We thank the referee for her/his review and comments.

Referee: This study conducts Observation System Simulation Experiments for an optimal observation system of surface ocean pCO2. The results give an important information to future observational strategy. The manuscript is well organized and the description is easy to follow. Followings are just a few points to be possibly improved.

The authors mentioned the analysis period is from 2008 to 2010. But they also mentioned data for the period 2001-2010 from SOCAT were used (line 96 and Table 1). How the data in 2001-2007 were used?

In Figure 1, the same color bar in all panels would be better to easily compare data numbers from respective data source and to avoid unnecessary white space.

Labels on horizontal axes in Figures 7g, 7h, 8g and 8h are different from those in the other panels.

Authors: We thank the reviewer for the positive evaluation of our study.

The SOCAT data for a particular month or even year are very sparse. Here we use SOCAT data from a larger period (2001-2010) than the period targeted by the reconstruction (2008-2010) for three reasons: 1. to benefit from a larger data set for Neural Network calibration; 2. to capture interannual variability from a long historical record of SOCAT data; 3. to see how SOCAT data can be enhanced by adding other observational platforms to the effort. While the data from 2001 to 2010 are used in training, the reconstruction focuses only on the years 2008 to 2010. We clarified this in the manuscript, lines 104-107 in the revised version.

Thank you for the comment on Fig. 1. We changed the colour bar.

Thank you to letting us know that labels in Fig. 7 and 8 need to be corrected. We modified it in the revised version.

We thank Luke Gregor for his interesting and important comments.

Please find below our responses to the comments.

Luke Gregor (LG): DISCUSSION

I feel that the study could be more strengthened by adding a discussion section where the implications of the study and limitations of the approach are addressed. Some points that would be good to add here are (not in order of importance):

Higher resolution compared with previous studies: This is a considerable step up from the typically used 1° by monthly resolution. The high temporal resolution is particularly interesting, especially considering the inclusion of mooring data, which could provide data at a daily resolution. An interesting question is then related to scale, are these improvements realized when the experiment is run at lower resolution? This is perhaps not a question for this study, but it would be interesting to test at some point.

Authors (A): The experiments were carried out using the outputs of a global ocean model NEMO/PISCES with a nominal resolution of ¼°. Usually in the literature we can find the reconstruction resolution 1°. With the Argo coverage of 1 profile per 3° box we believe that the results are not significantly affected by spatial resolution.
The temporal resolution could be one month if Argo profiles provide measurements for a longer period. As the FFNN is built to reconstruct each month, we suppose that our results will be still valid on monthly resolution.

LG: Sampling frequency of floats and representation error: In your study the sampling frequency is 5 days. Typically, Argo floats sample at 10 days. Further, Argo floats that sample at a 10-day resolution are only briefly at the surface. This topic is discussed in Monteiro et al. (2015, https://doi.org/10.1002/2015GL066009). In a dynamic region, this might introduce representation error (where the in-situ measurement is not representative of the 5-day mean). This is perhaps a limitation of the study that should be mentioned.

A: Thank you for this interesting comment. It is true that Argo floats provide measurements every 10-days and it can be a limitation of our study to apply it to the real observations as it does not represent an average value over 5 days. In this study we paid more attention to the spatial distribution and we believe that with Argo measurements for a longer period our results can be applied to one-month time steps. In this case, 3 monthly measurements can be representative of a monthly mean. We added in the manuscript (line 118-124 in the revised manuscript): "It is worth noting that Argo floats provide measurements every 10 days. Floats dive to a depth of 2000 m and then rise to the surface by measuring vertical profiles of ocean variables. In this study we use a 5-day time step (see below section b)) which can be a limitation to apply our results to real observations as it does not represent an average value over 5 days. We paid more attention to the spatial distribution, and we believe that with Argo measurements recorded over a longer period our results can be applied to one-month time steps. In this case, 3 monthly measurements can be representative of a monthly mean."

