

*We would like to thank the handling editor and reviewers for highly constructive and helpful comments and feedback. We have carefully addressed each question/comment and made changes where we agree that this would improve the manuscript. We have provided an itemized list below detailing our responses (in italic font) to the reviewer's suggestions. **Line numbers mentioned in this document refers to the tracked changes version of the manuscript.** All new supplementary figures in the revised version of the manuscript were added based on the reviewers' feedback, and the justification and explanation of the additional figures are provided within the responses to the reviewers below.*

Response to RC1

I have some reservations about the way the authors have used their method.

I would expect a more quantitative estimate of the magnitude of the differences potentially detected between the different experiments performed.

The results and discussion of the decadal variability of the oceanic CO₂ sink need to be improved (see specific comments).

Specific comments:

- 3) To calculate the residual pCO₂, the authors used the equation in line 160. In this equation, the term pCO₂^{mean} came from “surface ocean pCO₂ [...] using all 1°x1° grid cells from the testbed” (line 162). But in the original publication of the methodology (equation 2 in Bennington et al., 2022a), the term pCO₂^{mean} comes from an initial reconstruction of pCO₂. Therefore, in this submitted manuscript, by using model outputs instead of an initial reconstruction of pCO₂ fields, the authors assumed that their method would be able to perfectly reconstruct the long-term average pCO₂ at each grid cell. It seems to me that to obtain a more accurate evaluation of their method (i.e., more accurate observing system simulation experiments) the authors should follow the steps as they were originally

published. If this is not possible, could the authors explain why and prove that this assumption does not influence their results.

The reviewer raises a valid question that the method used here varies slightly in this way from the method presented in Bennington et al. (2022), given the testbed approach utilized here. However, previous work has found that this small difference in methodology does not have a large impact on the result. Bennington et al. (2022) did a sensitivity test of the pCO₂-Residual reconstruction to the source of mean pCO₂, by experimenting using the Takahashi pCO₂ climatology (Takahashi et al., 2009) as well as the mean pCO₂ of the SeaFlux observation-based products (Fay et al., 2021; Gregor & Fay, 2021). They found that alternative sources of the initial pCO₂ map had little influence on the reconstruction. For this reason, we have chosen to increase the efficiency of our data processing pipeline by using the full model field as our mean pCO₂ to calculate pCO₂-T and pCO₂-Residual.

2) To calculate the net sea–air CO₂ flux, authors used (line 272): “EN4.2.2 salinity (Good et al., 2013), SST and ice fraction from NOAA Optimum Interpolation Sea Surface Temperature V2 (OISSTv2) (Reynolds et al., 2002), and surface winds and associated wind scaling factor from the European Centre for Medium-Range Weather Forecasts (ECMWF ERA5 sea level pressure (Hersbach et al., 2020)”. But as mentioned in line 95: “The goal here is to assess the accuracy with which an ML algorithm can reconstruct the ‘model truth’”. Therefore, I would expect the model outputs (some of which have already been used for pCO₂ reconstruction) to be used to calculate the CO₂ flux, rather than observational data which may have different variabilities and/or trends to those simulated. The authors could then compare this calculated CO₂ flux to the simulated CO₂ flux (and not to a “model truth” CO₂ flux from simulated pCO₂ fields mixed with observational data). This is particularly important when the authors are discussing the ability of their method to reproduce CO₂ flux variability (see my next comment).

We completely understand the reviewer’s point of using model output instead of observational data to calculate flux. However, it is the winds that have the largest impact on flux calculations (Fay et al., 2021), and temporally high-resolution output is not available for the testbed. Only

monthly model output is available, and this is not sufficient for the flux calculation due to the square dependency of wind speed. We therefore used the ERA5 wind product, a choice consistent with Gloege et al. (2021) who also used the Large Ensemble Testbed to reconstruct pCO₂. Given the necessity to use observed winds, we also use observations for all necessary variables for the flux calculation (Fay et al., 2021), instead of mixing model output and observations.

Further, we wish to emphasize that the goal of this project is not to calculate real-world fluxes, but, instead, to better understand how sampling impacts the resulting pCO₂ fields and from pCO₂, the flux. For our study, the most important factor is to calculate consistently for all the experimental runs so that we can make direct comparisons. Therefore, using the same inputs to the flux calculation for each of the three models is also desirable to isolate this comparison. It would certainly be interesting to compare fluxes calculated by different methods (observations vs. model output), however this would be beyond the scope of this paper as we are not evaluating methods of flux calculation, but rather evaluating the impacts of sampling.

3) Authors wrote line 531: “The SOCAT baseline demonstrates a weakening of the global and Southern Ocean carbon sink in the 2000s (Figs. 10, S12), which is in agreement with various data products using real-world SOCAT data”. The weakening of the Southern Ocean carbon sink occurred in the 1990s (Le Quéré et al., 2007), while a reinvigoration of the sink was observed during the 2000s (Landschützer et al., 2015). The authors therefore need to revise their text. More importantly, this study focuses on the ability of the authors’ method to reproduce “model-truth” variability and not the “real-world” variability. Consequently, I would suggest calculating certain metrics of variability (for example, the size of decadal variability or trends) from simulated CO₂ fluxes (and not from recalculated “model-truth” CO₂ fluxes, see my previous comment) and comparing the values of these metrics with the values that would be obtained when reconstructed CO₂ fluxes are used. Because, otherwise, it assumes that all models perfectly reproduce the variability of the ‘real world’, which might not be the case.

We were referring to the distinct “peak” of the weakening of the sink that can be seen around the year 2000, however, we have re-phrased this sentence as suggested by the reviewer (lines 954-956):

“The ‘SOCAT-baseline’ demonstrates a weakening of the global and Southern Ocean carbon sink starting in the 1990s with a peak around year 2000 (Figs. 10, S18), which is in broad agreement with various data products using real-world SOCAT data (e.g., Gruber et al., 2019; Landschützer et al., 2015; Bushinsky et al., 2019; Bennington et al., 2022; Gloege et al., 2022)”.

We agree with the reviewer that diving deeper into understanding the flux variability, and comparing fluxes based on the testbed vs. observations would be valuable and we appreciate their suggestion. We believe however that this deserves a more in-depth discussion that will be best presented as an individual paper, and we are planning to explore this further in a future study (this is mentioned in the discussion (lines 1009-1010): “we will further explore this issue in future work”). To avoid a lengthy discussion, we would like to restrict the main focus of this study to assessing the impacts of sampling by using the testbed.

4) The LET has 75 members (i.e., simulations). For each experiment, the values given in the manuscript and in the figures are for the most part averages calculated over the 75 members of the ensemble. But no information is given on the dispersion (or confidence interval) around these averages. It is therefore not possible to assess whether the differences mentioned between the experiments are significant or not.

For example:

- The interpretation of Figure 5 (line 335): “The ‘one-latitude’ ‘high-sampling’ run ‘x13_10Y_J-A’ (44,250 observations) show similar bias or is outperformed by all ‘zigzag’ runs as well as the ‘one-latitude’-runs that restrict sampling to southern hemisphere winter months (i.e., ‘x5_5Y_W’ and ‘x13_10Y_W’).” How similar or superior is the performance? Is it true for all members?
- Line 346: “Run ‘Z_x10_5Y_W’, which has the lowest bias out of the ‘zigzag’ runs (Fig. 5), shows improvement even further back in time, until the beginning of the testbed period (Fig. S6).” Is it really significant?

I would therefore suggest not only reporting the averages over the 75 members, but also taking advantage of the study of the spread around these averages.

We thank the reviewer for this suggestion, and in the revised version we have included additional supplementary figures showing the spread amongst ensemble members (Figs. S8, S10, S14, S16 – these are shown below). Since we are comparing several experiments, it would be difficult to interpret figures showing the spread of 75 members of 10 different experiments, so we chose to keep the figures showing the testbed mean in the main text. It is important to note that in order to fairly compare sampling experiments, it is critical to compare the same ensemble member for each experiment. By that we mean that performance metrics must be calculated based on the same member's 'reconstruction vs. truth pair' for each of the 10 sampling experiments. For example, the 'reconstruction vs. truth pair' for CESM member 001 for experiment 1 must be compared to the 'reconstruction vs. truth pair' for CESM member 001 for experiment 2 and so on. There are 75 members in our testbed, and thus, for each experiment, there are 75 'reconstruction vs. truth pairs'. As shown by our supplementary figures (and additional figures below), overall, the mean calculations reflect the majority of individual members in terms of how the different experiments compare to each other.

*However, we agree with the reviewer that it is important to show the spread. We have tried to make it more clear throughout the text that we are comparing mean values, but that there is a spread. We added this sentence to **Section 2.3** (Statistical Analysis in the Testbed, **lines 250-253**): “We focus our discussion on the mean across 75 members of the testbed for bias and RMSE. The spread across testbed ensemble members is non-negligible and will be the focus of future work; here, we present the testbed spread primarily in the **Supplement**”.*

*Further, a recent study by Hauck et al. (2023) performed similar sampling experiments, but used a different type of reconstruction method and testbed (i.e., a single hindcast model), and show that additional autonomous sampling leads to a weakened Southern Ocean sink, which is the opposite to our findings. This study was not published when we submitted our initial manuscript, but in the revised version we have added a paragraph to the discussion which touches upon the potential importance of the testbed spread (**lines 940-953**):*

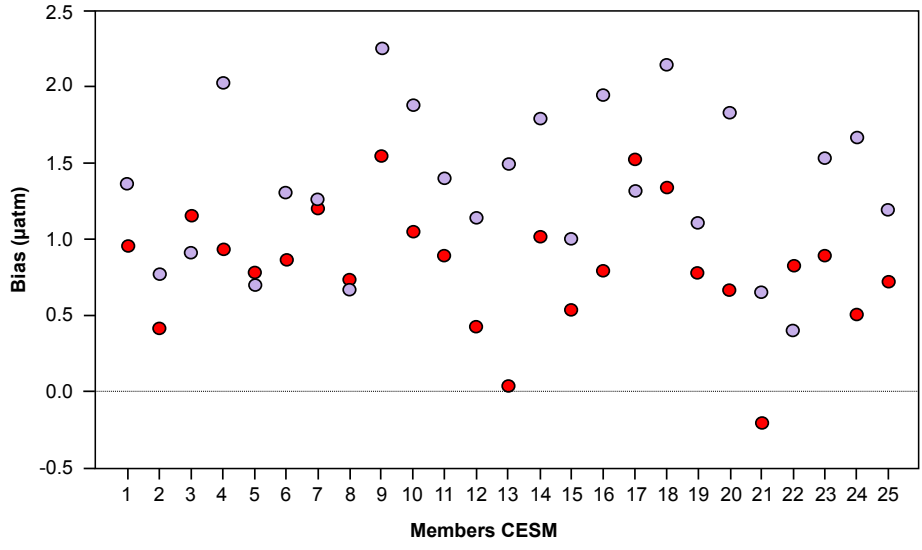
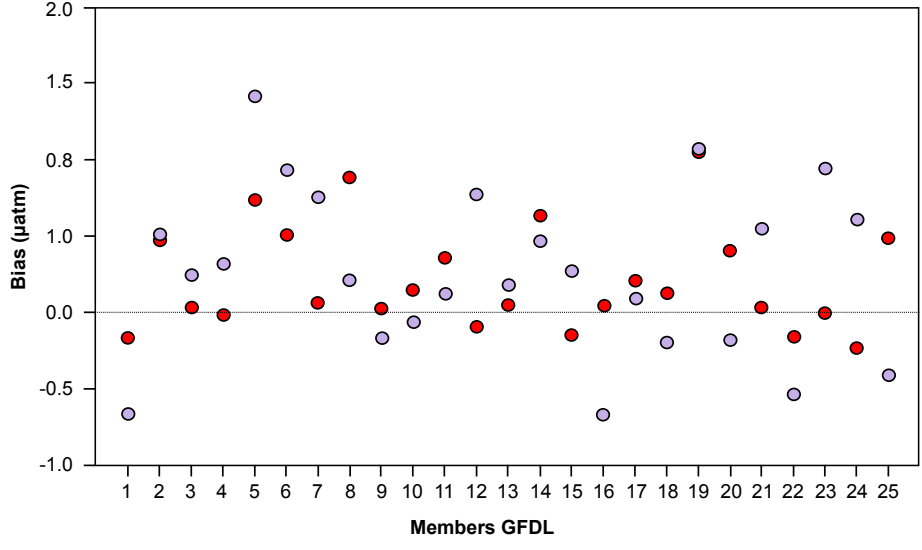
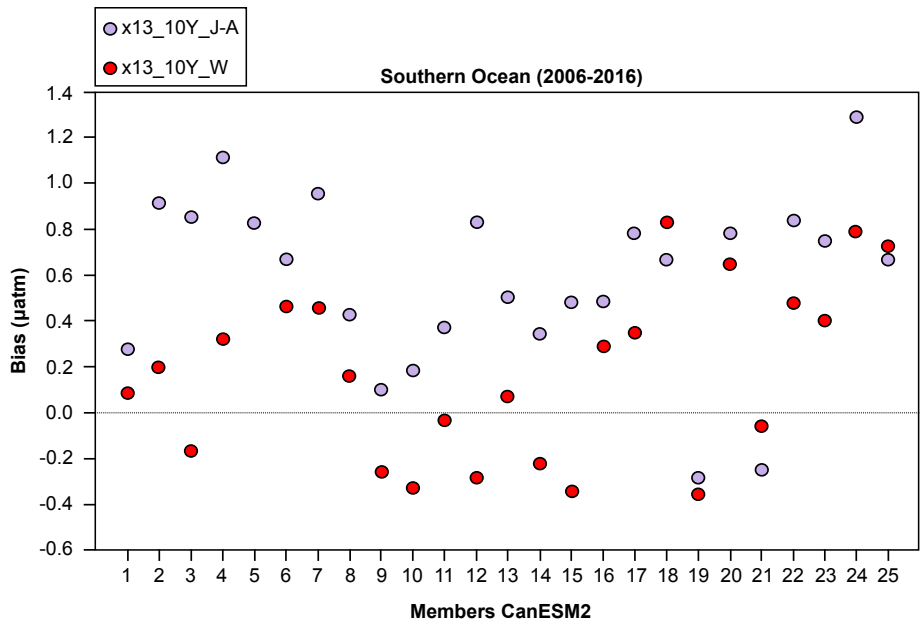
“Bushinsky et al. (2019) and Hauck et al. (2023) performed similar sampling experiments as presented here, by comparing ML surface ocean pCO₂ reconstructions based on SOCAT vs.

additional SOCCOM or ideal virtual floats. These studies showed that SOCAT sampling alone overestimates the CO₂ uptake in the Southern Ocean, and that additional floats reduce this overestimation, leading to a decreased (weakened) ocean carbon sink. In contrast, we find that the pCO₂-Residual method underestimates the CO₂ uptake with only SOCAT sampling, and that adding USVs increased (strengthened) the Southern Ocean and global ocean sink by up to 0.1 Pg C yr⁻¹ (Figs. 10, S18; Table S2).

