We would like to express our gratitude to both reviewers and the editor for the second reviews of this manuscript. We have thoroughly addressed all specific comments below. Line numbers refer to the "tracked changes" version of the manuscript. Changes from the previous round of revisions are marked in red. Changes for this second round of revisions are highlighted in yellow. The main changes to the manuscript are:

- *We included a sentence regarding the source of mean pCO_2 for the pCO_2-Residual calculation (lines 217-219).*
- *We added a statement explaining why we use observations (and not testbed output) to calculate fluxes (lines 384-388).*
- *In the introduction, we now highlight the new knowledge contributed in our study and how our work complements previous studies (lines 131-133 and 135-140)*
- *We mention the additional importance of sampling masks when comparing testbed studies reconstructing surface ocean pCO_2 (line 879 and 212).*
- *We have replaced "observation-based data products" with "mapping methods" throughout the manuscript (line 50 and 209).*
- *We discussion overfitting in the main text (lines 417-423) and in the Supplementary Material (**Supplementary Text A**) and added a supplementary figure (**Fig S6**).*
- *We added a statement which explains that reconstructions using pCO_2-Residual as the target variable as opposed to pCO_2 leads to higher skill (lower RMSEs) (lines 208-211).*

Response to Reviewer 1

I would like to point out that the responses to my comments could have been done better. In my first comment on the changes to the methodology, the authors replied that the required sensitivity analysis had been carried out as part of an earlier publication. But the authors didn't mention any value in their response to my comment and wrote "had little influence on the reconstruction", leaving me not knowing how "little" that is.

We are sorry that the reviewer feels our responses were insufficient. Our answer refers directly to experiments performed by Bennington et al. (2022), and reported in that publication. Here is the full text in the last paragraph of section 2.3 of that paper. "We tested the sensitivity of the reconstruction to the source of mean pCO_2 used in the calculation of pCO_2-T with Equation 2, which is then input to the pCO_2-Residual calculation in Equation 3. Reconstructions using the Lamont-Doherty Earth Observatory (LDEO) pCO_2 climatology (Takahashi et al., 2009) and the mean pCO_2 of the SeaFlux observation-based products (Fay et al., 2021). The alternative sources of mean pCO_2 did not significantly impact reconstructed pCO_2 or resulting air-sea CO_2 exchange, so we maintain our own method for the initial reconstruction of pCO_2." To be more precise on "little influence" from this work - the test statistics, as reported in Table 3 of Bennington et al. (2022), were not sensitive to the choice of initial pCO_2 field.

Furthermore, Bennington et al (2022) have already demonstrated the skill of the pCO_2-Residual approach using real-world data in their comparisons to independent data at BATS and HOT, and from GLODAP and from the LDEO pCO_2 database (i.e., points not in SOCAT). Together, Bennington et al. (2022) provide evidence that this approach performs better, admittedly

marginally, compared to other observation-based products. Thus, to use the approach in this Large Ensemble Testbed study focused on the impact of sampling distribution is reasonable without further sensitivity studies on the method itself.

In the revised manuscript we added a sentence stating that alternative sources of mean pCO_2 have been assessed by Bennington et al. (2022) and that they did not significantly impact the test statistics or reconstructed pCO_2 (lines 217-219):

"Alternative sources of mean pCO_2 were assessed by Bennington et al. (2022a), but they found no significant impact on the test statistics or reconstructed pCO_2."

In the reply to my second comment, the authors wrote "high temporal resolution output is not available for the test bench". Why would the authors need high-resolution model outputs, when the pCO2 reconstruction performed by the method is carried out with the same spatio-temporal resolution as the model outputs (1°x1°, monthly temporal resolution). This argument seems irrelevant, whereas the second part of their reply, which mentions that their aim was not to calculate the real-world fluxes, is more understandable.

We are sorry that this was unclear. Our goal was to convey the impact on our ability to calculate air-sea CO_2 fluxes given the lack of temporally high-resolution output of winds. Because of the square dependence of the flux on winds, one needs high-resolution (3 or 6 hourly) winds to calculate fluxes. Since only monthly model output for the winds is available, we cannot use model-based winds for the flux calculation.

