Response to reviewers on "Reconstructing ocean carbon storage with CMIP6 models and synthetic Argo observations"

To the editor and reviewer:

Thank you for your review to our updated manuscript. The reviewer's comments have been very helpful, and we appreciate how they have considered the methodology as our reconstruction has moved further into the ocean interior. These considerations have made our study more rigorous and a better step towards a real-world carbon analysis.

We have updated the figures to reflect the reviewer's point on de-drifting the data before creating an analysis. There are some improvements in the deeper carbon fields after de-drifting. Overall, the main messages of the previous manuscript remain, as most carbon changes are within the upper part of the water column.

We have attached our responses to the reviewer's comments below.

Best regards,
Katherine Turner

Authors' response to review
Reviewer's text in black
Authors' response in blue

I want to thank the authors for their extensive review, their detailed answers to all major and minor comments, and their patience with me when I did not understand everything. The extension of the analysis to the upper 2000 m are a great plus and after the response, I entirely agree that it is best to apply the method to the ARGO data in a separate study.

However, I have two last outstanding major question with respect to the Methods that need to be clarified before publication and that could substantially improve the results. Although I have clicked 'major revisions', I do not want you to see this as a 'traditional' major revisions. The manuscript is in great shape but I think the two major points would make it much better. The first point would have to be adressed in text if the model drift was accounted for and needs additional analysis if it was not. The second point could be adressed by changing the text, but I believe an additional analysis might substantially strengthen the study.

1) I could not find information if the model output was detrended using the pre-industrial control output. Trends are usually small to negligible close to the ocean surface due to exchanges with the atmosphere but can be substantial in the subsurface ocean. If such drifts exist and vary across the models in the ensemble, it might be almost impossible to obtain the optimal weights from these models. Such a drift may well cause the bipolar optimal coefficient pattern for T and S that is shown in Figure 7. If the drift is not accounted for, I think it has to be accounted for. The best way is probably to fit a spline to the pre-industrial control run and to remove it from the historical simulation. In this way, the inter-annual or decadal variability, which are likely similar in the piControl and historical run, is not removed with the drift.

Thank you for this suggestion. We have now included drift removal in the current manuscript as it was indeed lacking from our previous analysis. In the manuscript we have included the line:

> "The drifts in temperature, salinity, and DIC were calculated and removed by subtracting linear trends at each ocean grid cell in the piControl runs." (Section 2.1.1)

Instead of fitting a spline to the piControl runs, we have opted to take the linear trend in ocean temperature, salinity, and DIC at every oceanic grid cell before regridding and integrating vertically into our depth horizons. This drift removal makes minimal assumptions about the nature of drift for any particular model.

We have updated the figures to include this de-drifting, although we do note (from the qualitative similarity between the previous manuscript figures and this manuscript's figures) that model drift does not have a large impact on our results. We speculate that the uncertainties included in the model ensemble outweigh the uncertainties provided by the long-term drifts, and that the process of integrating carbon, temperature, and salinity within relatively thick layers may allow for some compensation between drifts at different depths.

2) It is somehow concerning that problems arise for MPI-ESM1.2-LR because of the higher spatial resolution. I believe that a mistake was made. Following Séférian et al. (2020) (https://link.springer.com/article/10.1007/s40641-020-00160-0), MPI-ESM1.2-LR has one of the coarsest resolutions. If the MPI-ESM1.2-LR resolution is indeed coarser, the argument would have to be altered. At the moment, the weak performance of the DIC reconstruction under MPI-ESM1.2-LR is discussed essential weakness of this method when being applied to observations and regions with lots of mesoscale processes. However, this might be wrong if MPI-ESM1.2-LR was not highly resolved. From my experience, MPI-ESM1.2-LR is more likely an outlier and may not affect the performance of the methods when applied to observations. I believe an adjustment of the Discussion is absolutely necessary. However, I believe the authors could significantly increase trust in the reconstruction method, and I believe the method merits it, by including GFDL-ESM4 (1/2° resolution) and GFDL-CM4 (1/4° resolution). Both model versions are highly resolved compared to other models and GFDL-ESM4 is the best performing model with respect to historical carbon uptake (Annex in Terhaar et al., 2022). It would take a large amount of work and I hence cannot ask for it but I think it would be a worth it.

MPI-ESM1.2-LR is indeed a relatively low-resolution model for the CMIP6 ensemble – thanks for pointing this out. Our statements are incorrect and were originally made for a reconstruction that used the high-resolution MPI ESM. We have modified the text to correct this error.

We have refrained from using the GFDL models in this proof-of-concept work as the models have only one realization. For this step we restricted our study to only models that had multiple realizations (5) in order to include uncertainties from both model architecture and initial conditions/ the phasing of climate modes of variability. As we continue our work and move towards creating a real-world reconstruction, we will be updating our ensembles to include other models such as those from GFDL and the NCC (as NorESM is provided as a test case but not included in the original model ensemble).

