Response to RC1

We would like to thank the reviewer 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 especially think the added discussion and figures regarding the testbed spread improved the manuscript. We have provided an itemized list below detailing our responses (in italic font) to the reviewer's suggestions.

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 CO_2 sink need to be improved (see specific comments).

Specific comments:

1) To calculate the residual pCO_2, the authors used the equation in line 160. In this equation, the term pCO_2^{mean} came from "surface ocean pCO_2 […] 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 pCO_2^{mean} comes from an initial reconstruction of pCO_2. Therefore, in this submitted manuscript, by using model outputs instead of an initial reconstruction of pCO_2 fields, the authors assumed that their method would be able to perfectly reconstruct the long-term average pCO_2 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 pCO_2-Residual reconstruction to the source of mean pCO_2, by experimenting using the Takahashi pCO_2 climatology (Takahashi et al., 2009) as well as the mean pCO_2 of the SeaFlux observation-based products (Fay et al., 2021; Gregor & Fay, 2021). They found that alternative sources of the initial pCO_2 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 pCO_2 to calculate pCO_2-T and pCO_2-Resisudal.

2) To calculate the net sea–air CO_2 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 pCO_2 reconstruction) to be used to calculate the CO_2 flux, rather than observational data which may have different variabilities and/or trends to those simulated. The authors could then compare this calculated CO_2 flux to the simulated CO_2 flux (and not to a "model truth" CO_2 flux from simulated pCO_2 fields mixed with observational data). This is particularly important when the authors are discussing the ability of their method to reproduce CO_2 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 pCO_2. 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 pCO_2 fields and from pCO_2, 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 CO_2 fluxes (and not from recalculated "model-truth" CO_2 fluxes, see my previous comment) and comparing the values of these metrics with the values that would be obtained when reconstructed CO_2 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:

*"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: "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): "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****".*

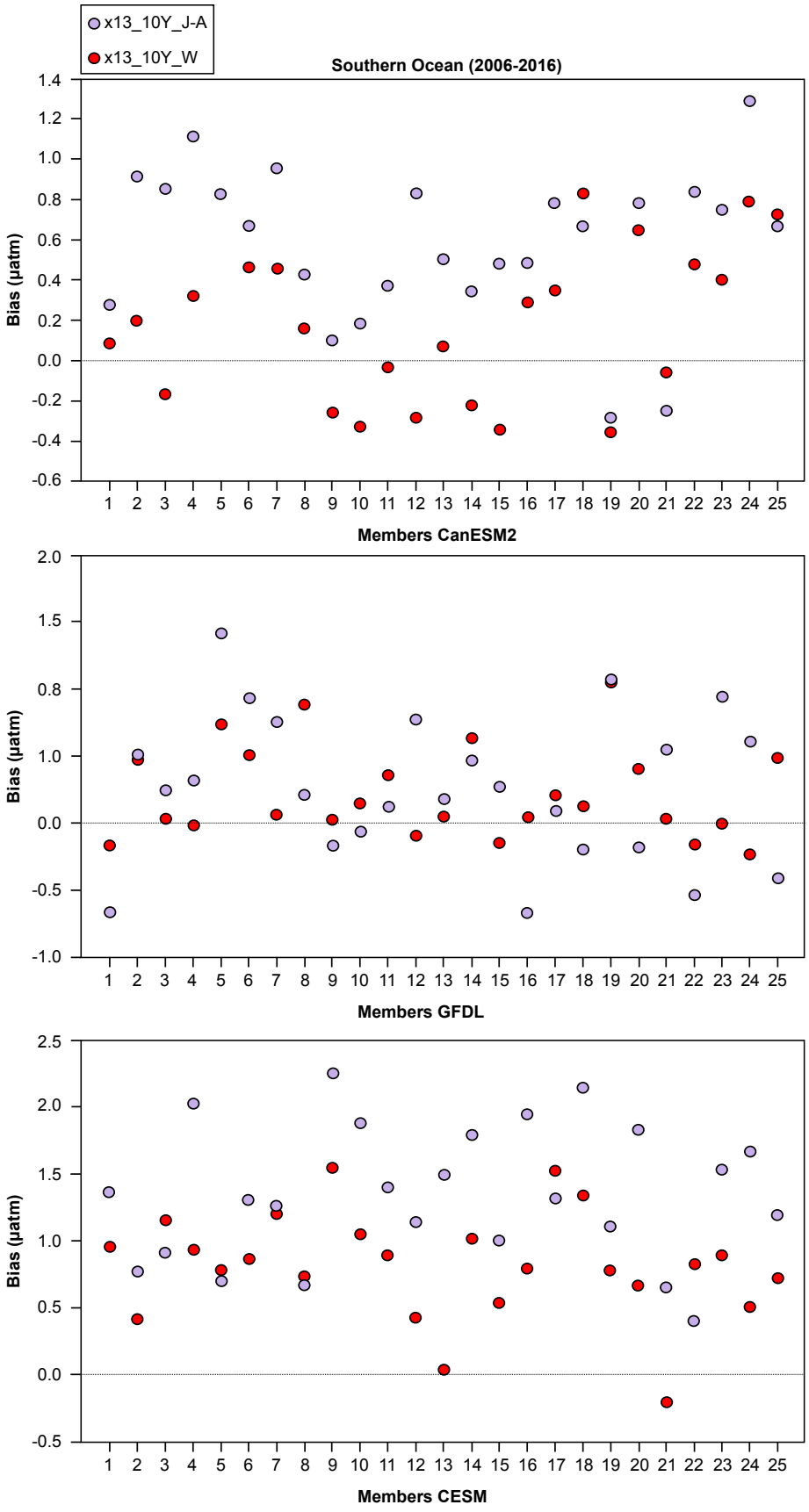
Further, a recent study by Hauck et al. (2023) performed similar sampling experiments, but used a different type of reconstruction method and testbed (i.e., a single hindcast model), and show that additional autonomous sampling leads to a weakened Southern Ocean sink, which is the opposite to our findings. 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:

*"Bushinsky et al. (2019) and Hauck et al. (2023) performed similar sampling experiments as presented here, by comparing ML surface ocean pCO_2 reconstructions based on SOCAT vs. additional SOCCOM or ideal virtual floats. 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. In contrast, we find that the pCO_2-Residual method underestimates the CO_2 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****).*

