Review of Wu et al. (essd-2024-309)

General comments

Wu et al. describe a new data product that reconstructs sea surface *p*CO_2 in the North American Atlantic coastal ocean margin over nearly thirty years. The authors rely on the gridded Surface Ocean CO_2 Atlas (SOCAT) dataset as the baseline observations for this data product, and they reconstruct *p*CO_2 using a two-step random forest regression (RFR) + linear regression (LR) approach. They find that their data product (ReCAD-NAACOM-*p*CO_2) effectively captures coastal features and variability along the North Atlantic coastal margin. The authors report a region-wide R^2 of 0.83 and RMSE of 18.64 µatm in comparison to observations. Overall, this manuscript describes a useful product that has value for those engaged in studies of ocean acidification and air-sea CO_2 flux in the region. There are some areas, however, where more detailed explanations and thorough analyses would make this a stronger contribution.

General responses: We sincerely appreciate the reviewer's recognition of the value of our work. We have carefully addressed the reviewer's suggestions for strengthening our contribution through more detailed explanations and thorough analyses. The manuscript has been extensively revised to incorporate feedback from both reviewers. Major improvements include:

1. Clarified the usage and distinction between *f*CO_2 and *p*CO_2 throughout different sections of the manuscript
2. Enhanced the presentation and explanation of Mean Bias Error (MBE) in Section 3.2 and Fig. 5 to avoid potential confusion
3. Substantially revised Section 3.3 to better explain why product-estimated *p*CO_2 and SOCAT observations show discrepancies in the northern areas, attributing these differences to limited observational coverage
4. Restructured Section 3.4 and Fig. 7 to clearly differentiate between previously documented regional variations and newly identified phenomena revealed by our product

Additionally, we have made an important revision regarding model evaluation. In our previous version, we reported the model outputs for the training dataset (80% of X1) using results from 10-fold cross-validation. We have now updated our methodology to use direct predictions from the final trained model [y = *f*(X1)] for these data points. This revision aligns with machine learning best practices, as the final data product should utilize predictions from the complete trained model rather than intermediate cross-validation results. The cross-validation metrics remain valuable for model evaluation during the development phase, while the final product benefits from the full model trained on the entire training dataset. For uncertainty quantification, we maintain the use of validation set RMSE, as it aligns well with 10-fold cross-validation results and provides a more comprehensive assessment of model uncertainty. Noted that this revision did not essentially change the results of this work.

All revisions are highlighted in red in the manuscript and are detailed in this response letter.

R1C1. The strategy of adjusting RFR estimates with an LR is a unique and straightforward way to mitigate possible biases in the RFR estimates. However, this aspect of the methodology could use more explanation, in particular with respect to why this correction might be needed and how it improves the product ReCAD-NAACOM-$p\mathrm{CO_2}$ relative to not implementing the LR step. If, as indicated, the LR serves to "mitigate potential systematic biases in RFR-derived $f\mathrm{CO_2}$ values [that] arise from spatiotemporal heterogeneities in the SOCAT observational dataset", I envision a figure like Fig. 4c before and after applying the LR would emphasize the added value of this methodological step.

> **Response:** We thank the reviewer for this constructive suggestion. To demonstrate the value of the LR calibration step, we have added two new figures comparing the performance before and after LR calibration across six sub-regions in **Appendix A**. These figures show both monthly climatology and $p\mathrm{CO_2}$ trends. While the LR calibration yields modest improvements in monthly climatology, it significantly enhances the representation of monthly $p\mathrm{CO_2}$ anomalies (deseasonalized), as evidenced by improved R^2 values and reduced RMSE. We have incorporated these findings into Results Section 3.1, lines 271-277:
>
> "Our product employs a two-step RFR+LR algorithm to retrieve $p\mathrm{CO_2}$. The initial RFR step accurately captures most seasonal and decadal $p\mathrm{CO_2}$ variations across all six sub-regions (**Appendix A**). When comparing only at matching grid cells where SOCAT measurements are available, the differences (N = 12) in monthly mean climatology between SOCAT and RFR-derived $p\mathrm{CO_2}$ are less than 2 µatm on average with standard deviations below 5 µatm across all sub-regions (**Fig. A1**). However, the RFR-derived $p\mathrm{CO_2}$ shows lower accuracy in capturing long-term $p\mathrm{CO_2}$ changes in the GoMe and SAB. The subsequent LR calibration improves the performance significantly: R^2 values increase from 0.69 to 0.81 in the GoMe and from 0.83 to 0.93 in the SAB, while RMSE decreases from 12.43 to 10.51 µatm in the GoMe and from 10.83 to 8.12 µatm in the SAB (**Fig. A2**)."