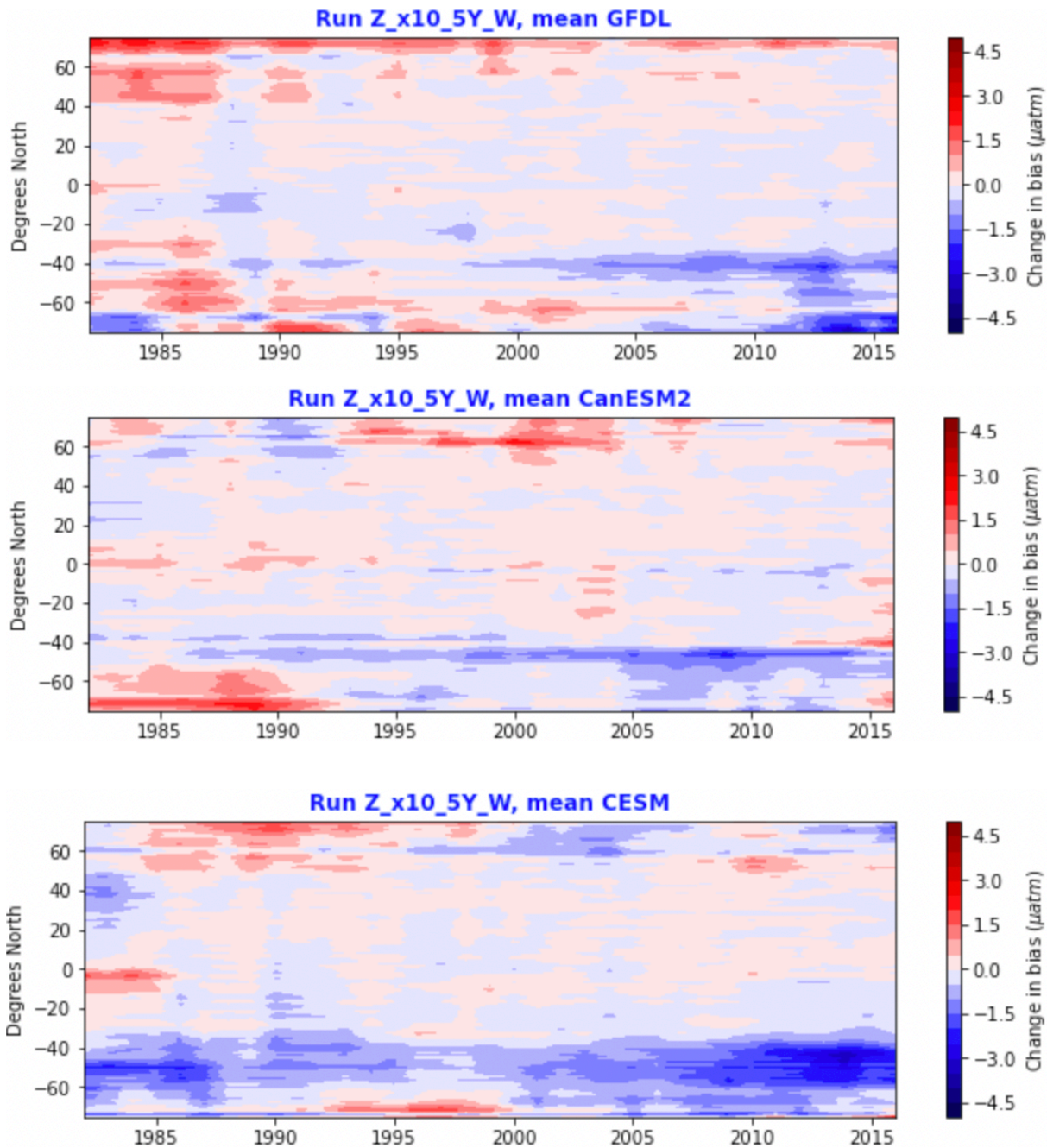
Going forward, additional studies are needed to better understand why these results suggest a different direction of the sink change with additional sampling. These differences could stem from the use of different reconstruction methods assessed. Hauck et al. (2023) used the MPI-SOM-FFN and CarboScope/Jena-MLS reconstruction methods, while we use the pCO₂-Residual method. Another substantial difference between the studies is the models and numbers of ensemble members used as the testbed. Hauck et al. (2023) use a single hindcast model, while we use 25 members each from three Earth System Models. We find substantial spread across these 75 members (Figs. S8, S10, S14, S16), indicating that model structure and internal variability significantly impact results. Our study and Hauck et al. (2023) use different approaches for the calculation of fluxes, which could also be a factor. Targeted, coordinated studies using multiple reconstruction approaches with consistent testbed structures and experimental approaches are clearly needed (Rödenbeck et al., 2015). Despite this need for this additional work, studies do agree that additional Southern Ocean observations could significantly improve reconstructions of air-sea CO₂ fluxes”.

Answers to the reviewer's specific questions above:

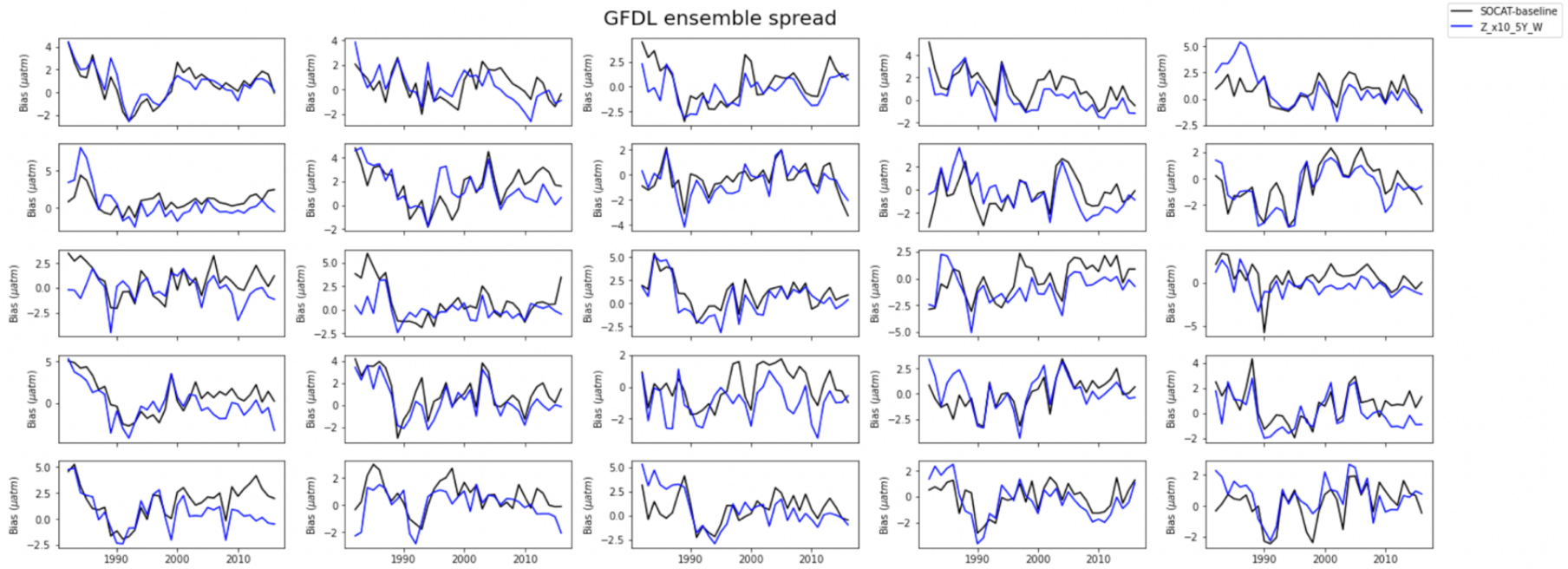
*1. Below, we show the bias (over the Southern Ocean for the period of 2006-2010) of each individual member of the models in the testbed, comparing the high-sampling run 'x13_10Y_J-A' with the equivalent run that restricts sampling to southern hemisphere winter months ('x13_10Y_W'). As shown by the figure below, the majority (~ 80%) of members for run 'x13_10Y_W' (winter sampling) outperform (i.e., have a bias closer to zero) those of run 'x13_10Y_J-A' (Jan-Aug sampling), reflecting the ensemble means shown in **Figure 5**.*

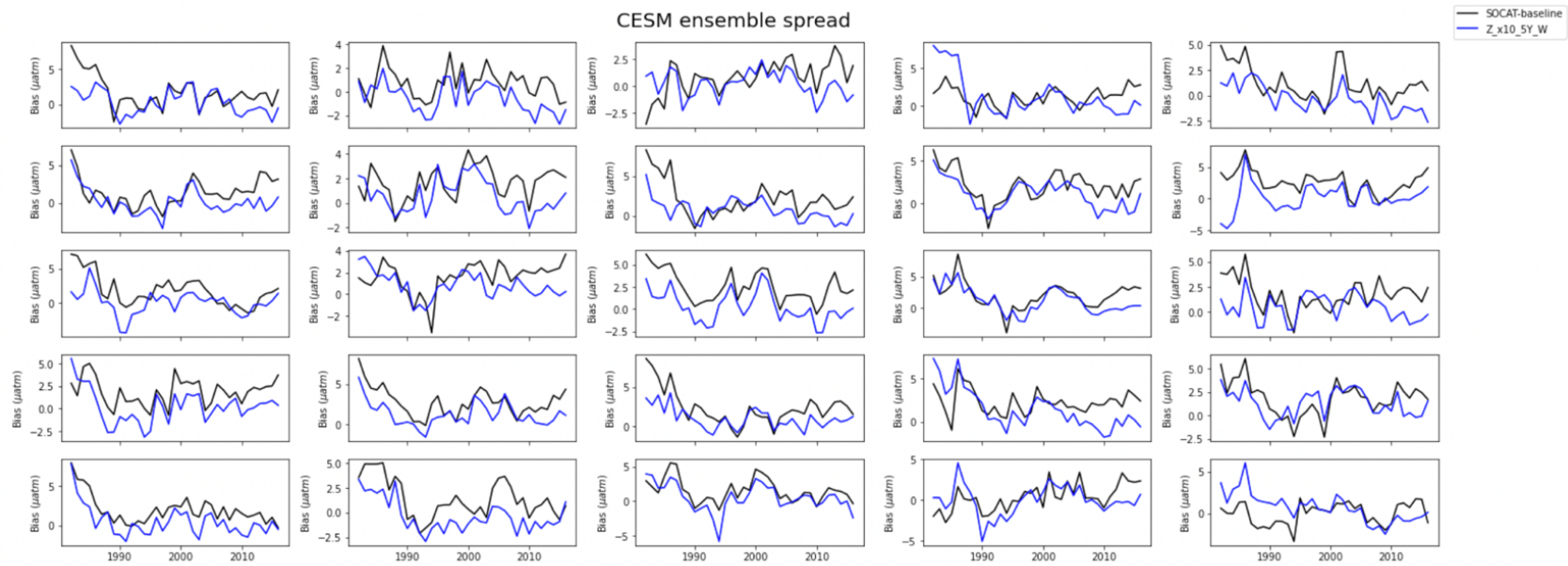
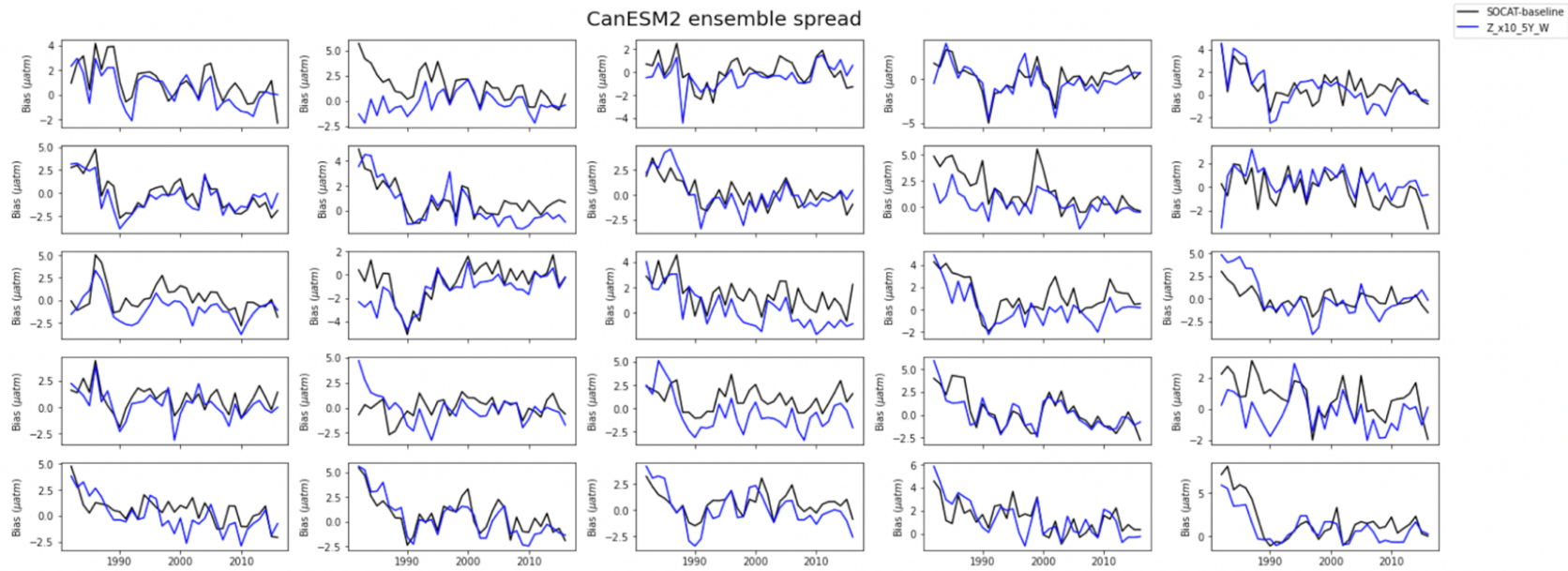


2A. Below, we present zonal annual mean Hovmöller plots showing the change in bias when comparing run 'Z_x10_5Y_W' to the 'SOCAT-baseline'. As shown by the figure below, all models show improvement back in time beyond the additional sampling duration (2012-2016), reflecting the ensemble mean shown in **previous Figure S6 (new Figure S9)**, but there is less improvement for GFDL members compared to CESM and CanESM2.

