*Response to reviewer 2 of "**The sensitivity of pCO2 reconstructions in the Southern Ocean to sampling scales: a semi-idealized model sampling and reconstruction Approach**"*
by Djeutchouang, Chang, Gregor, Vichi, Monteiro

*First of all, we would like to take the opportunity to thank you for the thoughtful comments and suggestions. Meanwhile, we have revised the title of the study as follows: "**A semi-idealized model sampling and reconstruction approach across a Southern Ocean front: the sensitivity of pCO2 reconstructions to sampling scales**". We will respond (in italics) to each of your specific comments as follows.*

Summary and overall impression

The paper addresses an important question for the global carbon cycle community: how to reduce the uncertainties and biases of machine-learning-based mapping approaches in the Southern Ocean, a data-sparse but globally important region. The authors create synthetic data by subsampling a high-resolution model in a subregion of the Southern Ocean over 1 year. The synthetic data resembles different observational platforms in terms of the typical temporal resolution of the different platforms. These platforms include ship, float, Windglider, and Saildrone data. They then run two different machine learning mapping approaches with these synthetic observations and compare the mapped reconstruction of the seasonal cycle to the actual model field to estimate the biases and uncertainties. They run the method multiple times with different subsets of synthetic data to highlight how sampling in different seasons, as well as with different types of observational platforms affect the uncertainty and bias. They find that the addition of wintertime ship data would greatly reduce the errors in the reconstructions. They also find that Saildrones are an optimal platform to address both the large-scale spatial and high-resolution temporal sampling and have the most effective impact on reducing the uncertainties and biases of the seasonal and annual mean reconstructions of air-sea CO2 fluxes in the Southern Ocean.

This paper addresses a crucial gap in our current knowledge and provides suggestions on how the carbon cycle community can improve current estimates of the Southern Ocean carbon fluxes, through an improved sampling strategy. I very much support the method of using synthetic data to create a sampling strategy, the paper is well-written and has clear figures that support the findings. However, I have some major comments, which I believe should be addressed before publication. In my review, I mostly focus on the Methods section, as requested by the editor.

Thank you.

General comments

• My main concern is that the machine-learning approach used to reconstruct the model fields is very different to the common methods (e.g., by Landschuetzer, and Gregor…) and thus, I am not convinced that the lessons learned from the authors' approach can be directly translated to these methods. Specifically, the established mapping methods use training data from quite large regions (from a clustering step), which are a lot bigger than the region of this study. Thus, more

data flows in, and they might be more robust to be able to reconstruct the seasonal cycle despite the sparsity in winter data. In addition, there are zonal differences and hot spots of in the Southern Ocean, and the subregion might not be representative for the Southern Ocean as a. whole. I do think we can still learn from this current study, but this issue should be discussed thoroughly. A follow-up study could later focus on the Southern Ocean as a whole (or even globally within e.g., the clusters by Landschuezter or Gregor et al.).

Response: Thank you for taking the time to provide these important comments. We did look into the two commonly-used reconstruction methods by Landschutzer et al. (2014) and Gregor et al. (2019) that both adopt a two-step machine learning (ML) approach in which the first step consists of clustering the reconstruction domain whereas the second step applies ML regression and mapping in each cluster generated. We are aware of the necessity of this clustering step to overcome the spatial and temporal limitations of observations. In Fig. (1a), we illustrated the Southern Ocean Fay and McKinley (2014) biomes, one of the clustering methods used by Gregor et al. (2019) This figure helps to understand the motive of skipping the clustering step in this particular study as it shows that the clustering step was not necessary given the size of the study domain.

This study domain (black box, Fig. 1a, of this note) was not only spatially and temporally coherent but also big enough to reflect the spatial and temporal variability necessary to provide sufficient sensitivity to the different sampling strategies. We also recall, as we did with reviewer 1, that while our selected domain does not resolve all Southern Ocean scales, it is representative of the scale variability we aimed to address. This sub-region is roughly 50% STSS/SPSS of the Southern Ocean (Fig. 1a-b, of this note). Further, it is divided by the Sub-Antarctic Front (SAF) and thus also overlaps both the Sub-Antarctic Zone (SAZ) and Polar Frontal Zone (PFZ) which are relatively the two most sampled regions of the Southern Ocean.