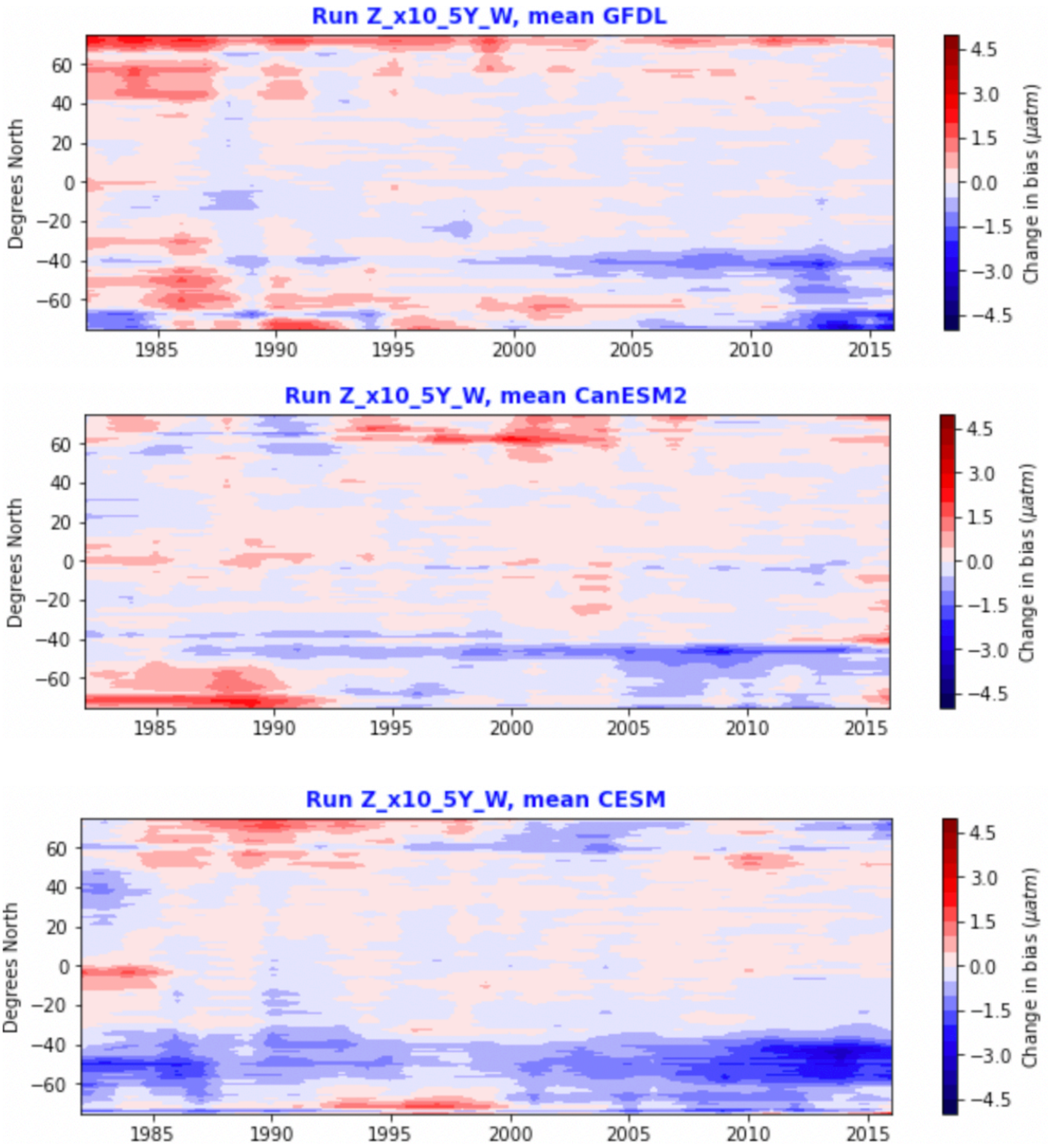
*Going forward, additional studies are needed to better understand why these results suggest a different direction of the sink change with additional sampling. These differences could stem from the use of different reconstruction methods assessed. Hauck et al. (2023) used the MPI-SOM-FFN and CarboScope/Jena-MLS reconstruction methods, while we use the pCO_2-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 CO_2 fluxes".*

Answers to the reviewer's specific questions above:

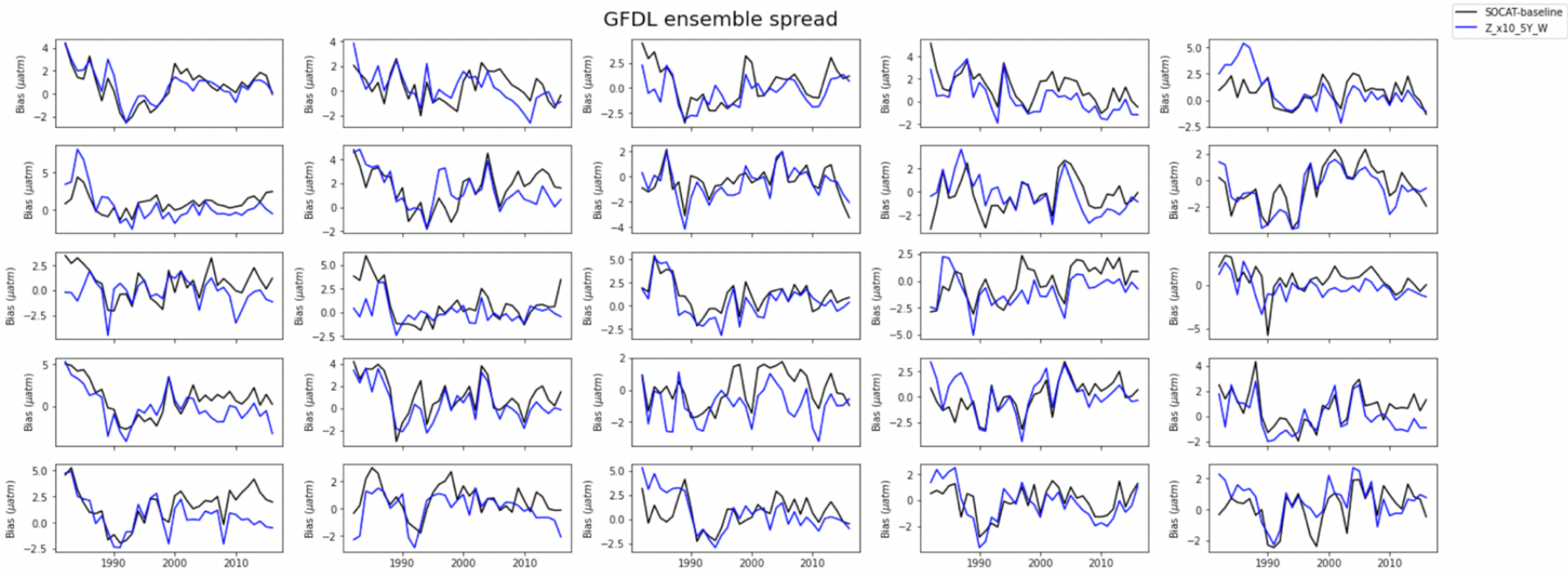
*1. Below, we show the bias (over the Southern Ocean for the period of 2006-2010) of each individual member of the models in the testbed, comparing the high-sampling run 'x13_10Y_J-A' with the equivalent run that restricts sampling to southern hemisphere winter months ('x13_10Y_W'). As shown by the figure 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**.*

