General comments

I have read and reviewed the manuscript titled "*Reconstructing Global Monthly Ocean Dissolved Oxygen (1960–2023) to Nearly 6000 m Depth Using Bayesian Ensemble Machine Learning"* by Mingyu Han and Yuntao Zhou. In the article, the authors present a new data product (BEM-DOR) of monthly dissolved oxygen concentrations in the global ocean, from the surface to 5902 m depth, built using an ensemble of decision-tree machine learning models. The models are trained on a combined dataset of *in situ* dissolved oxygen observations from the World Ocean Database 2023 and Argo floats (target) with ORAS5 model output for oceanographic variables such as temperature, salinity, zonal velocity, meridional velocity and geospatial coordinates (features).

While the product represents an advancement compared to the data products already available, as it expands the vertical coverage of dissolved oxygen products built using machine learning, I have several concerns on the methodology and results presented in the paper. Moreover, the discussion of the new BEM-DOR product lacks deep contextualisation and comparison with the products already available.

Additionally, the assets available for review only include the final dataset in netcdf format. In line with good practices in the field of machine learning, I would like the authors to make their code available for reproducibility testing.

Overall, the text requires some substantial re-writing and the authors need to provide additional evidence to some of their claims. The language used to describe results and discuss the BEM-DOR data product's reconstructions compared to other available products is at times misleading and needs to be changed. Only after the authors have addressed my comments and feedback, I will be happy to reconsider this manuscript for publication.

Specific comments

L103-105: provide a quantitative definition of unrealistically high or low. Also, the arbitrary exclusion of casts where any reading is below 10 μmol/kg would exclude areas of severe hypoxia and low oxygenated waters. If the authors followed an established methodology, I would want to see a reference to it. Otherwise, I would suggest their method to be at least partially reconsidered.

L110: what is the rationale behind the inclusion of zonal and meridional velocities as environmental predictors for dissolved oxygen concentrations?

L145: I understand from L111 and the documentation of ORAS5 that the data is gridded at 0.25° x 0.25° resolution. Where is this 1° x 1° grid coming from?

L157: why did the authors decide to use six models in the ensemble? And why are all the algorithms tree-based? Please explain further in the text.

L164-165: why did the authors include CatBoost in the ensemble if there are no categorical features in the framework proposed (BEM-DOR)?

L210: how were the hyperparameters to be tuned chosen? And how was the search range identified / selected?

L228: where would all predictions be missing? On land? Or at locations where no observations are available in the validation split during cross-validation? Please clarify further.

L246-250: it is unclear to me how this temporal cross-validation differs from the cross-validation done for hyperparameter tuning. First, I would like to have a more detailed explanation of what years formed the test set and what the training set, as the expression provided in line 247 is not clear. Additionally, I would like to see a detailed clarification of the differences between hyperparameter-tuning cross-validation and temporal cross-validation and the rationale behind cross-validating twice in model development.

L275-294 (Sect. 3.2) and then 337-366 (Sect. 4.1): Did the authors make sure that the observations they validate against in GLODAPv2 are not also included in the World Ocean Database 2023? Otherwise, they might validate against the same observations they are using to train the model. Similarly, the GOBAI-O_2 product (Sharp et al., 2023) is built using GLODAPv2 observations as training data, and the product of Ito et al. (2024) is built on World Ocean Database 2018 data. How did the authors ensure that their validation data were not included in the training of these two models as well?

L338: could the authors please provide a detailed description of how the comparison was performed, as it is unclear in the text?

L368: why do the authors not include the product of Roach & Bindoff (2023) in their comparison in Sects. 4.2 and 4.3, especially as that product is available up to depths of 6800 m?

L377-384: the exact difference between the lines is hard to quantify from the graph, but the authors claim that the difference between their work / WOA2023 and Ito is 2-5 umol/kg between 800-1000m when the lines seem to overlap. At the same time, they say that the difference between their study and WOA23 is 2-3 umol/kg at deeper depths down to 5902 m, while the graph clearly shows the lines diverging. This paragraph needs to be revised and the discrepancy in analytical interpretation addressed.

L417-429 (Sect. 5.1): this section does not add much to what is already known from a scientific perspective about large scale dissolved oxygen distribution. I suggest the authors delve deeper into some specific features of the data product that are novel compared to what is already available in the literature to provide additional evidence of why their data product is valuable.

