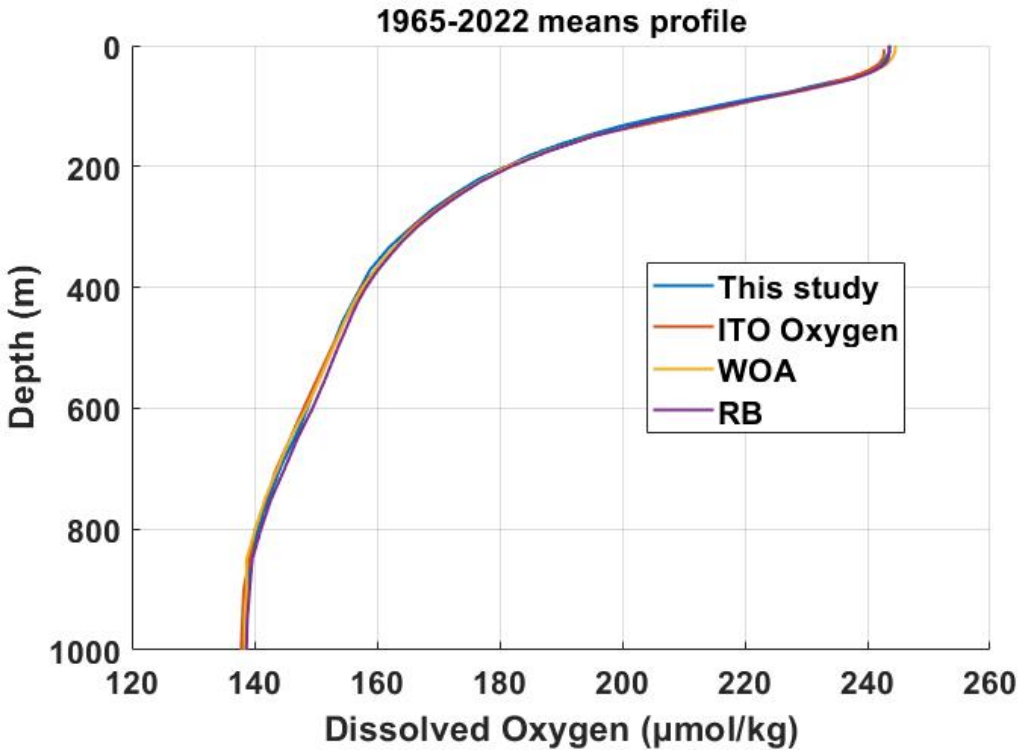

Figure R3. Global mean vertical profiles of different dissolved oxygen products (1965 – 2022). Solid lines show our reconstruction (blue), Roach & Bindoff's reconstruction (purple), ITO Oxygen (orange) and WOA23 climatology (yellow), plotted from the surface down to 1000 m.