We do acknowledge that the sub-region is smaller compared to the clusters and that might have some impacts. That is why a follow-up study is being conducted to extrapolate and test the idea in the entire Southern Ocean while including in the method a clustering step like in Landschutzer et al. (2014) and Gregor et al. (2019) in order to overcome the spatial and temporal limitations of observations.

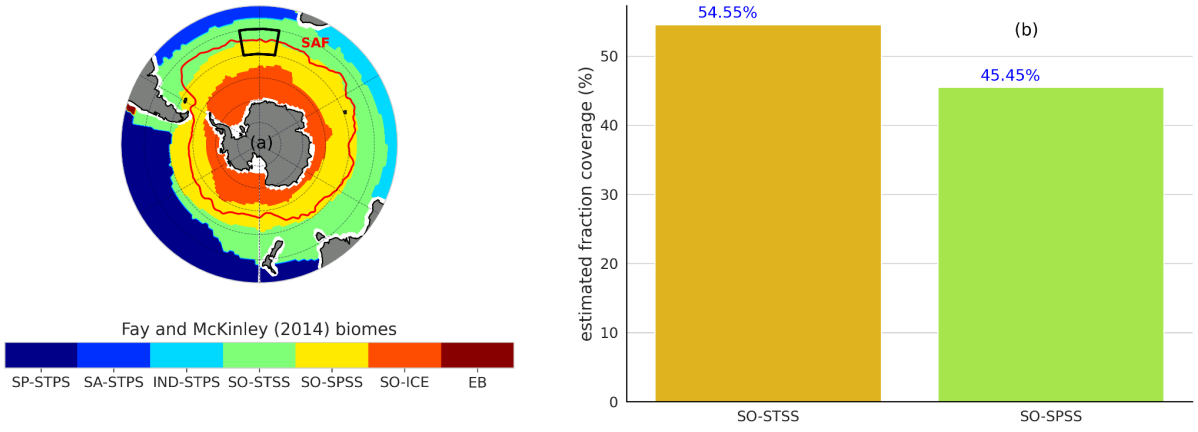

Figure 1: *Panel **(a)** is the Southern Ocean regions or biomes (Fay and McKinley, 2014) as extended and used in Gregor et al. (2019) on which are added the Sub-Antaractic Front (SAF) (red line) and the study domain (black box); and panel **(b)** show the fraction coverage estimates (%) of the two most sampled regions: STSS and SPSS biomes relatively to the area of our box. EB biome stands for Eastern Boundaries (Gregor et al., 2019). For other biome abbreviations (below the colour bar), see Fay and McKinley (2014).*

• I think it's great that the authors use an ensemble of two ML-based approaches. However, I would appreciate a short analysis of how the two estimates differ, to understand how robust the findings are.

Response: Thank you for this comment. Indeed, we reported in the Supplementary Materials (Table S3) a short analysis of the in-sample errors of the two members of the ensemble for all the different sets of experiments we performed.

• The authors say that there is very limited data to allow for both training and testing/validation data. But how do the authors then know that the outcome is not overfitted? As the data is synthetic, could one not add more synthetic data that would then allow for both training and testing data?

Response: Thank you for these questions. Indeed, given the size of the study region and the idea of mimicking as much as possible the sampling scales of observing platforms, the sample sizes were relatively small to afford splitting the simulated observations for both training and testing.

However, to answer the question about overfitting was addressed as follows. To better control the overfitting, we incorporate a K-fold cross-validation (CV) during training in order to find the set of hyper-parameters that enable a better generalisation of ML2. The K-fold CV is applied identically to each of the two-member algorithms (like in Gregor et al., 2019) and the tuning of hyper-parameters was achieved using Bayesian optimization instead of the standard grid-search CV. The optimal values of hyper-parameters used were reported at the end of the model training and are included in the revised Supplementary Information for reproducibility.

