Comments on the manuscript entitled "Assessing improvements in global ocean $p\mathrm{CO}_2$ machine learning reconstructions with Southern Ocean autonomous sampling"

November 15, 2023

This study exploits Large Ensemble Testbed (LET) experiments targeting to prompt meridional and winter samples by Saildrone USVs in the Southern Ocean to improve the reconstruction of surface seawater partial pressure of CO_2 ($p\mathrm{CO}_2$) and air-sea fluxes. For LET, 75 Earth System Models (ESM) have been selected to provide input ($p\mathrm{CO}_2$ and potential driver variables) for a machine learning (ML)-based mapping approach. Two primary exercises have been conducted: ML-based reconstructions of $p\mathrm{CO}_2$ with only SOCAT baseline and with Saildrone USVs sampling tracks added. Data reconstructions are evaluated with the model truth. In this manuscript, the authors have demonstrated that the reconstructions with additional USV data allow reducing the errors in $p\mathrm{CO}_2$ and flux estimates. 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 $p\mathrm{CO}_2$ 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.

1. 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 pCO_2. Here, the algorithm is assisted by removing this known temperature effect, and it must therefore only learn the pCO_2 impacts from biogeochemical drivers*": there exist many other ML approaches [Friedlingstein et al., 2022] which do not separate the SST-effects from others on $p\mathrm{CO}_2$ but succeeds in estimate $p\mathrm{CO}_2$. The major concerns are how to assess the uncertainty derived from SST effect removal and impacts on the experiment outputs.

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 $p\mathrm{CO}_2$-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.

Editorial and specific comments:

1. Lines 11-12: "anthropogenic" can be removed. The SO has taken up atmospheric CO_2 without specifying natural or anthropogenic sources.

2. Line 37: "$f\mathrm{CO}_2$" is not defined. "uncertainty of < 5 μatm": this holds only for the measurements chosen to provide gridded SOCAT datasets.

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

4. Line 45: "These data products" $\longrightarrow$ "These methods".

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

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

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!

8. Line 57: "indirect $p\mathrm{CO}_2$ estimates": can you define this term? Are they computed from float measurements of other carbonate variables?

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 $p\mathrm{CO}_2$ but to draw this statement, it requires to provide the availability of USVs to sample $p\mathrm{CO}_2$ by showing the sampling frequency and data coverage area over the SO?

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

11. Lines 89-90: "*in an ESM, surface ocean pCO_2 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?

12. Lines 161-162: "*where pCO_2 mean and SST mean is the long-term mean of surface ocean pCO_2 and temperature, respectively, using all 1°x1° grid cells from the testbed*": $p\mathrm{CO}_2$ mean is different regionally, why you don't compute a global map of $p\mathrm{CO}_2$ mean?

13. Lines 165-168: Please clarify. The authors have excluded $p\mathrm{CO}_2$-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 $p\mathrm{CO}_2$-Residual values can not 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 $p\mathrm{CO}_2$ values higher than 816 μatm (e.g., over permanently or seasonally strong upwelling regions: Eastern Equatorial Pacific, Western Arabian Sea, Benguela, etc).

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.

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 model-based data excluding all used in ML training. Specifically, the evaluation should not consider 'zigzag+one-latitude' ('SOCAT+USV') $p\mathrm{CO}_2$ data.

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?

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.

18. Line 386: ''Z_x10_5Y_YR

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 $p\mathrm{CO}_2$ 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.

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 eduction 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 increases 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.

21. Lines 552-554: "*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, 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.

References

S. M. Bushinsky, P. Landschützer, C. Rödenbeck, A. R. Gray, D. Baker, M. R. Mazloff, L. Resplandy, K. S. Johnson, and J. L. Sarmiento. Reassessing southern ocean air-sea co2 flux estimates with the addition of biogeochemical float observations. *Global Biogeochemical Cycles*, 33(11):1370–1388, 2019.

A. Denvil-Sommer, M. Gehlen, and M. Vrac. Observation system simulation experiments in the atlantic ocean for enhanced surface ocean pco_2 reconstructions. *Ocean Science*, 17(4):1011–1030, 2021. doi: 10.5194/os-17-1011-2021. URL https://os.copernicus.org/articles/17/1011/2021/.

P. Friedlingstein, M. O'Sullivan, M. W. Jones, R. M. Andrew, L. Gregor, J. Hauck, C. Le Quéré, I. T. Luijkx, A. Olsen, G. P. Peters, W. Peters, J. Pongratz, C. Schwingshackl, S. Sitch, J. G. Canadell, P. Ciais, R. B. Jackson, S. R. Alin, R. Alkama, A. Arneth, V. K. Arora, N. R. Bates, M. Becker, N. Bellouin, H. C. Bittig, L. Bopp, F. Chevallier, L. P. Chini, M. Cronin, W. Evans, S. Falk, R. A. Feely, T. Gasser, M. Gehlen, T. Gkritzalis, L. Gloege, G. Grassi, N. Gruber, O. Gürses, I. Harris, M. Hefner, R. A. Houghton, G. C. Hurtt, Y. Iida, T. Ilyina, A. K. Jain, A. Jersild, K. Kadono, E. Kato, D. Kennedy, K. Klein Goldewijk, J. Knauer, J. I. Korsbakken, P. Landschützer, N. Lefèvre, K. Lindsay, J. Liu, Z. Liu, G. Marland, N. Mayot, M. J. McGrath, N. Metzl, N. M. Monacci, D. R. Munro, S.-I. Nakaoka, Y. Niwa, K. O'Brien, T. Ono, P. I. Palmer, N. Pan, D. Pierrot, K. Pocock, B. Poulter, L. Resplandy, E. Robertson, C. Rödenbeck, C. Rodriguez, T. M. Rosan, J. Schwinger, R. Séférian, J. D. Shutler, I. Skjelvan, T. Steinhoff, Q. Sun, A. J. Sutton, C. Sweeney, S. Takao, T. Tanhua, P. P. Tans, X. Tian, H. Tian, B. Tilbrook, H. Tsujino, F. Tubiello, G. R. van der Werf, A. P. Walker, R. Wanninkhof, C. Whitehead, A. Willstrand Wranne, R. Wright, W. Yuan, C. Yue, X. Yue, S. Zaehle, J. Zeng, and B. Zheng. Global carbon budget 2022. *Earth*

System Science Data, 14(11):4811–4900, 2022. doi: 10.5194/essd-14-4811-2022. URL https://essd.copernicus.org/articles/14/4811/2022/.

J. Hauck, C. Nissen, P. Landschützer, C. Rödenbeck, S. Bushinsky, and A. Olsen. Sparse observations induce large biases in estimates of the global ocean co2 sink: an ocean model subsampling experiment. *Philosophical Transactions of the Royal Society A*, 381 (2249):20220063, 2023.

P. Landschützer, T. Tanhua, J. Behncke, and L. Keppler. Sailing through the southern seas of air–sea co₂ flux uncertainty. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2249):20220064, 2023. doi: 10.1098/rsta.2022.0064. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2022.0064.