*We add to the revised manuscript a statement that we do not have high-resolution output of **winds** (lines 384-388):*

"Winds have the largest impact on flux calculations (Fay et al., 2021), and temporally high-resolution output is not available for the LET. Monthly output is available, but this is not sufficient for the flux calculation due to the square dependency of wind speed (Wanninkhof, 2014). Given the necessity to use observed winds, for consistency, we use observations for all necessary variables for the flux calculation."

Finally, in their response to my 4th specific comment, the authors wrote three paragraphs that were primarily aimed at the second reviewer's comments. These paragraphs did not address my comment and appear to have been poorly copied and pasted. Therefore, additional care needs to be taken.

We are sorry that the reviewer does not feel like we appropriately responded to this valuable comment that helped us to significantly improve the manuscript. It led us to include testbed spread comprehensively across the manuscript. The first 2 paragraphs in the previous response do directly address Reviewer 1's comments. The following three paragraphs were, indeed, part of an answer to reviewer 2. We believe these amplify the reviewer's point about the necessity of showing the testbed spread. We are sorry Reviewer 1 finds these additional paragraphs unnecessary.

Response to Reviewer 2

General comment:
The point is not to highlight the use of any specific type of pCO2 measurements over the others for the estimation of global maps of pCO2. For instance, float-based data provides indirect observations of pCO2 and thus high uncertainty for pCO2 estimates. However, the suggestions learned from the previous works [Bushinsky et al. (2019), Denvil-Sommer et al., 2021, Djeutchouang et al., 2022, Hauck et al., 2023, Landschützer et al., 2023] are to obtain more accurate (precise) estimates of pCO2 by extending the observing systems or considering additional data sources available in space and time. Besides, many of the existing works have exploited the sensitivity of pCO2 and flux estimates to the data sparsity over the Southern Ocean. However, I agree that Thea Hatlen Heimdal et al have contributed a new finding about different USV sampling strategies to the global reconstruction of pCO2. It's worth to add few sentences in the last paragraph in Section Introduction to bold the new contributions as complements to the previous works. A summary of Section Methods would be enough: e.g. one-latitudes and zigzag sampling, … which differ from the SOCAT+SOCCOM or Argo-float ideal sampling over the global ocean by Hauck et al., 2023).

Thank you for this clarification. We have added some sentences in the last paragraph in the introduction to highlight how our work complements these previous studies (lines 131-133 and 135-140):

"We test the impact of two different USV Southern Ocean sampling schemes, the first based on a sampling campaign completed in 2019 (Sutton et al., 2021), and the second on logistically feasible potential future meridional sampling."

"Combined, the sampling patterns tested here complements previous studies exploring the impact of additional sampling in the Southern Ocean based on idealized full global coverage of floats, and float observations from recent deployments, including the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project, moorings and sailboats (Bushinsky et al., 2019; Denvil-Sommer et al., 2021; Djeutchouang et al., 2022; Hauck et al., 2023; Behncke et al., 2024; Landschützer et al., 2023)."

Specific comments:
I do not support the following arguments of the authors in their responses to the reviewers:

"We do find the study by Hauck et al. (2023) interesting, but note that it was not published when we submitted our initial manuscript. In the revised version we have added a paragraph discussing this study and comparing their results to ours (lines 933-954). A key point made is that both Bushinsky et al. (2019) and Hauck et al. (2023) show an overestimation of the ocean sink with current sampling, while we show the opposite – an underestimation of the ocean sink."

First, I am not aware whether the initial manuscript was submitted to other journals or not. But as tracking the MS record in Biogeosciences, this study first appeared for review in September 2023 while Hauck et al. (2023) was published in March 2023.

We apologize for our oversight here. We should have said in our first response to reviewers that we were not aware of this publication when we submitted the first version of this manuscript.

Second, it has a level of confidence of an overestimation of pCO2 based on present-SOCAT sampling as tested by Bushinsky et al. (2019) and Hauck et al. (2023).