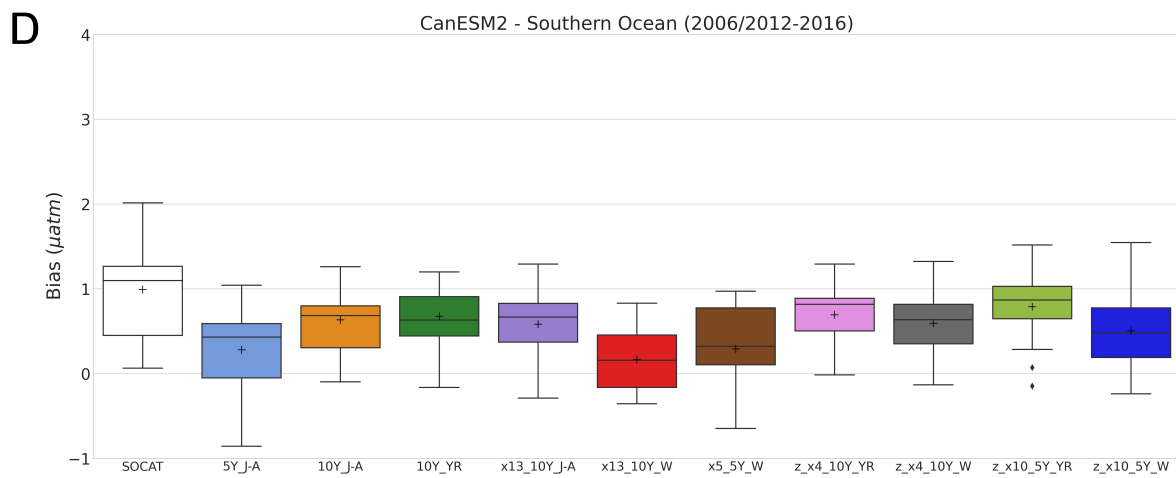
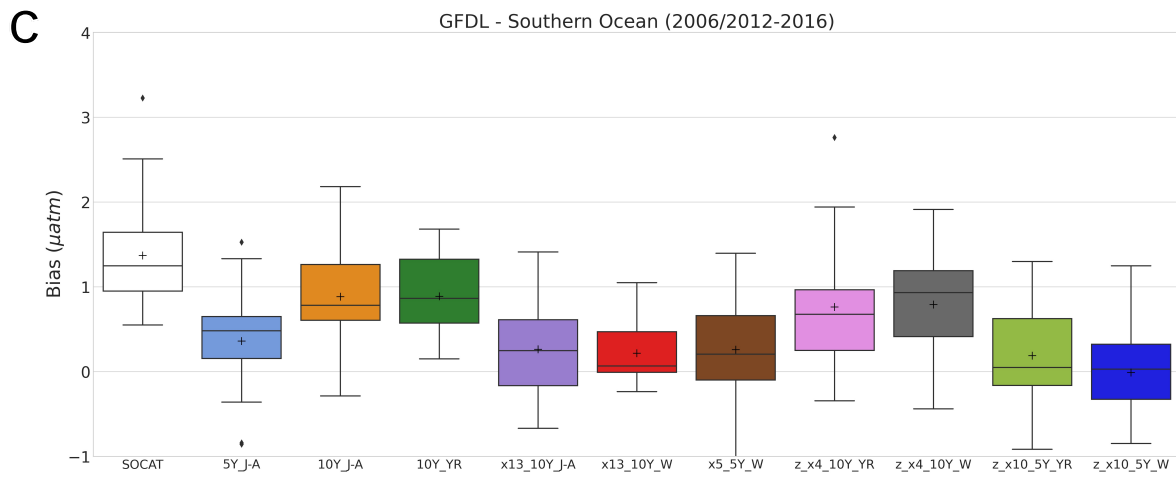
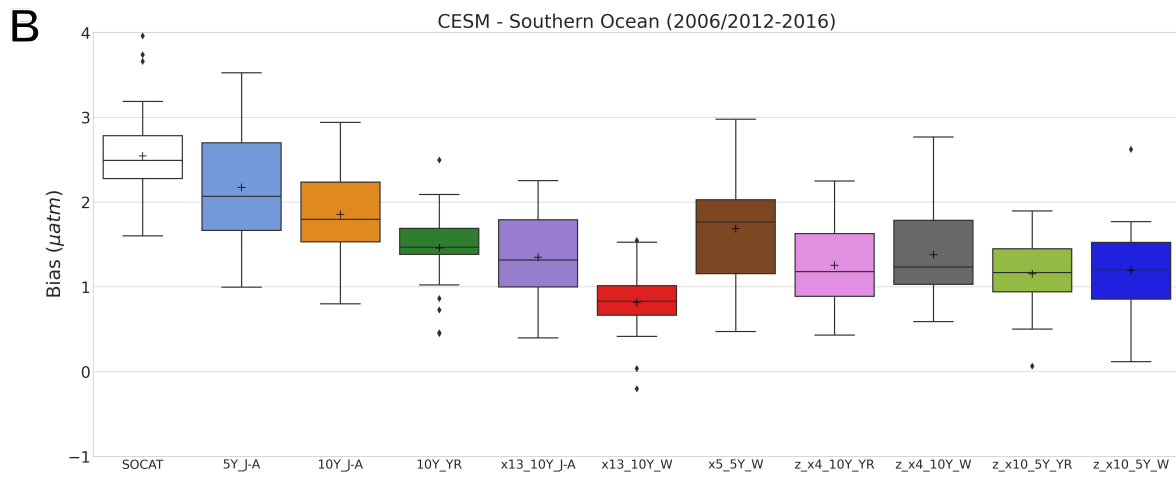
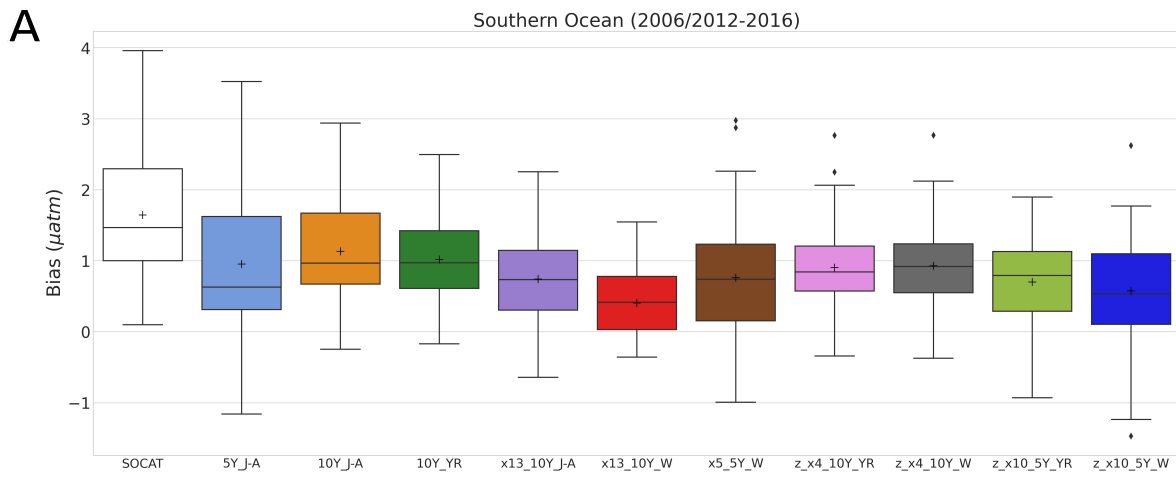


2B. To examine individual members, we plot time series of bias for run 'Z_x10_5Y_W' and the 'SOCAT-baseline' averaged over the area of highest improvement shown in **previous Fig. S6** (between 50°S and 35°S ; **new Figure S9**). These figures show improvement in bias compared to the 'SOCAT-baseline' already in the beginning of the testbed period for the majority of members, but more so for CESM and CanESM2 compared to GFDL.

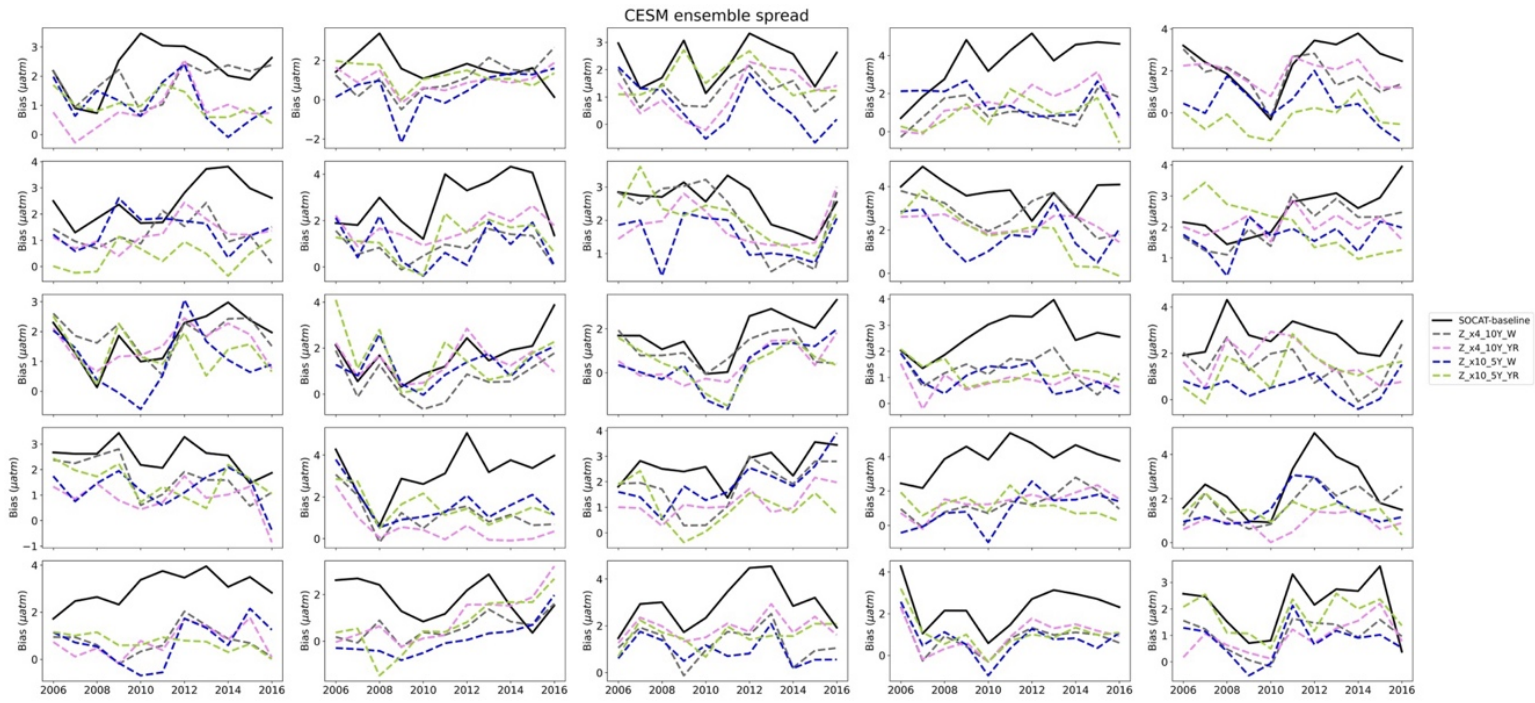




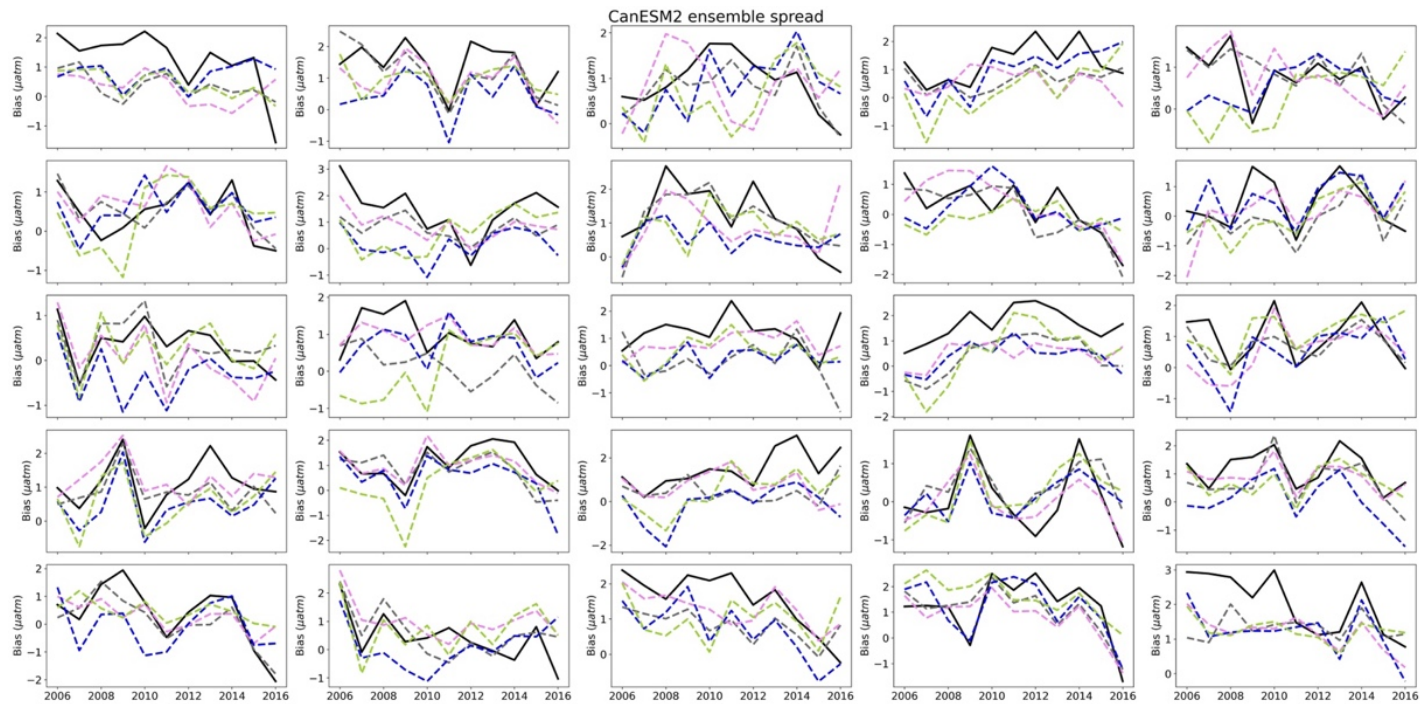
Overview of new supplementary figures showing the ensemble spread (S8, S10, S14, S16):



