Response to Referee #2

General comments:

Zhang et al. present a global monthly surface ocean pCO_2 dataset (SJTU-AViT) and corresponding air-sea CO_2 fluxes spanning 1982-2023 at 1° resolution, developed using a Vision Transformer-based deep learning model. The approach combines SOCAT observation, and observations of climate data with multiple ocean biogeochemical models and incorporates physical-biogeochemical constraints. The authors show that their product successfully captures the spatial and temporal variations of observed pCO_2 patterns, from seasonal cycles to interannual variability. The product shows more realistic small-scale spatial variability and temporal interannual variability than previous pCO_2 products. The resolved air-sea CO_2 fluxes agree with other estimates based on pCO_2 observations. The paper is well written, the methodology is robust, and the line of thought is mostly clear to me. I only have minor comments regarding some of the technical details and presentation.

We thank the reviewer for the helpful and constructive feedback. We have revised the manuscript to address all of these comments. Overall, the reviewer's main concerns focused on the transparency and robustness of the model training strategy, the contribution of physical-biogeochemical constraints, the adequacy of uncertainty estimation method, and several issues related to data processing and presentation. In response to these concerns, we have made the following revisions.

- Clarify and validate the two-stage training framework, and quantify the contributions of its components. We elaborated the physical motivations for CMIP6 pre-training, MOM6 constraints, and SOCAT fine-tuning, and added ablation experiments to demonstrate their respective roles in improving convergence, generalization, and accuracy (major comment #1).
- Revise the uncertainty estimation framework. We replaced the observation-dependent u_{map} with an algorithm-based uncertainty estimate ($u_{algorithm}$) derived from synthetic sampling experiments, and integrated the complete workflow and quantitative results into the Methods and Results sections (major comment #2).
- Enhance diagnostic analyses and visualization. We improved the calculation of seasonal variability by applying linear detrending prior to analysis, and added seasonal-phase diagnostics and peak–minimum month difference maps (minor comments #7, #9).
- Revise minor edits and clarifications (minor comments #1-11).

Please see our detailed point-by-point responses to each comment below.

Major comments:

1. The description of methodology is overall complete. However, certain technical details are still missing. It is not clear how pre-training on CMIP6 models contributes to the final model. It is not clear what the fine-tuning of MOM6 really does. Are your results sensitive to the choice of CMIP6 models and the fine-tuning? How do SOCAT

data fold into your refinement? For the physical-biogeochemical constraints, are you only using what is derived from MOM6, or also from CMIP6 models as well? How are your results, particularly on the seasonal cycle, impacted by these physical-biogeochemical constraints? In other words, if you exclude these constraints, how is the representation of the seasonal pCO_2 cycle affected?

We thank the reviewer for raising this comprehensive question. We have structured our response into six corresponding parts for clarity. The revisions include 5 ablation experiments, with summary findings presented in the main text (section 4, lines 601-612) and full experimental details reported in the supplement (section S5.2-S5.6).

(1) How pre-training on CMIP6 models contributes to the final model?

To quantitatively assess the impact of CMIP6-based pretraining on the reconstruction, we conducted two controlled experiments that were identical in all settings except for the use of CMIP6 pretraining.

(a) Test 1 (with CMIP6 pretraining): The model was first pretrained on CMIP6 simulation outputs, allowing it to learn from CMIP6 model results. It was then jointly fine-tuned using MOM6 and SOCAT observational data.

(b) Test 2 (without CMIP6 pretraining): Under the same conditions, the model relied solely on MOM6 and SOCAT data.

The ablation experiments reveal a substantial impact of CMIP6 pretraining on the results. When pretrained on CMIP6 (Test 1), the model achieved an RMSE of 7.44 μatm on the validation set. Without CMIP6 pretraining (Test 2), RMSE increased to 17.13 μatm. Thus, CMIP6 pretraining reduced RMSE by 9.69 μatm, corresponding to a relative decrease of approximately 56.57%. The spatial map (Fig. R1) indicates that the largest improvements occur in regions with sparse observations (particularly at high latitudes) and areas with pronounced low-frequency or interannual variability.

CMIP6 pretraining provides the model with a physically meaningful initialization. By learning from temporally and spatially complete simulation fields, the model can first capture large-scale spatial patterns and low-frequency signals, enabling faster convergence during fine-tuning, reducing overfitting in observation-sparse regions, and achieving better generalization at interannual scales. Although CMIP6 simulations may contain biases, these are effectively corrected during the subsequent fine-tuning with MOM6 and SOCAT, ensuring the final reconstruction remains consistent with observations. The substantial RMSE improvement (a reduction of 9.69 μatm, ~56.57%) demonstrates that this two-stage training strategy achieves an optimal balance between physical consistency and empirical accuracy.

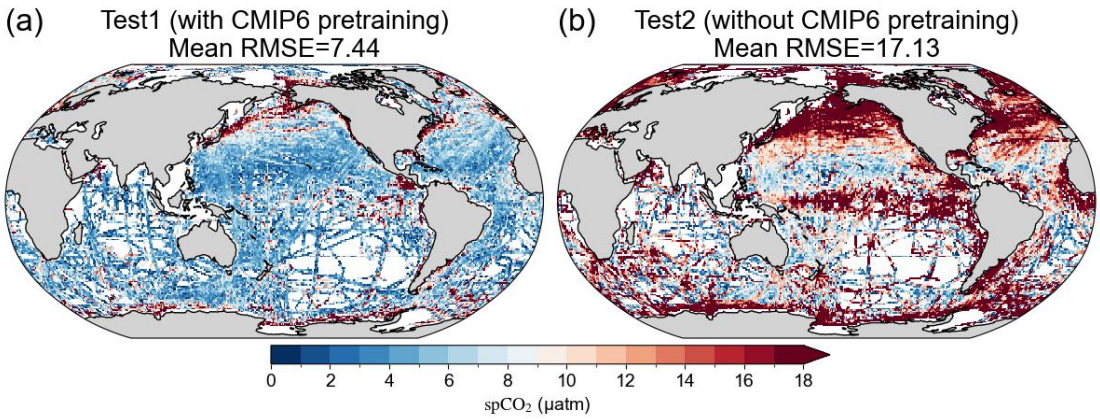

Figure R1 (Figure S9 in supplement section S5). Impact of CMIP6 pre-training on reconstructed $spCO_2$ fields. (a) Test 1 (with CMIP6 pretraining): CMIP6 pre-training followed by MOM6 & SOCAT fine-tuning; (b) Test 2 (without CMIP6 pretraining): no CMIP6 pre-training, trained only on MOM6 & SOCAT. Inclusion of CMIP6 pre-training reduces validation RMSE by 9.69 µatm (~56.57% relative reduction), justifying the two-stage training strategy.

(2) What the fine-tuning of MOM6 really does?

To assess the role of MOM6 fine-tuning in our reconstruction framework, we designed two comparative experiments while keeping all other model settings identical:

(a) Test 1 (with MOM6 in fine-tuning): The model was first pretrained on CMIP6 outputs and then fine-tuned using both MOM6 simulation outputs and SOCAT observations. MOM6 provides continuous, physically consistent global fields, while SOCAT supplies essential observational constraints.

(b) Test 2 (without MOM6 in fine-tuning): The model was pretrained on CMIP6 data as in Test 1 but fine-tuned solely with SOCAT observations, without incorporating MOM6 outputs.