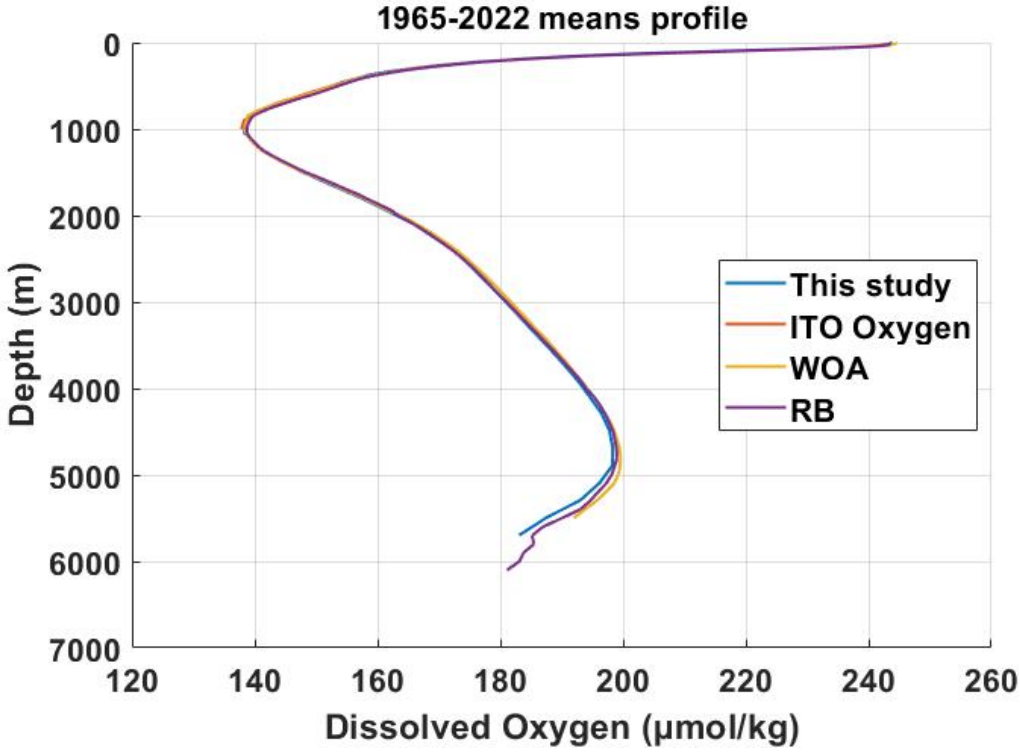

Figure R4. Global mean vertical profiles of different dissolved oxygen products (1965 – 2022). Solid lines show our reconstruction (blue), Roach & Bindoff's reconstruction (purple), ITO Oxygen (orange) and WOA23 climatology (yellow), plotted from the surface down to 6000 m.

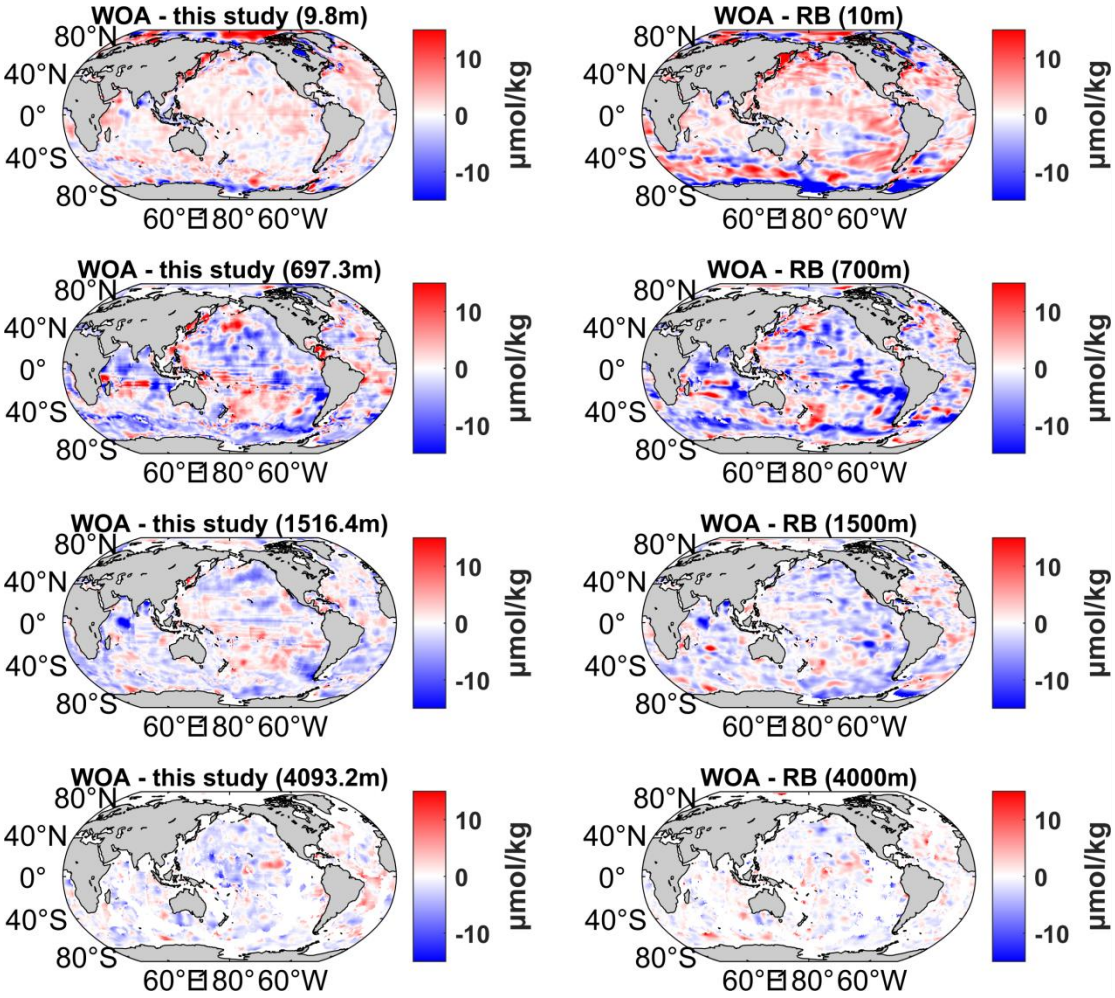

Figure R5. Spatial differences from WOA23 at four representative depths for this study and the Roach and Bindoff (RB) product. Left panels show WOA23 minus this study at 9.8, 697.3, 1516.4, and 4093.2 m. Right panels show WOA23 minus RB at 10, 700, 1500, and 4000 m. Units are μmol kg^{-1}.

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).

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.

Reference

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.

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.

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.