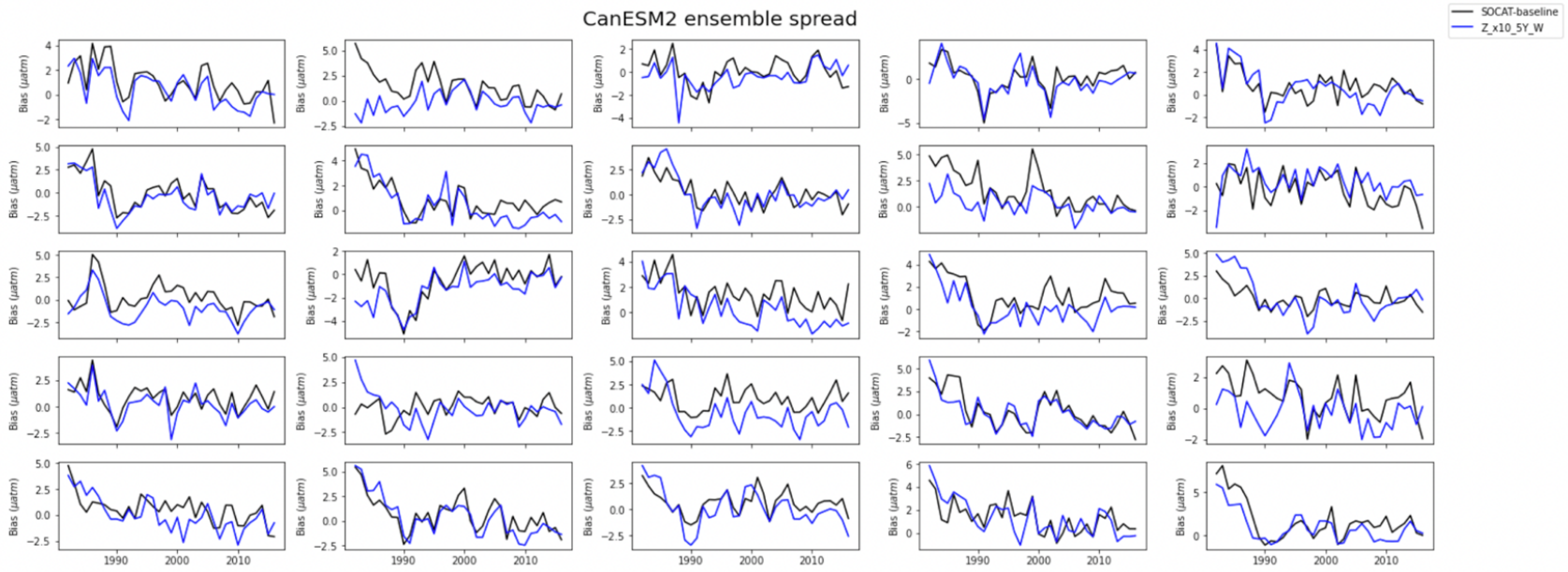

x13_10Y_J-A
x13_10Y_W
Southern Ocean (2006-2016)
Bias (µatm)
Members CanESM2
Members GFDL
Members CESM

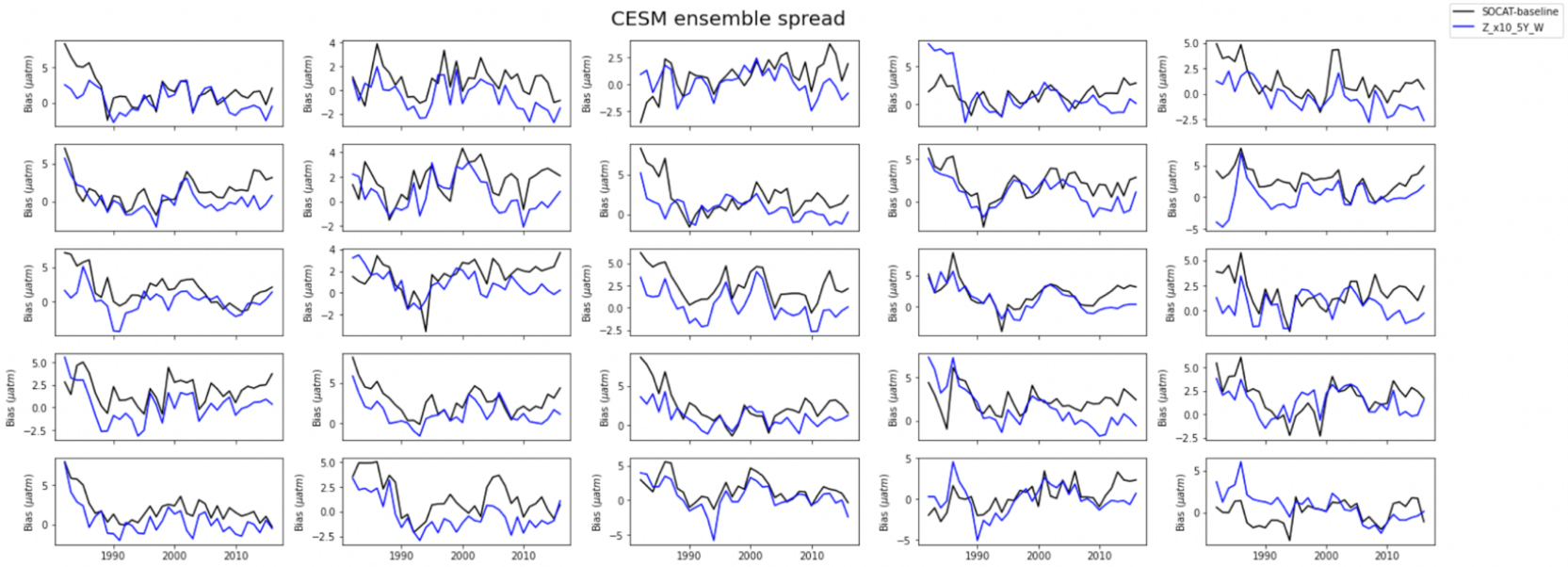
2A. Below, we present zonal annual mean Hovmöller plots showing the change in bias when comparing run 'Z_x10_5Y_W' to the 'SOCAT-baseline'. As shown by the figure below, all models show improvement back in time beyond the additional sampling duration (2012-2016), reflecting the ensemble mean shown in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



