This is an interesting and timely study looking at sub-monthly pCO2 sampling uncertainty that is often overlooked at the monthly 1°×1° scale. I like the use of high-resolution synthetic observations to ask if the measurements within a grid cell can really represent its monthly mean state. I have a few comments and suggestions that may help strengthen the interpretation and improve the clarity of the manuscript.

- **Links between sampling error and wind speeds**. The manuscript attributes part of the negative pCO2 sampling error to under-sampling of high-wind conditions. Figure 10 clearly shows a wind sampling bias, but the authors may want to provide a more direct link by plotting the wind sampling difference (U_sampled - U_full) against the pCO2 sampling error for each monthly, 1°×1° grid cell. This would provide more direct support for this interpretation. However, I was a little surprised by the wind speeds shown in Figure 10, especially the JRA mean wind speeds in panel a, which appear quite high (> 20 m s-1) and close to the maximum values in panel b. The Southern Ocean often has high winds, but it is surprising that the wind is always at ~ 20 m s-1. Maybe I misunderstand how the gray and yellow bars are calculated. It would help to clarify this. It may also be worth briefly discussing why BGC Argo sampling appears biased much toward low wind conditions, since I would expect that the Argo should sample the data randomly.
- **Model dependence and interpretation of the 14 µatm bias.** The authors acknowledge that one of the limitations of this study is that the subsampling bias may depend on the model itself (i.e., if the model testbed is good enough in simulating the true sub-month variability of pCO2). One way to improve the confidence is to directly compare the SOCTA data with the subsampled model pCO2. Additionally, only half of the SOCAT data are used for the assessment due to the focus on open ocean. The authors have stated that quantitative results in this study are sensitive to the data reduction in this second step. Therefore, the authors may want to make it clear that the 14 uatm bias estimate is specific to this E3SM-based experiment, especially in the abstract and conclusions.

I have three further suggestions rather than comments, which I leave it for the authors to consider.

- **Analyze the pCO2 distribution and spectra.** One hypothesis this article proposed is that the SOCAT (+ BGC Argo) under-represent the high-pCO2 events, which results in the subsampling error. To support this, it may be useful to directly examine this by inspecting the distribution of pCO2. For example, for the monthly 1°×1° cells sampled by SOCAT, the authors could compare the distribution of high-frequency modeled pCO2 with that of the synthetic observations at SOCAT sampling times. This could provide a more direct view of which part of the pCO2 distribution is missed by the historical sampling, and may also help support the proposed link between high wind under sampling and the negative pCO2 sampling error. For the BGC Argo data, a spectra analysis may also interesting. For instance, picking up a region with dense Argo data, comparing the power spectra of the subsampled synthetic pCO2 time series and the high-frequency modeled pCO2 (a method similar as Fig s8 and Fig 3 in a previous study: 10.1126/sciadv.adn5781). It may directly show if the BGC Argo sampling under-represent the variability in a specific frequency.

- **Test the spatial matching distance threshold.** The SOCAT and BGC Argo observations are matched to synthetic E3SM observations with distances up to ~250 km. Since the study shows large sub-grid variability, this spatial mismatch may itself introduce some representativeness error. It would be useful to show a simple sensitivity test using a smaller matching distance.
- **Clarify the BGC Argo improvements.** Although adding the BGC Argo improves the agreement between subsampled and the full data, but it still gives me the impression that the improvement is limited, especially in the mean subsampling error. I am wondering if this is because apples-and-pears are compared. The mean sampling errors for SOCAT and SOCAT + BGC Argo are calculated from quite different numbers of grid cells. Since BGC Argo adds many new grid cells, especially in winter, when we say the Argo data only slightly reduce the mean sampling error, we implicitly assume that the potential bias in these SOCAT cells is similar as those in the GGC Argo cells. This is likely unfair. The unique winter cells occupied by the BGC Argo data may naturally suffer from larger variability due to the storm. I think it would be helpful to also compare the two datasets using only the common sampled grid cells. This would better separate the improvement within already sampled regions from the benefit of increased coverage.

Minor comments:

- **Line 34-36:** You may find this eddy covariance study useful to support the argument here, in Line 106, and in the Discussion. It provides direct observational evidence that small-scale, high CO_2 flux events occur frequently in the Southern Ocean and can substantially affect the mean flux. Importantly, these events were mainly driven by variations in ΔpCO_2 rather than wind speed.10.1126/sciadv.adn5781
- **Line 38:** In fact, SOCAT only collects direct pCO_2 observations, while the derived pCO_2 from pH, DIC, TA are not accepted by the community right now.
- **Figs 6 and 7**: Not sure why specifically show the summer season. The previous comparison shows good agreement in summer (Fig. 5). How about change panel b from summer to winter to show the storm effect?