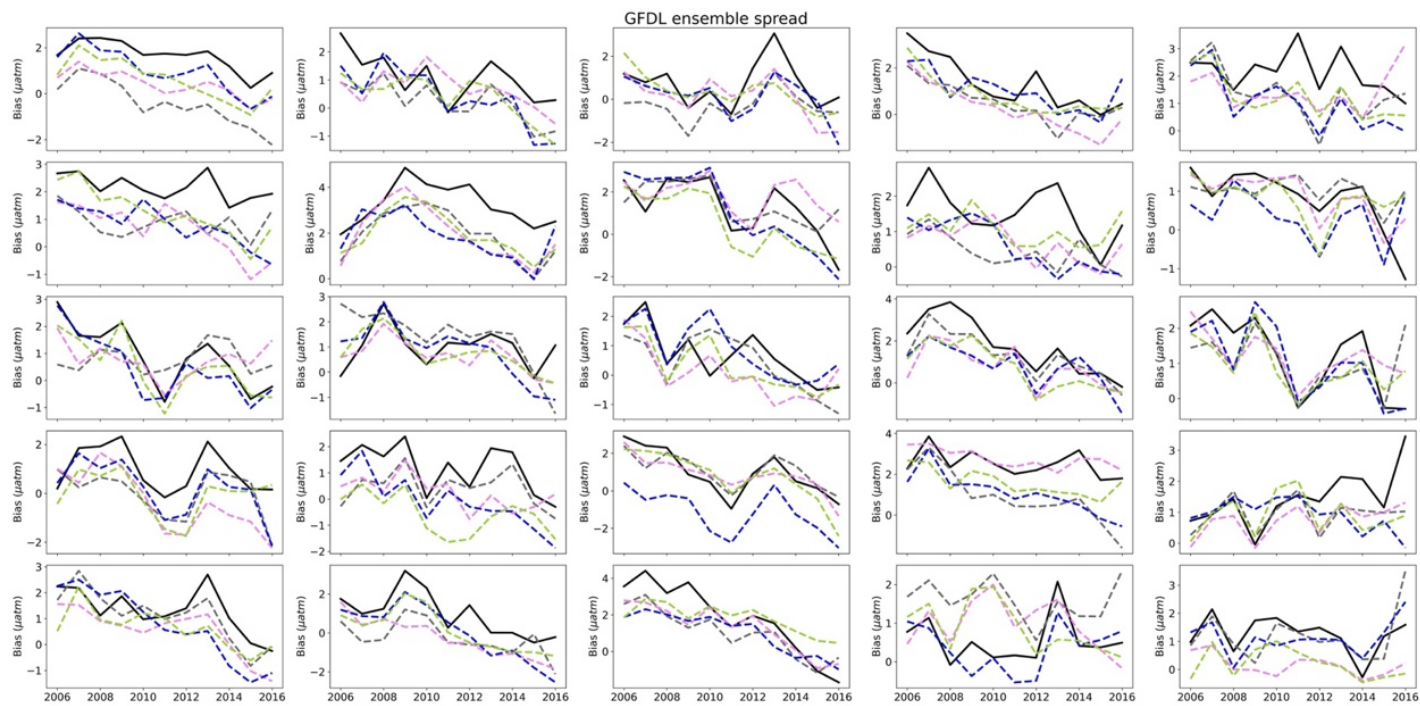
New Fig S8



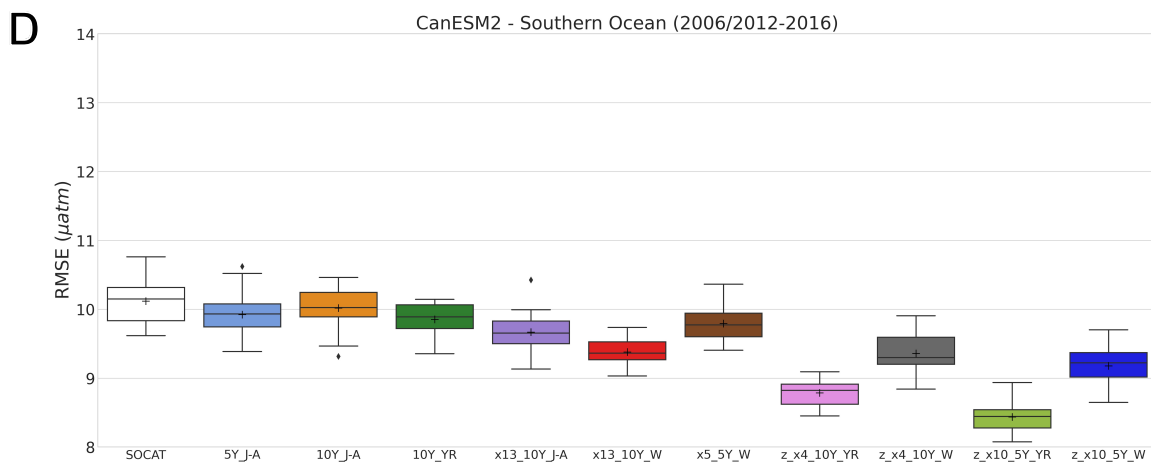
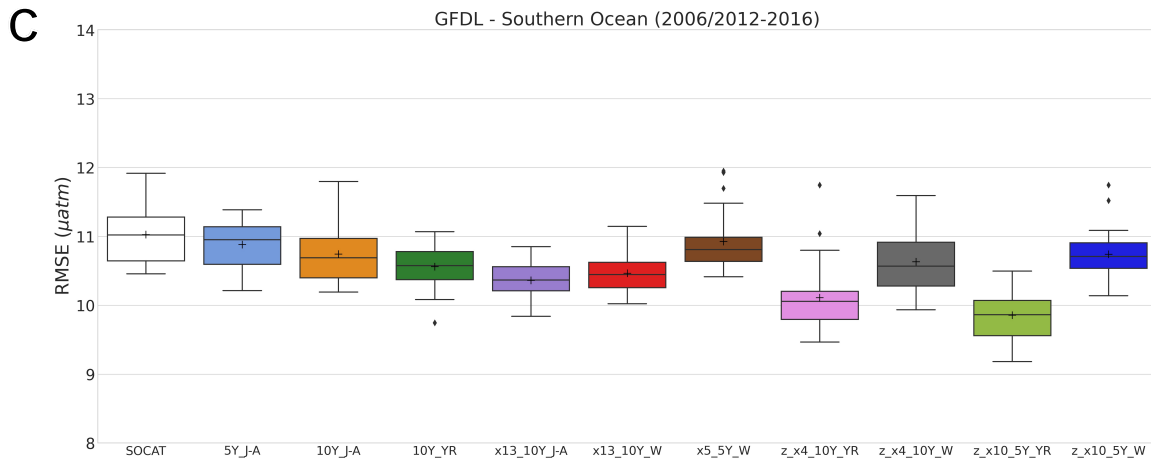
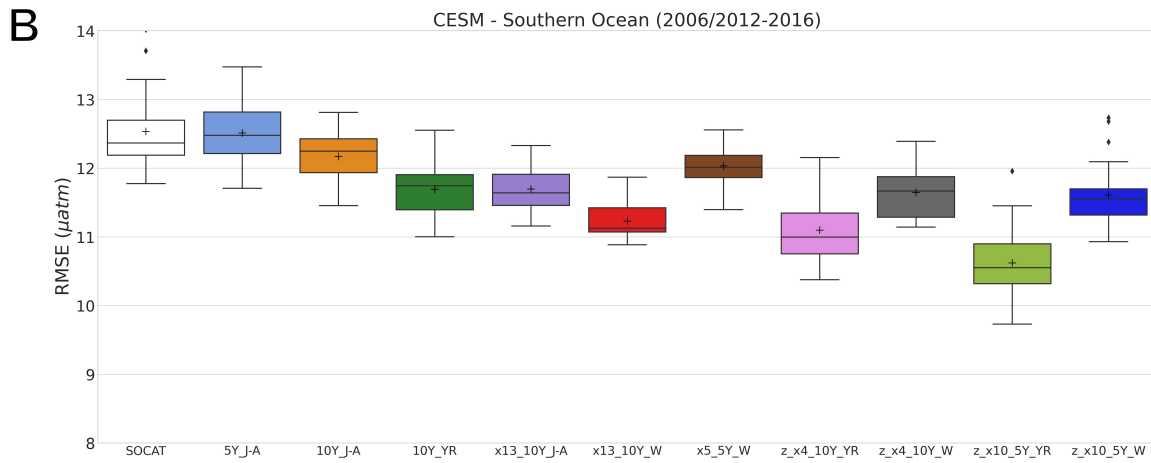
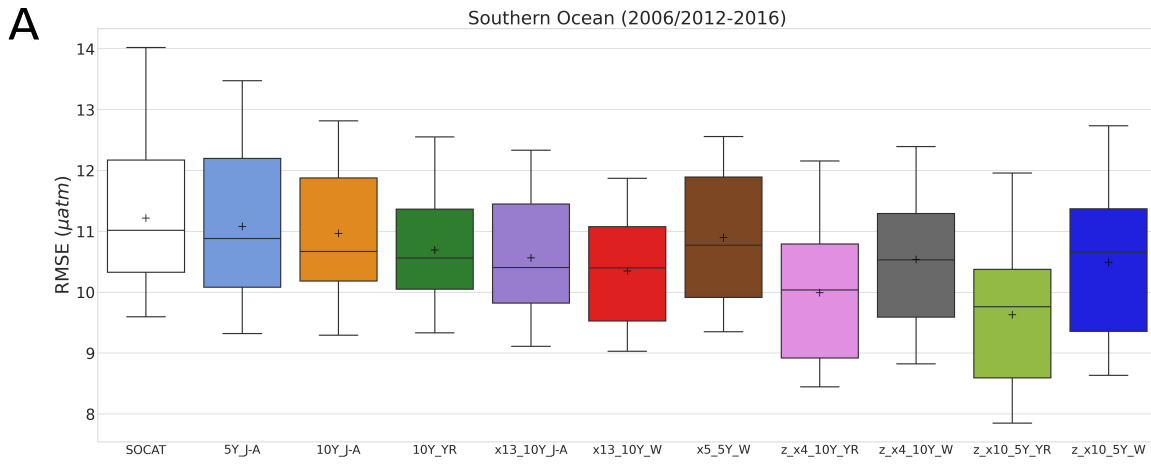
New Fig S10

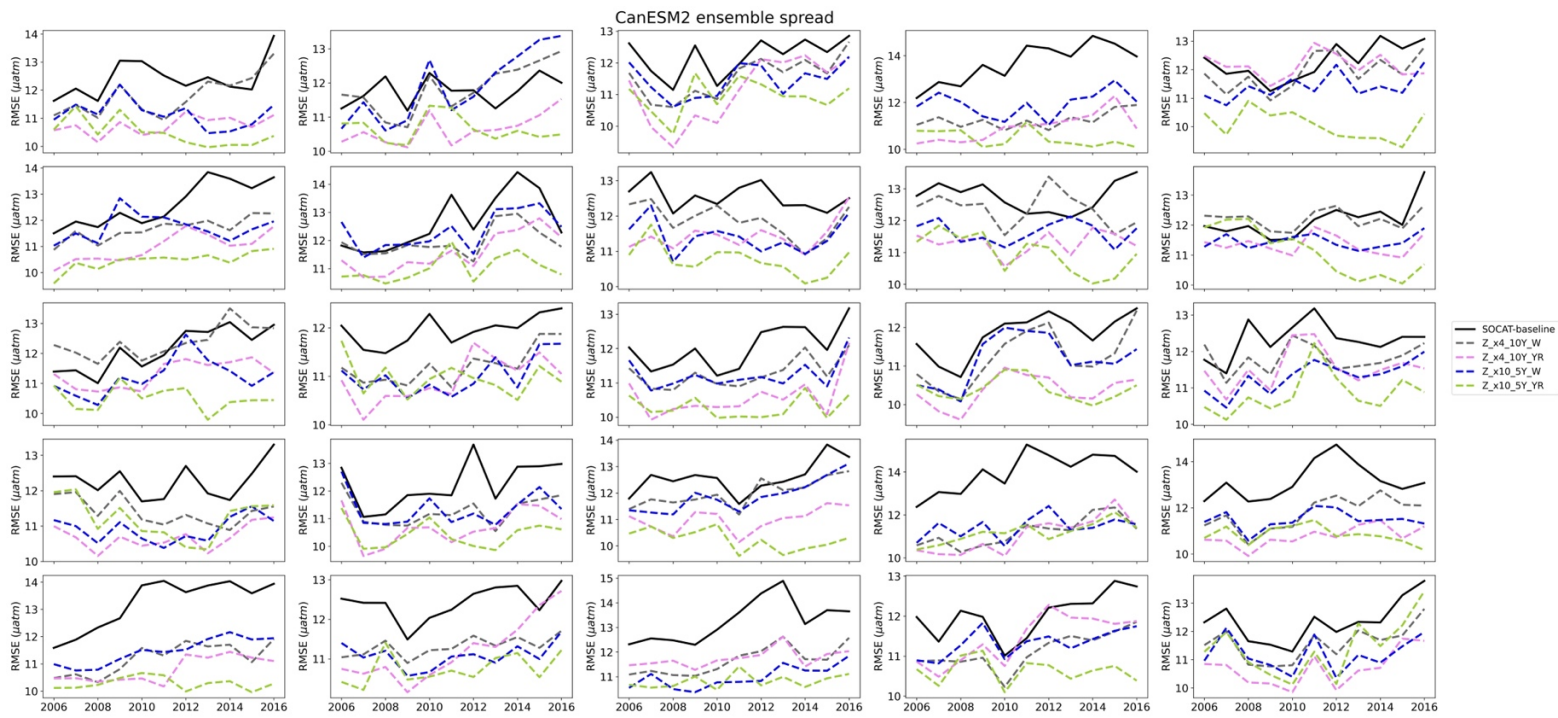


New Fig S10 cont.

