Earth System
Science
Data

Supplement of

A surface ocean pCO_2 product with improved representation of interannual variability using a vision transformer-based model

Xueying Zhang et al.

Correspondence to: Enhui Liao (ehliao@sjtu.edu.cn) and Shiyu Liang (lsy18602808513@sjtu.edu.cn)

Section S1 Monte Carlo method for uncertainty estimation

Estimating u_{inputs} requires quantifying the uncertainties of nine input variables. A conservative approach was adopted, using the highest reported uncertainty for each variable when available. The total uncertainty from input variables was calculated as:

40
$$u_{input} = \sqrt{u_{SST}^2 + u_{SSS}^2 + u_{MLD}^2 + u_{Chl-a}^2 + u_{pco_{2air}}^2 + u_{\frac{\partial spCO_2}{\partial SST}}^2 + u_{\frac{\partial spCO_2}{\partial SSS}}^2 + u_{\frac{\partial spCO_2}{\partial DIC}}^2 + u_{\frac{\partial spCO_2}{\partial ALK}}^2}$$

For SST, we used a global mean standard deviation of 0.24 °C from the OISST dataset. SSS uncertainty was set to 0.23 psu, based on the global mean standard deviation reported in the Hadley Centre EN.4.2.2 dataset. For MLD, we adopted a value of 7.06 m, derived from the global mean grid-level standard deviation in the WOCE Global Data Version 3.0. Chl-a uncertainty was set to 0.25 mg m^{-3}, represented as the RMSD of log_{10}-transformed chlorophyll-a concentration in seawater

45   provided by the ESA CCI Ocean Colour dataset. Lastly, the uncertainty in pCO_{2air} was taken as 0.22 ppm, based on the global mean uncertainty of xCO_2. The uncertainties of $\frac{\partial spCO_2}{\partial SSS}$, $\frac{\partial spCO_2}{\partial SST}$, $\frac{\partial spCO_2}{\partial DIC}$, $\frac{\partial spCO_2}{\partial ALK}$ are estimated using the standard deviations derived from monthly climatological data, with corresponding values of 0.16, 0.32, 0.06, and 0.05, respectively. These values were used in the Monte Carlo simulation to propagate input uncertainties through the pCO_2 estimation process.

We estimated uncertainty by individually perturbing each input variable. For a given input x_i, we generated 100 sets of

50   random perturbations $\varepsilon_i \sim N(0, u_i)$, where u_i is the assumed uncertainty of x_i. The perturbed inputs $x_i + \varepsilon_i$ were used to recompute spCO$_2$, and the differences Δ_i between the original and perturbed outputs were calculated. The standard deviation of Δ_i at each grid cell was taken as the uncertainty contribution of that input variable to the reconstructed spCO$_2$.

Section S2 The data description and climate mode selection

The stations used in this study include the Bermuda Atlantic Time-Series Study (BATS), Hawaii Ocean Time-series (HOT),

55   Eastern Pacific Ocean (Papa station), Irminger Sea Station, California Current Ecosystem (CCE1), Bay of Bengal Ocean Acidification (BOBOA), Iceland Station, Tropical Atlantic Ocean (TAO), and the European Station for Time-Series in the Ocean (ESTOC). The detailed locations are shown in Fig. S2a. These stations are strategically located across different ocean basins, covering regions such as the tropical and subtropical zones, high-latitude oceans, and coastal upwelling areas, each with its own distinct physical and biogeochemical properties.

60   Air-sea CO$_2$ flux data are available for 17 Earth System Models (ESMs) from the CMIP6 ensemble at the Lawrence Livermore National Laboratory node. From these 17 models, we selected a subset of 7 ESMs based on the availability of download access through our cluster and the availability of environmental variables. As detailed in Table S1, the selected models are: CESM2, CESM2-FV2, CESM2-WACCM, CESM2-WACCM-FV2, GFDL-ESM4, NorESM2-MM, and NorESM2-LM. For ease of data analysis, the output data from these models were regridded from their native horizontal grids

65   to a regular 1° x 1° grid using a bilinear remapping method (xESMF, python package, https://doi.org/10.5281/zenodo.1134365).