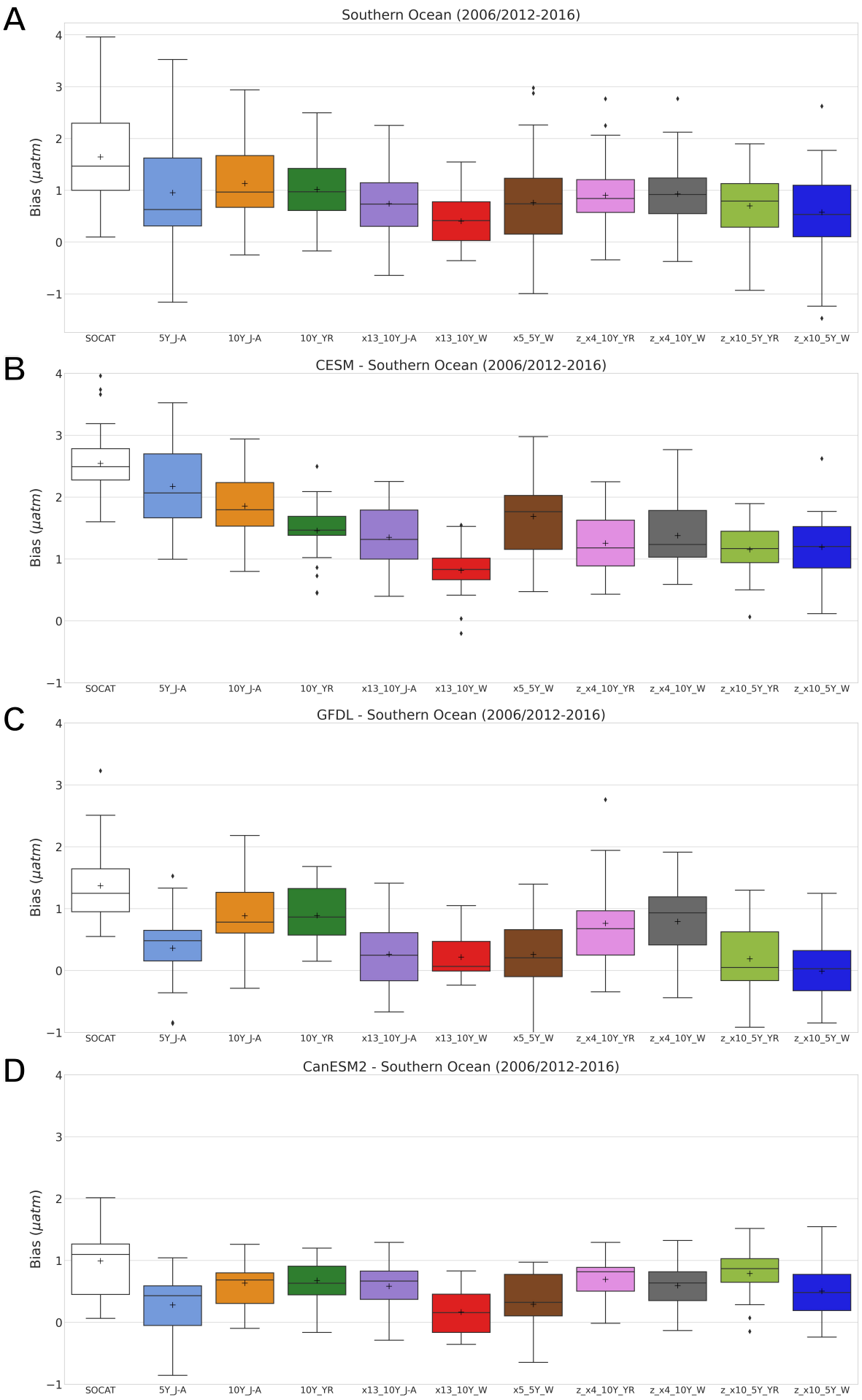
2B. To examine individual members, we plot time series of bias for run 'Z_x10_5Y_W' and the 'SOCAT-baseline' averaged over the area of highest improvement shown in ***Fig. S6*** *(between 50° S and 35° S). These figures show improvement in bias compared to the 'SOCAT-baseline' already in the beginning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



