

Insufficient observational coverage is one of the major challenges in quantifying the Southern Ocean carbon sink, while its impact on monthly gridded $p\text{CO}_2$ estimates has not been effectively quantified. This manuscript addresses this important gap by using high-frequency synthetic observations from E3SM model to assess how incomplete spatial and temporal sampling within monthly $1^\circ \times 1^\circ$ grid cells affects estimates of surface $p\text{CO}_2$. The study has a clear scientific value, and the manuscript contains several interesting and potentially important results.

However, I have a few concerns about whether the current methodology cleanly isolates the sampling error that the study aims to quantify. In particular, the matching of historical observations to nearby synthetic floats may introduce an additional spatial representation error. I also find the attribution of the predominantly negative sampling error to high-wind undersampling insufficiently demonstrated, and several broader conclusions regarding real-world Southern Ocean CO_2 uptake appear stronger than can be supported directly by the present experiment. Below are the detailed comments.

Major comments

1. Spatial matching between historical observations and synthetic floats.

Historical SOCAT and BGC-Argo observations are not sampled directly from the model at their actual locations. Instead, each observation is matched to a nearby synthetic E3SM float observation. Although the temporal mismatch is limited to approximately two hours, the spatial mismatch can reach approximately 250 km and matches within a 2° radius are accepted. The authors also acknowledge that the quantitative results are sensitive to the data reduction associated with this step. This is potentially important because the quantity being investigated is the sampling error within a monthly $1^\circ \times 1^\circ$ grid cell. A displacement of tens to hundreds of kilometres is comparable to, or larger than, the spatial scale of the target grid and may cross fronts or mesoscale gradients where $p\text{CO}_2$ differs substantially. I wonder why the model was not sampled or interpolated directly at the historical observation locations and times. Besides, sensitivity tests are needed. The analysis could be repeated using smaller matching-distance thresholds or limiting the historical and synthetic observations to fall within the same $1^\circ \times 1^\circ$ grid cell. This would help understand how much of the reported error arises from incomplete sampling and

how much result from the spatial displacement between real and synthetic sampling location.

2. **Attribute the negative sampling error to high-wind undersampling.** The Abstract states that the large negative sampling errors are “primarily caused by undersampling during high-wind events”. Figure 10 clearly suggests that high-wind conditions are undersampled. However, demonstrating a wind sampling bias is not equivalent to demonstrating that this bias is the primary cause of the negative $p\text{CO}_2$ sampling error. The manuscript currently provides only indirect evidence for this explanation. A more direct analysis would substantially strengthen the manuscript. For example, the authors could relate the wind sampling bias to the $p\text{CO}_2$ sampling error for individual grid-months. The apparent low-wind bias of BGC-Argo sampling also deserves further explanation. Unlike ships, autonomous profiling floats would not be expected to deliberately avoid storms. It would be interesting to further investigate which factors are responsible for this pattern.
3. **Model dependence of the -13 to $-14 \mu\text{atm}$ estimate.** This quantitative result strongly depends on whether E3SM realistically represents the sub-monthly and sub-grid-scale $p\text{CO}_2$ variability of the Southern Ocean. The authors make an effort to compare simulated high-frequency variability with available observations, which is useful, but the observational validation remains geographically sparse, especially in regions where the model simulates the largest variability. Only a fraction of the original SOCAT observations is retained after the open-ocean and matching criteria are applied, and the authors explicitly acknowledge that the quantitative results are sensitive to the data reduction during matching. The reported -13 to $-14 \mu\text{atm}$ should be presented consistently as an E3SM-based estimate of sampling bias, rather than a directly established value for the real Southern Ocean. Additional validation of the modelled $p\text{CO}_2$ variability against SOCAT or other high-frequency observations would strengthen confidence in the quantitative magnitude. Even if an exact point-by-point comparison is not possible, comparison of regional or seasonal distributions and variability statistics would be useful.
4. **SOCAT-only versus SOCAT+BGC-Argo comparison.** The SOCAT-only and merged datasets contain substantially different sets of grid-months: 5,171 for SOCAT and 18,076 for SOCAT+BGC-Argo. This means that the two averages do not necessarily represent the same locations and months. Differences such as the reported 29% reduction in winter sampling-error magnitude may reflect both: 1.

improved sampling within grid-months that were already observed by SOCAT; 2. the addition of entirely new grid-months sampled only because BGC-Argo observations are available. These are scientifically different effects. I recommend adding a paired comparison using only the grid-months common to both datasets. This would isolate the effect of additional BGC-Argo sampling within otherwise identical grid-months. The newly sampled BGC-Argo-only grid-months could then be analysed separately to quantify the benefit of increased spatial and seasonal coverage.