In addition, I have one comment that is neither major nor minor:

1) Section 3.1 seems to be complicating the message. The co-evolving trends of atmospheric CO2 and warming lead to correlations that are not related to the effect of warming on the ocean system, i.e., changes in circulation and solubility. In fact, the ensemble coefficients in Figure 4 show this as the positive correlation from temperature is entirely accounted for by the increase in pCO2 and the pCO2 coefficient, whereas the T coefficient accounts for the solubility. I believe that Figure 4 really shows the strength of the approach and merits to be a centerpiece of the results.
I am not an author of this study but would seriously consider removing section 3.1, as it confused me more than it helped. However, I am only one person and others may disagree. And as nothing is wrong with the section, I only wanted to voice my opinion and do not want to ask for changes.

We have opted to include the sections on the correlations and the breakdown of the correlations between DIC and temperature to tie in more clearly to our future work. These synthetic reconstructions assume that observations taken from the model output are perfect. Therefore, the least squares solutions shown in Figure 4 is the optimal solution. The optimal coefficients will change when including imperfect observations, but the covariance fields used by the Ensemble Optimal Interpolation algorithm will remain the same if the ensemble makeup is the same. Thus, while the inclusion of the correlation fields makes the paper longer, we feel it is an important step to show for future work. We have included a statement in the beginning of 3.1 that reflects our logic in including the covariance fields first.

Please find below some minor comments:

1) Figure 8: It seems that the reconstruction performs best, where most carbon is stored (North Atlantic and Southern Ocean) and performs poorly where little to no carbon is stored, for example in the deep North Pacific. Thus, the maps in Figure 8 might suggest that the reconstruction performs poorer than it does. Maybe it would be better to show differences between the reconstructed and the true DIC, like Figure 9.
The statistics in Figure 8 have been provided to show areas where uncertainties in the models could lead to conservative estimates of DIC as well as regions that remain difficult to predict due to errors in the covariance fields. Because of the multiple models used, it would be difficult to isolate specific difference plots. However, we have included an additional figure in the Supplementary that shows the standard deviation of DIC' across the ensemble, as well as the ensemble average RMSE from the individual reconstructions. This figure clearly reflects the ideas you have mentioned, where the model does well in regions with high DIC' changes (and thus a high standard deviation), while the regions with error increases generally have low DIC' changes.
2) Line 3 of the abstract: I would maybe highlight here that the temperature and salinity coverage is much less sparse than the carbon observation coverage. The contrast between both coverages is what this is all about.
Thank you, we have extended the sentence to emphasize this point.
3) Line 20: Should probably be Terhaar et al. (2022) and not 2020.
We have revised the manuscript to reflect the correct citation.
4) Line 28: would replace 'sufficient' by 'enough'. The coverage of, 1.5% is likely not sufficient (L. Gloege, G. A. McKinley, P. Landschützer, A. R. Fay, T. L. Frölicher, J. C. Fyfe, T. Ilyina, S. Jones, N. S. Lovenduski, K. B. Rodgers, S. Schlunegger, Y. Takano, Global Biogeochem. Cycles, in press, doi:https://doi.org/10.1029/2020GB006788.)
We have updated this sentence to describe the order of magnitude of pCO2 observations rather than qualifying with "sufficient."
5) Lines 38 to 47: This paragraph should probably also include reference to the application of the eMLR* method by Gruber et al. (2019)
(https://www.science.org/doi/10.1126/science.aau5153)
We have included this work with an emphasis that these reconstructions are concerned with the evolving anthropogenic carbon inventory.

6) Lines 56-57: What exactly are these well-understood relationships? Maybe worth to elaborate a bit more.

We have split up this sentence to more explicitly describe the inverse relationship between temperature and DIC and the direct relationship between salinity and DIC.

7) Line 59: Maybe worth to reference Weiss et al. (1974). (https://www.sciencedirect.com/science/article/pii/0304420374900152)

We have included this citation in the first part of the expanded relationships section above.

8) Line 234-235: I believe that this sentence is not correct. Fig. 2 is rather a combination of the parallel increase in atm CO2 and temperature and the solubily effect of T on pCO2.

We have altered the sentence to read: "Thus the heterogenity found in the overall correlation between DIC and temperature in Figure 4 can be understood as the sum of an emissions-driven undersaturated response that correlates to but is not driven by warming, and a temperature-driven solubility response."

9) Line 336: How many regions are 'most regions'.

We have extended the sentence to read "Across the sensitivity tests and outside of the Southern Ocean and small regions in the North Atlantic, North Pacific, and equatorial Pacific, carbon can be reconstructed with a relative RMSE reduction of at least 50%."