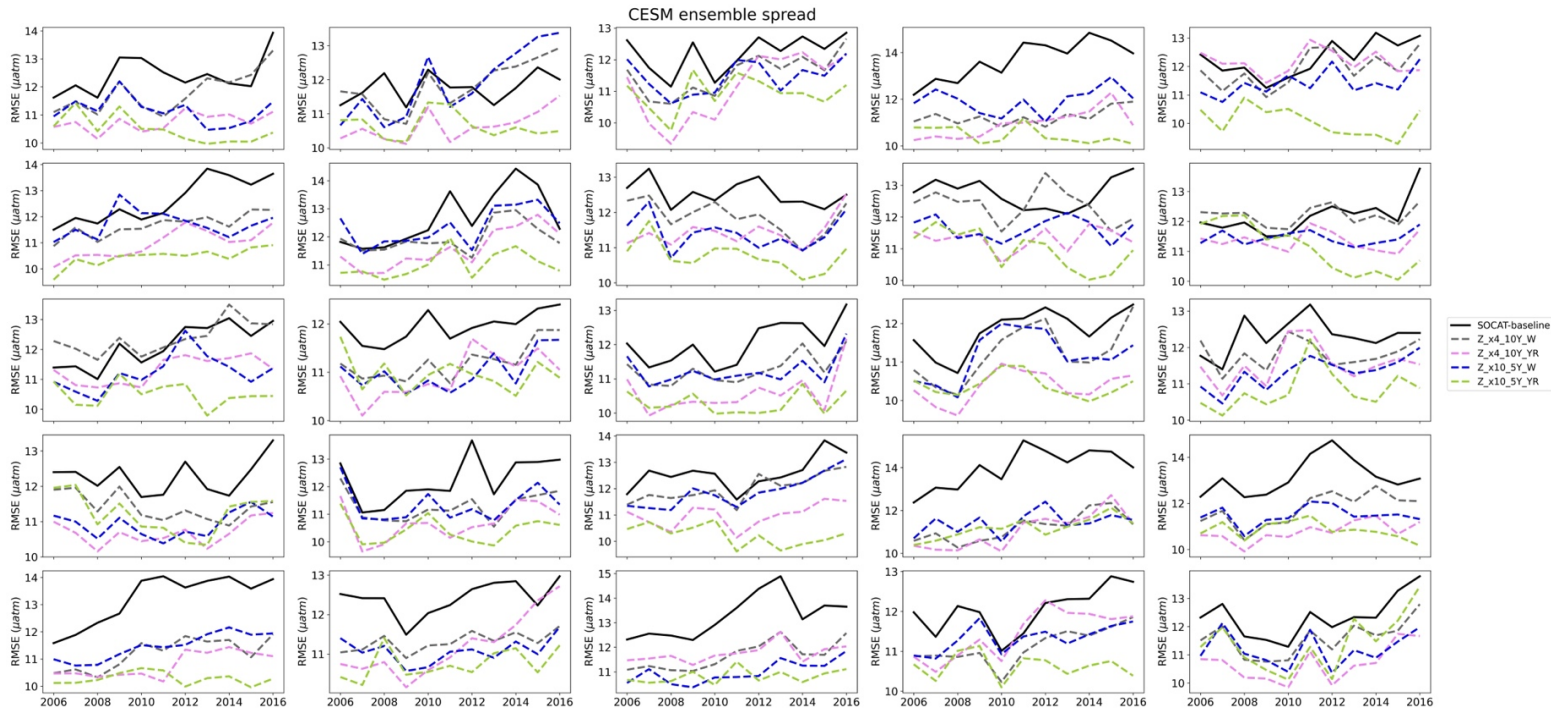


New Fig S10 cont.

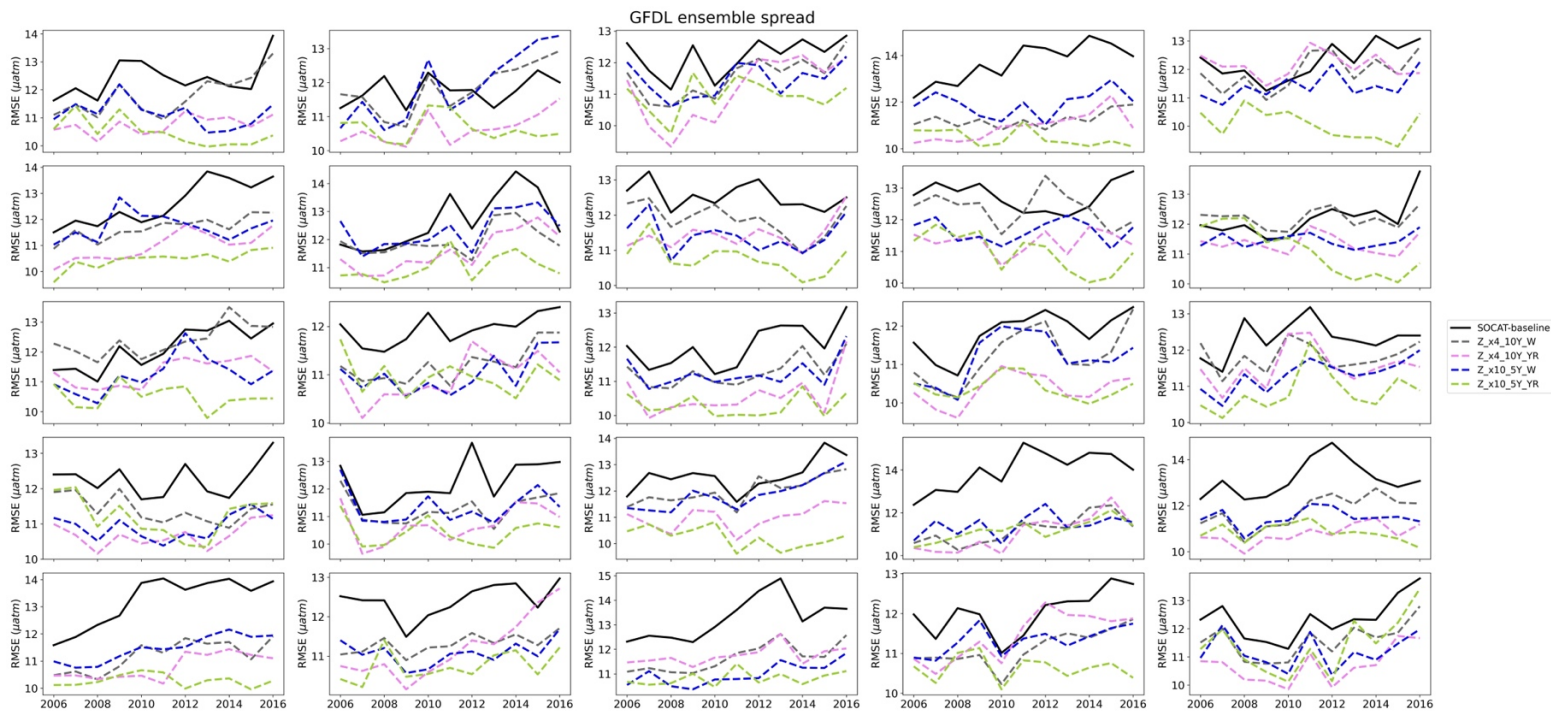




New Fig S16



New Fig S16 cont.



New Fig S16 cont.

Technical corrections:

5) Line 37, Please explain the acronym "fCO₂". Note that the term "pCO₂" is also used in the manuscript. Although understandable to researchers working on this topic, it is less clear to a wider audience. Therefore, authors should be more careful about the terms they use, especially in the Abstract and Introduction sections.

In the revised version we have defined fCO₂ (line 46), which is the fugacity of carbon dioxide (fCO₂) as opposed to pCO₂ which is the partial pressure of CO₂ in the ocean. The fCO₂ is equal to the pCO₂ corrected for non-ideality of CO₂ solubility in water using the virial equation of state (Weiss 1974). The fugacity correction for surface water is 0.996 and 0.997 at 0 °C and 30 °C respectively (Dickson et al. 2007), or 0.7 to 1.2 μatm lower than the corresponding pCO₂, and depends primarily on temperature for the conversion, although pressure is also included in the conversion equation. It is common practice in the observational community to report values as fCO₂ as this is what is released in the SOCAT database, but model output is typically reported as pCO₂ which is why we have chosen to go with that variable in this study.

6) Line 178, Why aren't the number of decision trees and depth levels different for each reconstruction?

The depth levels and decision trees are fixed, which we have now stated in the main text (line 222). The depth levels and decision trees used represent the optimized parameters for this type of reconstruction. The dominating input for all experiments is based on the SOCAT coverage, and the different USV experiments represent a small increase in the data density. Further, increasing the maximum depth level would make each decision more complex, making the final algorithm less generalizable. Adding more trees is not necessarily going to improve the overall algorithm. Finally, 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.

7) Line 188, After reading this sentence, I wasn't sure whether the authors were always going to use the "unseen" values. Could the authors be clearer?

This should have been communicated more clearly, and we have now revised this sentence and added some more information (lines 231-236): “Here, we calculate error statistics based on the full reconstruction (pCO₂ from all 1°x1° grid cells of the testbed, except for those masked or filtered out). In the full reconstruction, ~ 99 % of the data do not correspond to SOCAT or Saildrone USV observations used to train the algorithm (Fig. S1). Training data would ideally be removed before performance evaluation, but since the training data represent only ~ 1 %, the impact of not removing them is negligible (Fig. S2)”.

8) Line 203: “2) potential future meridional USV observations (‘zigzag’ track)”. Are they realistic? I found some elements of response later in the text, but it would be good to know here whether all the experiments are realistic or not.

*The reviewer raises an important question, and as pointed out, we touch upon this in **Section ‘2.4.2 Zigzag runs’** and in the discussion. The potential future meridional USV track has been developed in collaboration with experts from the ocean observing community to test realistic sampling. Due to the USV technology, Saildrones can sample meridional gradients, as opposed to other autonomous platforms. Further, we account for limiting incoming solar radiation to power the Saildrone below 55° S. **Section 2.4** is meant to provide an overview of the different type of experiments we have performed. This section already provides a lot of information, and in order not to exhaust the reader with details, we chose to focus on the details under **Section ‘2.4.2 Zigzag runs’** instead. However, we added the word “realistic” (line 259) and refer to further information in **Section 2.4.2 (line 260)**. We also added some more information under section **‘2.4.2 Zigzag runs’ (lines 290-294)**: “Saildrone USVs can operate at a speed capable of covering the spatial extent of meridional gradients in the Southern Ocean (Djeutchouang et al., 2022). However, Saildrone USVs are solar powered, and thus their range is restricted by the availability of solar radiation. To account for this and maintain a realistic sampling scenario, sampling occurs only to a maximum latitude of 55° S in these experiments”.*

9) Table 1: I suggest replacing table 1 with table S1. This is because the information in table 1 is repeated in table S1, and table S1 contains important values that the reader should be able to access easily.

This has been replaced in the revised manuscript.

10) Line 268, Why not use the same method across all models to calculate $p\text{CO}_2^{\text{atm}}$? Do all the values obtained take into account the contribution of water vapor pressure?

*The reviewer raises a valid question. The reason for this is that the GFDL model output that we have access to includes the $p\text{CO}_2^{\text{atm}}$ variable, while for CanESM2 and CESM we do not have this output variable. Therefore, the atmospheric value had to be calculated for these two models. Each individual model defines its own atmosphere concentration, and some models account for water vapor pressure and others do not when running their model. In GFDL and CESM, the contribution of water vapor pressure is taken into account, but this is not the case for CanESM2. Thus, when calculating $p\text{CO}_2^{\text{atm}}$ for CanESM2 and CESM, the contribution of water vapor pressure was taken into account for only CESM. We now specify that “the contribution of water vapor pressure was corrected for in CESM and GFDL” (**lines 380-381**).*

11) Line 293: “where algorithm generally overestimates $p\text{CO}_2$ ”. This is not the case for the Atlantic sector of the Southern Ocean.

*With this statement we were just trying to convey that, overall, $p\text{CO}_2$ is generally overestimated in the Southern Ocean, however, the reviewer is correct that parts of the Atlantic section show an underestimation. We have revised this sentence (**lines 407-409**): “RMSE is highest in the Eastern Tropical and Southeastern Pacific Ocean and in the Southern Ocean, where the algorithm generally overestimates $p\text{CO}_2$ (i.e., positive bias; **Fig. 3a**), with some exceptions in the Atlantic section”.*

12) Figure 3, colour scale: The colour scales need to be harmonised. In panel a, a white colour means a good value, whereas in panel b, it means a bad value.

*We agree with the reviewer, and we tested several different colormaps, however, if we switch the colors in **Fig. 3** (i.e., dark color equals “worse”), we would have the same problem in our maps showing our main results (**Figs. 4, 6, 7, 9**). These maps do not show RMSE for each USV experiment, but rather the difference in RMSE between the experiments and the ‘SOCAT-baseline’.*

*We could choose a completely different colormap for RMSE in **Fig. 3**, but for consistency, chose to use the same range of colors for RMSE (and bias) throughout the paper.*

13) Figure 3, line 301 to 307: All this information is already present in the text. Please write shorter figure captions. This is a general comment, not just on figure 3.

Noted, and revised.

14) Line 318 and wherever necessary in the text: "...where the baseline reconstruction..." Please, use the expression "SOCAT baseline" that was introduced in the method section.

Noted, and revised.

15) Line 384, Please delete the reference to "bias". This was introduced in the previous section.

Noted, and revised.

16) Line 493, why not excluding the hypothetical data points that would be covered by sea ice?

The seasonal ice coverage in high latitudes varies, and the sea-ice fraction is uncertain. We chose to show the map of the global sea-ice extent as defined by the SeaFlux product, which is from NOAA OISSTv2 (Reynolds et al., 2002) as an example. Since the sea-ice fraction is uncertain and varies by month, we chose to show where reconstructions could significantly improve regardless of potential ice coverage. If current/future technology allows for sampling in these high-latitude areas it is important to know the extent of the potential improvement.

17) Figure 10: The figure starts in 1985 and not 1982, why?

*The flux calculations begin in 1985 because this corresponds to the earliest SeaFlux inputs. We now add mention of the 1985 start in **Section 2.5 (line 369)**.*

18) Figure S3: Because you focused on the open-ocean (line 123), non-open-ocean data should be removed as they were not use for the training, is it right? Does this drastically modified the data availability and explain why better results are obtained from 1990?

Testbed output for coastal areas, the Arctic Ocean and marginal seas were removed before training in all experimental runs, and also when comparing the experiments to the testbed truth when calculating bias, RMSE and air-sea flux. As shown in **Figs. 3, 4 and 7** (and equivalent figures in the supplement) the white areas represent areas of no data as this was removed. Better results are likely obtained from 1990 because, as shown by **previous Fig. S3 (new Fig. S5c in the revised version)**, SOCAT observations start to drastically increase from these times. This was mentioned in the manuscript (**lines 494-496**): “Considering the change in bias from year-to-year, the ‘SOCAT-baseline’ shows positive bias at all latitudes in the beginning of the testbed period, before improvement occurs around 1990 (**Fig. 6a**). This is consistent with increasing SOCAT sampling with time for the period considered here (i.e., up to 2016; **Fig. S5c**)”.

References

Bennington, V., Galjanic, T., and McKinley, G. A.: Explicit Physical Knowledge in Machine Learning for Ocean Carbon Flux Reconstruction: The pCO₂-Residual Method, *Journal of Advances in Modeling Earth Systems*, 14(10), <https://doi.org/10.1029/2021ms002960>, 2022.

Dickson, A. G., Sabine, C. L., & Christian, J. R. (Eds): Guide to best practices for ocean CO₂ measurement, Sidney, British Columbia, North Pacific Marine Science Organization, 191 (PICES Special Publication 3; IOCCP Report 8), <http://dx.doi.org/10.25607/OBP-1342>, 2007.