The fine-tuning strategy that included MOM6 data (Test 1) achieved a validation RMSE of 7.44 µatm. In contrast, excluding MOM6 during fine-tuning (Test 2) resulted in a substantially higher RMSE of 12.27 µatm. Thus, incorporating MOM6 during fine-tuning reduced RMSE by 4.83 µatm, corresponding to a relative decrease of approximately 39.36%. The spatial map (Fig. R2) indicates that the largest improvements occur in regions with sparse observations, particularly at high latitudes, and in areas with pronounced low-frequency or interannual $spCO_2$ variability, highlighting the crucial role of MOM6 in enhancing reconstruction accuracy.

In our framework, MOM6 outputs are incorporated alongside SOCAT observations during the fine-tuning stage. SOCAT provides the essential observational constraint, but its spatial and temporal coverage is sparse and uneven. MOM6 complements this by supplying continuous global fields that embed large-scale physical consistency, thereby stabilizing the training process and enhancing generalization, particularly in data-poor regions. Mechanistically, MOM6 fine-tuning serves three key functions: (i) it exposes the network to continuous, globally coherent background fields (e.g.,

large-scale gradients, seasonal cycles, and interannual variability), thereby reducing overfitting to the sparse and uneven SOCAT distribution; (ii) it aligns model weights with physically plausible oceanographic relationships, mitigating the direct transfer of structural biases from heterogeneous CMIP6 pre-training and avoiding abrupt or unrealistic weight corrections during SOCAT anchoring; (iii) it supplies realistic background variability, enabling the model to learn coherent patterns prior to adjustment with pointwise observations, which strengthens generalization in data-limited regions. In summary, MOM6 fine-tuning functions as a physically consistent bridge between synthetic CMIP6 pre-training and sparse SOCAT observations, significantly improving the stability, robustness, and reliability of the reconstruction, especially in regions with limited observational coverage.

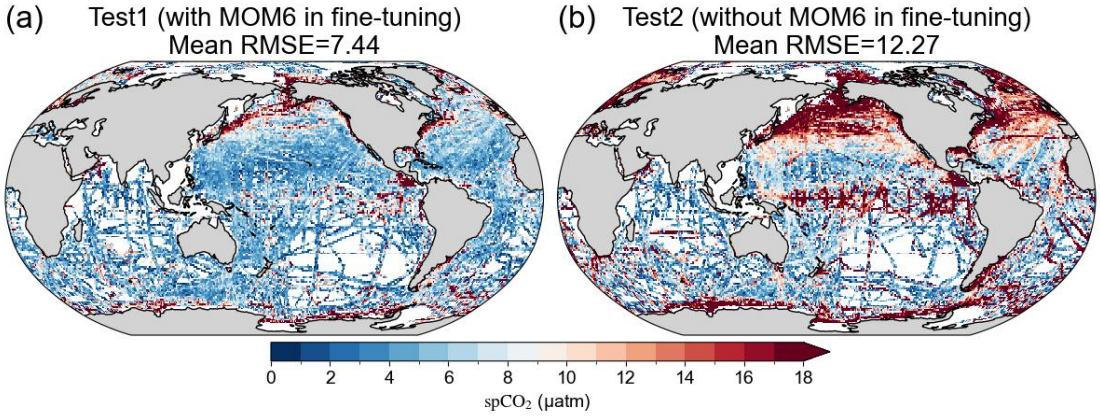

Figure R2 (Figure S10 in supplement section S5). Impact of MOM6 fine-tuning on reconstructed $spCO_2$ fields. (a) Test 1 (with MOM6 in fine-tuning): CMIP6 pre-training followed by MOM6 & SOCAT fine-tuning; (b) Test 2 (without MOM6 in fine-tuning): CMIP6 pre-training, fine-tuning only on SOCAT. Inclusion of MOM6 fine-tuning reduces validation RMSE by 4.83 µatm (~39.36% relative reduction), highlighting the crucial role of MOM6 in enhancing reconstruction accuracy.

(3) Are your results sensitive to the choice of CMIP6 models and the fine-tuning?

To assess the sensitivity of our reconstruction to the choice of CMIP6 models and the fine-tuning strategy, we conducted two comparative pre-training experiments while keeping all other model settings identical:

(a) Test 1 (3-model CMIP6 pre-training): The model was pre-trained on a subset of three CMIP6 simulations (GFDL-ESM4, NorESM2-LM, NorESM2-MM) and then fine-tuned with the same MOM6 and SOCAT data.

(b) Test 2 (4-model CMIP6 pre-training): The model was pre-trained on a different subset of four CMIP6 simulations (CESM2, CESM2-FV2, CESM2-WACCM, CESM2-WACCM-FV2) and fine-tuned using the same MOM6 and SOCAT data.

The ViT reconstruction using the 3-model subset (Test 1) achieved a validation RMSE of 10.48 µatm, while the 4-model subset (Test 2) yielded a slightly lower RMSE of 9.54 µatm. Both are higher than the RMSE obtained using all seven CMIP6 models (7.44 µatm), indicating that the total amount of pre-training data can influence reconstruction performance. Nevertheless, the difference between the two subsets is

small (RMSE difference of 0.94 µatm, ~8.97%), and deviations from the 7-model pre-training result are modest (~2-3 µatm).

Overall, these results indicate that, as long as multiple CMIP6 models are included to capture diverse large-scale oceanic patterns, the reconstruction is largely robust to the specific choice of pre-training models. The two-stage training framework effectively stabilizes reconstruction performance, corrects model-specific biases, and reliably integrates observational information. To further strengthen robustness, CMIP6 models were carefully selected based on the evaluation framework of Liao et al. (2021), ensuring that the chosen models accurately represent key oceanic carbon dynamics. Through multi-model pre-training combined with carefully designed fine-tuning strategies, our approach maintains stable and reliable reconstruction performance, effectively capturing large-scale patterns, low-frequency variability, and regional details across different spatial and temporal scales.

The reconstruction results are robust to reasonable systematic changes in key fine-tuning hyperparameters (such as learning rate, batch size, patch size, and Transformer block number) though extreme changes (e.g., reducing Transformer blocks from 10 to 5) can substantially affect performance. Fine-tuning data are crucial: MOM6 provides physically consistent global fields to stabilize training and enhance generalization (see response 1.2), while SOCAT observations correct local and regional biases (see response 1.4), together ensuring stable, reliable, and physically coherent $spCO_2$ reconstructions across both well-observed and data-sparse regions.

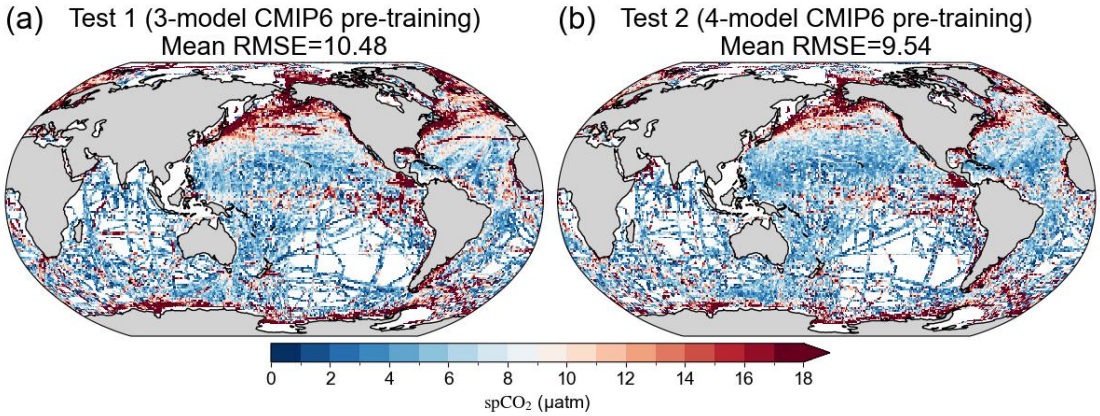

Figure R3 (Figure S11 in supplement section S5). The sensitivity of reconstructed $spCO_2$ fields to the choice of CMIP6 models. (a) Test 1 (3-model CMIP6 pre-training): three CMIP6 simulations (GFDL-ESM4, NorESM2-LM, NorESM2-MM) pre-training followed by MOM6 & SOCAT fine-tuning; (b) Test 2 (4-model CMIP6 pre-training): four CMIP6 simulations (CESM2, CESM2-FV2, CESM2-WACCM, CESM2-WACCM-FV2) pre-training followed by MOM6 & SOCAT fine-tuning.