El Niño and La Niña events are identified based on the Niño 3.4 index, which is the 3-month running mean sea surface temperature (SST) anomaly for the Niño 3.4 region (5°N–5°S, 120°W–170°W). These events are defined as five consecutive overlapping 3-month periods with SST anomalies at or above +0.5°C for El Niño (warm) events, and at or below -0.5°C for

70   La Niña (cool) events (for more details, see https://ggweather.com/enso/oni.htm). The selected El Niño and La Niña events are listed in Table S2. The Indian Ocean Dipole (IOD) is defined by the Dipole Mode Index (DMI). IOD events are determined as the three-month running mean DMI is +0.4ºC or above (-0.4ºC or below) for at least three consecutive months between June and November (see details in https://ds.data.jma.go.jp/tcc/tcc/products/elnino/iodevents.html). The selected positive IOD events are also shown in Table S2.

Section S3 Table S1-S2

Table S1. List of the CMIP6 Earth system models used in this study.

Model	Ocean component	Ocean biogeochemical module	Ocean resolutions (lonxlat, levels)	Data DOI	Members labels
CESM2	POP2	MARBL	320x384, 60 levels	(Danabasoglu, 2019a; b; c)	r1i1p1f1
CESM2-FV2	POP2	MARBL	320x384, 60 levels	(Danabasoglu, 2019d)	r1i1p1f1
CESM2-WACCM	POP2	MARBL	320x384, 60 levels	(Danabasoglu, 2019e; f; g)	r1i1p1f1
CESM2-WACCM-FV2	POP2	MARBL	320x384, 60 levels	(Danabasoglu, 2019h)	r1i1p1f1
GFDL-ESM4	MOM6	COBALTv2	720x576, 75 levels	(John et al., 2018; Krasting et al., 2018)	r1i1p1f1
NorESM2-LM	MICOM	HAMOCC	360x384, 70 levels	(Seland et al., 2019a; b; c)	r1i1p1f1
NorESM2-MM	MICOM	HAMOCC	360x384, 70 levels	(Bentsen et al., 2019a; b; c)	r1i1p1f1

Table S2. List of selected El Niño, La Niña, and positive IOD events from 1985 to 2018.

Event	El Niño		La Niña		Positive IOD	
Event No.	Start Date	End Date	Start Date	End Date	Start Date	End Date
1	1986-12-01	1987-03-01	1988-12-01	1989-03-01	1994-09-01	1994-12-01
2	1987-12-01	1988-03-01	1995-12-01	1996-03-01	1997-09-01	1997-12-01
3	1991-12-01	1992-03-01	1998-12-01	1999-03-01	2006-09-01	2006-12-01
4	1994-12-01	1995-03-01	1999-12-01	2000-03-01	2007-09-01	2007-12-01
5	1997-12-01	1998-03-01	2007-12-01	2008-03-01	2012-09-01	2012-12-01
6	2002-12-01	2003-03-01	2010-12-01	2011-03-01	2015-09-01	2015-12-01
7	2009-12-01	2010-03-01	2011-12-01	2012-03-01	2017-09-01	2017-12-01
8	2015-12-01	2016-03-01	\	\	2018-09-01	2018-12-01

Table S3. List of selected independent test stations with long-term observations.

Station	Coordinates	Time range	Number of samples	URL
BAT	31.67°N, 295.83°E	10/1991-6/2022	324	https://bios.asu.edu/bats
HOT	22.75°N, 202°E	10/1988-12/2023	325	https://hahana.soest.hawaii.edu/hot/hotco2
ESTOC	29.07°N, 344.17°E	10/1995-11/2009	115	https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/Coastal/ESTOC.html
CCE1	33.50°N, 237.50°E	11/2008-12/2023	144	https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/Moorings/Pacific.html
TAO	-0.51°N, 189.98°E	2/2010-8/2016	45	https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/Moorings/Pacific.html
BOBOA	15°N, 90°E	11/2013-11/2018	53	https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/Moorings/Indian.html
Papa	50.13°N, 215.17°E	6/2007-4/2023	168	https://www.pmel.noaa.gov/co2/story/Papa
Iceland	68°N, 347.40°E	2/1985-11/2021	158	https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/Moorings/Atlantic.html
Irminger	64.30°N, 332°E	3/1983-11/2012	99	https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/Moo

Table S4. Skill metrics of the reconstructed spCO₂ by ocean basin.