Fay, A. R., Gregor, L., Landschützer, P., McKinley, G. A., Gruber, N., Gehlen, M., Iida, Y., Laruelle, G. G., Rödenbeck, C., Roobaert, A., and Zeng, J.: SeaFlux: harmonization of air–sea CO₂ fluxes from surface pCO₂ data products using a standardized approach, *Earth Syst. Sci. Data*, 13, 4693–4710, <https://doi.org/10.5194/essd-13-4693-2021>, 2021.

Gloege, L., McKinley, G. A., Landschützer, P., Fay, A. R., Frolicher, T. L., and Fyfe, J. C.: Quantifying Errors in Observationally Based Estimates of Ocean Carbon Sink Variability, *Global Biogeochemical Cycles*, 35(4), <https://doi.org/10.1029/2020gb006788>, 2021.

Gregor L., & Fay, A. R.: SeaFlux data set: harmonised sea-air CO₂ fluxes from surface pCO₂ data products using a standardised approach (2021.04, Data set: Zenodo. <https://doi.org/10.5281/zenodo.5148460>, 2021).

Hauck, J., Nissen, C., Landschützer, P., Rödenbeck, C., Bushinsky, S., and Olsen, A.: Sparse observations induce large biases in estimates of the global ocean CO₂ sink: and ocean model subsampling experiment, *Philosophical Transactions Of the Royal Society A*, 381:20220063, <https://doi.org/10.1098/rsta.2022.0063>, 2023.

Takahashi, T., Sutherland, S.C., Wanninkhof, R., Sweeney, C., Feely, R.A., Chipman, D.W., Hales, B., Friederich, G., Chavez, F., Sabine, C. and Watson, A.: Climatological mean and decadal

change in surface ocean pCO₂, and net sea–air CO₂ flux over the global oceans. Deep Sea Research Part II: Topical Studies in Oceanography, 56(8-10), pp.554-577, 2009.

Weiss., R.: Carbon dioxide in water and seawater: the solubility of non-ideal gas, Marine Chemistry, 2(3), 203-215, [https://doi.org/10.1016/0304-4203\(74\)90015-2](https://doi.org/10.1016/0304-4203(74)90015-2), 1974.

Response to RC2

Despite appreciating the author's efforts in this study, Reviewer has not been convinced by its originality. Based on ESM output, numerous existing research works have shown additional data sampling (e.g., bgcArgo, SOCCOM, Sailboat,...) critical for error reduction in pCO₂ and flux estimation over the Southern Ocean and/or the global ocean [Bushinsky et al., 2019, Denvil-Sommer et al., 2021, Hauck et al., 2023, Landschützer et al., 2023]. One suggestion that would add value to the manuscript's findings is an analysis of spatial and temporal variations of flux estimates: to what extent their variability changes subject to the additional data. Some other major concerns are listed below.

Our study presents new findings that provide more insight into the number of additional samples and spatial pattern, consistent with current technology, that could reduce uncertainty in the ocean carbon sink, particularly in the Southern Ocean. There is no other study quantifying the impacts of meridional sampling by comparing different USV sampling tracks (also taking winter vs. summer sampling into account) in the Southern Ocean by using a Large Ensemble Testbed. Bushinsky et al. (2019) base their experiments on real-world SOCCOM float observations and use the SOM-FFN product for reconstruction. This is an important contribution. However, float-based estimates of pCO₂ are not incorporated into the SOCAT database and there are concerns about bias. It is therefore important to test the impact of realistic USV sampling, that can take direct pCO₂ observations with low uncertainties, can cover meridional gradients in the Southern Ocean, and are already incorporated into the SOCAT database.

The study by Hauck et al. (2023) uses GOBM output from one single model and reconstructs using two reconstruction methods (SOM-FFN and CarboScope), while we use ESM output from 75 different members and the pCO₂-Residual method. We also test a very different sampling pattern

compared to the “idealized” sampling in Hauck et al. (2023). 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. This further suggests that our study complements previous studies and adds value to this pertinent topic of ocean carbon research. It is important to present studies with different types of testbeds and reconstruction methods, so that we can better understand the impact of adding autonomous observations.

The study by Denvil-Sommer et al. (2021) is different to ours as it assesses sampling in the Atlantic Ocean, whereas our study focuses on sampling in the Southern Ocean and we show global reconstructions. Further, their study uses a different reconstruction method and assumes sampling from floats, not USVs.

Lines 149-153: "To build reconstruction algorithms through the data-driven training that occurs in ML, the statistics in all other algorithms developed to date must identify a function that disentangles these competing effects of SST on pCO₂. Here, the algorithm is assisted by removing this known temperature effect, and it must therefore only learn the pCO₂ impacts from biogeochemical drivers": there exist many other ML approaches [Friedlingstein et al., 2022] which do not separate the SST effects from others on pCO₂ but succeeds in estimate pCO₂. The major concerns are how to assess the uncertainty derived from SST effect removal and impacts on the experiment outputs.

Our study is not an evaluation of different ML approaches, but rather an assessment of how sampling impacts pCO₂ reconstructions. An evaluation of the method itself has already been performed by Bennington et al. (2022). They demonstrated that the pCO₂-Residual method performs better compared to other products when evaluating against independent data. They also showed improved skill when using pCO₂-Residual as the target variable as opposed to pCO₂. We want to assess how different sampling patterns affect the pCO₂ reconstruction. As we use the same

method for all experiments, we can directly compare them and evaluate how sampling impacts the reconstructions.

2. Figure 3: Relatively small bias and RMSE values have shown their imprints on the SOCAT track compared to "unseen" model truth. This evidences the problems of model overfitting. The authors can double-check whether model overfitting comes from the cross-validation technique or the pCO₂-Residual method. As the key findings of this manuscript are based on the data reconstruction results, Reviewer suggests the authors to carefully verify their methods and solve the problems of model overfitting before further consideration for publication.

*We would argue that the global mean bias and RMSE for the SOCAT reconstruction is comparable to values shown for pCO₂ reconstructions using other methods (e.g., Stamell et al., 2020; Gregor et al., 2019). For example, as shown in **Figure 3**, bias generally ranges between -10 to +10 μatm , which is comparable to the study by Hauck et al. (2023). However, after carefully evaluating our calculations following the reviewer's feedback, we noticed an error in our code that calculates the RMSEs. After fixing this error, the mean RMSE values increased by $\sim 3\text{-}4 \mu\text{atm}$.*

Editorial and specific comments:

1. Lines 11-12: "anthropogenic" can be removed. The SO has taken up atmospheric CO₂ without specifying natural or anthropogenic sources.

The Southern Ocean actively cycles natural and absorbs anthropogenic carbon. Gruber et al. (2009) demonstrate that the Southern Ocean is a source for natural carbon. The ocean sink for anthropogenic carbon is what we wish to focus on in this discussion.

2. Line 37: "fCO₂" is not defined. "uncertainty of $< 5 \mu\text{atm}$ ": this holds only for the measurements chosen to provide gridded SOCAT datasets.

*Noted and revised (**lines 43-46**): "The Surface Ocean CO₂ Atlas (SOCAT; Bakker et al., 2016) is the largest global database of surface ocean CO₂ observations, with data starting in 1957. The*

main synthesis and gridded products contain over 33 million high-quality direct shipboard measurements of $f\text{CO}_2$ (fugacity of CO_2) with an uncertainty of $< 5 \mu\text{atm}$ (Bakker et al., 2022)”.

3. Line 42: "Observation-based data products" \rightarrow "Data mapping methods".

We wish to use the term ‘observation-based data products’ consistently following recent literature (e.g., Fay et al., 2021; Crisp et al., 2022; Friedlingstein et al., 2023).

4. Line 45: "These data products" \rightarrow "These methods".

See above comment.

5. Lines 46-47: please remove or change ";" in the brackets to facilitate reading. You can use "-" instead. Line 47: " $x\text{CO}_2$; atmospheric CO_2 " \rightarrow "atmospheric CO_2 - $x\text{CO}_2$ "

Noted and revised.

6. Line 48: "where these are co-located" \rightarrow "where their available data are colocated".

We chose to keep the original sentence.

7. Lines 50-51: "Since the data products rely on observations to train the algorithms and thus produce these relationships": please rephrase this sentence. Data products do not train algorithms and produce relationships, but the ML-based methods themselves estimate the function between predictors and target data!

Noted and revised (lines 59-61): “Since the data products rely on $p\text{CO}_2$ observations to estimate functions between the target and driver variables, data sparsity remains a fundamental limitation to this technique”.

8. Line 57: "indirect pCO₂ estimates": can you define this term? Are they computed from float measurements of other carbonate variables?

Noted and revised. We added this sentence (lines 83-86): "These large uncertainties and biases arise when pCO₂ is not measured directly as in the observations included in SOCAT, but is rather estimated using measurements of pH combined with a regression-derived alkalinity estimate (Williams et al., 2017; Gray et al., 2018). SOCAT includes only direct pCO₂ observations".

9. Lines 67-68: "Such improvements in sampling are critically important in the undersampled Southern Ocean": USVs with low measurement uncertainty would prompt to be employed for observing network systems of pCO₂ but to draw this statement, it requires to provide the availability of USVs to sample pCO₂ by showing the sampling frequency and data coverage area over the SO?

*Additional high-accuracy observations from the sparsely sampled Southern Ocean, such that can be obtained by USVs, are key to provide further constraints on the ocean carbon sink and air-sea flux. We do not believe it is necessary to go into detail about the data coverage over the Southern Ocean, as we reference studies such as Bakker et al. (2016, 2022) describing the SOCAT coverage (which includes the Saildrone observations from Sutton et al. (2021) in the latest version). We also mention that the SOCAT coverage is shown in supplementary **previous Fig. S3 (new Fig. S5 in the revised version)**.*

10. Line 86: "actual observations": should be clarified. If you used the SOCAT gridded data tracks in your LET experiments, please change to "SOCAT observation-based data" or "SOCAT gridded data".

We have revised the sentence (lines 115-118): "However, instead of using real-world observations, we sample the target (i.e., surface ocean pCO₂) and driver variables (i.e., SST, SSS, MLD, Chl-a and xCO₂) from our Large Ensemble Testbed (LET) of Earth System Models (ESMs) (e.g., Stamell et al., 2020; Gloege et al., 2021; Bennington et al., 2022a)".

11. Lines 89-90: "in an ESM, surface ocean pCO₂ is known at all times and locations": not precise enough. It depends on which approximations and computational resources. So far, the models have been derived at 1 ° or 0.25° and monthly resolutions?

We are just aiming to convey that an ESM will not have huge gaps like in the real ocean. We have revised the sentence (lines 119-120): "First, in an ESM, the surface ocean pCO₂ field is provided precisely at all model times and 1°x1° points". The models used in our study have a 1°x1° resolution, which is stated multiple times throughout the manuscript (lines 120, 152, 154, 174, 198, 232, 263, 324, 449, 450).

12. Lines 161-162: "where pCO₂ mean and SST mean is the long-term mean of surface ocean pCO₂ and temperature, respectively, using all 1°x1° grid cells from the testbed": pCO₂ mean is different regionally, why you don't compute a global map of pCO₂ mean?

We do compute a mean of pCO₂ globally (see line 197), which is the pCO₂^{mean} and this is used to calculate the residual.

13. Lines 165-168: Please clarify. The authors have excluded pCO₂-Residual which have values below -250 μatm or over 250 μatm. They mention that such outliers correspond to model values higher than the maximum SOCAT data (816 μatm) and that do not reflect reality. It is not correct. First, both negative and positive pCO₂- Residual values cannot represent the upper bound of SOCAT data. Second, SOCAT only covers a tiny portion of the global ocean at a monthly time scale, and there might exist unobserved pCO₂ values higher than 816 μatm (e.g., over permanently or seasonally strong upwelling regions: Eastern Equatorial Pacific, Western Arabian Sea, Benguela, etc).

*We are not saying that both negative and positive pCO₂-Residual values represent the upper bound of SOCAT data. Our statement is "These pCO₂-Residual values **generally** correspond to high pCO₂, above the maximum value in SOCAT (816 μatm)". By this we mean that the majority of the pCO₂-Residual values that have been filtered out represent pCO₂ values that are larger than 816 μatm. However, since this seemed to be unclear, we have re-phrased this sentence (lines 208-210):*

“Prior to algorithm processing, $p\text{CO}_2$ -Residual values $> 250 \mu\text{atm}$ and $< -250 \mu\text{atm}$ from the testbed were filtered out targeting values that are not representative of the real ocean. The majority of the $p\text{CO}_2$ -Residual values that were filtered out correspond to high $p\text{CO}_2$, above the maximum value in SOCAT (816 μatm ; Stamell et al., 2020)”.

14. Lines 310-311: "Our presentation of global maps is limited to runs 'x5_5Y_W' (5022 observations) and 311 'Z_x4_10Y_YR' (7600 observations)". The information of gridded data used in the experiments should be declared in addition to the number of observations by USVs.

We revised the sentence (lines 449-450): “Our presentation of global maps is limited to runs 'x5_5Y_W' (5,022 monthly $1^\circ \times 1^\circ$ observations) and 'Z_x4_10Y_YR' (7,600 monthly $1^\circ \times 1^\circ$ observations)”.

15. Lines 319-321: How did the authors compute Bias (and RMSE) over the global ocean? In order to fairly compare the results of two or more runs (e.g., zigzag vs one-latitude, SOCAT vs SOCAT+USV), error statistics are computed on modelbased data excluding all used in ML training. Specifically, the evaluation should not consider 'zigzag+one-latitude' ('SOCAT+USV') $p\text{CO}_2$ data.

The reviewer is correct - the training data should ideally be removed before computing error statistics. When using actual observations, one would evaluate the reconstruction based on the test set alone. However, since we are using a model testbed, we have the opportunity to evaluate against $p\text{CO}_2$ values from “unseen” grid cells as well. In our study, we compute error statistics based on the full reconstruction, however this should have been communicated more clearly. The training data represents only about 1% of the full reconstruction (see new Fig. S1 below). Below, we show the 75-member testbed spread in bias and RMSE calculated based on the full reconstruction (what we present in our study) vs. ‘unseen’ grid cells for the ‘SOCAT-baseline’. The difference in mean bias and RMSE between the full and ‘unseen’ reconstruction is only 0.01 μatm and 0.08 μatm , respectively (see new Fig. S2 below). The results from the different runs can therefore be compared even though the full reconstruction is taken into account. We agree however with the reviewer that the training data should have been removed. Considering that we would have to re-run all

experiments, and it would not change the error statistics significantly or change our conclusions, we chose not to move forward with this for this study. However, for future studies using the testbed, the training set will be removed before calculating statistical metrics.