LG: Other gap-filling methods: the method used here, FFNN, uses latitude and longitude as feature variables. Do the authors think that this could also apply to other approaches such as SOMFFN which uses "remote" information through the clustering approach?

A: FFNN showed its good capacity to reconstruct pCO_2 over the global ocean. The present study builds on our previous findings to improve the results of FFNN in the Southern part of the Atlantic basin.

In principle, the SOMFFN method can also be applied to the type of experiments presented here. However, SOM introduces discontinuities between clusters. We wanted to design a method that can be applied globally.

LG: The comparability of the simulated environment to reality: One has to make the assumption that the model output is some representation of reality in OSSE work. However, it would be good for the authors to make some statement about the model's ability to represent the seasonality and variability of pCO2.

A: NEMO/PISCES has been widely used for carbon cycle studies. The configuration used in this study corresponds to the one described in Gehlen et al. (2020) and Terhaar et al. (2019). Gehlen et al. (2020) presents an assessment of global fields of SST and air-sea fluxes of CO_2. The mean state and the seasonal cycle are compared to observations. Terhaar et al. (2019) focuses on the Arctic ocean. Both papers are open access and we rather point to these papers than dublicating the analysis.

We added the information to the manuscript: "Information on the simulation is given in Gehlen et al. (2020) and Terhaar et al. (2019), including the evaluation of the modelled mean state and the seasonal cycle of sea surface temperature and air-sea fluxes of CO_2 (Gehlen et al., 2020).

LG: The uncertainty of measurements: The authors state that they will address this in their next study. A bit more depth on this topic would really show the need for the study that they are planning to publish.

A: We added in the manuscript's conclusion: "The real measurements contain instrumental and representation errors. The inclusion of errors in pseudo-observations will help to estimate the impact of observations on the reliability of OSSEs presented in this work… The consistent introduction of error estimates for each predictor will provide this information."

LG: Introduction. The introduction could be bolstered with a paragraph dedicated to previous work on OSSEs. See the following literature:

Kamenkovich, I., Haza, A., Gray, A. R., Dufour, C. O., & Garraffo, Z. (2017). Observing System Simulation Experiments for an array of autonomous biogeochemical profiling floats in the Southern Ocean. Journal of Geophysical Research: Oceans, 122(9), 7595–7611. https://doi.org/10.1002/2017JC012819

Majkut, J. D., Sarmiento, J. L., & Rodgers, K. B. (2014). A growing oceanic carbon uptake: Results from an inversion study of surface pCO2 data. Global Biogeochemical Cycles, 28(4), 335–351. https://doi.org/10.1002/2013GB004585
Lenton, A., Bopp, L., & Matear, R. J. (2009). Strategies for high-latitude northern hemisphere CO2 sampling now and in the future. Deep-Sea Research Part II: Topical Studies in Oceanography, 56(8–10), 523–532. https://doi.org/10.1016/j.dsr2.2008.12.008

Scheel Monteiro, P. M., Schuster, U., Hood, M., Lenton, A., Metzl, N., Olsen, A., Rodgers, K. B., Sabine, C. L., Takahashi, T. T., Tilbrook, B., Yoder, J. A., Wanninkhof, R. H., & Watson, A. J. (2010). A Global Sea Surface Carbon Observing System: Assessment of

Changing Sea Surface CO2 and Air-Sea CO2 Fluxes. Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, 1, 702–714. https://doi.org/10.5270/OceanObs09.cwp.64

A: Thank you for pointing to these four additional articles related to our study. We added them to introduction. However, we cited "Majkut, J. D., Carter, B. R., Frölicher, T. L., Dufour, C. O., Rodgers, K. B., and Sarmiento, J. L.: An observing system simulation for Southern Ocean carbon dioxide uptake, Philos. Trans. R. Soc. A, 372, https://doi.org/10.1098/rsta.2013.0046, 2014." instead of Majkut, J. D., Sarmiento, J. L., & Rodgers, K. B. (2014). A growing oceanic carbon uptake: Results from an inversion study of surface pCO2 data. Global Biogeochemical Cycles, 28(4), 335–351. https://doi.org/10.1002/2013GB004585

We added "Monteiro et al., 2010" on Line 42.