R1C2. I find the analysis presented in Section 3.3 to be somewhat lacking. While the similarities in large-scale climatological patterns between the raw observations and ReCAD-NAACOM-$p\mathrm{CO_2}$ is encouraging, more interesting is where, when, and why the two datasets differ, and how those differences speak to the value added by the gap-filled product. In particular, I see much higher wintertime $p\mathrm{CO_2}$ in the observations compared to the product in the northern region in Fig. 6. Is this result due to preferential observational coverage of high-$p\mathrm{CO_2}$ areas in that season, as potentially indicated by Fig. 2d? This type of analysis I think is more interesting to readers, and more effective at communicating the utility of the new product.

> **Response:** We thank the reviewer for this constructive and valuable suggestion. Following both reviewers' comments, we have expanded Section 3.3 to include a more detailed analysis of the differences between observations and our gap-filled product, particularly focusing on regional and seasonal variations. We have added new discussions about the sampling limitations in the northern regions and how our product addresses these gaps. For convenience, we attached the revisions below (lines 339-353):
>
> " One of the primary objectives of this product is to capture the seasonal cycle of $p\mathrm{CO_2}$ across the NAACOM region. **Figure 6** showcases the applicability of the product in capturing the $p\mathrm{CO_2}$ seasonal cycles across the southern and northern areas of

NAACOM (red and blue boxes in **Fig. 2**). The comparison of monthly climatologies between the gap-filled product and SOCAT observations reveals strong agreement in the southern regions, despite of the coverage difference, with product-estimated monthly means being only 3.05 ± 5.60 µatm higher than SOCAT (**Fig. 6a**), suggesting that our product effectively captures the seasonal cycle where data are abundant.

In the northern regions where SOCAT data are sparse, the gap-filling ability of the product is also well demonstrated. In the northern region, the area-average monthly *p*CO_2 climatology calculated from the continuous reconstructed product are 22 ± 11.12 µatm lower than SOCAT observations, which can be attributed to limited observational coverage in this area. This area is characterized by sparse sampling, with observation density approximately 50% lower than in the southern regions (**Fig. 2**) due to the smaller area and limited cruise coverage. For instance, the GStL region only has one summer cruise in SOCAT database (**Fig. 2b**), and the SS and GoMe have particularly sparse winter observations (**Fig. 2d**). The higher latitudes typically exhibit larger seasonal amplitudes in *p*CO_2, making the limited sampling from SOCAT particularly problematic for accurate characterization. Our gap-free product provides comprehensive spatial and temporal coverage, enabling more robust analysis of *p*CO_2 patterns and variability in these historically under-sampled regions."

R1C3. The comparisons to global products detailed in Section 3.4 would benefit from some quantitative results to be presented alongside the qualitative interpretation of the annual mean climatological figures. The authors assert, for example, that compared to ULB_SOMFFN_coastal_v2 "the ReCAD-NAACOM-*p*CO_2 product exhibits closer values to the observations", but provide no evidence outside visual inspection of Fig. 7. Instead, by comparing (for instance) the average and RMSE of differences between the gridded SOCAT observations and corresponding values from the products within specific regions, the authors could more clearly emphasize the level of improvement provided by ReCAD-NAACOM-*p*CO_2.

Response: We thank the reviewer for catching this. Regarding Figure 7 in Section 3.4, our previous work (Wu et al., 2024) indicated that existing *p*CO_2 products did not adequately meet our requirements for regional analysis. Therefore, the objective of Fig.7 is to show the capability of this product in capturing these regional variations.

We agreed that Section 3.4 could be more quantitative and revised the relative descriptions. The revised section now presents two distinct components: (1) validation of previously confirmed regional variations, and (2) discussion of novel patterns revealed by our product that warrant future investigation. The expanded Section 3.4 is provided below (lines 380-398):

"The ReCAD-NAACOM-*p*CO_2 product demonstrates superior alignment with SOCAT observations in capturing these regional features that have been reported in previous observation-based studies (**Fig. 7b**) ….