*We now add mention of this (**lines 231-236**): “Here, we calculate error statistics based on the full reconstruction (pCO₂ from all 1°x1° grid cells of the testbed, except for those masked or filtered out). In the full reconstruction, ~ 99 % of the data do not correspond to SOCAT or Saildrone USV observations used to train the algorithm (**Fig. S1**). Training data would ideally be removed before performance evaluation, but since the training data represent only ~ 1 %, the impact of not removing them is negligible (**Fig. S2**)”.*

***Figs. S1 and S2** are shown below.*

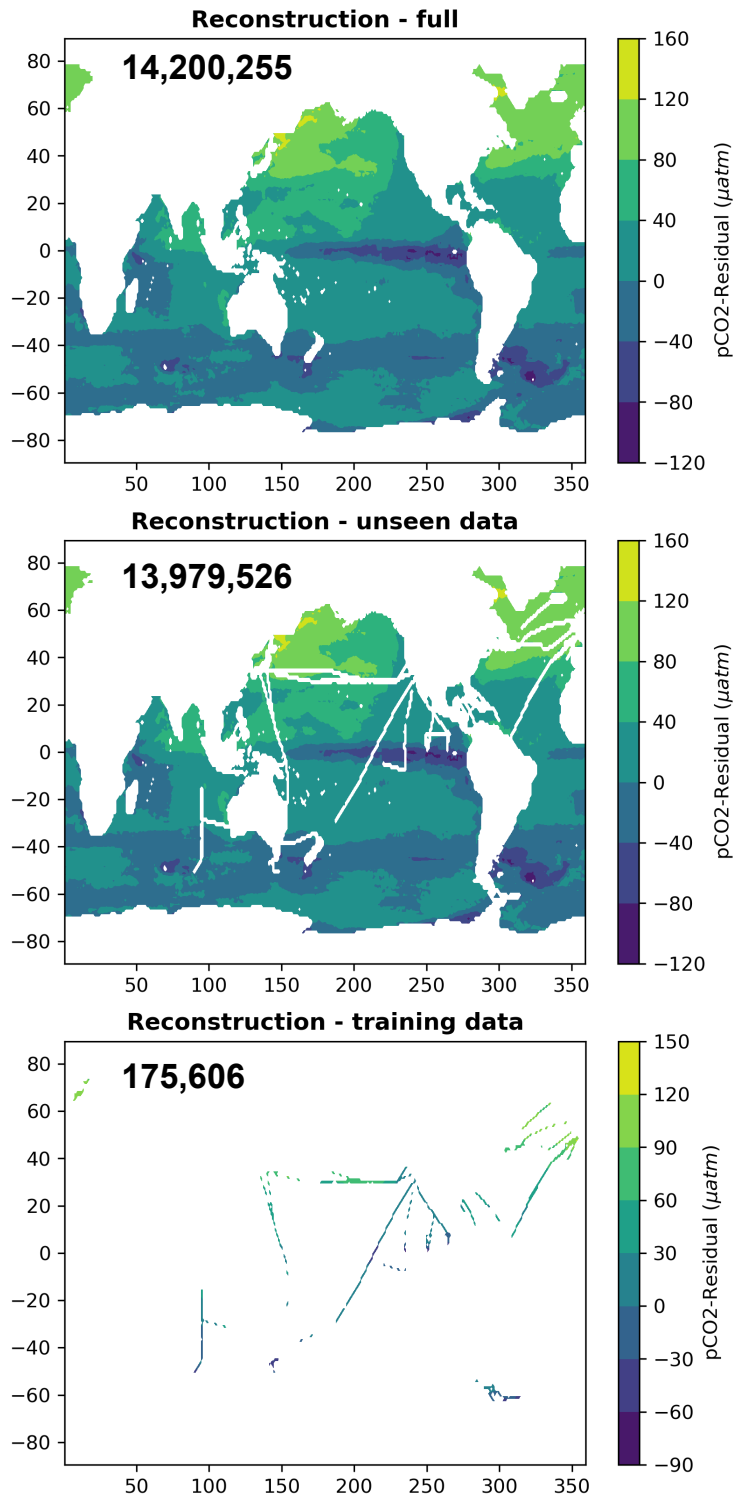


Figure S1: Maps of the full pCO₂-Residual reconstruction (all 1°x1° grid cells of the testbed, except for those masked or filtered out; see **Section 2.1** and **2.2**), ‘unseen’ reconstruction (all 1°x1° grid cells that do not correspond to SOCAT observations), and training data from the testbed. The maps show data from CESM member 001 for the month of March 2016 for the ‘SOCAT-baseline’.

Numbers on panels represent the total monthly $1^\circ \times 1^\circ$ grid cells for the entire testbed period (1982-2016) for each group of data.

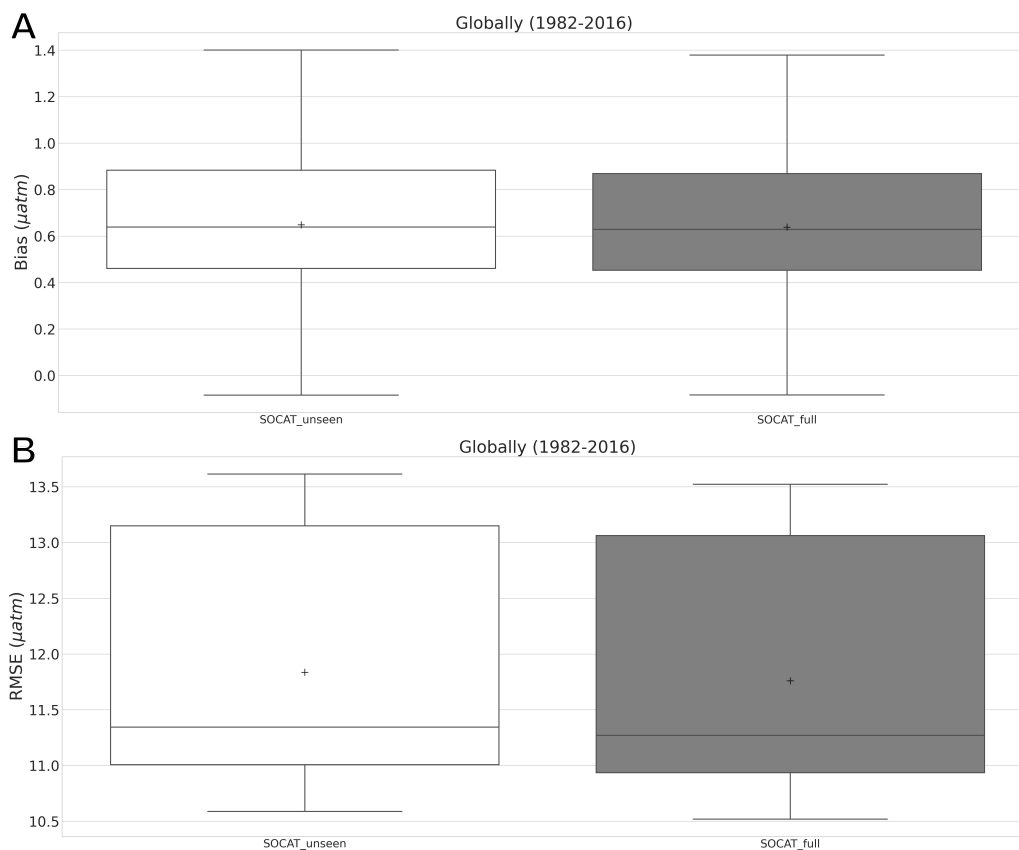
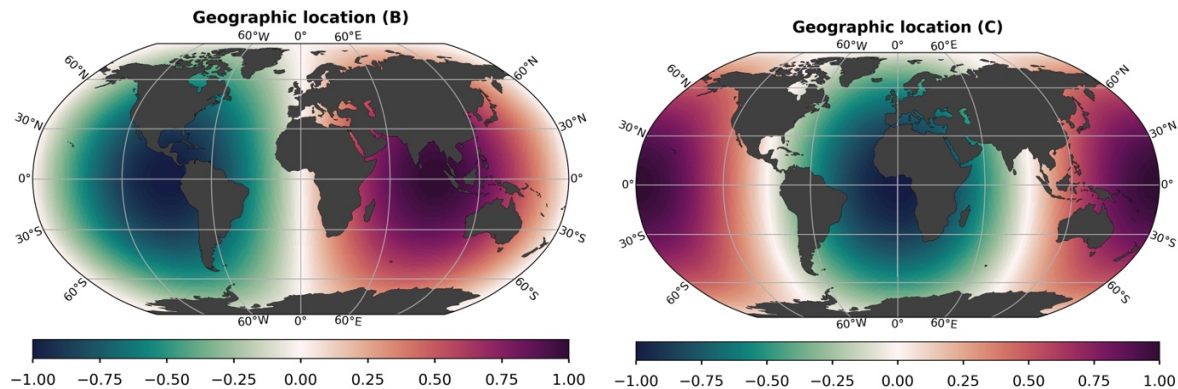


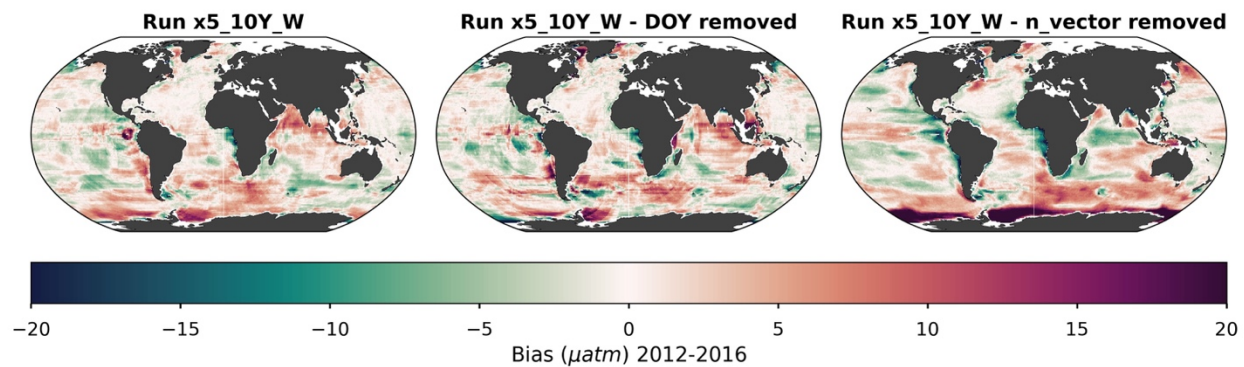
Figure S2: Spread of bias (a) and RMSE (b) for the 75 members of the Large Ensemble Testbed for the ‘unseen’ and full reconstruction for the ‘SOCAT-baseline’. The ‘unseen’ reconstruction represents independent data, i.e., all $1^\circ \times 1^\circ$ grid cells that do not correspond to SOCAT or Saildrone USV observations, and is not part of the training set.

16. Figures S4 and S5 show cyclic marks (it would be exposed clearly if the authors use a discrete colormap with a low number of colors). Would they be imprints of a driver variable?

These “cyclic marks” are likely imprints of the three-component n -vector that replaces the longitude and latitude coordinates to continuous values between 0 and 1 (i.e., to avoid the algorithm interpreting 0 and 360 degrees to be far apart; see figure below).



Bennington et al. (2022) present global maps (their Fig. 4) of the feature importance of various driver variables used in the surface ocean $p\text{CO}_2$ reconstruction (MLD, SST, Chl- a , location and day of year). Such “cyclic marks” are apparent for “geographic location” and “day of year”, but none of the other drivers. We did two test runs (using only one member from the testbed), removing day of year (DOY) and geographic location (n -vector; A, B and C) as inputs for the reconstruction. As shown by the figure below, the “cyclic” marks disappear when the n -vector is removed. When removing the n -vector transformation, however, the reconstruction shows significantly higher bias in the Southern Ocean, so we chose to keep these driver variables.



17. Figures 5 and 8: The author should report the number of data gridded from USV observations used in ML training. And the error statistics must be computed on the evaluation data (i.e., model-truth-based data excluding all the training data). Figure 8’s caption: The mean of RMSEs here is computed with respect to space or time? Instead, the author should compute the mean of squared errors over the global ocean and the periods of interest and then report its square root.

*The number of monthly 1°x1° observations for each experiment is described in **Table 1** as well as shown on the x-axis of **Figure 5** and **8**. This was specified in the **Table 1** caption, but we now specify this in the figure captions as well (**lines 538-539 and 661-662**): ‘# additional observations’ = number of monthly 1°x1° USV observations in addition to SOCAT. We state in the manuscript that (**lines 217-218**): “The test and validation set each account for 20 % of the data, leaving 60 % for training”. For both **Fig. 5** and **8**, the mean is computed with respect to both space (top figure shows global and bottom figure shows Southern Ocean, which in our study is defined as south of 35° S) and time, which is 2006-2016 (for the 10-year sampling) and 2012-2016 (for the five-year sampling). This information is stated in the figure headlines.*

Regarding comment about error statistics, please see answer #15.

18. Line 386: ‘Z_x10_5Y_YR

Noted and revised.

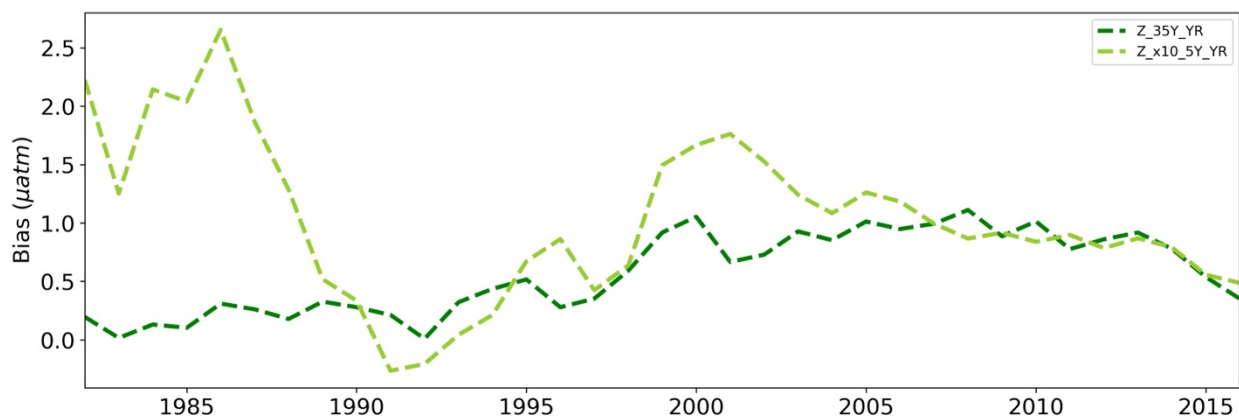
19. Lines 497-499: "Although run ‘x13_10Y_W’ demonstrates the highest reduction in bias out of all runs, the ‘zigzag’ runs still reduce bias in the Southern Ocean by 44-65 % (vs. 77 % for run ‘x13_10Y_W’)". The evaluation should not put high confidence on the bias reduction since this statistic is computed as the mean of negative and positive differences between pCO₂ estimates and model truth. Reviewer agrees that the bias can be used to assess model over- or underestimation but RMSD is a better metric for an overall evaluation.

