September 9th, 2023

<u>Author Comments to Reviewer #2</u>

We appreciate the encouragement, constructive criticisms and thoughtful comments from the Reviewer #2. Below we copy the original comments in black, and we provide point-to-point responses in red.

Ito et al. provide a good try to use synthetic data to understand an objective interpolation method (i.e. Ito et al. approach). They find that the global O2 change might be under-estimated using Ito's objective interpolation approach, which makes sense because the approach infills climatology (no data, no signal) into data-sparse regions. The use of synthetic data is also a good approach I believe. Generally, the paper can be published. But I have some major concerns, mainly on the interpretation and presentation of the results. I hope these concerns can be addressed before publication.

We appreciate the reviewer's support for our use of synthetic data to evaluate a statistical gap-fill approach. Our mapping approach is admittedly simple with a Gaussian covariance function with a constant e-folding scale. This choice has certain benefits, for example, that the results from a simple method are easy to understand, and that it is also easy to notice and to correct mistakes. It can be replicated by other groups relatively easily. If the ocean deoxygenation has a wide-spread, large-scale signal as well as regional hotspots, we anticipate that a simple method should, at least, capture majority of the large-scale component and some regional features. There are some drawbacks too, as correctly pointed out by the reviewer #2. It tends to smooth out spatial gradients, and it may not represent regional signals very well. It is expected that it will underestimate the signal in data poor regions, which we are trying to quantify in this work. Overall, we are in the same opinion with the reviewer #2 that the revised manuscript should clearly state the limitations of this study and the potential for improvements in future studies.

Major:

1. The first concern is that the results are all specific to the particular OI approach proposed by Ito et al., and can not be generalized to all OI approach. This is very important because different groups (even all using OI) have different settings and considerations/assumptions, such as influencing radii, covariance etc., and the performance would be fundamentally different. Therefore, I would strongly

insist on being more specific in the paper title and in the Abstract that the "underestimation" is for Ito et al. OI approach.

We agree that the results presented in this study is based on a specific implementation of the optimal interpolation method. This single study cannot be generalized to all variants of optimal interpolation approaches. The concern can be addressed in the abstract and the main text, as well as some modification to the title as suggested by the reviewer #2.

2. The sub-sampling strategy has to be more clearly introduced, e.g. how do you re-sample data if a 1X1 grid box has more than 1 observation? How do you deal with the difference in land-ocean masks (models differ from the real-world, for sure)? How do you construct the climatology: are you doing this based on re-sampled data over which time period?

   In the original manuscript, there was only one sentence explaining our sampling strategy in L189-191, which was not enough. It would be good to expand and clarify the points raised by the reviewer #2. Below are our answers to the specific questions.

   The sub-sampling strategy assumes that a grid box is sampled if, at least, 1 observation exists within the grid cell at a particular year/month. If so, we retain model data in the sub-sampled dataset. In reality, there could be multiple samples within the same grid and/or there could be significant variability in oxygen within the same grid box and within the same year/month. Multiple casts and/or variability within a single cell are not represented in our sub-sampling strategy.

   There are slight differences in the land-ocean masks between models, and we use the model topography as they are provided. This could cause some small discrepancies in the ocean volume, but we do not make correction for this specific effect.

   The analysis period is 1965-2014. The climatology is constructed for this period. Sampling pattern affects the representation of climatology, and this effect is included in our analysis. The models' climatological O2 fields are separately calculated for full and sub-sampled model data. These different climatological O2 fields are used to define anomalies for the full and sub-sampled models.

   These are all good questions and will be explicitly stated in the revision.

3. The over-simplification of the approach has to be stated clearly in the abstract: e.g. no sub-grid variability is considered (observations do have all scales of variability, but your resampled profiles only contain variability >1 grid and 1 month), and no instrumental errors and potential biases are considered. Thus, this current approach is more “conceptual” than you can definitely quantify the “underestimation”. I believe the current study is over-confident in the quantitative numbers of the overestimation. I strongly suggest the authors take this more seriously.

   We appreciate the reviewer #2’s concern and expanding the points raised in the first comment. The concern of over-generalization is well taken, and it can be addressed primarily in the abstract and the main text.

   The sub-grid scale variability that are not represented by earth system models could exist in the real observations, and our current analysis lacks this variability. If sub-grid scale variability were included, it would lead to a lower signal-to-noise ratio and a larger uncertainty. We are aware of this issue and our viewpoint is explained in the sentences starting L469 where we discussed the lack of mesoscale eddies in the model, but this discussion can be broadened to address the instrumental error raised.