Line 51-58: "Majkut et al. (2014) and Kamenkovich et al. (2017) reported on observing system simulations with autonomous biogeochemical profiling floats in the Southern Ocean that improve estimates of carbon dioxide uptake and biogeochemical variables. While Majkut et al. (2014) used a coarse-resolution model and fixed floats, Kamenkovich et al. (2017) extended this work to a more realistic case with moving floats and high-resolution numerical simulations. Based on a coupled climate carbon model and observations, Lenton et al. (2009) proposed sampling strategies to obtain large-scale integrated CO_2 fluxes in the North Pacific and North Atlantic. They show that regular sampling of ocean surface pCO_2 with a 3-month time step and every 6° in latitude and 10° in longitude is sufficient to capture more than 80% of total CO_2 flux variability."

LG: FIGURES

I like the choice of figures, but I feel that they could be improved by keeping the following in mind:

- Increase the axes label and legend text size
- Improve subplot title and number placement
- Captions should be more descriptive. There is very little information at the moment.
- The data to ink ratio is sometimes skewed too heavily toward the latter. The data should always be prioritized. Some pointers in this regard:

  Remove excess axes lines (e.g. target plots have a lot of empty space)

  If axes limits are shared, only the leftmost and bottom figures would need axes labels.

  For maps, don't draw rivers – they are not important for your study.

  Scale the color bars proportionately to the figures.
- I've noticed that markers for OSSEs 1, 3, 4, and 10 are in bold. Be explicit why these are bold.

A: We increased the axes label in Fig. 4, 7, 8, 9 and legend text size in Fig. 3, 4, 7, 8 and 9.

We improve subplot titles and number placement in Fig. 4, 7, 8 and 9.

We added more information to each caption. Please find a description of modifications for each figure at the end of this document.

Target plots were modified to remove empty space. We removed the shared labels in Fig. 5, 6, 7, 8 and 9. We removed rivers form the continents in Fig. 1, 2, 5 and 5. Colour bars were re-scaled on Fig. 1, 2, 5 and 6.

We plotted OSSEs 1, 3 and 10 in bold to emphasise them as these three OSSEs represent the OSSEs retained of a detailed comparison and discussion. OSSE 4 was erroneously plotted in bold, it was corrected. We added it to Figures 4 and 9, Tables 3 and 4.

Comments in text.

Line 97 (line numbers correspond to the initial manuscript): **LG:** Do you mean the raw data that is used to make the gridded data? If so, this data is not daily, but saved at the native sampling resolution (on the order of minutes).
A: Thank you for this remarque, yes, it is raw data referred to as synthesis SOCAT v5. We modified the sentence and clarified what we meant by "daily measurements", line 108 in the revised manuscript.

Line 101: **LG:** What do you mean by it here? The argo network, or the product that you used? Can you give a link to the data?
A: By "it" we meant the Argo network defined in the study by Gasparin et al. (2019). We replaced "it" by "This network" (line 112 in the revised manuscript). The data are available from Gasparin et al. (2019). We also added the citation of the original Argo product and its DOI (line 111).

Line 102: **LG:** Can you comment on how this coarser sampling pattern would influence the results? Is the profiler always at the same point in this 3° box?
A: The distribution of Argo measurements is 1 measure per 3° box per 10 days. We added "per 10 days" in the text (line 113 in the revised manuscript). It corresponds to the real Argo trajectories that are mentioned in the text. It means that one profile can leave the box and another one can come.
We would not say that this sampling pattern is coarse or coarser. It improves the coverage by SOCAT, and we have already seen that the SOCAT data distribution works well for Machine Learning in the Northern Hemisphere. Thus, we believe that the distribution of 1 profile per 3° box per 10 days is a good starting point. Also, as it was mentioned in the manuscript, this distribution corresponds to the BGC Argo target: "The target for BioGeoChemical Argo (1/4 of ARGO coverage) (Bittig et al., 2018) was derived from this distribution."