We agree with the reviewer, and that is why we report both bias and RMSE. Our conclusions do not fully rely on bias alone, as is shown throughout the paper. For example, we conclude that the zigzag-runs perform best overall, even though run ‘x13_10Y_W’ demonstrates a higher reduction with respect to mean bias.

20. Lines 536-541: "To better understand this discrepancy, we performed an additional experiment based on run 538 ‘Z_x10_5Y_YR’, but assumed sampling every year for the entire testbed period

(i.e., 1982-2016). The results from this experiment show a significant reduction in the temporal variability of reconstruction bias; with the additional USV sampling, the reconstructed Southern Ocean air-sea CO₂ flux closely matches the ‘model truth’ for the entire testbed duration (Fig. S14).". Here biases increase in the last two decades that do not reflect the increase in the number of SOCAT (SOCAT+USV) data as shown in the previous results.

As shown by the figure below, run ‘Z_x10_5Y_YR’ (shown in **Fig. 6** in main text) and ‘Z_x10_35Y_YR’ (shown in **previous Fig. S14** in supplement; in the revised version, this is now **Fig. S20**) show similar variability the last five years when the sampling is identical. For run ‘Z_x10_5Y_YR’, USV observations have been added only for the last five years of the testbed, while for run ‘Z_x10_35Y_YR’, USV observations have been added for the whole testbed period (35 years). The bias decreases more significantly in the earlier decades for run ‘Z_x10_5Y_YR’ because there are no additional USV observations at this time, and there are significantly less SOCAT observations in this period compared from 1990 and onwards (see **new Fig. S5c**; this was **Fig. S3c** in the previous version).



21. Lines 552-554: "Further, we find that this modest amount of additional Sairdrone USV sampling increases the global and Southern Ocean air-sea CO₂ flux by up to 0.1 Pg C yr⁻¹, 25% of the uncertainty in the ocean carbon sink ". The increase in global ocean CO₂ sink estimated by the LET testbed can not be compared with the uncertainty derived from the GCB’s quantification [Friedlingstein et al., 2022]. First, they are two different statistics. Second, the GCB’s uncertainty

is computed based on the ensemble of different data mapping and modeling methods, and thus the value might be significantly larger than the one estimated by each method itself.

These values can be compared as they are in the same units. We wish to demonstrate that 0.1 Pg C/yr is a significant reduction. Following the reviewer's comment, we revised the sentence: "Further, we find that this modest amount of additional Saildrone USV sampling increases the global and Southern Ocean air-sea CO₂ flux by up to 0.1 Pg C yr⁻¹, a quantity equivalent to 25 % of the uncertainty in the ocean carbon sink".

References

Bakker, D. C. E., Pfeil, B., Landa, C. S., Metzl, N., O'Brien, K. M., Olsen, A., Smith, K., Cosca, C., Harasawa, S., Jones, S. D., Nakaoka, S., Nojiri, Y., Schuster, U., Steinhoff, T., Sweeney, C., Takahashi, T., Tilbrook, B., Wada, C., Wanninkhof, R., Alin, S. R., Balestrini, C. F., Barbero, L., Bates, N. R., Bianchi, A. A., Bonou, F., Boutin, J., Bozec, Y., Burger, E. F., Cai, W.-J., Castle, R. D., Chen, L., Chierici, M., Currie, K., Evans, W., Featherstone, C., Feely, R. A., Fransson, A., Goyet, C., Greenwood, N., Gregor, L., Hankin, S., Hardman-Mountford, N. J., Harlay, J., Hauck, J., Hoppema, M., Humphreys, M. P., Hunt, C. W., Huss, B., Ibánhez, J. S. P., Johannessen, T., Keeling, R., Kitidis, V., Körtzinger, A., Kozyr, A., Krasakopoulou, E., Kuwata, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lo Monaco, C., Manke, A., Mathis, J. T., Merlivat, L., Millero, F. J., Monteiro, P. M. S., Munro, D. R., Murata, A., Newberger, T., Omar, A. M., Ono, T., Paterson, K., Pearce, D., Pierrot, D., Robbins, L. L., Saito, S., Salisbury, J., Schlitzer, R., Schneider, B., Schweitzer, R., Sieger, R., Skjelvan, I., Sullivan, K. F., Sutherland, S. C., Sutton, A. J., Tadokoro, K., Telszewski, M., Tuma, M., van Heuven, S. M. A. C., Vandemark, D., Ward, B., Watson, A. J., and Xu, S.: A multi-decade record of high-quality fCO₂ data in version 3 of the Surface Ocean CO₂ Atlas (SOCAT), *Earth System Science Data*, 8, 383–413, <https://doi.org/10.5194/essd-8-383-2016>, 2016.

Bakker, D. C. E., Alin, S. R., Becker, M., Bittig, H. C., Castaño-Primo, R., Feely, R. A., Gkritzalis, T., Kadono, K., Kozyr, A., Lauvset, S. K., Metzl, N., Munro, D. R., Nakaoka, S., Nojiri, Y., O'Brien, K. M., Olsen, A., Pfeil, Benjamin, P., Denis, S., Tobias, S., Kevin F., Sutton, A. J., Sweeney, C., Tilbrook, B., Wada, C., Wanninkhof, R., Willstrand W. A., Akl, J., Apelthun, L. B., Bates, N., Beatty, C. M., Burger, E. F., Cai, W., Cosca, C. E., Corredor, J. E., Cronin, M., Cross, J. N., De Carlo, E. H., DeGrandpre, M. D., Emerson, S. R., Enright, M. P., Enyo, K., Evans, W., Frangoulis, C., Fransson, A., García-Ibáñez, M. I., Gehrung, M., Giannoudi, L., Glockzin, M., Hales, B., Howden, S. D., Hunt, C. W., Ibánhez, J. S. P., Jones, S. D., Kamb, L., Körtzinger, A., Landa, C. S., Landschützer, P., Lefèvre, N., Lo Monaco, C., Macovei, V. A., Maenner J. S., Meinig, C., Millero, F. J., Monacci, N. M., Mordy, C., Morell, J. M., Murata, A., Musielewicz, S., Neill, ., Newberger, T., Nomura, D., Ohman, M., Ono, T., Passmore, A., Petersen, W., Petihakis, G., Perivoliotis, L., Plueddemann, A. J., Rehder, G., Reynaud, T., Rodriguez, C., Ross, A. C., Rutgersson, A., Sabine, C. L., Salisbury, J. E., Schlitzer, R., Send, U., Skjelvan, I., Stamataki, N., Sutherland, S. C., Sweeney, C., Tadokoro, K., Tanhua, T., Telszewski, M., Trull, T., Vandemark, D., van Ooijen, E., Voynova, Y.

G., Wang, H., Weller, R. A., Whitehead, C., Wilson, D.: *Surface Ocean CO₂ Atlas Database Version 2022 (SOCATv2022) (NCEI Accession 0253659)*, NOAA National Centers for Environmental Information [dataset], <https://doi.org/10.25921/1h9f-nb73>, 2022.

Bennington, V., Galjanic, T., and McKinley, G. A.: *Explicit Physical Knowledge in Machine Learning for Ocean Carbon Flux Reconstruction: The pCO₂-Residual Method*, *Journal of Advances in Modeling Earth Systems*, 14(10), <https://doi.org/10.1029/2021ms002960>, 2022.

Bushinsky, S. M., Landschützer, P., Rödenbeck, C., Gray, A. R., Baker, D., Mazloff, M. R., Resplandy, L., Johnson, K. S., and Sarmiento, J. L.: *Reassessing Southern Ocean air-sea CO₂ flux estimates with the addition of biogeochemical float observations*, *Global Biogeochemical Cycles*, 33(11), 1370-1388, <https://doi.org/10.1029/2019GB006176>, 2019.

Crisp, D., Dolman, H., Tanhua, T., McKinley, G. A., Hauck, J., Bastos, A., Sitch, S., Eggleston, S., & Aich, V.: *How well do we understand the land-ocean-atmosphere carbon cycle?* *Reviews of Geophysics*, 60(2), e2021RG000736, [doi:10.1029/2021RG000736](https://doi.org/10.1029/2021RG000736), 2022.

Denvil-Sommer, A., Gehlen, M., & Vrac, M.: *Observation system simulation experiments in the Atlantic Ocean for enhanced surface ocean pCO₂ reconstructions*, *Ocean Science*, 17(4), 1011-1030, <https://doi.org/10.5194/os-17-1011-2021>, 2021.

Fay, A. R., Gregor, L., Landschützer, P., McKinley, G. A., Gruber, N., Gehlen, M., Iida, Y., Laruelle, G. G., Rödenbeck, C., Roobaert, A., and Zeng, J.: *SeaFlux: harmonization of air-sea CO₂ fluxes from surface pCO₂ data products using a standardized approach*, *Earth Syst. Sci. Data*, 13, 4693–4710, <https://doi.org/10.5194/essd-13-4693-2021>, 2021.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Bakker, D. C. E., Hauck, J., Landschützer, P., Le Quéré, C., Luijkx, I. T., Peters, G. P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni, P., Barbero, L., Bates, N. R., Becker, M., Bellouin, N., Decharme, B., Bopp, L., Brasika, I. B. M., Cadule, P., Chamberlain, M. A., Chandra, N., Chau, T.-T.-T., Chevallier, F., Chini, L. P., Cronin, M., Dou, X., Enyo, K., Evans, W., Falk, S., Feely, R. A., Feng, L., Ford, D. J., Gasser, T., Ghattas, J., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Hefner, M., Heinke, J., Houghton, R. A., Hurtt, G. C., Iida, Y., Ilyina, T., Jacobson, A. R., Jain, A., Jarníková, T., Jersild, A., Jiang, F., Jin, Z., Joos, F., Kato, E., Keeling, R. F., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Körtzinger, A., Lan, X., Lefèvre, N., Li, H., Liu, J., Liu, Z., Ma, L., Marland, G., Mayot, N., McGuire, P. C., McKinley, G. A., Meyer, G., Morgan, E. J., Munro, D. R., Nakaoka, S.-I., Niwa, Y., O'Brien, K. M., Olsen, A., Omar, A. M., Ono, T., Paulsen, M., Pierrot, D., Pockock, K., Poulter, B., Powis, C. M., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T. M., Schwinger, J., Séférian, R., Smallman, T. L., Smith, S. M., Sospedra-Alfonso, R., Sun, Q., Sutton, A. J., Sweeney, C., Takao, S., Tans, P. P., Tian, H., Tilbrook, B., Tsujino, H., Tubiello, F., van der Werf, G. R., van Ooijen, E., Wanninkhof, R., Watanabe, M., Wimart-Rousseau, C., Yang, D., Yang, X., Yuan, W., Yue, X., Zaehle, S., Zeng, J., and Zheng, B.: *Global Carbon Budget 2023*, *Earth Syst. Sci. Data*, 15, 5301–5369, <https://doi.org/10.5194/essd-15-5301-2023>, 2023.

Gloege, L., McKinley, G. A., Landschützer, P., Fay, A. R., Frolicher, T. L., and Fyfe, J. C.: *Quantifying Errors in Observationally Based Estimates of Ocean Carbon Sink Variability*, *Global Biogeochemical Cycles*, 35(4), <https://doi.org/10.1029/2020gb006788>, 2021.

Gray, A. R., Johnson, K. S., Bushinsky, S. M., Riser, S. C., Russell, J. L., Talley, L. D., Wanninkhof, R., Williams, N. L., and Sarmiento, J. L.: *Autonomous biogeochemical floats detect significant carbon dioxide outgassing in the high-latitude Southern Ocean*, *Geophysical Research Letters*, 45(17), 9049-9057, <https://doi.org/10.1029/2018GL078013>, 2018.

Gregor, L., Lebehot, A. D., Kok, S., and Monteiro, P. M. S.: *A comparative assessment of the uncertainties of global surface ocean CO₂ estimates using a machine-learning ensemble (CSIR-ML6 version 2019a) – have we hit the wall?*, *Geoscientific Model Development*, 12, 5113-5136, <https://doi.org/10.5194/gmd-12-5113-2019>, 2019.

Gruber, N., Gloor, M., Mikaloff Fletcher, S. E., Doney, S. C., Dutkiewicz, S., Folows, M. J., Gerber, M., Jacobson, A. R., Joos, F., Lindsay, K., Menemenlis, D., Mouchet, A., Muller S. A., Sarmiento, J. L., & Takahashi, T.: *Oceanic sources, sinks, and transport of atmospheric CO₂*, *Global Biogeochemical Cycles*, 23(1), <https://doi.org/10.1029/2008GB003349>, 2009.

Hauck, J., Nissen, C., Landschützer, P., Rödenbeck, C., Bushinsky, S., and Olsen, A.: *Sparse observations induce large biases in estimates of the global ocean CO₂ sink: and ocean model subsampling experiment*, *Philosophical Transactions Of the Royal Society A*, 381:20220063, <https://doi.org/10.1098/rsta.2022.0063>, 2023.

Stamell, J., Rustagi, R. R., Gloege, L., and McKinley, G. A.: *Strengths and weaknesses of three Machine Learning methods for pCO₂ interpolation*, *Geoscientific Model Development Discussions[preprint]*, doi:10.5194/gmd-2020-311, 22 October 2020.

Sutton, A. J., Williams, N. L., and Tilbrook, B.: *Constraining Southern Ocean CO₂ flux uncertainty using uncrewed surface vehicle observations*, *Geophysical Research Letters*, 48(3), e2020GL091748, <https://doi.org/10.1029/2020GL091748>, 2021.

Takahashi, T., Sutherland, S.C., Wanninkhof, R., Sweeney, C., Feely, R.A., Chipman, D.W., Hales, B., Friederich, G., Chavez, F., Sabine, C. and Watson, A.: *Climatological mean and decadal change in surface ocean pCO₂, and net sea–air CO₂ flux over the global oceans*. *Deep Sea Research Part II: Topical Studies in Oceanography*, 56(8-10), pp.554-577, 2009.

Williams, N. L., Juranek, L. W., Feely, R. A., Johnson, K. S., Sarmiento, J. L., Talley, L. D., Dickson, A. G., Gray, A. R., Wanninkhof, R., Russell, J. L., Riser, S. C., and Takeshita, Y.: *Calculating surface ocean pCO₂ from biogeochemical Argo floats equipped with pH: An uncertainty analysis*, *Global Biogeochemical Cycles*, 31(3), 591-604, <https://doi.org/10.1002/2016GB005541>, 2017.