   We are in agreement with the Reviewer #2 that our model-based analyses lack some sources of variability, and that our analysis might be too optimistic to capture the signal due to the lack of such variability in the models we used. We indeed take this issue seriously, and have suggested some possible solution for future studies in the discussion section, which can be expanded further in the revision.

4. Comparison between model results, and subsampled results with “Observations of Lto et al” (black line in Fig. 4) should be more careful. The so-called “observation” itself is biased by the conservative error (as investigated in this study) and it is more problematic because it also contains sub-grid variability and instrumental problems. Thus, please to be more careful to compare the models and subsampled results with this observation. It makes more sense to remove the black line in Fig.4, and also combines the right and left columns (e.g. full-model in solid lines and subsampled models in dashed lines at the same panels so they can be directly compared). So the focus is on the sampling issue. I have similar concerns about Fig. 5 and 6. I don’t even believe “observation” (as you argued in this study, it is problematic in many places because of the sampling/interpolation issues)

These are great suggestions, and we would be very happy to revise Figure 4-6. The suggested change of removing the black line from Fig 4 and combining left and right column can easily be done and will put stronger emphasis on the sampling issue and the comparison between full model and subsampled and gap-filled model output. Fig 5 and 6 can also be revised as suggested to provide more detailed spatial structure.

5. 7: because of the area difference with latitudes, doesn't it make more sense to define the coverage by area instead of number of cells?

   We appreciate the suggestion. Indeed, it makes more sense to convert the number of cells to areal coverage, and it can easily be revised.

6. 5/6, I would be more keen to see the real spatial pattern (1X1 grid trend), instead of the box averages. It will be more straightforward to show the dynamic regimes.

   We appreciate the suggestion. Fig 5 and 6 can also be revised as suggested to provide more detailed structure.

7. 8/9 and section 3.4, with these analyses, I guess the authors want to have an "empirical" correction to the OI approach. But as I said in my 1-3 major points, the quantifications are useless because of the oversimplification of the approach. I don't see the value of doing this analysis.

   We appreciate the reviewer #2's concern. There are indeed limitations in relying on a single, relatively simple implementation of optimal interpolation approach. However, we believe that there still is a value in estimating the bias caused by this simple implementation of optimal interpolation. This type of comparative analysis was never performed collaboratively by modelers and observationalists using synthetic data as a testbed. For the future studies, we would be excited to invite others to join and collaborate including different and perhaps better gap fill approaches.

   Since the primary reason for the objection is mainly comes from the simplicity of our specific gap-fill method, we suggest that it can be addressed with more careful statements about its limitations. This includes revisions in the title, abstract as well as the main text that, this is for a specific, relatively simple implementation of the optimal interpolation.

Overall, I do see the value of this study, but the interpretation and presentation of the results need to be revised in a substantial way to make it a more rigorous study (also leave room for more improvements in the future).

We appreciate the encouragement and support here. There is indeed much room left for more improvements in the future. Potential areas of future improvements are included in the original manuscript, and it can be broadened in the revision.

Minor

1. Abstract "more than 80%", I don't think it is trustable, because the values are apparently model-dependent, and also depend on subgrid-variability and instrumental errors.

   This estimate is subject to the specific implementation of optimal interpolation for this study. As discussed earlier, it will be spelled out more clearly in the revision about its limitation and weakness. Having said this, these estimates are internally consistent results under the assumption of this study. The number could potentially become smaller if we correct the signal-to-noise ratio to a lower value due to unaccounted sub-grid scale variability. It would be important to revisit with more sophisticated estimation approaches in the future study.

2. Line 105: it is a strange choice, because if you remove Argo because of the precision, then CTD and OSD have different accuracy as well. To me, including Argo is valuable because you just want to test the sampling issue. You do nothing with the precision/accuracy in this paper with a synthetic data approach. I understand the authors might not do the things all over again, so a clear statement on the caveats is a necessary.

   BGC Argo is an important data source especially after mid-2010s. Since the period of trend analysis is from 1967 to 2012, it can impact on the last period. Our decision was to stay on the cautionary side and not to include the ARGO data at this time, but it will be an area for the future study and improvement.

3. Line 185: just to confirm, is IPSL-CM6A-LR a Earth System Model?

   IPSL-CM6A-LR is the latest version of the IPSL earth system model. In addition to the physical atmosphere-land-ocean-sea ice model based on the LMDz, ORCHIDEE, NEMO (including the LIM and PISCES subcomponents) models. This model includes a representation of the global biogeochemical cycling including carbon, oxygen and nutrients. Further description of the IPSL-CM6A-LR climate model is available through the reference paper listed in Table 1 of the original manuscript.