In “Reconstructing Global Monthly Ocean Dissolved Oxygen (1960-2023) to Nearly 6000 m Depth Using Bayesian Ensemble Machine Learning”, the authors use an ensemble of six machine learning models to reconstruct 4D fields of oxygen going back to 1960. They train their models on a mix of Argo oxygen and shipboard data in the World Ocean Database, then create a composite reconstruction that tries to weight the final output toward the models that perform better in a given location and time.

This work is timely, as many groups are trying to make use of the recent increases in oxygen data to create temporally- and spatially-resolved oxygen maps. It is important, as oxygen is an essential ocean variable and can provide a significant amount of information on its own as well as provide useful context to a wide range of oceanographic studies. However, I have several concerns about important details of their work and conclusions regarding ocean deoxygenation rates. Unfortunately, given the apparent use of overlapping datasets for validation and for training, I do not think the manuscript is suitable for publication in its current state and the analysis would need to be redone entirely. I therefore recommend that this manuscript be rejected, though I do hope that it is resubmitted after these issues are dealt with.

Sincerely,

Seth Bushinsky, University of Hawai‘i at Mānoa

Main issues:

World Ocean Database and GLODAP are not independent of one another. There is overlap between the datasets and it is not clear from the text that this was considered and the overlapping cruises removed. Given that no mention of this is made in the text, my guess is that the “independent” evaluation is actually just evaluating the oxygen product with a subset of the training data, in which case it is no surprise that your product performs so well. Please clarify if you did, in fact, remove GLODAP cruises from WOD, and if not, then I would rethink your approach to training and validation.

Another point is that much of this manuscript seems to be written for machine learning researchers to understand, not chemical oceanographers who use oxygen. To an extent, that makes sense, but it would be very beneficial if you could explain the choices made in the methods (especially 2.2.3 and 2.2.4) so that those who do not run ML models can understand their import and your rationale.

Uncertainty estimates: Currently you only are including in individual measurements (i.e. sensor precision) but ignoring the impact of potential biases, as we and others have

recently shown exist in the float oxygen dataset (Gouretski et al 2024, Bushinsky et al 2025, etc). You also assume that your observations within a given grid cell adequately capture the range of uncertainty, when this ignores the fact that a few samples within a 1x1 monthly grid cell are unlikely to capture the range in environmental variability. I am glad that you do include an uncertainty estimate, but with some adjustments it might do a better job of reflecting true uncertainty in your reconstructions.

Global deoxygenation rates: Your conclusion of dramatic increases in deoxygenation since 2010 need to be re-examined in light of biases in the Argo oxygen dataset. In addition to the ones mentioned above, optode-based oxygen sensors, which represent roughly half of the Argo oxygen data, have a relatively slow response time and are therefore known to be biased in the thermocline (typically toward low values) / regions of strong oxyclines. This may be part of the strong rate of change in oxygen seen in the blue line in Figure 5 between ~50m and 500m and described in the text. You cite Bittig's paper on this topic and mention it in your conclusions but need to consider the importance when evaluating the apparent decreases in oxygen content that you describe. Also, some assessment of the uncertainty in your rates should be included.

Other comments:

Line 47 – sentence fragment needs to be removed

Line 104 – why would you assume oxygen below 10umol/kg is an issue with unit descriptions? What about oxygen deficient/minimum zones? If there was an incorrect unit (i.e. should have been ml/l) that would surely affect the whole profile, not just an individual measurement. This filter does not make sense to me.

Line 125 – 5 dimensional fields? What is the fifth dimension?

Line 126 – what is CV?

Line 145-147 – How do you deal with the different vertical sampling resolutions of bottle and, CTD, and float profiles? If you are just averaging all points you would be weighting bottle measurements the least, which doesn't make sense as those are the highest accuracy data that you have.

Table 1 – title mentions 19 environmental factors, text says 11 predictors, but 3 seem to be time and 3 seem to be position, so I'm not sure whether either descriptor is correct.

Line 252- I think this should read Tables **2-5**.

Figure 2. It is difficult to see the differences that you discuss in the text. I would recommend adding a subplot to the right that shows the differences relative to WOA as a function of depth.

Lines 395-413: It seems like in this paragraph you both say that it is good when you match WOA but then say that when you disagree it means that your product is better. Both cannot be true without more of a rationalization for that conclusion. Also, why assume that WOA is best? Maybe ITO is better when there are disagreements, or maybe your product is. Please include the rationale for trusting the WOA maps.

Data availability – no mention of what version of WOA, GLODAPv2 (what year?), WOD, etc. Need to be included.