(4) How do SOCAT data fold into your refinement?

To evaluate the role of SOCAT observations in the fine-tuning stage, we designed two comparative experiments while keeping all other model settings identical:

(a) Test 1 (with SOCAT in fine-tuning): The model, pretrained on CMIP6 and optionally fine-tuned with MOM6 fields, was further fine-tuned using SOCAT in situ

pCO_2 observations. SOCAT provides high-quality pointwise constraints that correct model biases and ensure alignment with real-world ocean conditions.

(b) Test 2 (without SOCAT in fine-tuning): The same pretrained model was fine-tuned without using SOCAT data, relying solely on MOM6 fields for spatial coverage and physical consistency.

Incorporating SOCAT observations during fine-tuning (Test 1) yielded a validation RMSE of 7.44 µatm. In contrast, excluding SOCAT (Test 2) resulted in a dramatically higher RMSE of 26.87 µatm. Thus, the inclusion of SOCAT reduced RMSE by 19.43 µatm, corresponding to a relative decrease of approximately 72.31%. This large improvement demonstrates the critical role of SOCAT observations in aligning the reconstructed $spCO_2$ field with real-world measurements.

SOCAT data act as a supervisory signal that corrects local and regional biases in the model, ensuring the fine-tuned reconstruction reproduces observed variability while retaining large-scale spatiotemporal patterns learned during CMIP6 pretraining and MOM6 fine-tuning. Without SOCAT, the model cannot accurately capture local pCO_2 variations, leading to substantial errors. Proper integration of SOCAT with MOM6 fields balances the influence of sparse observational points and physically consistent background patterns, enhancing overall predictive skill, particularly in regions with limited observations.

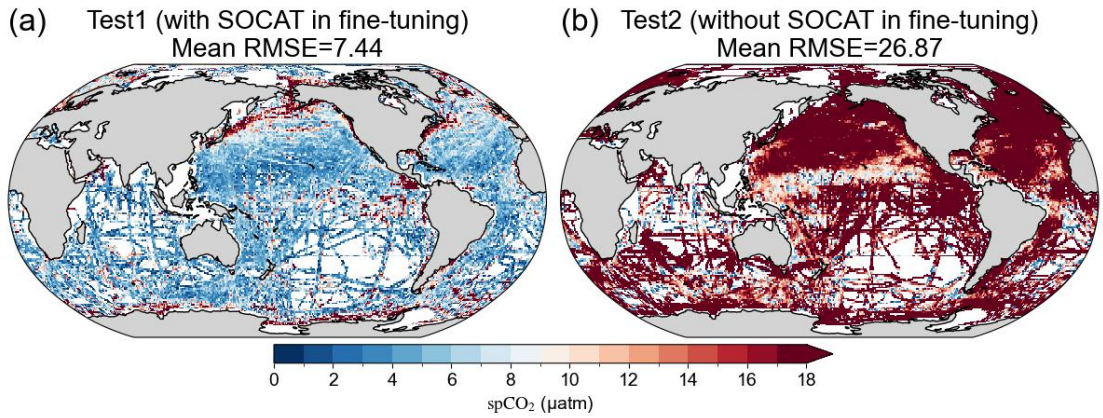

Figure R4 (Figure S12 in supplement section S5). Impact of SOCAT observations on the fine-tuning of the reconstructed $spCO_2$ field. (a) Test 1 (with SOCAT in fine-tuning): CMIP6 pre-training followed by MOM6 & SOCAT fine-tuning; (b) Test 2 (without SOCAT in fine-tuning): CMIP6 pre-training, fine-tuning only on MOM6. Inclusion of SOCAT observations reduces validation RMSE by 19.43 µatm (~72.31% relative reduction), demonstrating the pivotal role of SOCAT in achieving accurate $spCO_2$ reconstruction.

(5) For the physical-biogeochemical constraints, are you only using what is derived from MOM6, or also from CMIP6 models as well?

In our study, the physical-biogeochemical constraints incorporated in the ViT model are derived exclusively from MOM6 simulations. MOM6 provides high-resolution, rigorously validated ocean-driven fields that more accurately represent conditions relevant to the surface carbon system (Stock et al., 2020; Liao et al., 2020).

Evaluations such as those by Liao et al. (2021) indicate that CMIP6 model outputs contain relatively large biases in space and time, which could reduce the reliability of any constraints derived directly from CMIP6. Therefore, MOM6 is used as the sole source for physical-biogeochemical constraints to ensure accuracy, consistency, and physical realism in the model refinement stage. In addition, we have conducted comparison experiments in the manuscript between models trained with and without these constraints, demonstrating that incorporating MOM6-derived information significantly improves predictive skill in data-sparse regions and high-latitude oceans.

(6) How are your results, particularly on the seasonal cycle, impacted by these physical-biogeochemical constraints? In other words, if you exclude these constraints, how is the representation of the seasonal pCO_2 cycle affected?

To assess the impact of MOM6-derived physical-biogeochemical constraints on the seasonal cycle of $spCO_2$, we conducted two comparative experiments while keeping all other model settings identical:

(a) Test 1 (with physical-biogeochemical constraints): The SJTU-AViT model reconstruction incorporated MOM6-derived constraints during training, enforcing physically and biogeochemically plausible relationships among environmental variables.

(b) Test 2 (without physical-biogeochemical constraints): The SJTU-AViT reconstruction excluded these constraints, allowing the model to rely solely on observational and CMIP6-derived information.

The constraints systematically improve model performance across all seasons (Fig. R5), as reflected in reduced RMSE values: MAM decreases from 11.66 to 11.35 μatm (~2.66%), JJA from 12.31 to 11.93 μatm (~3.09%), SON from 13.67 to 12.51 μatm (~8.49%), and DJF from 10.32 to 10.18 μatm (~1.36%). On average, the inclusion of constraints reduces RMSE by ~3.90% across the four seasons.

These improvements are systematic and physically meaningful rather than random fluctuations. The MOM6-derived constraints anchor the model to physically and biogeochemically plausible relationships, enhancing the accuracy and robustness of the seasonal $spCO_2$ representation. The constraints are particularly effective in regions with sparse observational coverage, where purely data-driven reconstructions may be prone to larger errors. Overall, the results demonstrate that including physical-biogeochemical constraints play a substantial and reliable role in improving the seasonal cycle representation of $spCO_2$, rather than merely introducing stochastic or localized enhancements.