Line 139: **LG:** Cite Kamenkovich et al (2017) and Majkut et al. (2014) here.
A: We cited these articles in introduction. Here we explain our motivation to add Argo float only in the Southern Hemisphere to improve the pCO_2 reconstruction over the Atlantic basin.

Line 170: **LG:** What do you mean by regularly? Randomly or spaced at regular intervals?
A: Each third grid point was kept for evaluation of FFNN model, each fourth for its validation. We added this information in the text, line 189 in the revised manuscript.

Line 171: **LG**: Do you mean weights or nodes here?
A: Parameters are weights. We added in the text "parameters/weights", line 190.

Line 172: **LG:** Perhaps rephrase this to read something like this:
“we limited the number of nodes/weights to the number of training points ÷ 10.”
A: We modified the sentence: that suggests limiting the number of parameters to the number of training data points divided by 10 to avoid overfitting (line 191 in the revised manuscript).

Line 174: **LG:** What is the size of each hidden layer?
A: The input layer has 15 input nodes and 20 output nodes that represent the input for the first hidden layer. The first hidden layer has 25 output nodes and the second hidden layer – 10 output nodes. We added this information in the text, lines 193,194 and 196 in the revised manuscript.

Line 221: **LG:** What is a shadow cloud?! I just saw this in the figures. I would refer to this as shading, rather than shadow cloud.
A: We replaced “shadow cloud” by “shading” throughout the text.

Line 230: **LG:** I don’t have an issue with this since your comparison is internally consistent. But for future work, NCEP 1 and 2 have some issues in the Southern Ocean (and possibly in other regions) according to Swart et al. (2015). ERA5, JRA55/CCMP/ERA5 are all good alternatives.
A: Thank you for this comment. As you mentioned, our comparison is internally consistent. We agree that alternative atmospheric forcing products could be used in the future.

Line 240: **LG:** this part needs to be rephrased. What does "others" refer to?
A: We rephrased it: Over the regions with poor observational coverage the results from OSSE 1 lie at a distance from results of all other OSSEs (line 261).

Line 243: **LG:** what does X mean?
A: Our objective is to remind the reader of the marker symbol for OSSE 3. We added “” and explained it when the marker symbol was first introduced in the text (line 259 in the revised manuscript).

Line 258: **LG:** would be more clear to replace and with "/" since this is how you present the metrics
A: Thank you for suggestion, we made the modification.

Line 274: **LG:** While it is not large and is thus less likely to have a large impact on fluxes, I noticed that biome 12 has a positive bias when SOCOM-like floats are included. This is an interesting element that should be discussed.
A: We expanded the discussion in the next paragraph (lines 312-313 in the revised manuscript): “Error compensation also contributes to positive biases computed for OSSEs 6-11 for biome 12 (Table 4). Additional data from Argo floats correct the negative bias in the southern part of the biome close to the African coast (Fig. 5c). Thus, the strong positive bias in the northern part becomes dominant and results in a total positive bias.”

Line 343: **LG:** It would be good to highlight that biome 13 is the SH equivalent of biome 11. Really highlight the fact that low data coverage is a strong driver of the differences here. Further, this region should be relatively easy to resolve as pCO2 would be strongly driven by temperature.
A: We added, lines 369-370 in revised manuscript: “. This region has a dynamic similar to biome 11 in the Northern Hemisphere, however the data coverage in biome 13 represents only 15% of data coverage in biome 11 (Fig. S5).” We added a new figure S5 to the

Supplementary material that shows the seasonal data distribution per biome. Figures S5-S17 have been renumbered accordingly.