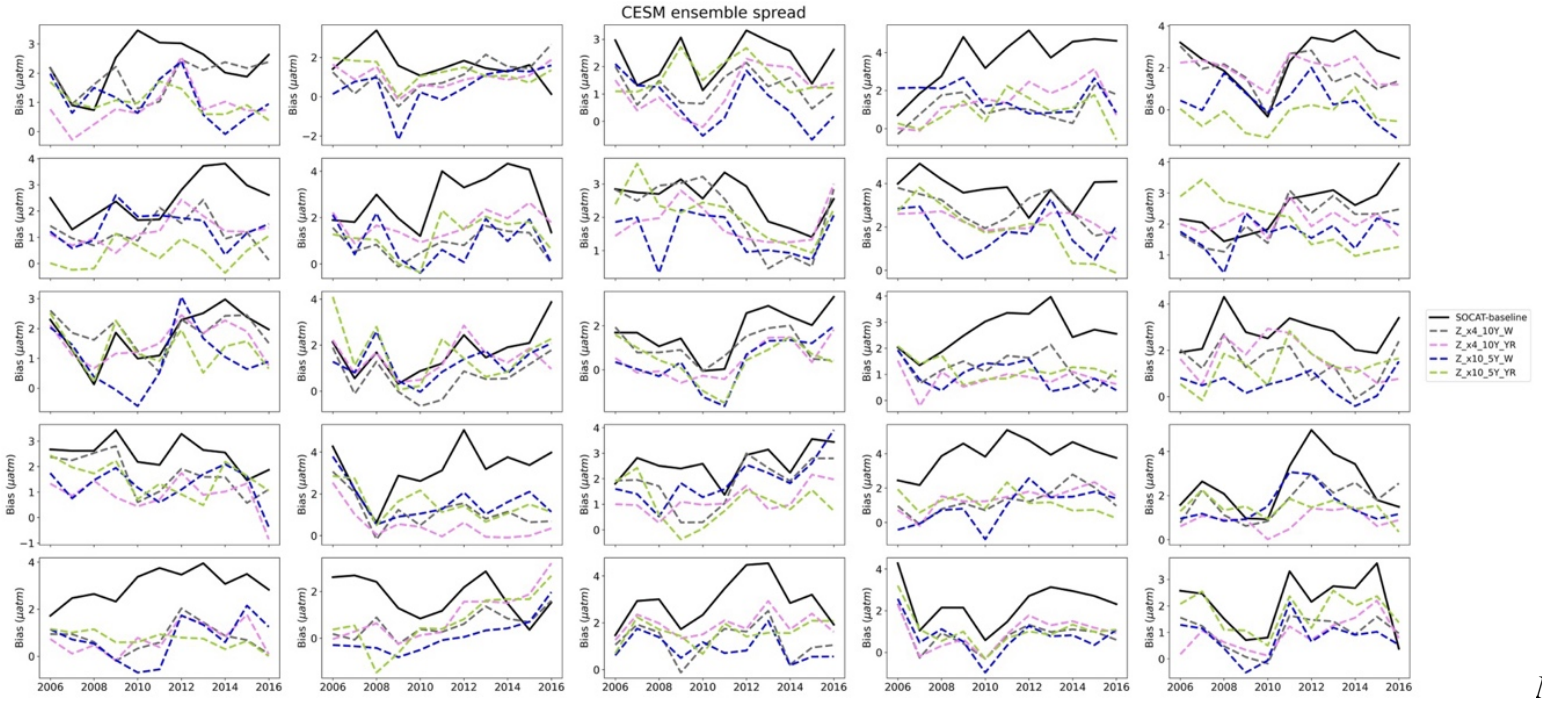




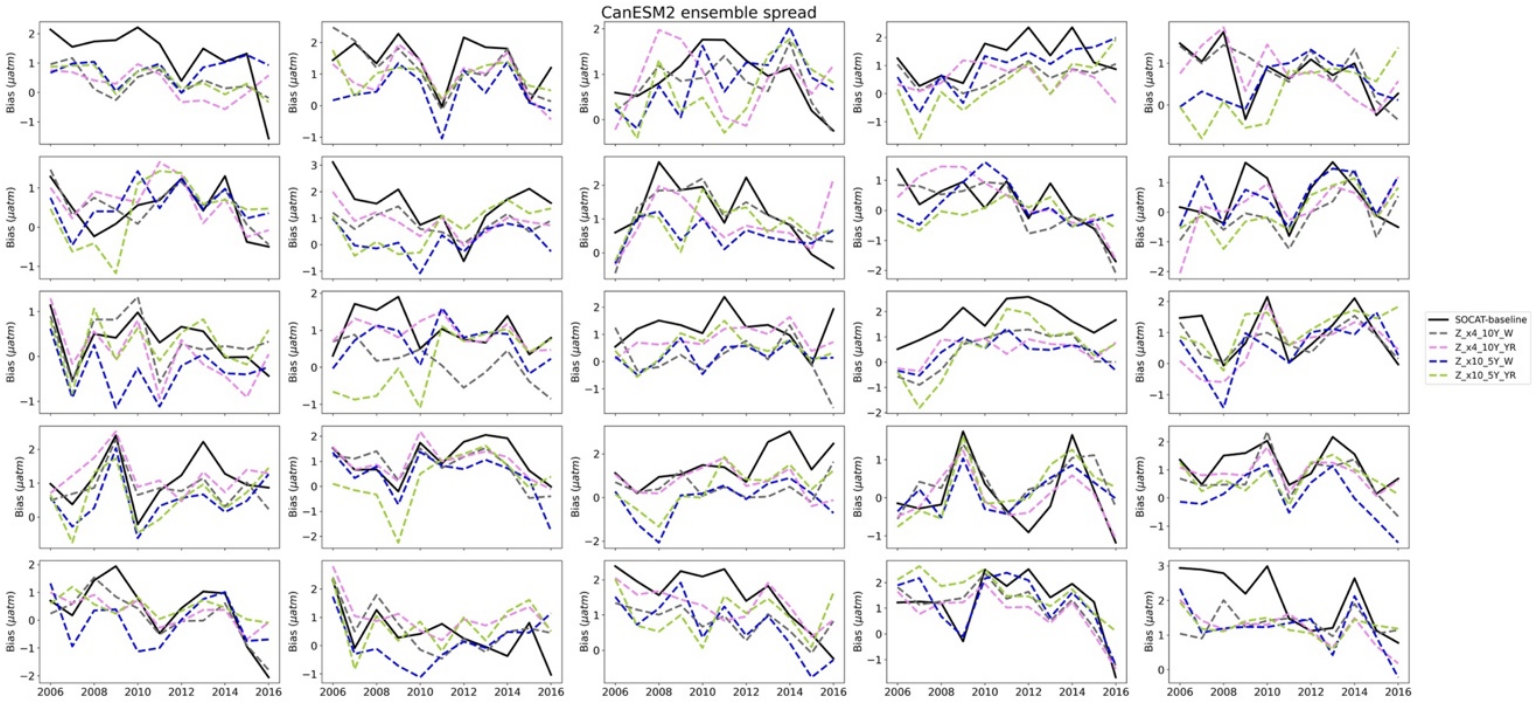
*Overview of new supplementary figures showing the ensemble spread (**S8, S10, S14, S16**):*


A
Southern Ocean (2006/2012-2016)
B
CESM - Southern Ocean (2006/2012-2016)
C
GFDL - Southern Ocean (2006/2012-2016)
D
CanESM2 - Southern Ocean (2006/2012-2016)
Bias (μatm)
SOCAT
5Y_J-A
10Y_J-A
10Y_YR
x13_10Y_J-A
x13_10Y_W
x5_5Y_W
z_x4_10Y_YR
z_x4_10Y_W
z_x10_5Y_YR
z_x10_5Y_W