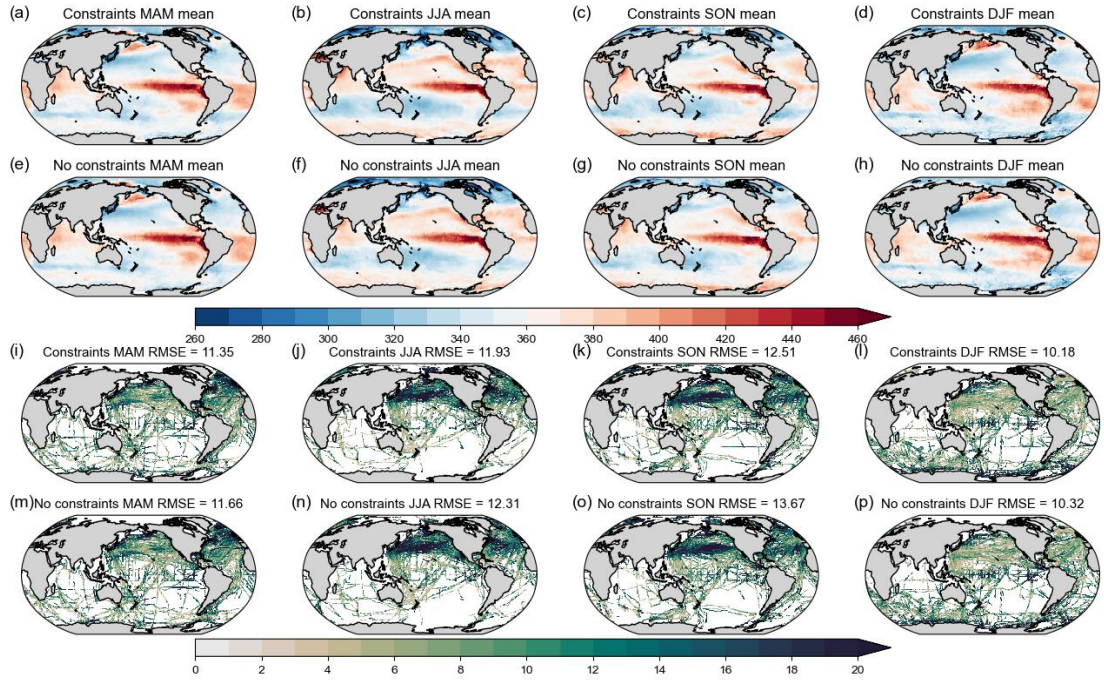

Figure R5 (Figure S13 in supplement section S5). Seasonal comparison of SJTU-AViT $spCO_2$ means and RMSE with and without physical-biogeochemical constraints. (a-d) Test 1 (with physical-biogeochemical constraints): seasonal mean $spCO_2$ from SJTU-AViT with physical-biogeochemical constraints for MAM (March-May), JJA (June-August), SON (September-November), and DJF (December-February). (e-h) Test 2 (without physical-biogeochemical constraints): seasonal mean $spCO_2$ from SJTU-AViT without constraints. (i-l) Test 1 (with physical-biogeochemical constraints): seasonal RMSE of $spCO_2$ between SJTU-AViT and SOCAT with constraints. (m-p) Test 2 (without physical-biogeochemical constraints): seasonal RMSE of $spCO_2$ between SJTU-AViT and SOCAT without constraints. For RMSE calculations, SJTU-AViT $spCO_2$ was interpolated to SOCAT observation locations and times.

The results are summarized in the main text (lines 601-612) as *"In addition, we evaluated the contributions of CMIP6 pre-training, MOM6 fine-tuning, SOCAT observations, and MOM6-derived physical-biogeochemical constraints within the SJTU-AViT framework. CMIP6 pre-training substantially improved model initialization and skill, reducing validation RMSE by ~56.57% versus random initialization by supplying large-scale structure and low-frequency variability. MOM6 fine-tuning further stabilized the model—especially in observation-sparse regions—lowering RMSE by ~39.36% and enforcing physically plausible relationships. Including SOCAT during fine-tuning was critical for local and regional accuracy, reducing RMSE by ~72.31% through high-quality pointwise constraints. Sensitivity tests indicate the reconstruction is largely robust to the specific choice of CMIP6 pre-training subsets, provided multiple models are used to capture diverse large-scale patterns. Finally, adding MOM6-derived physical constraints improved overall performance (MAE from 7.15 to 5.95 μatm) and reduced seasonal RMSE by*

1.36-8.49%, with the largest gains in high-latitude and data-sparse regions. Collectively, these results confirm that CMIP6 pre-training followed by MOM6- and SOCAT-constrained fine-tuning with physically informed constraints yields a robust, reliable, and physically consistent reconstruction of $spCO_2$ across spatial and temporal scales.".

2. The uncertainty quantification might benefit from more detail. For u_{map}, what if there are no observations in one grid? How do you then quantify u_{map} there? Have you conducted an analysis on the spatial heterogeneity of the dominant source of uncertainty? In addition, I think it would be more appropriate to replace u_{map} with "algorithm uncertainty." Perhaps this can be done by generating a large ensemble of $spCO_2$ Alternatively, this can be done by using synthetic data. You might consider subsampling SOCAT data from one of your models and then applying the ML model to subsampled model fields to generate an $spCO_2$ map. Then you can compare the absolute differences between pCO_2 from the ocean model and the ML reconstruction.

We acknowledge that the traditional u_{map} approach depends directly on observational coverage and may underestimate uncertainty in regions with sparse or missing SOCAT data. To address this limitation, we performed an additional experiment using synthetic data to provide a more robust estimate of algorithm uncertainty. Specifically, we used the RECCAP2 simulation from the Scott Doney group (hereafter SD data) as an independent reference "truth," which the ViT machine learning model had never seen before. The SD data were divided into two subsets:

- SD_SOCAT: SD outputs sampled at the spatiotemporal locations of SOCAT observations.
- SD_nonSOCAT: the remaining SD outputs.

Following our standard workflow (CMIP6 pretraining, MOM6 fine-tuning, and SD_SOCAT fine-tuning), we reconstructed $spCO_2$ and quantified three RMSE values:

(a) RMSE_SD_SOCAT = 5.58 µatm. This is bias at training locations, indicating good consistency with data the model has seen.

(b) RMSE_SD_nonSOCAT = 7.40 µatm. This is bias at independent validation points, demonstrating generalization to unseen data.

(c) RMSE_SD_all = 7.39 µatm. This is bias over the full SD dataset, reflecting the model's overall performance.

These results show that the training error is slightly lower, as expected, and the validation and overall errors are nearly identical. This indicates that the ViT model does not overfit and that its uncertainty estimates are robust across different spatial domains. The close agreement also demonstrates that algorithm uncertainty captures the spatial heterogeneity of errors, particularly in high-latitude or data-sparse regions where u_{map} cannot be defined.

Based on this analysis, we adopt the RMSE_SD_all = 7.39 µatm as a quantitative measure of algorithm uncertainty ($u_{algorithm}$), and have updated the manuscript accordingly in section 2.5 (lines 273-274) and section 3.6 (lines 580-581). Specifically, it is now stated as: *"$u_{algorithm}$ is evaluated as the RMSE between the reconstructed and reference ocean model $spCO_2$ field."* (lines 273-274) and *"with the*

dominant contribution arising from the algorithm uncertainty ($u_{algorithm}$), which reaches 7.39 µatm." (lines 580-581). The full experimental details have been reported in the supplement (section S5.7).