About the question regarding adding more synthetic data to allow both training and testing data, in fact, that is what we are doing by comparing our results with the model data (known truth) that were not involved in the synthetic platforms simulations.

Specific and minor comments to the text:

Introduction:

• I found the introduction a bit misleading. After reading the introduction, I expected that the paper would include the interannual to decadal variability, but it "only" focuses on the seasonal cycle based on data from one year. Consider rephrasing this to not disappoint the reader.

Response: Thank you for this suggestion. We have carefully taken this into consideration and the revised introduction refocuses on the seasonal cycle based on one year-round data in a semi-idealized subdomain of the model within the Southern Ocean. It also emphasizes the two most observed regions of the Southern Ocean; that is, the Sub-Antarctic Zone (SAZ) and Polar Frontal Zone (PFZ) whose the chosen subdomain represents the most.

• Similarly, the introduction should mention clearly that this study "only" focuses on a subregion within the Southern Ocean.

Response: Thank you for the suggestion. The revised introduction takes it into consideration.

• Gloege et al. 2021 did an in-depth analysis of the uncertainty of ML-based mapping approaches, using synthetic data at a global scale. I think the introduction should mention that study and be explicit about how this current approach differs and what's new about this study in comparison.

Response: Thank you for this suggestion. We are aware of the Gloege et al. (2021) study and the advances their work made in this area of research. The revised introduction makes a deeper and explicit connection with that study and explicitly mentions how our approach differs.

• L.114: It's mentioned later that how well the model matches the observations does not really matter in this context. However, please consider mentioning here already why using that model works (considering e.g., the Mongwe et al. 2018 study that showed how the CMIP models completely disagree on the phase and magnitude of the seasonal cycle).

Response: Thank you. Regarding this comment, the revised version takes it into consideration by giving further explanation around that statement as follows. Having a complete model data, the full domain-truth knowledge of pCO2 variability can be assumed to be known independently of that particular model constraints. Therefore,we think using any physics-biogeochemistry forced ocean model output (including any model from CMIP5 models) would work as long as the model can represent the range of temporal and spatial modes of variability that are necessary to mimic the sampling scales of interest.

The modes of variability that the forced model (NEMO-PISCES) captures are still responses of pCO2 that are in some way related to the changes in driver variables (or related proxies). Thus, for the purpose of this study, the "correctness" of the pCO2 response to the driver variable is not the most important. What we try to show is that the reconstruction is sensitive to the way one samples with different modes of variability.

• L.196: Is this really the case? I would assume that any differences between the model and the observations could matter. I.e., the model might be generally a lot smoother than reality and thus the sampling strategy might be less sensitive in the model than in the real world. I would appreciate a short discussion on that.

Response: Thank you for this question and suggestion. Actually, the forced coupled ocean model (NEMO-PISCES) used in the study does not represent the full processes driving the CO2 in the Southern Ocean. As for the machine learning (ML) methods used in this study; that is, the feed-forward neural network, and the tree-based method gradient boosting machine, they do not capture the mechanisms that actually drive the pCO2 (Holder and Gnanadesikan, 2021). Rather, these ML methods capture changes in the drivers and then the associated changes in the pCO2.

Therefore, we thank the reviewers for their suggestion of assuming that any differences between the model and the observations could matter because this could mean the pCO2 may be driven by a mechanism or process different from the real world. Indeed, this does seem like a plausible option given that model pCO2 is largely driven by the SST as in the real world.

• L.335: I think the explanation of error and uncertainty is the wrong way round.

Response: Thank you for this comment. We took it into consideration by revising our explanation of the two concepts as follows. "The pCO2 total uncertainty (E) is dealt with as in Gregor and Gruber (2021). The authors identified within the surface ocean carbonate system three main sources of errors that contribute to E. This included (1) the measurement (M), (2) representation (R), and (3) prediction (P) errors. Under the assumption that these three components are independent of each other in the pCO2 uncertainty space, E can thus equivalently be expressed as …"