Ocean basin	N	R^2	RMSE	MAE	MBE
Pacific ocean	159783	0.94	6.79	5.29	0.30
Atlantic ocean	111326	0.81	7.10	5.31	-0.31
Indian ocean	6354	0.95	5.31	4.75	-0.08
Arctic ocean	10316	0.90	8.80	7.58	-0.24
Southern ocean	48636	0.88	8.20	6.76	-0.55

Table S5. Skill metrics of the reconstructed spCO₂ by latitude band.

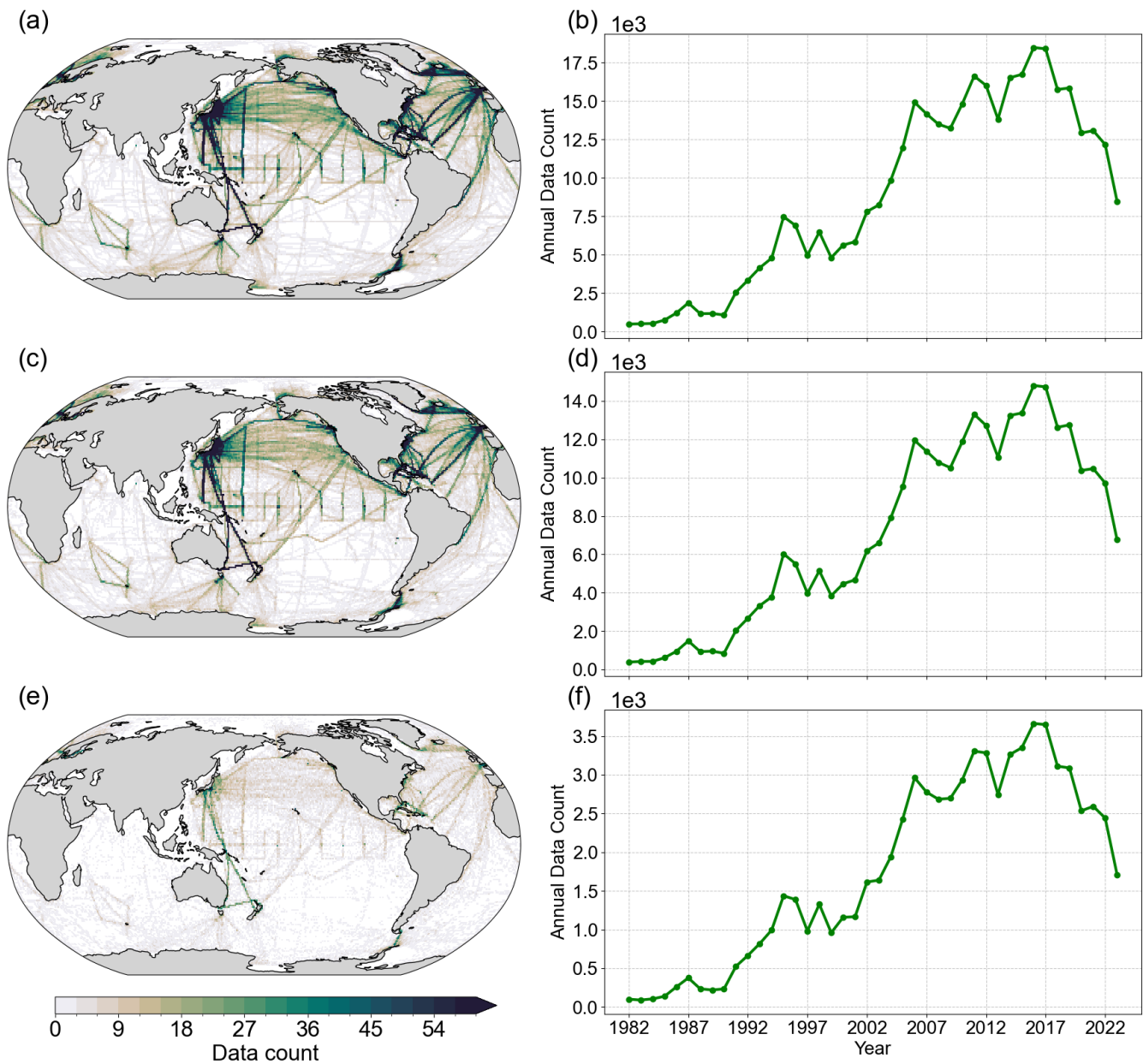| latitude band	N	R^2	RMSE	MAE	MBE
60°N-90°N	30802	0.92	9.23	7.58	-0.56
30°N-60°N	123357	0.91	9.13	6.40	0.07
0-30°N	96608	0.72	6.13	4.74	0.04
0-30°S	35804	0.97	5.70	4.96	-0.07
30°S-60°S	43497	0.90	6.13	5.29	-0.20
60°S-90°S	16602	0.86	11.80	9.29	-1.03