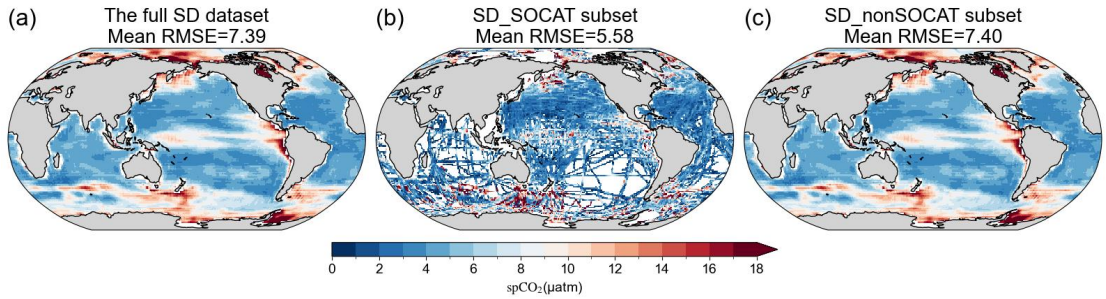

Figure R6 (Figure S14 in supplement section S5). Spatial distribution of RMSE (µatm) between the reconstructed $spCO_2$ field and the Scott Doney RECCAP2 simulation (SD data). (a) RMSE for the full SD dataset. (b) RMSE for the SD_SOCAT subset, i.e., SD data sampled at SOCAT observation locations and used in training. (c) RMSE for the SD_nonSOCAT subset, i.e., SD data at locations not sampled by SOCAT and reserved for independent validation. The mean RMSE value for each panel is indicated. The SD data is from Doney et al., (2009).

Minor comments:

1. L15-16: The statement that ocean surface partial pressure of $spCO_2$ directly determines the air-sea CO_2 flux is not exactly correct. It is the air-sea pCO_2 difference, which is modulated by surface wind speed and gas exchange velocity.

We agree that the original description was not accurate and have revised the corresponding sentence in the Abstract (lines 15-18) to read: *"The ocean plays a crucial role in regulating the global carbon cycle and mitigating climate change. Spatial and temporal variations of ocean surface partial pressure of CO_2 ($spCO_2$) influence the air-sea CO_2 flux through the difference between surface ocean and atmospheric pCO_2 (ΔpCO_2), which is further modulated by surface wind speed and gas exchange velocity."*

2. Introduction: Perhaps it is also worth mentioning that previous ML-interpolation of pCO_2 overly smooths the spatial patterns and interannual variability.

We have revised the Introduction accordingly (lines 67-69): *"Previous machine learning (ML)-based interpolations of pCO_2 may overly smooths the spatial patterns and interannual variability, which represents a potential limitation in capturing these features fully."*

3. L195: Is the interpolation based on inverse distance weighted average? How do you deal with the fine-resolution time (i.e., not monthly average)?

A similar question was also raised by the other reviewer. To clarify, we directly used the monthly 1° × 1° gridded product provided by the Surface Ocean CO_2 Atlas (SOCAT) for data construction. Therefore, no additional spatial interpolation was applied, and the temporal resolution is already monthly. To handle missing values, we

masked the corresponding reconstructed values at the same grid-time points before computing statistics, ensuring that all comparisons are made only where SOCAT provides valid data.

For independent validation at long-term stations, reconstructed values were extracted at the station locations using bilinear interpolation from the surrounding grid cells, rather than simply selecting the nearest grid cell. This approach yields smoother and more representative $spCO_2$ estimates. All datasets, including these station comparisons, were consistently processed as monthly averages, with no further temporal interpolation.

We have revised the section 2.3 (lines 212-220) accordingly. The new text reads: *"For comparison with SOCAT, we used the monthly 1° gridded SOCAT product and evaluated our SJTU-AViT reconstruction on the same grid, without applying any additional spatial interpolation. Reconstructed values were masked where SOCAT is missing, and all skill metrics were computed only at grid-time points with valid SOCAT data. For the independent test at long-term stations, reconstructed values were extracted at the corresponding station locations using bilinear spatial interpolation, which incorporates information from surrounding grid cells to provide smoother and more representative estimates, and skill metrics were subsequently computed to evaluate model performance. Detailed information for these stations, including their names, geographic locations, observation periods, number of samples, and data sources, is provided in supplement Table S3, and their locations are shown in supplement Fig. S2 to facilitate visual interpretation."*.

4. Figure 3: Systematic biases are clear at Iceland and Irminger, with SJTU-AViT underestimating the pCO_2. Any clues why?

These high-latitude regions are strongly influenced by processes such as seasonal sea-ice coverage and freshwater input from precipitation, which are not well captured in our machine learning model due to the lack of corresponding observational constraints. As a result, the model cannot fully resolve these pCO_2 variabilities, leading to the observed negative bias. This behavior is not unique to our product and similar biases have been reported in other reconstruction products under complex environmental conditions (e.g., Landschützer et al., 2016; Gregor et al., 2021). In future work, we plan to incorporate additional predictors, such as non-climatological mixed-layer depth (MLD), sea-ice coverage, precipitation, and chlorophyll, into the machine learning framework to improve reconstruction accuracy in high-latitude regions.

Modified sentence in the section 3.1 (lines 302-306): *"At the Irminger Sea and Iceland sites, the model exhibits large RMSE (35.24 and 21.82 μatm, respectively) and low correlations, with R^2 near zero. This suggests that the model has difficulty capturing rapid $spCO_2$ fluctuations or processes that are not well represented by the available input features. This discrepancy is likely due to high-latitude processes such as seasonal sea-ice variability and freshwater inputs, which are not fully represented in the current observational constraints."*

5. Figure 5: The negative bias would lead to an overestimation of global ocean CO_2 uptake through the bulk equation. Might be worth mentioning when you talk about the flux.

We have mentioned this potential bias in the updated manuscript in section 3.5. Specifically, while SJTU-AViT effectively reproduces the overall spatial patterns and mechanisms of air-sea CO_2 flux, negative $spCO_2$ biases remain in certain high-latitude regions (now is Fig. 6). These biases probably result from underestimation of pCO_2 in areas affected by seasonal sea-ice variability, freshwater inputs, and other high-latitude processes that are not fully captured by observational constraints.

The revised text now reads (section 3.5, lines 559-563): *"While SJTU-AViT effectively reproduces the overall spatial patterns and mechanisms of air-sea CO_2 flux, Figure 6 indicates that negative $spCO_2$ biases remain in certain high-latitude regions. The negative bias, likely associated with underrepresented high-latitude processes such as seasonal sea-ice variability and freshwater inputs, can lead to an overestimation of global ocean CO_2 uptake through the bulk equation and should be considered when interpreting the absolute flux magnitude."*.

6. Fig. 6b: Seems like the bias PDF is wider in certain years. Speculation?

It is noteworthy that the bias probability density function (PDF) exhibits interannual variability and even decadal trends (now is Fig. 7b). In the early years (1980s to mid-1990s), the bias distribution is relatively broad, reflecting larger uncertainties. This is probably attributable to the sparse SOCAT coverage during that period. Limited observational data constrained the model's ability to resolve local and temporal variations, leading to larger bias. Over time, as SOCAT coverage expanded, reconstruction accuracy improved in most regions. The bias distribution became narrower and more symmetric, with the PDF centered near zero, indicating reduced systematic bias.