We do not understand this comment. We cite directly the overestimation with SOCAT-only that has been reported by Bushinsky et al. (2019) and Hauck et al. (2023) in comparing to SOCAT+float sampling (and compared to the model truth):

- *Bushinsky et al. (2019): "The combined SOCAT+SOCCOM product yields a Southern Ocean sink that is 0.4 Pg C/yr weaker over 2015-2017 than that calculated from shipboard data alone" (page 1385, Section 3.4).*
- *Bushinsky et al. (2019): "…the SOCAT-only uptake 0.22 Pg C/yr stronger than the model and the SOCAT+SOCCOM uptake 0.14 Pg C/yr stronger than the true model flux…" (page 1383, Section 3.2)*
- *Bushinsky et al. (2019): "For SOSE, the neural network-derived SOCAT-only Southern Ocean uptake was 0.38 Pg C/yr stronger than the true model, while the SOCAT+SOCCOM uptake was 0.26 Pg C/yr stronger than the model…" (page 1384, Section 3.2)*
- *Hauck et al. (2023): "Both mapping methods overestimate the mean CO_2 uptake 2009-2018 and the trend 2000-2018 in the SOCAT sampling scheme. In the MPI-SOM-FFN method, the 12% overestimation of the mean in the SOCAT scheme is reduced to 9% in bgcArgo. The 9% overestimation in CarboScope (SOCAT) vanishes in the bgcArgo scheme" (page 9, "Air-sea CO_2 fluxes").*

Both these studies state that adding floats to SOCAT leads to a weaker mean sink. Our study shows the opposite - adding USV observations leads to a stronger mean sink.

In the manuscript we emphasize that it is the conclusion of these other studies that ML methods overestimate the CO_2 sink (lines 890-802)*: "**These studies showed** that SOCAT sampling alone overestimates the CO_2 uptake in the Southern Ocean, and that additional floats reduce this overestimation, leading to a decreased (weakened) ocean carbon sink."*

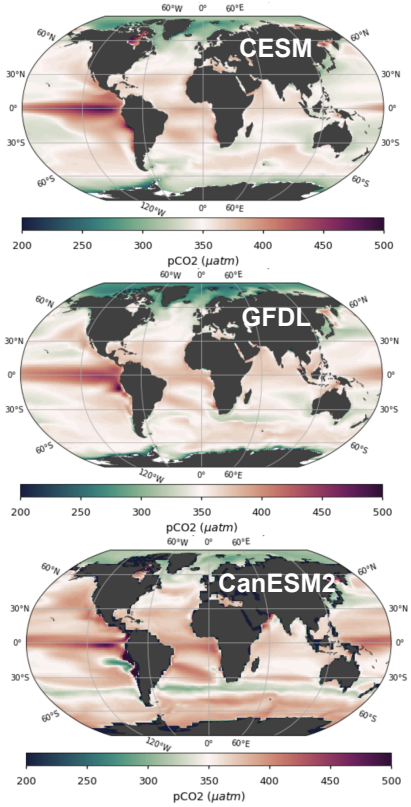
pCO2 generally increases over time and mapping methods tend to underestimate pCO2 (thence overestimate fluxes) based on sparse training datasets which have not covered the full range of realistic pCO2 values (many regions with high pCO2 values are unobserved). It's questioning about the distinction between the results in this study and the previous.

We agree that pCO_2 observations are sparse and do not fully cover the distribution of pCO_2 space. This is shown nicely by Hauck et al. (2023).

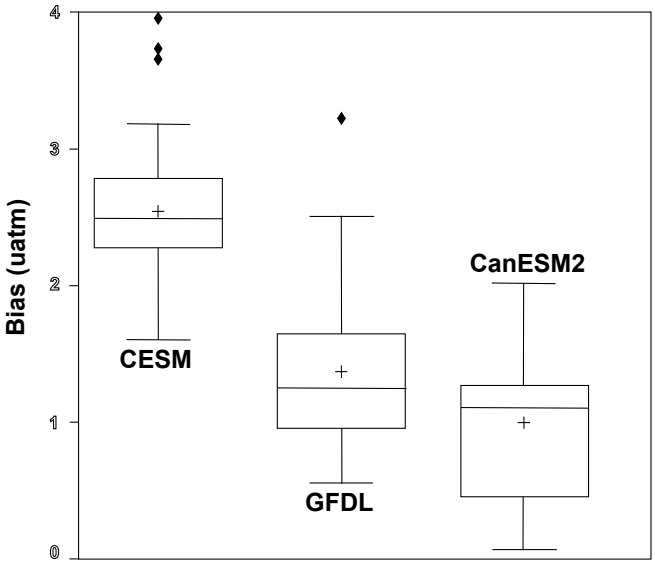
Whether the algorithm over- or underestimates pCO_2 appears to depend on both the type of reconstruction method used but also the type of testbed used (which models, which member, and whether conclusions are based on an ensemble or models or not). The figure below compares the 'model truth' pCO_2 field for the three individual models used in our testbed. As shown on left, the model mean pCO_2 fields differ. CanESM2 has higher pCO_2 in the Southern Ocean compared to CESM and GFDL. That CanESM2 is higher in the Southern Ocean may be related to its lower