In addition to these previously documented regional variations, our product reveals several notable features not previously captured by observations or other existing

> products. For instance, the GoMe displays intermediate pCO_2 levels around 380 µatm, distinctly higher than surrounding waters at comparable latitudes, a feature previously documented by a multiple linear regression reconstructed pCO_2 product (Signorini et al., 2013) and five-year (2004-2009) mooring and cruise data (Vandemark et al., 2011) but contradict to another two studies based on numerical models (Cahill et al., 2016; Rutherford et al., 2021). In the southern GStL (S.GStL, box 4 in **Fig. 7**), pCO_2 values are slightly higher compared to adjacent waters at similar latitudes, aligning with high nutrient concentrations typically observed in these river-influenced waters (Lavoie et al., 2021). These regional patterns could not be completely captured by the global products (**Fig. 7c and 7d**). Ability of the ReCAD-NAACOM-pCO_2 product in resolving such regional features demonstrates its potential value for investigating coastal carbon dynamics and their responses to local and regional forcing factors in the NAACOM."

Additional line-specific comments are provided below.

Specific comments

R1C4. 49: Aren't the fCO_2 data included in SOCAT from these cruises exclusively from underway measurements (not discrete)? In which case, perhaps this sentence should read "Underway measurements from these research cruises, combined with underway measurements from volunteer observing ships and buoy observations, …" or something to that effect.

> **Response:** We appreciate the reviewer's attention to detail regarding the types of measurements included in our study. In response, we have modified the sentence to read (lines 51-52):
>
> "Underway measurements from these cruises, combined with underway measurements from volunteer observing ships and buoy, are quality-controlled and compiled in the Surface Ocean CO_2 Atlas (SOCAT) database (Bakker et al., 2016),"

R1C5. Figure 1: The regional labels might be better displayed in orange rather than red. As they are now, one might understandably associate the red labels with the red 200m isobath, which can be confusing.

> **Response:** We thank the reviewer for this helpful suggestion. We have changed the regional labels from red to orange to avoid confusion with the 200m isobath. Additional revisions to the figure have also been made based on suggestions from another reviewer. The modified Figure 1 is attached for reference:

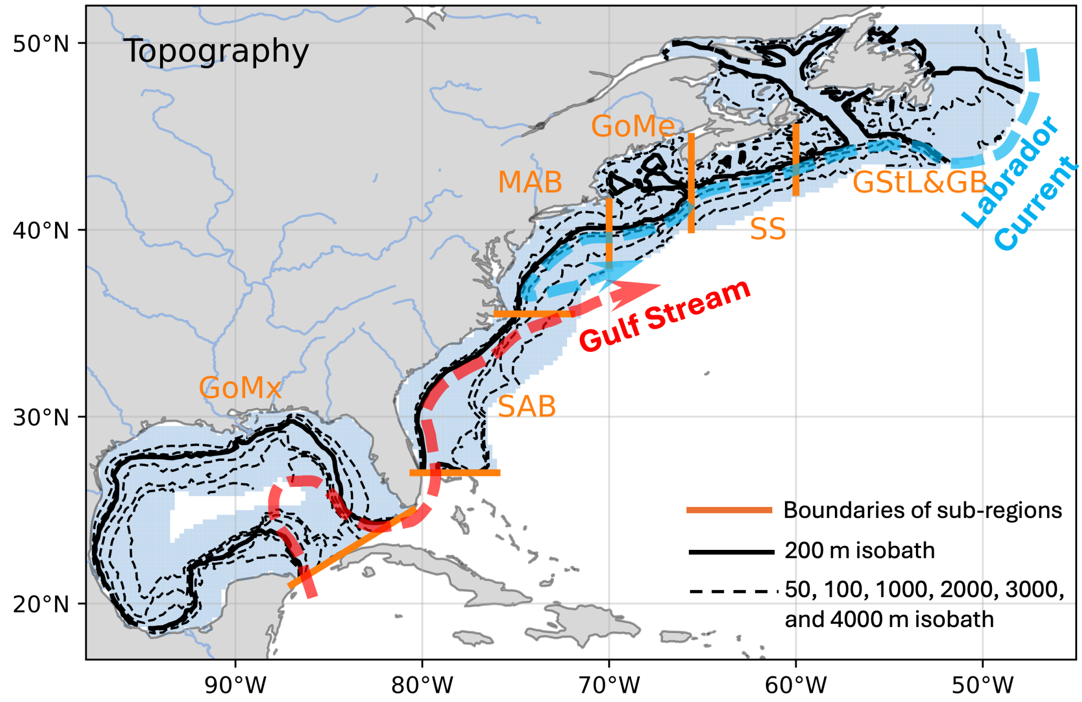

Figure 1. Topography (in meters) and large-scale circulation along the North American Atlantic Coastal Ocean Margin (NAACOM). The study region, defined as coastal areas extending 400 km offshore, is indicated by blue shading. The thick black line is the 200 m isobath, which roughly marks the shelf break and typically defines the continental shelf boundary. The Gulf Stream (thick red dashed line with an arrow) flows northward along the east coast of the United States before veering eastward into the open Atlantic Ocean around Cape Hatteras. The Labrador Current (thick light blue dashed line with an arrow) flows southward along the east coast of Canada before meeting the Gulf Stream. Following Fennel et al. (2019), the study region is divided into six sub-regions by straight orange lines: the Gulf of Mexico (GoMx), South Atlantic Bight (SAB), Mid-Atlantic Bight (MAB), Gulf of Maine (GoMe), Scotian Shelf (SS), and Gulf of St. Lawrence and Grand Banks (GStL&GB). Dashed contour lines indicate bathymetric depths of 50 m and 100 m on the shelf (from coastline to 200 m isobath), and 1000 m, 2000 m, 3000 m, and 4000 m from the shelf break to the open ocean.

R1C6. 75–77: These two sentences say essentially the same thing and could be combined.

Response: We thank the reviewer for catching this redundancy. We have removed the second sentence.

R1C7. 105: The word "enhanced" suggests a comparison for the spatial, seasonal, and decadal variability. It should be mentioned here that the capability of the product at resolving these variations is enhanced in reference to some other dataset. Global products? The gridded SOCAT observations?

Response: We agree that clarification is needed to specify the reference point for our product's enhanced capabilities. In response, we have modified the original sentence to read (line 108):

"… enhanced capability in resolving spatial variations and capturing the seasonal cycle and decadal trends of $p\mathrm{CO}_2$ better than those of the global products across different sub-regions along the NAACOM."

R1C8. 109: I find the "ground-truth data" terminology to be somewhat misleading. Ground-truth suggests data that is used to evaluate some model or remote-sensing measurement, but here the data is used not only as a ground-truth but also for training the model itself. Perhaps something like "observational data", "model training data", etc. might be more appropriate.

> **Response:** We appreciate the reviewer's suggestion regarding proper use of terminology. We initially adopted the term 'ground-truth data' from remote sensing and machine learning studies. However, we agree that in the context of our oceanographic research, "observational data" is a more appropriate and precise term. To address this, we have replaced three instances of "ground-truth data" with "observational data" in the manuscript.

R1C9. 118: Sampling density also looks to be particularly low in the western Gulf of Mexico.

> **Response:** We fully concur with the reviewer's observations. Indeed, the sampling density is also low in the western and southern Gulf of Mexico, as we previously noted in our Gulf of Mexico publication (Wu et al., 2024). We have modified the original sentence as follows (lines 121-122):
>
> "Observational data show lower sampling density in the areas north of Cape Cod and western and southern GoMx (blue box in **Fig. 2**)."

R1C10. Figure 3: My understanding is that the "Model" (light orange box with curved sides) is the same that is applied to all satellite and reanalysis data to construct the gridded product. As such, I'd recommend some modification to this flow chart. The arrow from "Model" to "Predictive model" is confusing if those two items are indeed the same.