In recent years, however, the bias range appears to increase. This widening is likely related to the extension of observational coverage into high-latitude, polar, and coastal regions, where conditions are more variable and extreme. In addition, recent ocean pCO_2 changes have exhibited enhanced seasonal and interannual variability, which the model may not fully capture, particularly under extreme or marginal conditions. These interpretations remain tentative, and more detailed analyses—such as targeted experiments and the incorporation of additional datasets—will be necessary to fully disentangle these drivers. We view this as an important avenue for future work and would welcome collaborations to further investigate these aspects.

This has been added to section 3.2 (lines 370-375) as *"However, we note that the absolute range of biases may increase in later years. This widening is likely due to a combination of factors, including the expansion of observational coverage to regions with more extreme or marginal conditions, which introduces a larger range of reconstructed values, as well as the enhanced seasonal and interannual variability that the model may not fully capture in some regions, leading to increased biases under local or extreme conditions. Overall, the temporal evolution of the bias distribution highlights both the influence of observational coverage and the*

challenges in capturing high-frequency or extreme variations.”.

7. L369-372: The section title is on the seasonal cycle, but the first few sentences focus on variability at all time scales. Might consider moving this to a later section. Also, the trend should be removed beforehand in calculating STD in Fig. 7.

We carefully considered splitting this section into two parts (full variability and seasonal variability). However, doing so would result in very little content for the full variability part and lead to an imbalance between subsections. Therefore, in the revised manuscript we chose to keep the two components together but revised the section title to *“3.3 Evaluation of full $spCO_2$ variability and seasonal cycle”* (line 399), which more accurately reflects the content.

In addition, as suggested, we removed the long-term trend prior to calculating the standard deviation (STD) in the figure (now is Fig. 8 and Fig. S5). After detrending, the STD values are slightly smaller than in the original calculation. But the relative magnitudes among different regions remain unchanged, and the spatial patterns of variability are still highly consistent with the observational data. Therefore, this adjustment does not affect our main conclusion that the model captures the full variability of $spCO_2$ well across most regions.

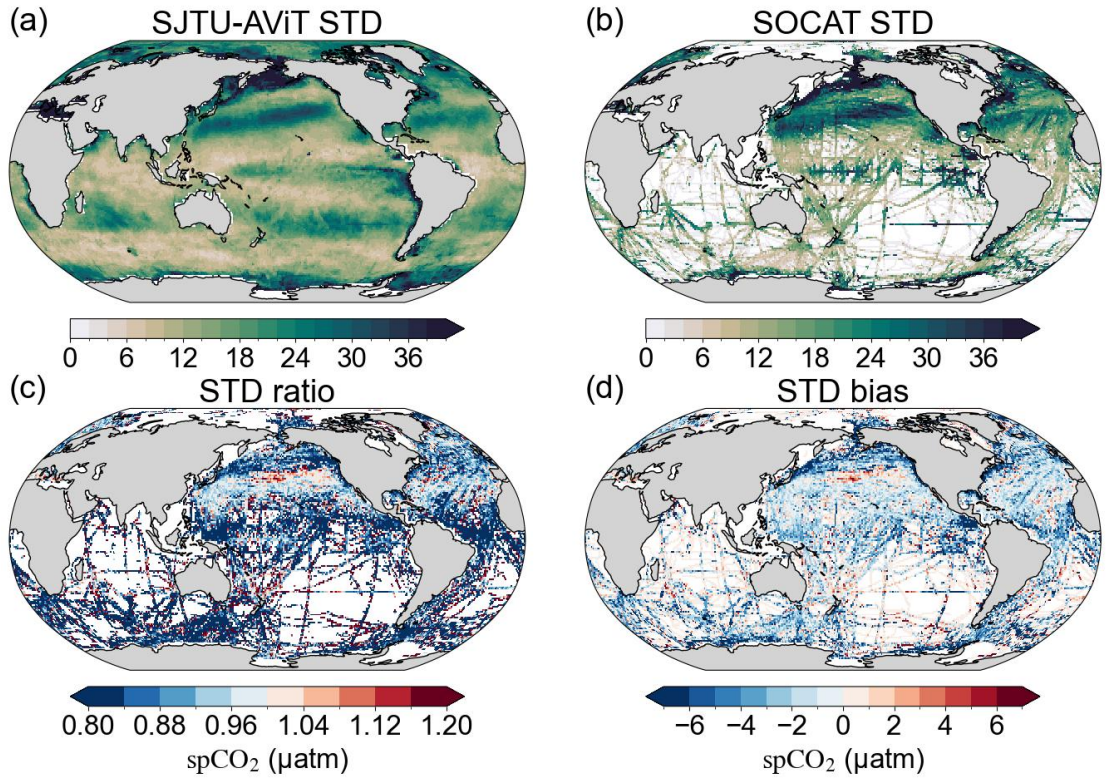

Figure R7 (Figure 8 in main text). Comparison of $spCO_2$ standard deviation from 1982-2023 between SJTU-AViT and SOCAT. (a) Standard deviation of $spCO_2$ from the SJTU-AViT reconstruction. (b) Standard deviation of $spCO_2$ from SOCAT data. (c) Standard deviation ratio, representing the ratio of SJTU-AViT to SOCAT standard deviation (SJTU-AViT divided by SOCAT). (d) Standard deviation bias, showing the difference between the SJTU-AViT and SOCAT standard deviations (SJTU-AViT minus SOCAT). The standard deviation (STD) is quantified as the standard deviation

of residuals after removing long-term trends. In the panels c and d, the SJTU-AViT values are interpolated to match the spatial and temporal locations of SOCAT observations (see detailed computation in section 2.3).