85

Figure S1. Data availability for spCO₂ reconstruction. (a) Spatial distribution of the number of all spCO₂ data points. (b) Annual data count of all spCO₂ data points over the period from 1982 to 2023. (c) Spatial distribution of the number of spCO₂ data points used for training. (d) Annual data count used for training over the period from 1982 to 2023. (e) Spatial distribution of the number of spCO₂ data points used for validation. (f) Annual data count used for validation over the period from 1982 to 2023.
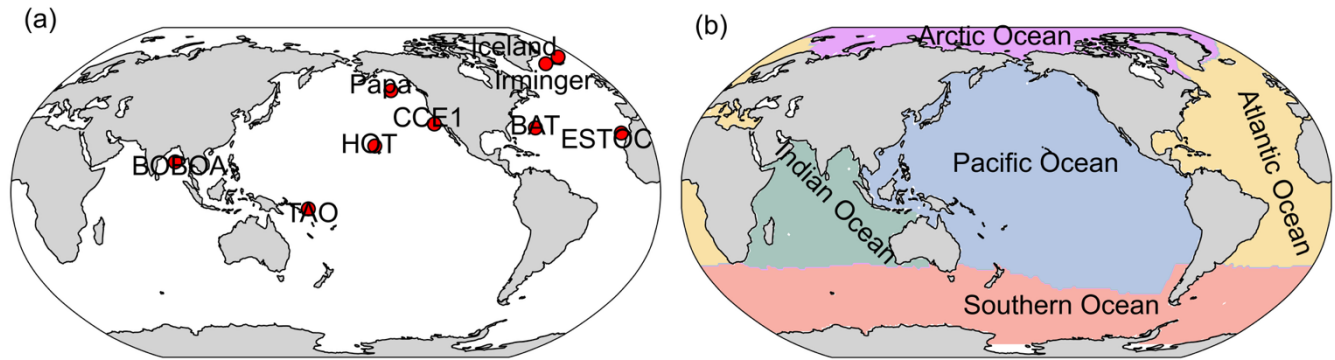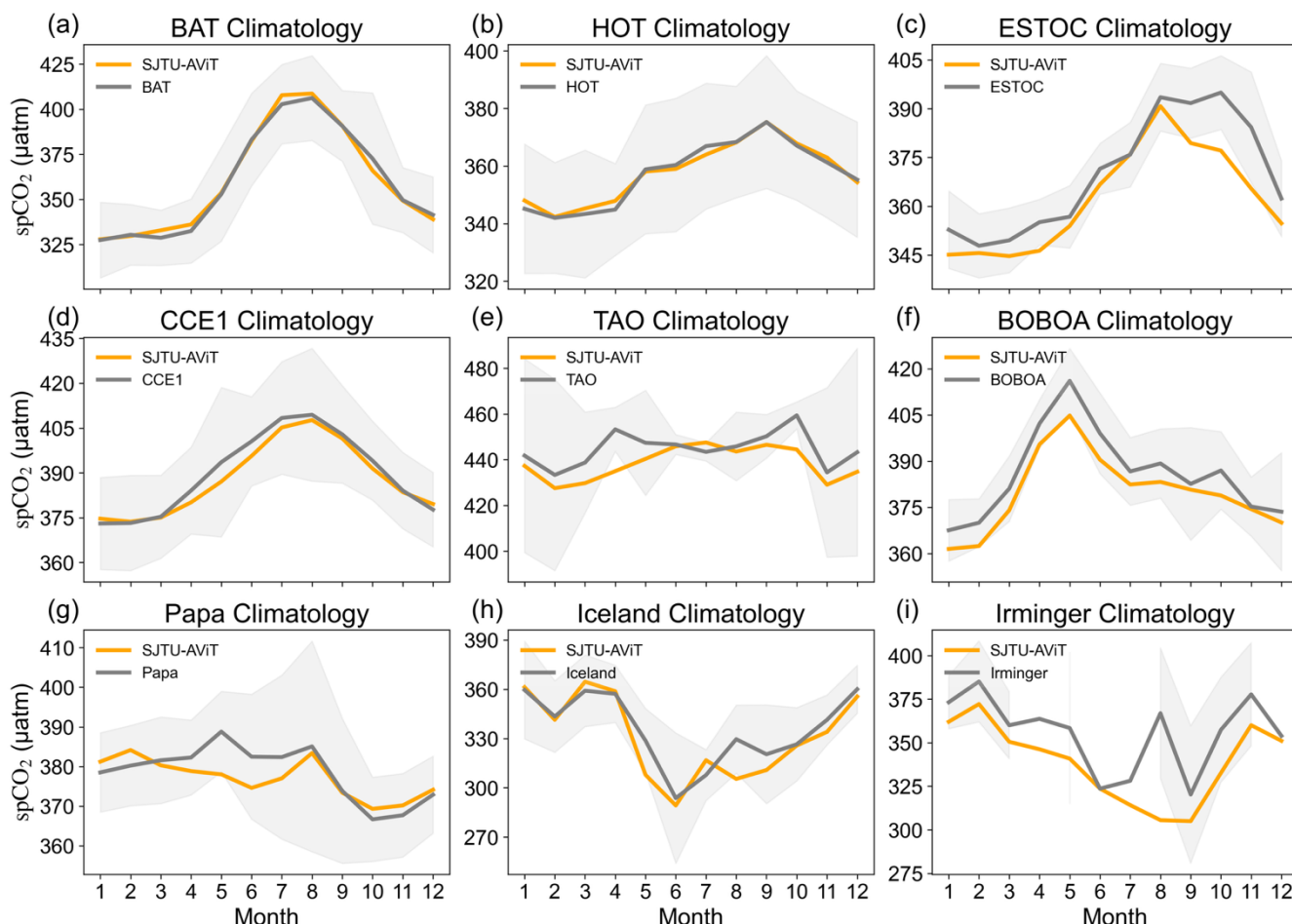