> **Response:** We sincerely appreciate the reviewer's comments regarding the clarity of our original flowchart. We acknowledge that the initial representation could have been clearer, and we have made three modifications to address this concern:
>
> 1) We have made the two "models" the same in the flowchart.
>
> 2) We have added a sentence in the caption explicitly stating that 'The two models in the orange boxes are identical.' to avoid any potential confusion.
>
> 3) Following another reviewer's valuable suggestion, we have included the model outputs in the flowchart. This addition clarifies the sequential nature of our approach: our machine learning model first outputs $f\text{CO}_{2\text{sea}}$, which is then converted to $p\text{CO}_{2\text{sea}}$ using OISST data.
>
> The revised flowchart with caption is attached:

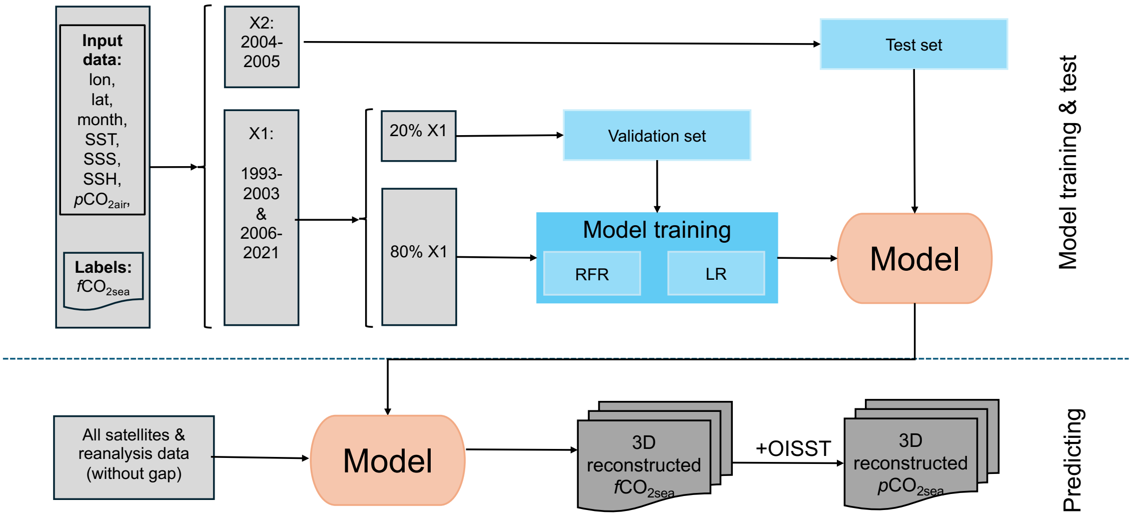

Figure 3. A flowchart of the two-step machine learning regression model for generating the reconstructed pCO$_2$ product. Grey boxes represent the input and output datasets, blue boxes illustrate the model training, validation testing, and independent test processes, and orange boxes represent the final trained model for predicting the reconstructed product. The two models in the orange boxes are identical. The training data, consisting of paired input variables (lon, lat, month, sea surface temperature (SST), sea surface salinity (SSS), sea surface height (SSH), and atmospheric pCO$_2$ (pCO$_{2air}$) and corresponding sea surface fCO$_2$ (fCO$_{2sea}$) labels), is divided into two sets: X1 (1993-2003 and 2006-2021) and X2 (2004-2005). X1 is further randomly divided into subsets for model training set (80%) and validation set (20%). The predictive model combines a random forest regression (RFR) and a linear regression (LR) algorithm. The trained and validated regression model is then applied to all satellite and reanalysis data (without gaps) to generate the 3D reconstructed fCO$_{2sea}$ product, which was then converted to pCO$_{2sea}$ with satellite SST data.

R1C11. 167–168: More explanation should be given here on exactly how the validation set is used to evaluate the model performance.

> **Response:** We appreciate the reviewer's suggestion for clarification regarding the validation set's role in model evaluation. We have expanded our explanation as follows (lines 176-182):
>
> "The validation set, comprising 20% of X1 randomly sampled from 1993-2003 and 2006-2021, serves as a critical monitoring step for model evaluation. This subset plays two key roles: first, it tests hyperparameter tuning by providing independent performance metrics on unseen data, and second, it helps detect potential overfitting by monitoring the divergence between training and validation performance. While the validation set itself cannot prevent overfitting, it enables detection of overfitting patterns when performance of the model improves on training data but deteriorates on validation data. Through this continuous evaluation process, the validation set ensures more robust model development and helps achieving better generalization capabilities."