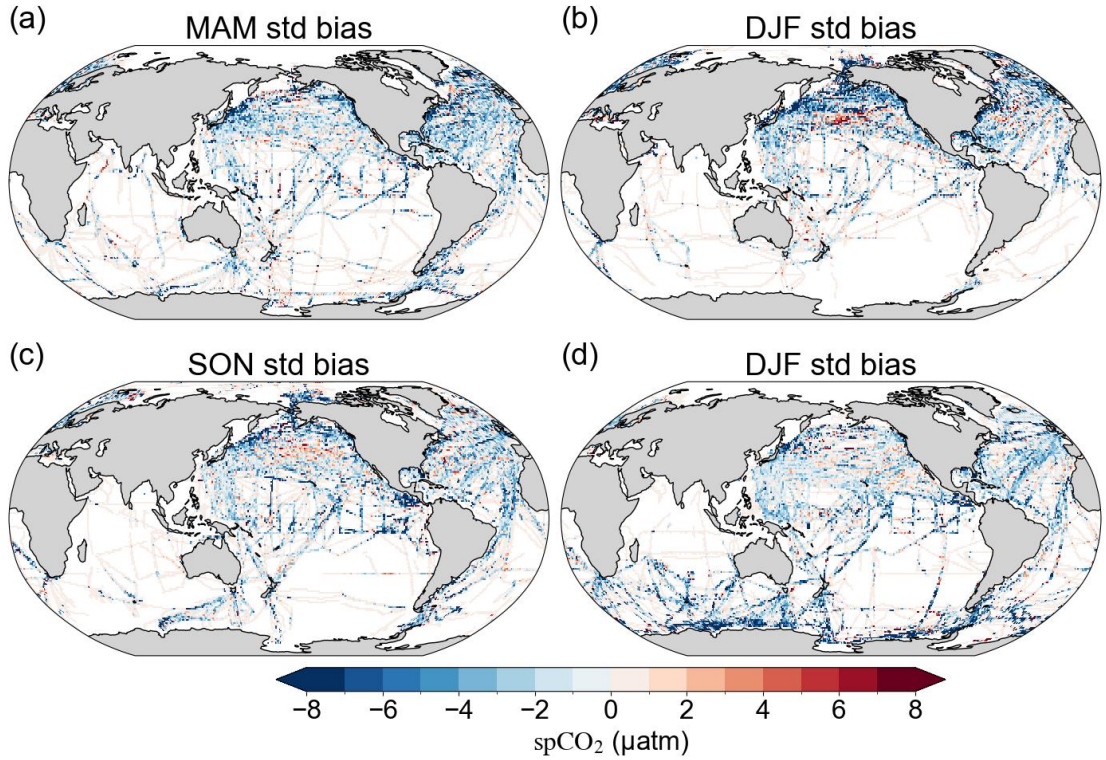

Figure R8 (Figure S5 in supplement section S4). Bias in the standard deviation of $spCO_2$ between SJTU-AViT and SOCAT at each season from 1982 to 2023. (a) MAM (March-May), (b) JJA (June-August), (c) SON (September-November), and (d) DJF (December-February). The standard deviation (STD) is quantified as the standard deviation of residuals after removing long-term trends. The bias is calculated as the difference between SJTU-AViT and SOCAT standard deviations at each season (SJTU-AViT minus SOCAT). Positive values (red) indicate overestimation of variability by SJTU-AViT, while negative values (blue) indicate underestimation. These seasonal biases highlight the model's performance across different seasonal periods and regions. The $spCO_2$ in SJTU-AViT is interpolated to match the SOCAT observation locations and times in the comparison (see detailed computation in section 2.3).

8. L391-396: A presentation issue. The seasonal changes are, physically, attributed to these factors you mentioned. This is based on our understanding of the ocean carbon dynamics rather than being directly learned from ML output. The sentences read like you confirm these dominant factors from your model output. Might consider making it clear that these are not model results. Or, indeed, you could do factor contribution analysis.

We agree that the physical attributions in the original text are based on established oceanographic understanding rather than direct causal inferences from the ML outputs. We have clarified the text in the revised manuscript (now is lines 422-427):

"Furthermore, the model reasonably reproduces seasonal increases in $spCO_2$ in the North Pacific and North Atlantic (40°-60°N) during Northern Hemisphere winter and early spring. This suggests that the model has likely captured underlying mechanisms, such as the deepening of the winter mixed layer and the entrainment of DIC-rich subsurface waters, which drive seasonal variations in surface ocean pCO_2 (Keppler et al., 2020). Conversely, a pronounced seasonal decrease in $spCO_2$ is simulated in the high-latitude Southern Ocean (south of 60°S) during the same period, indicating that the model may also have learned the influence of cooling-driven solubility changes and biological activity on ocean pCO_2."

9. Figure 9: I think what is missing here is to show whether the seasonal phases are consistent compared to SOCAT.

To evaluate whether our reconstruction can accurately capture the seasonal phase observed in SOCAT, we carried out additional analyses comparing the model results with SOCAT climatologies (new supplement section S5.9; see lines 448-453 in the revised manuscript). Specifically:

(a) Seasonal cycle comparison across ocean basins: We have evaluated the seasonal cycle month-by-month for the global ocean and five major basins, separately for the Northern and Southern Hemispheres. These comparisons demonstrate that the model well reproduces the seasonal cycle of $spCO_2$, with peak and minimum months largely consistent with SOCAT observations (Figs. R9-R10).

(b) Phase bias evaluation: We produced global maps of the difference in ocean pCO_2 peak month and minimum month between SJTU-AViT and SOCAT (in months, range ±6). Across most regions, the phase differences in both peak and minimum months are within ±1 month, with only ~5% of grid points exceeding this threshold (Fig. R11).

Together, these results indicate that the reconstruction reliably reproduces seasonal phasing. The corresponding text has been added in the revised manuscript section 3.3 (lines 448-453): *"To evaluate the accuracy of the SJTU-AViT in capturing the seasonal phasing of $spCO_2$, we compared it against SOCAT climatology (supplement Figs. S16-S18). Climatological seasonal cycles were evaluated for the global ocean and five major basins, separately for the Northern and Southern Hemispheres. The SJTU-AViT closely reproduces the timing of seasonal maxima and minima in $spCO_2$, generally aligning with SOCAT observations. Global maps of phase differences show that most regions deviate by less than ±1 month, with only ~5% of grid points exceeding this range. These results demonstrate that the reconstruction data reliably captures the observed seasonal phasing."*.

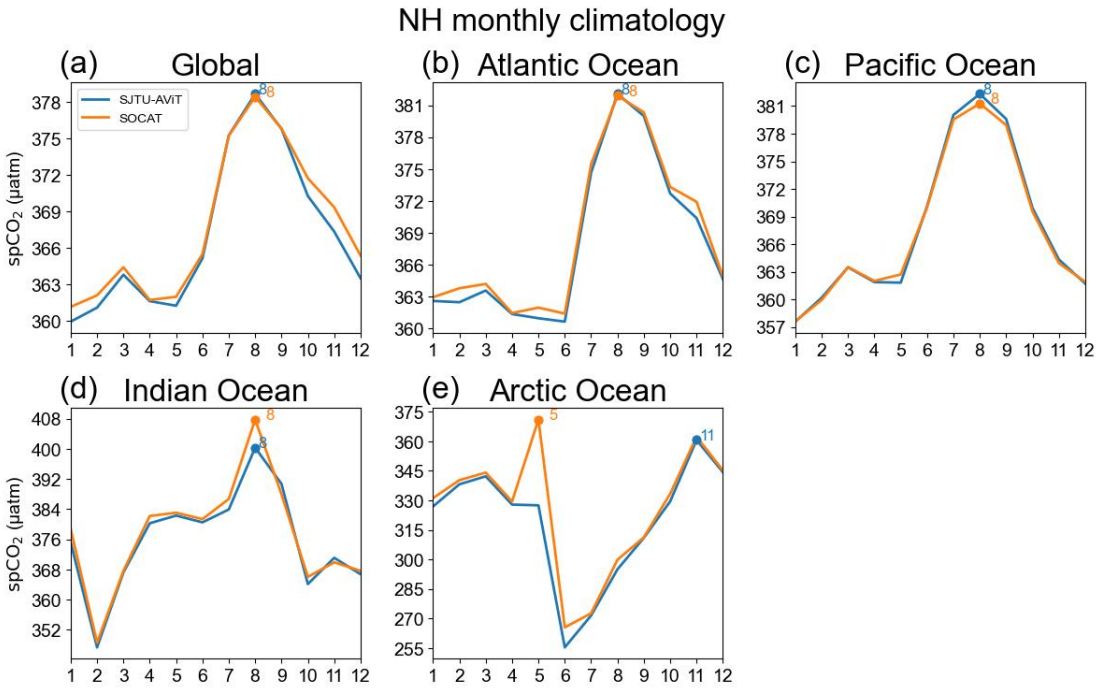

Figure R9 (Figure S16 in supplement section S5). Monthly $spCO_2$ regional time series for the Northern Hemisphere across different ocean regions from 1982 to 2023. Each panel shows the 12-month mean seasonal cycle for both the model (SJTU-AViT) and SOCAT observations. Peak months are indicated to allow direct comparison of seasonal phasing.