5. Comparison of sampling error with observational measurement uncertainty.

The manuscript emphasizes that the mean sampling error of 13-14 μatm exceeds the <5 μatm accuracy criterion for ship-based measurements and estimated systematic uncertainties of float-derived $p\text{CO}_2$. I am not sure these quantities are directly comparable. The sampling error estimated here is a grid-month representativeness error after averaging a particular set of synthetic observations, whereas the cited ship and float uncertainties generally refer to the accuracy or systematic uncertainty of individual observations or derived $p\text{CO}_2$ estimates. Random measurement errors may partially average out within a grid-month, whereas systematic errors may not. Therefore, the statement that sampling error “exceeds other sources of uncertainty” requires a common error-propagation framework.

Specific comments

- 1. Lines 39–40 and 102–105:** The citation to Zhang *et al.* (2026) could be described more precisely. This study identified a directional systematic bias in float derived $p\text{CO}_2$, rather than a general uncertainty. This distinction is important because systematic bias would not necessarily be reduced by increasing the number of observations.
- 2. Lines 98–110:** More information is needed on how the BGC-Argo-pH dataset was constructed.
- 3. Lines 120–125:** When multiple historical observations are matched to the same synthetic observation, that synthetic observation is included only once. Please report how many SOCAT and BGC-Argo observations are removed by this de-duplication step.

4. **Lines 138–140:** Please clarify whether the reported overall sampling errors of $-14.1 \mu\text{atm}$ for SOCAT and $-13.5 \mu\text{atm}$ for the merged dataset are area-weighted. Figure 5 explicitly states that differences in area coverage are accounted for when calculating monthly means, but it is not clear whether the headline -14.1 and $-13.5 \mu\text{atm}$ estimates are calculated in the same way.
5. **Figure 6(c):** Please explain how the observational ranges were calculated. The numbers in parentheses for each season in Fig. 6(c) should also be defined.
6. **Lines 164–169:** The median simulated diurnal $p\text{CO}_2$ range is $0.5\text{--}2.4 \mu\text{atm}$, whereas the observational estimate is $4.6\text{--}7.1 \mu\text{atm}$. These values suggest that the typical diurnal variability in E3SM may be noticeably smaller than observed, even though the extreme values are similar. Please discuss this difference.
7. **Figure 8:** The figure legend appears to label the light-colored points as “BGC Argo-pH”, whereas the caption states that the light colors represent the merged SOCAT + BGC Argo-pH dataset. Please check the legend and make the terminology consistent with what is plotted.
8. **Lines 190–193:** The ratio does not appear sufficient to identify the “dominant” source of variability. A ratio >1 only demonstrates that the range in daily-mean $p\text{CO}_2$ is smaller than the sampling error. Please clarify the interpretation of this diagnostic.
9. **Figure 10:** The comparison of maximum wind speed between historical sampling and the complete 3-hourly JRA record may be affected by unequal sample size.
10. **Lines 203–205:** The manuscript compares the overall mean sampling error of approximately $13\text{--}14 \mu\text{atm}$ with seasonal $p\text{CO}_2$ amplitudes of $\sim 30\text{--}40 \mu\text{atm}$ in seasonally ice-covered regions and $\sim 15 \mu\text{atm}$ in subpolar seasonally stratified regions. Since these seasonal amplitudes refer to specific Southern Ocean regimes, a more consistent comparison would use sampling errors calculated for the same regions.
11. **Line 269:** “biogeochemical” should be corrected to “biogeochemical”.