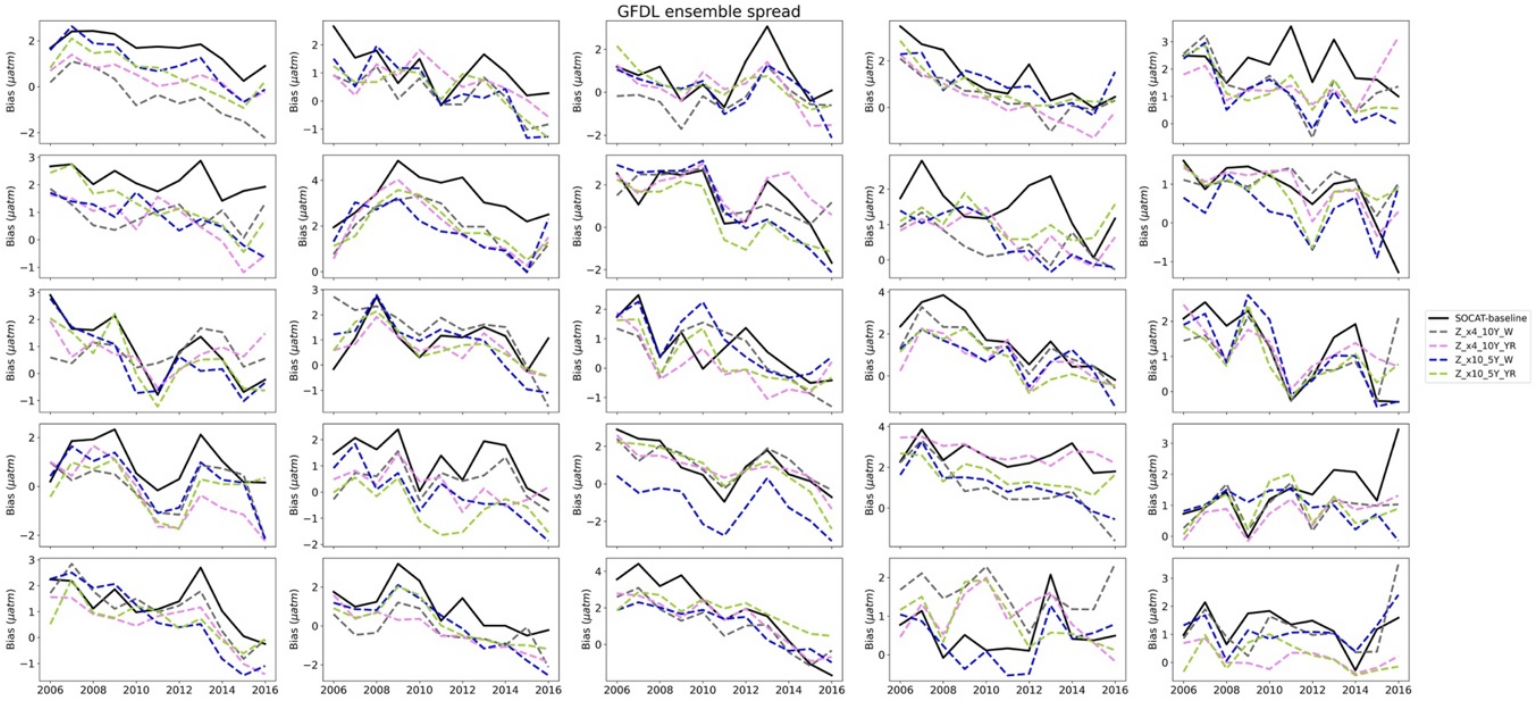
New Fig S8


CESM ensemble spread
Bias (µatm)
2006
2008
2010
2012
2014
2016
SOCAT-baseline
Z_x4_10Y_W
Z_x4_10Y_YR
Z_x10_5Y_W
Z_x10_5Y_YR