Figure S2. Spatial distribution of independent in situ observations and the definition of ocean basins used in this study.

Figure S3. Independent test of seasonal cycles of spCO₂ climatology between SJTU-AViT and in situ observations. These in situ data are independent data and are not used to train the model. The station description and location refer to supplement section S2 and Fig. S2. The spCO₂ in SJTU-AViT is interpolated to match the station locations and times in the comparison. The lines represent the monthly mean spCO₂ values, with the shaded regions indicating the standard deviation for the observed climatology. The SJTU-AViT data product demonstrates good agreement with the observed climatological spCO₂ patterns at each station.
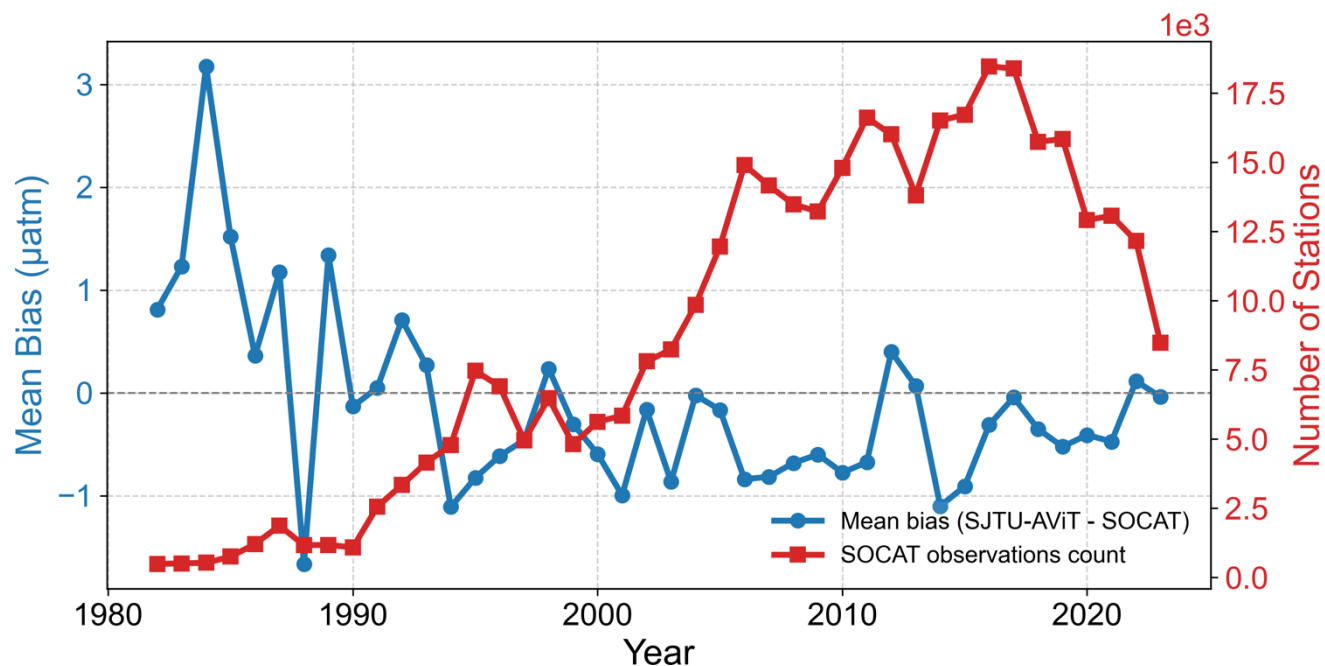
Figure S4. Temporal evolution of bias and SOCAT observation count from 1982 to 2023. The blue line represents the bias in long-term mean spCO₂ (SJTU-AViT minus SOCAT), while the red bars show the annual number of SOCAT observations contributing to the data. The increasing observation count over time correlates with a decrease in the mean bias, suggesting improvements in model performance as more observational data became available.
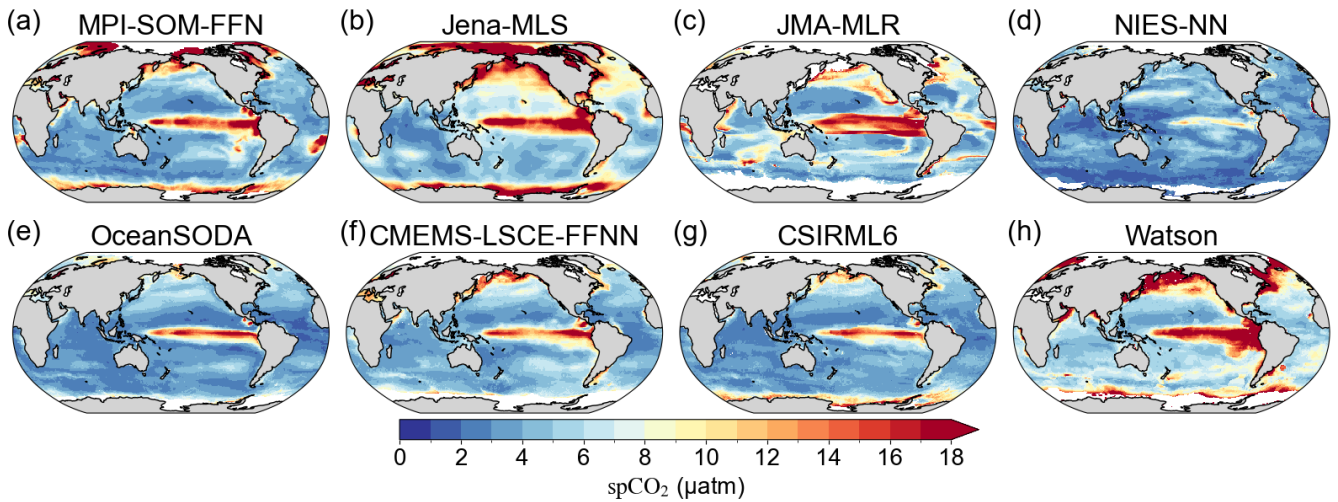
(a) MAM std bias　　(b) DJF std bias