bias (right), but more work will be required to confirm such a relationship. At this point, we can see that the choice of model and ensemble member in a testbed matters to reconstruction bias, and thus it is reasonable for us to discuss this potential impact on the comparison of our results to those of Hauck et al. 2023.

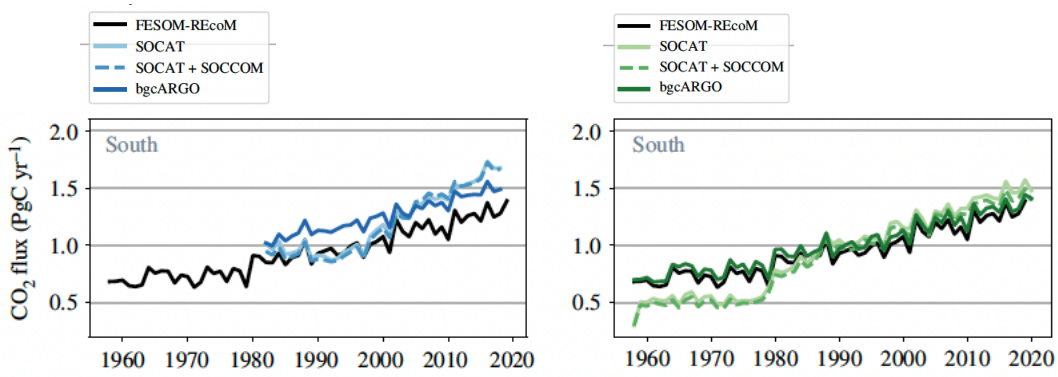
Model truth pCO_2 field (mean 1982-2016) for individual models of the LET



Ensemble spread of mean bias (over Southern Ocean, 2006-2016) for 'SOCAT-baseline' for individual models of the LET



Hauck et al. (2023) use a 'SOCAT', a 'SOCAT+SOCCOM' and an 'idealized float' sampling mask and one model as the "testbed" (FESOM-REcoM) in their study. They use two different reconstruction methods. As shown by the figure below (their figure 4; see below), the two reconstructions result in different fluxes. The MPI-SOM-FFN method predicts a larger carbon uptake compared to CarboScope. This tells us that the choice of reconstruction method matters.



Lines 846-854 (in the manuscript with track changes) will need to be revised (or excluded). For such sensitivity tests, one would not expect to see the comparison in performance of different mapping methods but of a fixed method to different sampling scenarios.

We fully agree with the reviewer. In order to directly compare our study and Hauck's, we would have to use the same sampling mask and testbed model(s), and also calculate the air-sea CO_2 flux in the same manner, and then compare the reconstructions using different ML methods. To resolve this, more experiments are needed and these would be beyond the scope of this study. This is the point we aim to convey in this paragraph, to which we have revised to add note of the additional importance of sampling masks (lines 904-907):

"Our study and Hauck et al. (2023) use ***different sampling masks*** *and approaches for the calculation of fluxes, which could also be a factor. Targeted, coordinated studies using multiple reconstruction approaches with consistent testbed structures,* ***sampling masks*** *and experimental approaches are clearly needed (Rödenbeck et al., 2015)."*

Even in this study, an ensemble of model output or the methods based on SST-removal effects from pCO2 would add more uncertainty to statistics such as bias, RMSD,... That's why I have suggested analyzing further differences in fluxes' variability (trends, seasonal cycles,...) with respect to different sampling strategies.

As mentioned in our response to Reviewer 1, through comparisons to independent data, Bennington et al. (2022) have demonstrated a marginally improved skill of this reconstruction approach compared to other published observation-based products (this is mentioned in lines 211-212). They also demonstrate lower RMSEs for reconstructions using pCO_2-Residual vs. pCO_2 (their figure S1), which indicates that the removal of temperature from the target variable enhances the performance of the method.