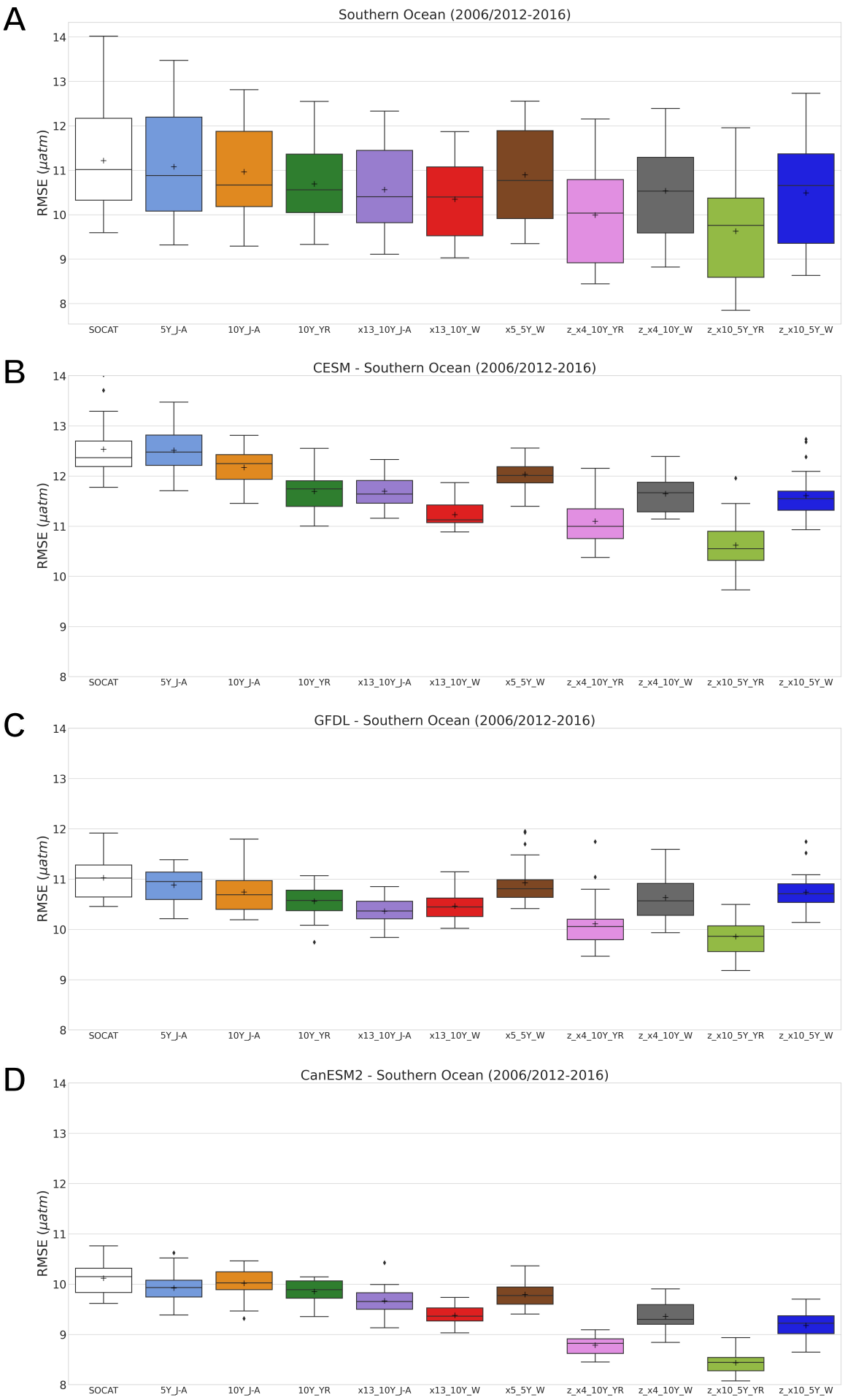
New Fig S10



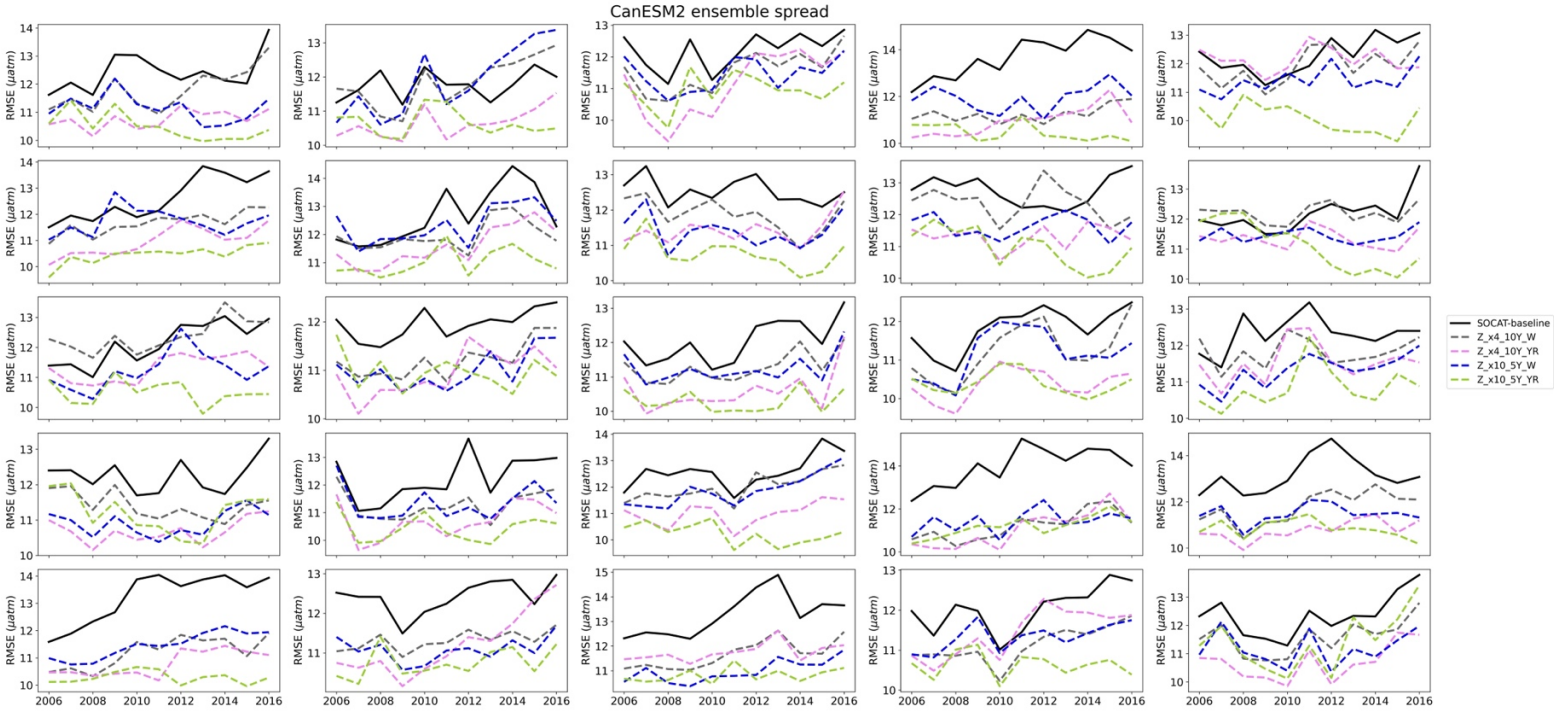
New Fig S10 cont.



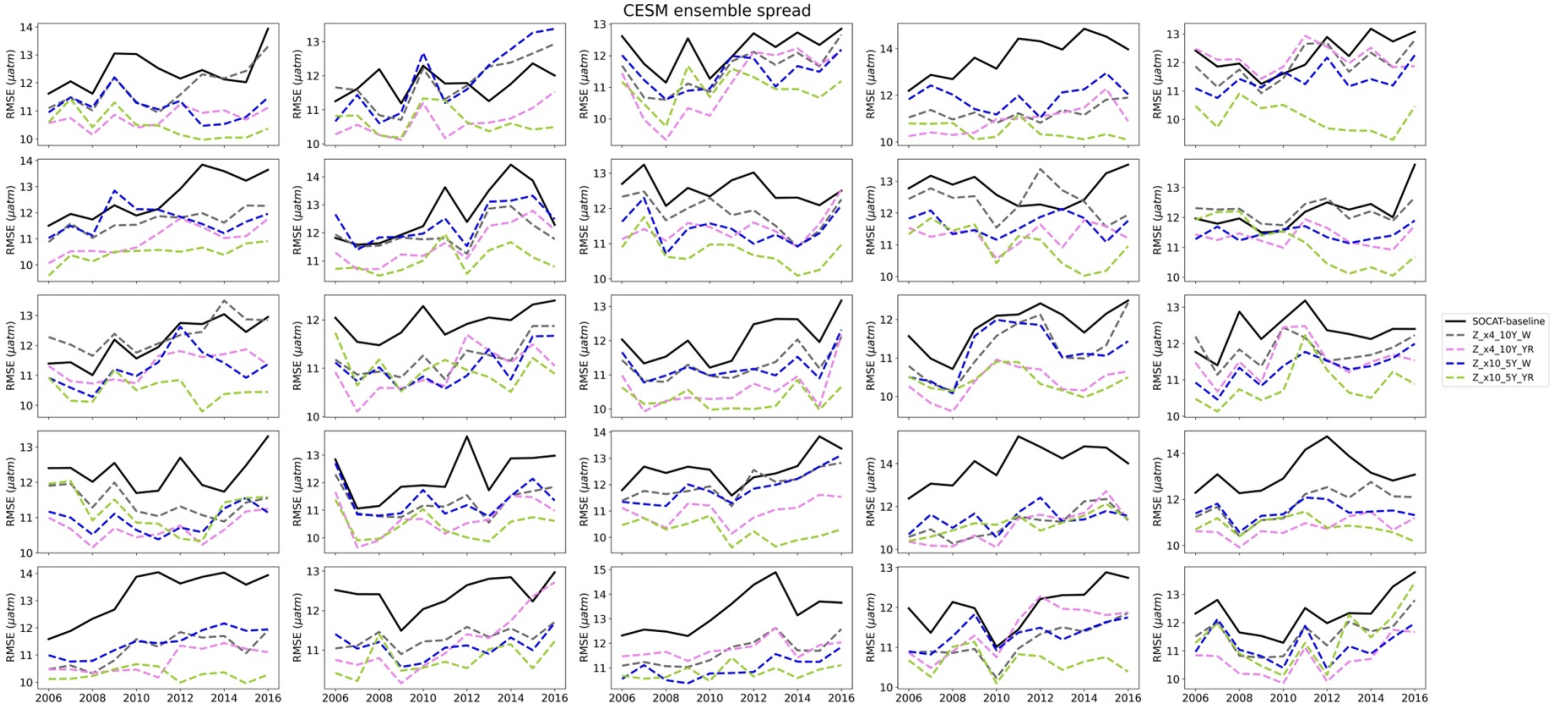
New Fig S10 cont.