(c) SON std bias　　(d) DJF std bias

spCO$_2$ (µatm)
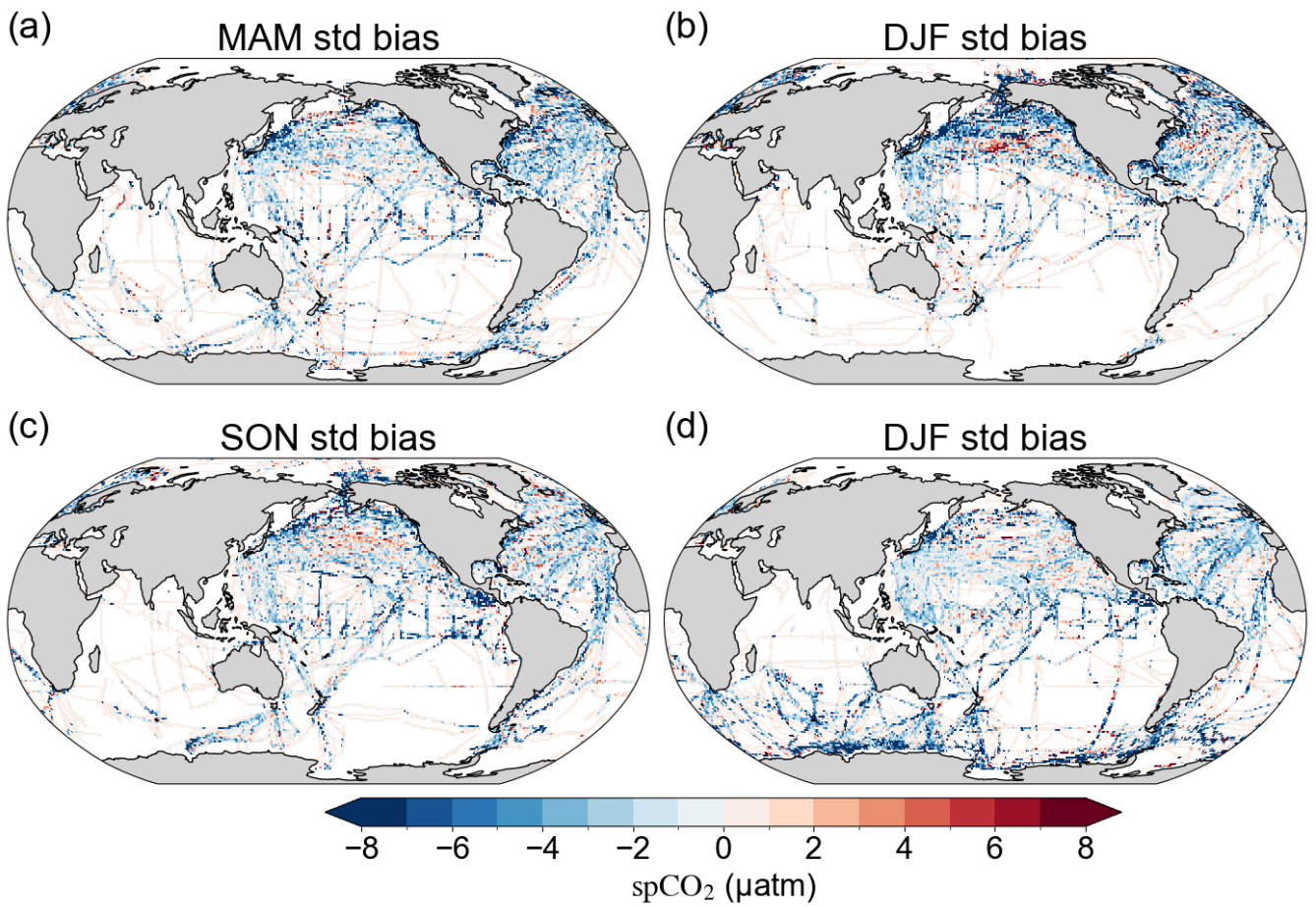
105

Figure S5. 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
110　**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).**


(a) MPI-SOM-FFN　(b) Jena-MLS　(c) JMA-MLR　(d) NIES-NN

(e) OceanSODA　(f) CMEMS-LSCE-FFNN　(g) CSIRML6　(h) Watson

spCO$_2$ (µatm)