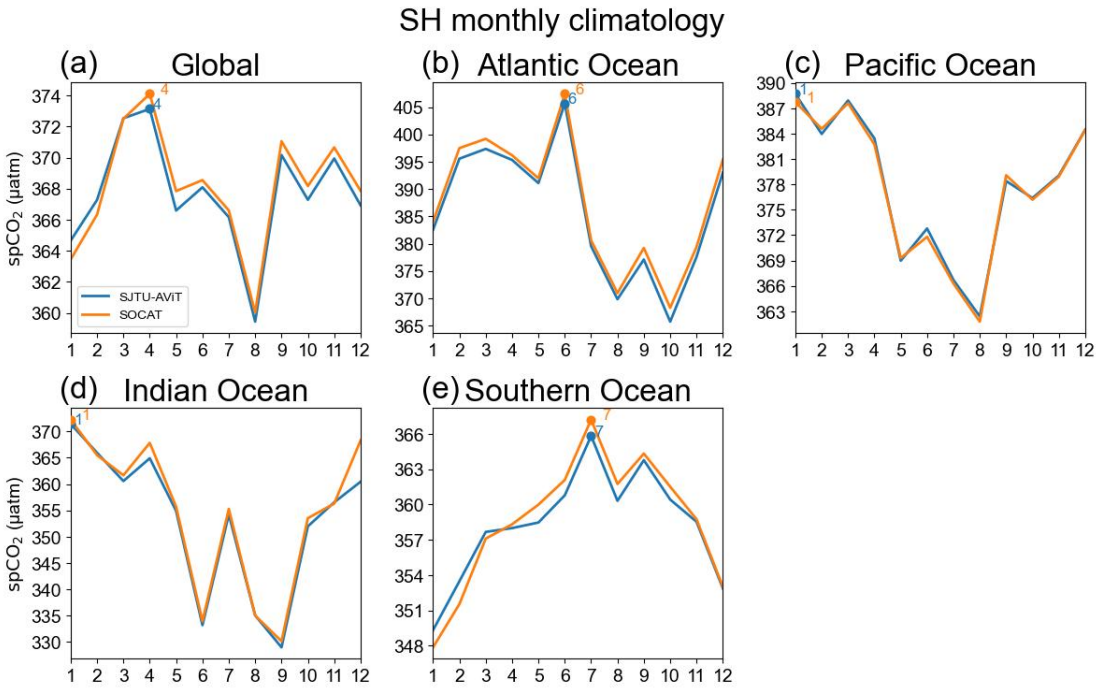

Figure R10 (Figure S17 in supplement section S5). Monthly $spCO_2$ regional time series for the Southern Hemisphere across different ocean regions from 1982 to 2023. Each panel shows the 12-month mean seasonal cycle for both the model (SJTU-AViT) and SOCAT observations. Peak months are indicated to allow direct comparison of

seasonal phasing.

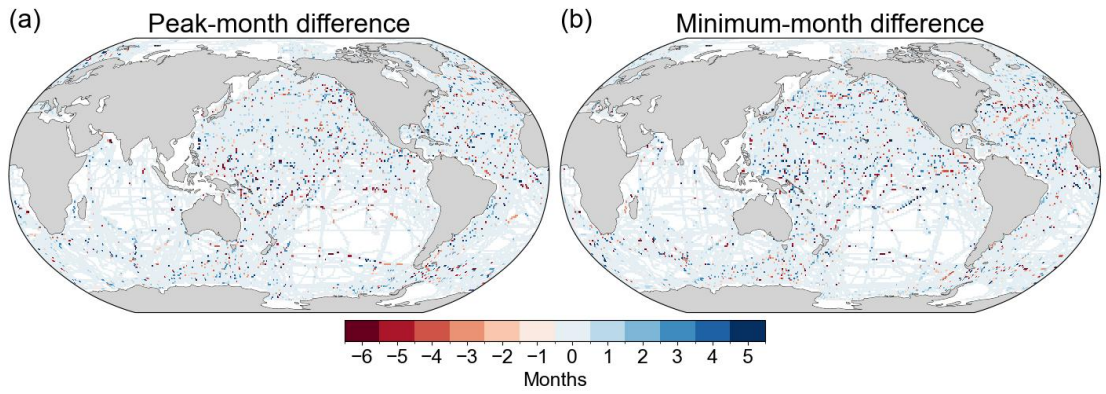

Figure R11 (Figure S18 in supplement section S5). Grid-scale maps of $spCO_2$ peak- and minimum-month differences (SJTU-AViT − SOCAT, in months, range ±6). For the peak-month difference map, positive values indicate that SJTU-AViT peaks later than SOCAT; for the minimum-month difference map, positive values indicate that SJTU-AViT minimums later than SOCAT. Regions with insufficient observational coverage are masked. These maps provide a spatial assessment of the model's ability to reproduce seasonal maxima and minima timing.

10. Figure 11: Linearly detrended $spCO_2$?

Yes, all $spCO_2$ data shown (now is Fig. 12) have been linearly detrended and deseasonalized. This processing ensures that the composite mean anomalies clearly highlight the typical $spCO_2$ responses associated with El Niño and La Niña events. We have updated the figure caption in the revised manuscript (lines 541-545), now as *"Figure 12. Comparison of $spCO_2$ anomalies during El Niño and La Niña events between SJTU-AViT and multiple data products. Panels (a) and (b) show the composite mean $spCO_2$ anomalies during eight El Niño and seven La Niña events, respectively, as reconstructed by the SJTU-AViT product. Panels (c) and (d) display the corresponding composite mean anomalies from the ensemble mean of eight $spCO_2$ data products. The eight El Niños and seven La Niñas are indicated in the supplement section S2 and S3. The $spCO_2$ anomalies are defined as residuals after removing both long-term trends and seasonal cycles."*.

11. L568-571: PDO-related SST patterns are used in your training; incorporating other indices (e.g., directly using PDO) would be double counting?

Indeed, the PDO signal is already implicitly embedded in the SST fields used as predictors, so directly adding the PDO index could raise concerns about double counting. However, machine learning models are not always efficient at extracting such low-frequency signals, particularly when the observational record is relatively short. In these cases, providing strong or even redundant cues can facilitate the machine learning model representation of decadal variability. In our additional experiments with physical constraints, we found that explicitly highlighting such kind of signals enabled the model to more effectively detect latent signals that are difficult

to capture, thereby improving reconstruction accuracy.

References not in manuscript:

Doney, S. C., Lima, I., Feely, R. A., Glover, D. M., Lindsay, K., Mahowald, N., Moore, J. K., Wanninkhof, R.: Mechanisms governing interannual variability in upper-ocean inorganic carbon system and air–sea CO_2 fluxes: Physical climate and atmospheric dust. Deep-Sea Res. Pt. II, 56, 640655. https://doi.org/10.1016/j.dsr2.2008.12.006, 2009.

Liao, E., Resplandy, L., Liu, J., and Bowman, K. W.: Future Weakening of the ENSO Ocean Carbon Buffer Under Anthropogenic Forcing, Geophys. Res. Lett., 48, e2021GL094021, https://doi.org/10.1029/2021GL094021, 2021.

Stock, C. A., Dunne, J. P., Fan, S., Ginoux, P., John, J., Krasting, J. P., Laufkötter, C., Paulot, F., and Zadeh, N.: Ocean Biogeochemistry in GFDL's Earth System Model 4.1 and Its Response to Increasing Atmospheric CO_2, J. Adv. Model. Earth Syst., 12, e2019MS002043, https://doi.org/10.1029/2019MS002043, 2020.