L436-464 (Sect. 5.2): similarly to the section (and comment) above, Sect. 5.2 only provides rather general and already well-known descriptions of the variations of mean dissolved

oxygen concentrations throughout the water column. Additionally, the mean dissolved oxygen concentration profile in Fuigure 5 is the same as the one plotted in Figure 2.

Dataset in netcdf format: the values of 'time' and 'depth' seem to be decoded incorrectly in the final version of the file. Time is only reported as timesteps (0 to 767; without any decodable information on month or year). When opening the file in 'ncview', depth is only readable as depth level (1 to 74, without any information on the depth value in meters). Lastly, latitude, longitude and depth are included in the dataset as variables instead of coordinates.

Technical corrections

L47: repetition of 'This sparse spatial coverage severely'.

L52 and throughout the text: *in situ* should be written in italics and without a dash.

L53-55: what do the authors mean with 'restricting'? This sentence is phrased awkwardly and needs clarifying.

L60: define what WOA23 is.

L65-66: very vague and broad sentence. Provide references in the context of Earth sciences and oceanography.

L68 and throughout the text: the reference to the GOBAI-O_2 product is incorrect. It is Sharp et al. (2023): *Sharp, J. D., Fassbender, A. J., Carter, B. R., Johnson, G. C., Schultz, C., & Dunne, J. P. (2022). GOBAI-O_2: temporally and spatially resolved fields of ocean interior dissolved oxygen over nearly two decades. Earth System Science Data, 2023, 15, 10, 4481-4518.*

L69 and throughout the text: it is O_2 rather than O2.

L73-75: provide references.

L86: Ito et al. (2024) is missing in the reference list.

L99: define what OSD and CTD mean.

L106: how many observations are there in the final dataset? What is their distribution in space and time? The latter question can be answered by providing an additional figure either in the text or Supplementary Information.

L108: what biological factors? None of the environmental factors included in the models represent biology.

L110: salinity is expressed as PSU and it is unitless.

L112: be more specific. Is it 5902 m?

L115-133 (Sect. 2.2): I find this paragraph quite hard to follow for the average reader, as there are many undefined abbreviations and technical terms. I would suggest simplifying it and referring the reader to the more detailed subsections that follow in the text.

L125: what does 'producing six complete five-dimensional DO fields' mean? Please clarify.

L151: could the authors provide a table (in Supplementary Information) of the correlation values?

L153: there are only 11 environmental factors in the table, while the caption mentions 19. Correct typo.

L157-180 (Sect. 2.2.2): This paragraph needs strong rewording. As it is, it provides very general descriptions of the models using algorithm-specific terminology that might not be familiar to the readers. Moreover, the descriptions are very surface-level, and the authors do not provide any references for the claims they make regarding the different algorithms.

L184-185: what do the authors mean with 'history of performance evaluations'? Please clarify.

L190-192: Please rewrite this sentence with more details in order to make it more understandable to the reader.

L196-200: what is the size of the training and validation sets defined for cross-validation?

L218 and 222: ensure consistency between alphabets used. In equation 6, w is not ω and, in equation 7, α is not α.

L237: name dimensions together with their sizes so the reader does not have to guess or do the math while reading.

L241: spell out what CF-compliant means.

L257: Table 2 should be labelled as table 3, and so on. There is already a table 2 at line 210. Additionally, I suggest clarifying further what years are included in each fold, either by providing a schematic representation or a more detailed explanation in the text. As they are, the tables are not very self-explanatory and do not add much to the results presented, so they could be moved to the Supplementary Information.

L293: 'catastrophic fold' is not scientific terminology. Please change accordingly.

L312: I would argue that for the readership of ESSD, it is unnecessary to include here the formula of standard deviation.

L397: correct typo. 'Light' should be 'slight'.

L398: based on plot A in Figure 3, differences at high latitudes seem much larger than ±2 µmol/kg. Please revise.

L409: change 'modest' with 'larger'.

L412: As for my comment above, the differences observed in plot G (Figure 3) seem higher than ±3 µmol/kg at some locations. Please revise.

L430: change subplot titles of Figure 4 to match the four depth levels mentioned in the text and caption.

L468: why is oxygen content plotted as a scatter plot when the values are completely independent of each other? Please change.

L502: this is a bold claim considering that the profiles nearly overlap in figure 2. Please tone it down.

Figures: all figures seem to be in low definition. Please provide higher-quality images.