We add note of this in the revised manuscript, with specific mention of Figure S1 of Bennington et al. (2022) (lines 208-210):

"Bennington et al. (2022a) demonstrate higher skill for reconstructions using pCO_2-Residual as the target variable as opposed to pCO_2 (Figure S1 in Bennington et al., 2022a), indicating that the removal of the temperature-driven component enhances the performance of the method."

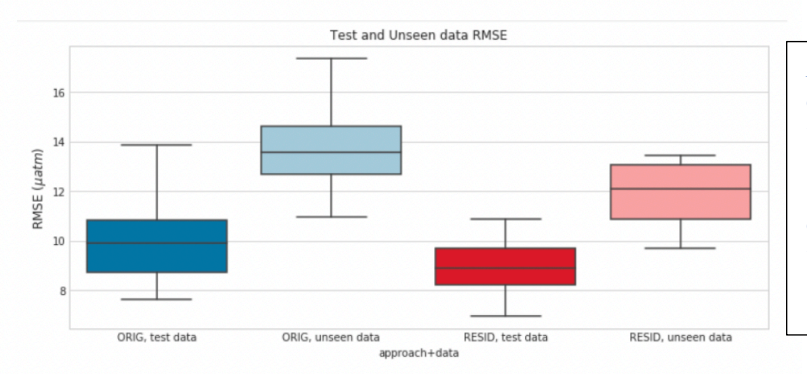


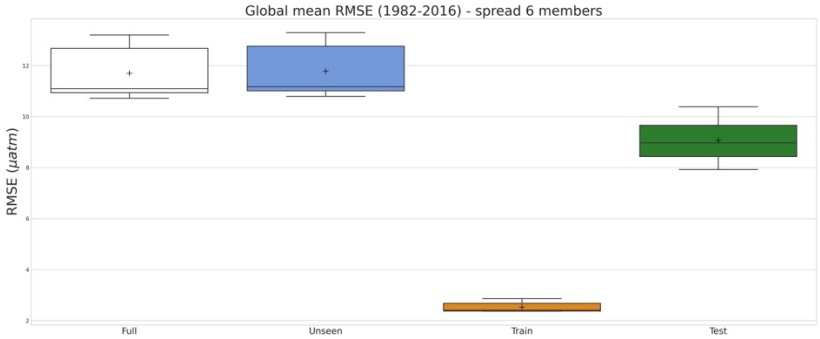
Figure S1 in Bennington et al. (2022), comparing unseen and test RMSE using the Large Ensemble Testbed. ORIG = reconstruction using pCO_2 (no removal of the pCO_2-T component). RESID = reconstruction using the pCO_2-Residual.

The discontinuity in Figures 3 and 7 still persists: we have obviously seen the gradients in RMSD at (SOCAT or zigzag) sampling tracks versus the "unobserved" areas. Therefore, I expect the authors to verify whether their mapping method put much higher weights on sampled locations than "unobserved" regions ('overfitting': i.e. over-exploitation of the entire available data for model training).

Figure 3 shows RMSE and bias for the 'SOCAT-only' experiment. As expected, bias and RMSE are higher at times/locations where SOCAT observations are scarce (e.g., in the Southern Ocean). Figure 7 shows the improvement in RMSE when USV observations are combined with SOCAT. RMSE improves mostly in the Southern Ocean where RMSE was initially high. New observations most improve predictions in similar areas because the augmented training set now contains these similar points. This is consistent with autocorrelation lengths for pCO_2 up to 400 km in the Southern Ocean (Jones et al. 2012).

Jones, S.D., Le Quere, C. and Rodenbeck, C.: Autocorrelation characteristics of surface ocean pCO2 and air-sea CO2 fluxes. Global Biogeochemical Cycles, 26, 2, https://doi.org/10.1029/2010GB004017, 2012.

Following the reviewer's comment, we further analyzed our algorithm to explore overfitting. Indeed, we find some evidence of this, i.e. a statistically significant difference between train and test set error (see figure below). This means that further tuning of the hyperparameters of our ML algorithm could increase generalization skill. But it is important to emphasize that this finding does not invalidate the test or unseen statistics that we present – it simply indicates that more tuning might further improve algorithmic skill.