R1C12. 171–172: This sentence is somewhat unclear.

> **Response:** We appreciate the reviewer's suggestion for clarification. We have modified our explanation as follows (lines 183-186):
>
> "The independent test set (X2), covering the years 2004-2005, serves as a critical evaluation period specifically designed to assess reliability of the model in predicting values for years that were completely excluded from both training and validation phases. Because we intentionally withhold these two years from model development, this approach directly tests capability of the model in generating reliable predictions and fill temporal data gaps for periods without observational data."

R1C13. 178: How is month treated in the model training? If you're only using 1–12 for the months of the year, there will be an unintended effect whereby months that should be treated as similar (e.g., January vs. December) will be treated as extremely different (1 vs. 12). See Sauzède et al. (2015) or Gregor et al. (2018) for information about transforming cyclical predictors using sine and cosine functions.

> **Response:** We used numerical values 1-12 to represent months in our algorithm. We agree with the reviewer's comment regarding the treatment of monthly data and the potential artificial discontinuity. Following this suggestion, we conducted additional analyses by implementing the suggested sinusoidal transformation [sin(month/12 * 2π)] and reran our complete modeling framework. The comparisons are shown in the table bellow. Our analysis revealed that this modification resulted in minimal differences in the model output matrix, with variations comparable to those stemming from random sampling algorithms.

Months	R^2	RMSE (μatm)	MAE (μatm)	MBE (μatm)
1-12	0.92	12.70	7.55	0.13
sin(month/12 * 2π)	0.90	14.59	8.70	-0.17

> These tests suggest that the seasonal cycle information in our study region is largely captured by other variables (SST, SSS, SSH, and $p\mathrm{CO}_{2\mathrm{air}}$), which inherently contain seasonal patterns. Based on these findings, we conclude that the minimal impact of the monthly representation method indicates our current conclusions remain robust.
>
> Given no improvements in model performance, we maintained the original methodology to preserve consistency in the manuscript's statistical analyses, as the Mente Carlo simulation. However, we acknowledge that implementing proper cyclical variable treatment is theoretically more appropriate. In our ongoing development of version 2 of this product, which includes reconstructed SSS fields for the entire Pacific and Atlantic margins (currently under validation), we plan to implement the sinusoidal transformation for monthly variables.

R1C14. 206: I believe P represents the total atmospheric pressure, not the "CO_2 atmospheric pressure".

> **Response:** We sincerely appreciate the reviewer's careful reading and attention to detail. Yes, the original text contained a typo. We have corrected this inaccuracy and revised

the text as follows:

"where P is the total atmospheric pressure on the sea surface, …"

R1C15. 315–324: I'm not sure this discussion is very valuable because the general features discussed here are evident in the product but also in the observations themselves. It might be more effective to discuss the seasonal cycle features in the product as they relate to the observations; what information is added by the gap-filled product?

Response: We thank the reviewer for this constructive comment. We have revised this paragraph to better highlight how our gap-free product enhances our understanding of seasonal cycles beyond what is visible in the raw observations. The revised discussion now quantifies the agreement between our reconstructed product and SOCAT observations, and explains regional differences in their monthly climatologies:

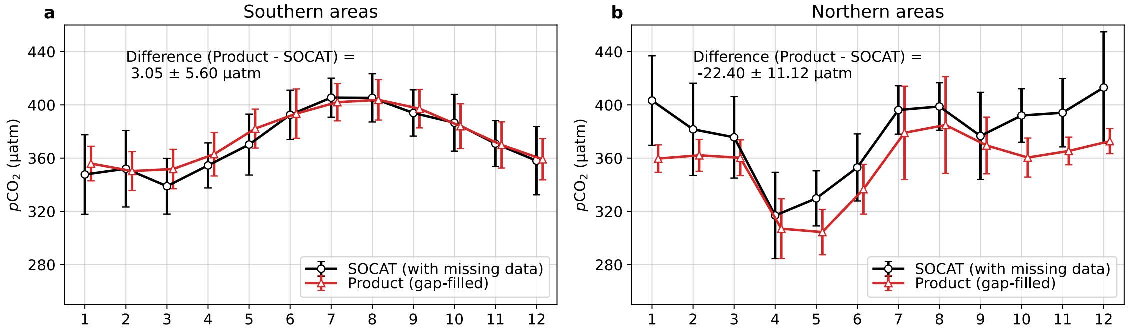

Figure 6. Monthly mean climatology of *p*CO₂ in the southern and northern areas of the NAACOM from 1993 to 2021. Sub-regions are **(a)** southern areas, the red box in Fig. 2, and **(b)** northern areas, the blue box in Fig. 2. Two data representations are shown: (1) SOCAT observations (black curves), which may be influenced by missing data; and (2) the complete gap-filled product output (red curves). Error bars denote one standard deviation of the monthly mean climatology of $p\text{CO}_2$. Numbers indicate the mean difference (± one standard deviation) between monthly climatological $p\text{CO}_2$ calculated from the two sources, with positive values indicating higher product estimates compared to SOCAT observations. The x-axis shows months (1-12, where 1 represents January), and the y-axis shows $p\text{CO}_2$ in µatm.

"One of the primary objectives of this product is to capture the seasonal cycle of $p\text{CO}_2$ across the NAACOM region. **Figure 6** showcases the applicability of the product in capturing the $p\text{CO}_2$ seasonal cycles across the southern and northern areas of NAACOM (red and blue boxes in **Fig. 2**). The comparison of monthly climatologies between the gap-filled product and SOCAT observations reveals strong agreement in the southern regions, despite of the coverage difference, with product-estimated monthly means being only 3.05 ± 5.60 µatm higher than SOCAT (**Fig. 6a**), suggesting that our product effectively captures the seasonal cycle where data are abundant.

In the northern regions where SOCAT data are sparse, the gap-filling ability of the product is also well demonstrated. In the northern region, the area-average monthly $p\text{CO}_2$ climatology calculated from the continuous reconstructed product are 22 ± 11.12 µatm lower than SOCAT observations, which can be attributed to limited observational coverage in this area. This area is characterized by sparse sampling, with observation density approximately 50% lower than in the southern regions (**Fig. 2**) due to the smaller area and limited cruise coverage. For instance, the GStL region only has one summer cruise in SOCAT database (**Fig. 2b**), and the SS and GoMe have particularly sparse winter observations (**Fig. 2d**). The higher latitudes typically exhibit larger seasonal amplitudes in $p\mathrm{CO_2}$, making the limited sampling from SOCAT particularly problematic for accurate characterization. Our gap-free product provides comprehensive spatial and temporal coverage, enabling more robust analysis of $p\mathrm{CO_2}$ patterns and variability in these historically under-sampled regions."

R1C16. 416: It should be clarified here that this uncertainty value for the North American Pacific Coastal Ocean Margin is specific to areas within 100km of the coastline and the uncertainty provided for ReCAD-NAACOM-$p\mathrm{CO_2}$ is for areas within 400km.

Response: We appreciate the reviewer's careful attention to detail. We have revised the original sentence to make it more precise (lines 454-456):

"Despite this conservative method, our calculated uncertainty for the Atlantic margins is comparable to the 43.4 µatm reported by Sharp et al. (2022) for areas within 100 km of the North American Pacific margins. suggesting a good product performance of our product."

Technical corrections

R1C17. 148: Should be "arising" or "that arise"

Response: We appreciate the reviewer's careful attention to detail. We have modified the original phrase "... arise from ..." to "... arising from ...". The revised sentence now reads (lines 154-157):

"… while the LR model is subsequently applied to mitigate potential systematic biases in RFR-derived $f\mathrm{CO_2}$ values arising from spatiotemporal heterogeneities in the SOCAT observational dataset **(Fig. 2)**."

R1C18. 424: Recommend changing wording here: "the performance…reduced" is somewhat awkward

Response: We thanks for pointing out this and revised the original sentence to (line 466):

"Furthermore, during the independent validation phase, the accuracy of the model predicted values reduced in the GoMe …"

References

Wu, Z., Wang, H., Liao, E., Hu, C., Edwing, K., Yan, X.-H., & Cai, W.-J. (2024). Air-sea CO2 flux in the Gulf of Mexico from observations and multiple machine-learning data products. *Progress in Oceanography*, *223*, 103244. https://doi.org/10.1016/j.pocean.2024.103244