A
Southern Ocean (2006/2012-2016)
B
CESM - Southern Ocean (2006/2012-2016)
C
GFDL - Southern Ocean (2006/2012-2016)
D
CanESM2 - Southern Ocean (2006/2012-2016)
RMSE (µatm)
SOCAT
5Y_J-A
10Y_J-A
10Y_YR
x13_10Y_J-A
x13_10Y_W
x5_5Y_W
z_x4_10Y_YR
z_x4_10Y_W
z_x10_5Y_YR
z_x10_5Y_W

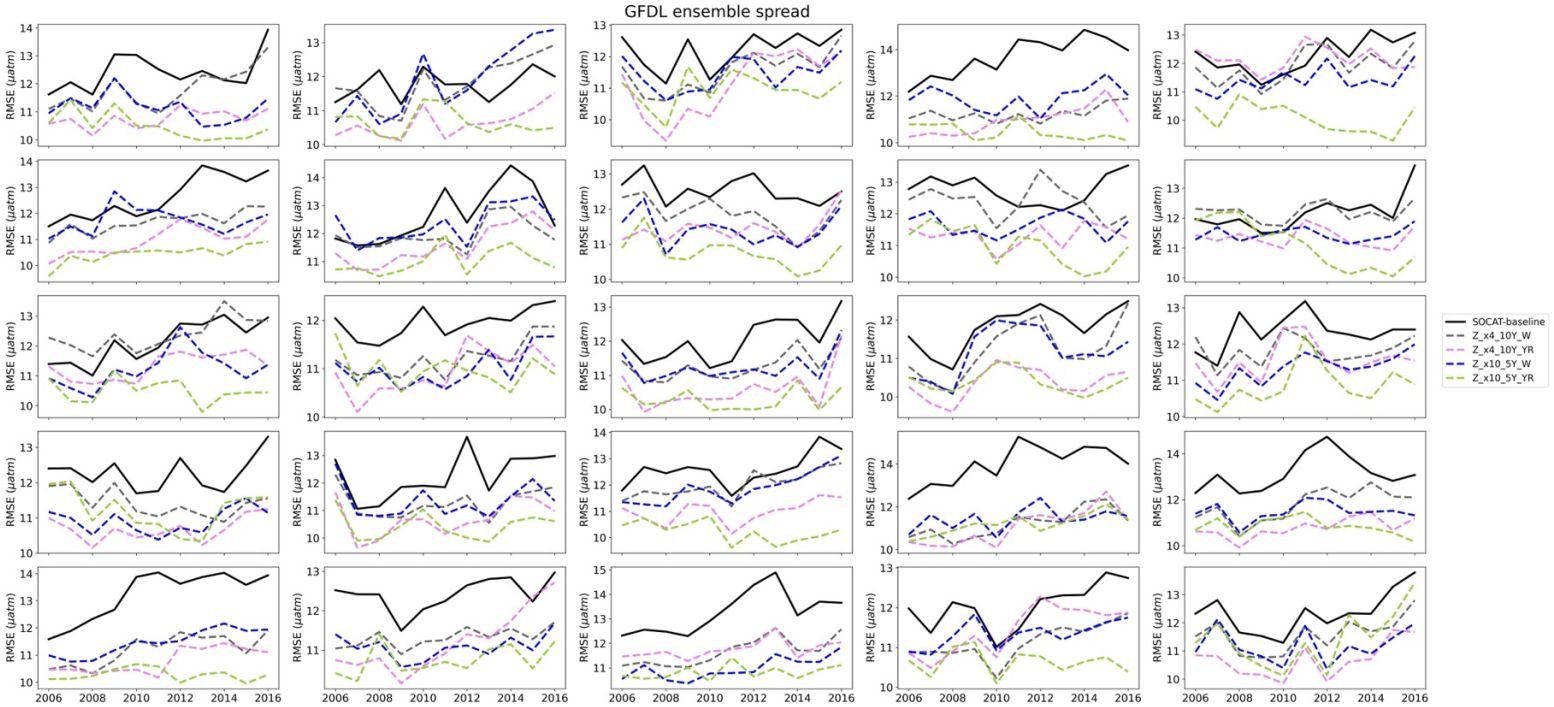
New Fig S14


CanESM2 ensemble spread
RMSE (µatm)
SOCAT-baseline
Z_x4_10Y_W
Z_x4_10Y_YR
Z_x10_5Y_W
Z_x10_5Y_YR

New Fig S16



New Fig S16 cont.



New Fig S16 cont.

Technical corrections:

5) Line 37, Please explain the acronym "fCO_2". Note that the term "pCO_2" 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 fCO_2, which is the fugacity of carbon dioxide (fCO_2) as opposed to pCO_2 which is the partial pressure of CO_2 in the ocean. The fCO_2 is equal to the pCO_2 corrected for non-ideality of CO_2 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 pCO_2, 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 fCO_2 as this is what is released in the SOCAT database, but model output is typically reported as pCO_2 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. 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: "Here, we calculate error statistics based on the full reconstruction (pCO_2 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" and refer to further information in **Section 2.4.2**. We also added some more information under section **'2.4.2 Zigzag runs'**: "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 pCO_2^{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 pCO_2^{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 pCO_2^{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".

11) Line 293: "where algorithm generally overestimates pCO_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, pCO_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: "RMSE is highest in the Eastern Tropical and Southeastern Pacific Ocean and in the Southern Ocean, where the algorithm generally overestimates pCO_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 <u>difference</u> 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****.*

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 ***Fig. S3*** *(****Fig. S5c*** *in the revised version), SOCAT observations start to drastically increase from these times. This was mentioned in the manuscript: "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 pCO_2-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 CO_2 fluxes from surface pCO_2 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 CO_2 fluxes from surface pCO_2 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, 1974.