Global mean RMSE (1982-2016) for full, unseen, train and test sets for the 'SOCAT-baseline' experiment. The boxplot shows the ensemble spread of six members of the LET.

The goal of this study is to explore how USV sampling added to the Southern Ocean would change skill with all else held equal. We use the same algorithmic approach (same hyperparameters, model is retrained) to reconstruct with only SOCAT or with SOCAT + USV sampling. Further fine-tuning of the algorithm, already shown to perform well by Bennington et al. (2022), is not required to test sampling patterns. Though we don't do the further tuning here, we will take advantage of this useful insight in future work with real-world observations and attempt to further optimize the algorithm to maximize skill.

It is worth noting that most ML studies have only the test data with which to estimate generalization skill. Here, we also have unseen data. The plot above shows that test and unseen errors are similar, and thus quoting the test error as a proxy for generalization error appears to be reasonable, if slightly optimistic, for real-world studies where unseen data are not available. More investigation of test and unseen statistics is warranted to better inform real-world uncertainty estimates.

*We added **Supplementary Text A** and Supplementary **Figure S6** (the figure above), and included the following text in the revised manuscript (lines 417-423):*

*The predicted pCO_2 is thus more accurate in areas similar to and surrounding the SOCAT "observations" (i.e., monthly 1°x1° grid cells equivalent to SOCAT coverage, but sampled from the LET). **Figure 3** shows mean bias and RMSE for the full reconstruction (see **Section 2.3**), but note that there is a statistically significant difference between the train and test set errors (**Fig. S6**). This indicates potential overfitting in our ML model (i.e., higher errors for the 'unseen' reconstruction), and that further tuning of the hyperparameters could increase generalization skill (see **Supplementary Text A**).*

Supplementary Text A:

*"The hyperparameters for the XGB algorithm used in this study were fixed for all experiments. As we are comparing how sampling impacts the reconstruction, changing the decision trees and depth levels for each experiment would make it difficult to assess whether or not potential changes in bias and RMSE are due to the different sampling strategies or the optimization process. However, **Figure S6** demonstrates a statistically significant difference between train and test set error for the 'SOCAT-baseline' experiment, which may indicate overfitting in our ML model. This suggests that further tuning of the hyperparameters of our ML algorithm might increase generalization skill, and thus reduce test and 'unseen' reconstruction errors.*

However, further tuning of the algorithm is not the purpose of this study nor is it necessary for the evaluation performed here. As the only factor we change between the experiments is additional Southern Ocean sampling (e.g., the SOCAT mask, algorithmic approach and hyperparameters are the same), we can compare the experiments and understand how different sampling patterns and strategies would change skill in pCO_2 reconstructions compared to SOCAT-sampling only.

***Figure 3** (in main text) shows that errors are higher in locations where SOCAT observations are scarce, such as in the Southern Ocean. The improvements in bias and RMSE when USV observations are combined with SOCAT generally occur at times/locations where errors were originally high, and in the Southern Ocean where the new USV "observations" originate from (e.g., **Figs. 4 and 7**). The additional USV observations thus most improve*

*predictions in surrounding areas because the augmented training set contains these similar points. However, note that the lower error values shown for the training set do not impact the final error metrics presented in our study, as ~ 99 % of the reconstruction consists of 'unseen' data points (**Figs. S1, S2**)."*

From a statistical point of view, different mapping methods learned on different model testbeds (i.e. different training data have different data ranges) probably result in different magnitudes of RMSE or Bias. It is not convincing to mention that their mapping method has error values in line with those in the previous study.

We agree that this comparison is not precise. We wanted to show that our error values were not significantly different compared to related studies.

Again, in the following sentence and others in the text, please be careful using the phrase "Observation-based data products". Precisely, "mapping methods" have been developed to estimate pCO2 and generate global "Observation-based data products". Lines 50-52 (in the manuscript with track changes): "Observation-based data products have been developed to estimate full-coverage surface ocean pCO2 across space and time by extrapolating to global coverage from these sparse *SOCAT observations."*

We have replaced "observation-based data products" with "mapping methods" (lines 50 and 212).