Line 353: **LG:** Could you also say something about the seasonality of the bias? For OSSE 1, are estimates better in summer where there is a greater data density?
A: Regarding biome 17, estimates are improved during the winter season. It is also coherent with the seasonal data distribution that shows more data available for ML in December-January (please refer to the new Fig. S5). We added, lines 381-383: "OSSE 1 underestimates the pCO$_2$ in this region over the full seasonal cycle. The maximum difference is obtained in September-October, which also corresponds to the months with the lowest number of available observations (Fig. S5)."

Line 369: **LG:** There is definitely a seasonal element to the biases (Fig 8f) - the FFNN systematically overestimates flux during winter. The other peaks in the time series are not as apparent.
A: We added, lines 401-403 in the revised manuscript: "The maximum differences between OSSE 1 and NEMO/PISCES are systematically found in January and June, the months with the lowest number of available observations for training (Fig. S5)."

Line 428: **LG:** This should be included to indicate that there is not a major loss in performance
A: We did not find to what this comment was related.

Line 440: **LG:** These could be pseudo mooring instruments - e.g. surface gliders, sail-drone, or sail buoy.
A: Thank you for this comment, we added, line 473: "as well as sail-drones and sail buoys".

Line 599: **LG:** Biome 12 is very interesting. The inclusion of floats results in a larger bias... Why is this?
A: It results from error compensation as explained in our answer to your comment to line 312 in the revised manuscript.

Figures' improvement:
Line 589: **LG:** Could you make the plots bigger. Very difficult to see the details. I would also not include the rivers on the land. This is an unnecessary detail. This applies to all maps.
A: Figure 1: Sub-plots were made bigger, and rivers were removed from maps.

A: Figure 2: Rivers were removed from the continents. Colour bar was rescaled and only presented biomes were kept.

Line 596: **LG:** More text would be useful. i.e. what is purple. It is in the text, but would be useful for the reader to see here.
For example, it would be useful to know that OSSE 1 is based on SOCAT. Perhaps you can add this to the text label above the figure (top right corner)
A: Figure 3: Figure captions were extended. Y-axis of Standard Deviation was dropped out in internal sub-plots to win more space.

Line 600: **LG:** I like the target plots, but I feel that they do not give the data priority - there is a lot of axes material compared to the data. For example, this could be achieved on a cartesian axes too? That would allow you to have only the positive uRMSD values. And without the circles, one could reduce the scale of the bias axes to -15:15, which would give the data a bit more space.

Lastly, this might be a bit nit-picky, but would it be possible to put the markers over axes lines?

Again, I think it would benefit the reader to have labels of the experiment set up above each plot, e.g., OSSE 1: SOCAT
A: Figure 4: We have zoomed the target diagrams to exclude the empty space. The caption has been supplemented with more details.

Line 603: **LG:** I recommend that you add the labels for the OSSE methods above each figure. Much easier for the reader. Great results BTW :) would it be possible to give the figures a bit more space? I also don't think LAT labels are necessary for each figure - this way each plot can be bigger.
These comments also apply for Figure 6
A: Figures 5 and 6: We made the main figures bigger compared to colour bars, also rivers were removed from continents. We added names of OSSEs for each column. More information was added in the caption. We kept only one y-axis on the first sub-plot.

Line 612: **LG:** The label positions of the biomes could be improved. You should have the dates for the x-axes in figures g and h.

Change shadow to shading.

It would be useful to see the SOCAT sampling density over time for these figures, particularly for biome 13, where the biases are likely due to a lack of sampling. This would really drive the point home.
A: Figures 7 and 8: We changed labels and legend size, also the x-axis is presented only in the bottom sub-plots to win more space. The captions have been supplemented with more details. We added a new figure S5 to the Supplementary material that shows a seasonal data distribution per biome.

Line 625: **LG:** I like these plots! They provide very valuable information about the sampling density. Though, the 0-line should extend the entire length of the axes. I thought it was a line denoting an average of something when first looking at the figure.

biome label placement could be better.
A: Figure 9: We added x and y cartesian axes, axis labels are made bigger as well as markers in figures. We dropped out common labels in internal sub-plots to win more space.

Corresponding modification have been made in the Supplementary material as well.