Figure S6. Spatial distribution of standard deviation in interannual time scale of reconstructed spCO$_2$ at multiple data products
115　**from 1985 to 2018. All the panels show the standard deviation of residuals after removing long-term trends and seasonal cycles. The color scale represents the magnitude of variability in spCO$_2$, with higher values (red) indicating greater variability.**
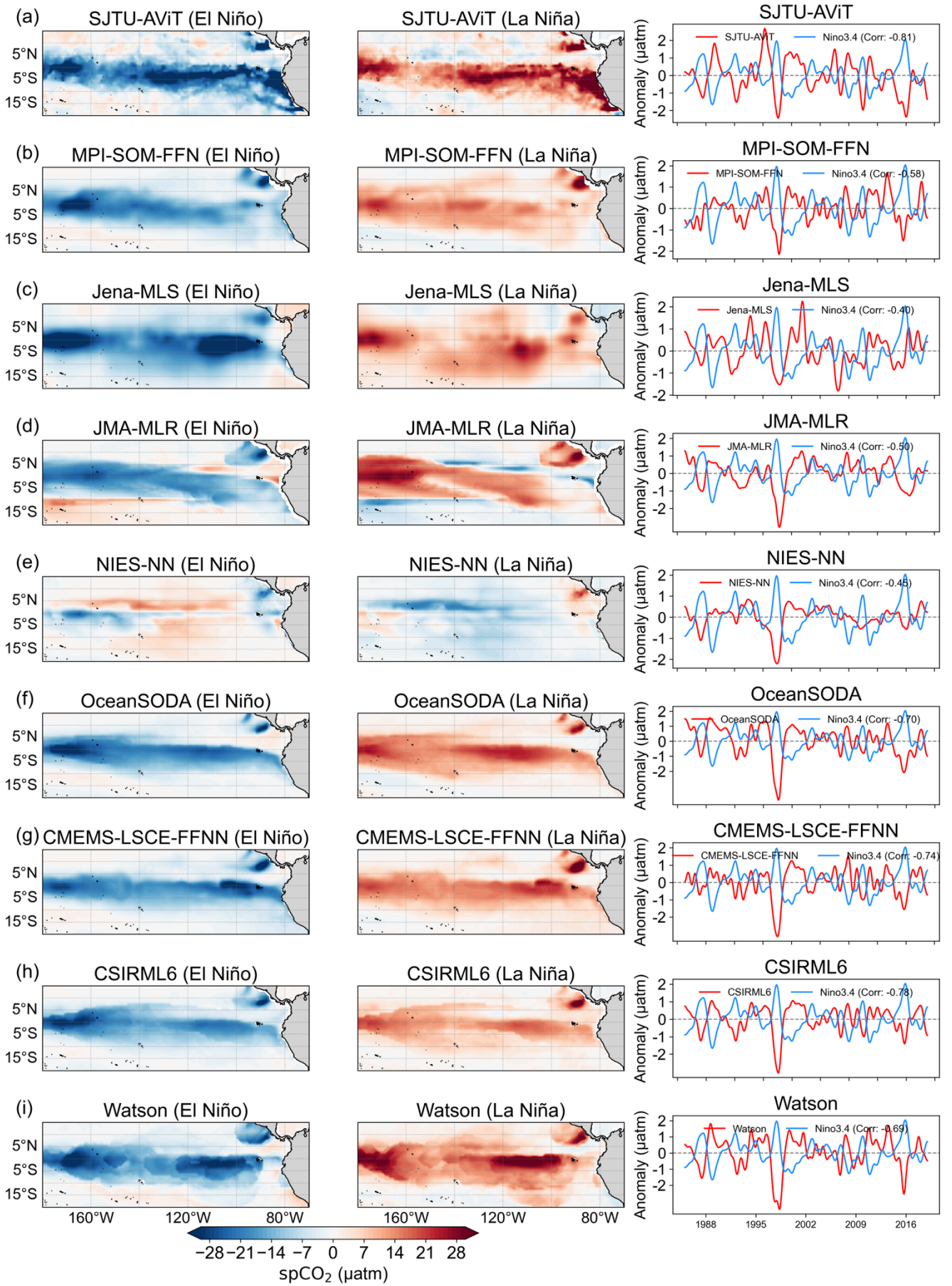
Figure S7. Spatial and temporal patterns of spCO₂ anomalies during ENSO events in the equatorial Pacific Ocean: comparison between SJTU-AViT and multiple data products. The left column shows composite spatial distribution of spCO₂ anomalies during eight El Niño events. The middle column shows composite spatial distribution of spCO₂ anomalies during seven La Niña events. The right column shows the time series of spCO₂ anomalies averaged over the equatorial eastern Pacific and their correlation with the Niño 3.4 SST index. The eight El Niños and seven La Niñas are indicated in the supplement section S2 and S3.

120
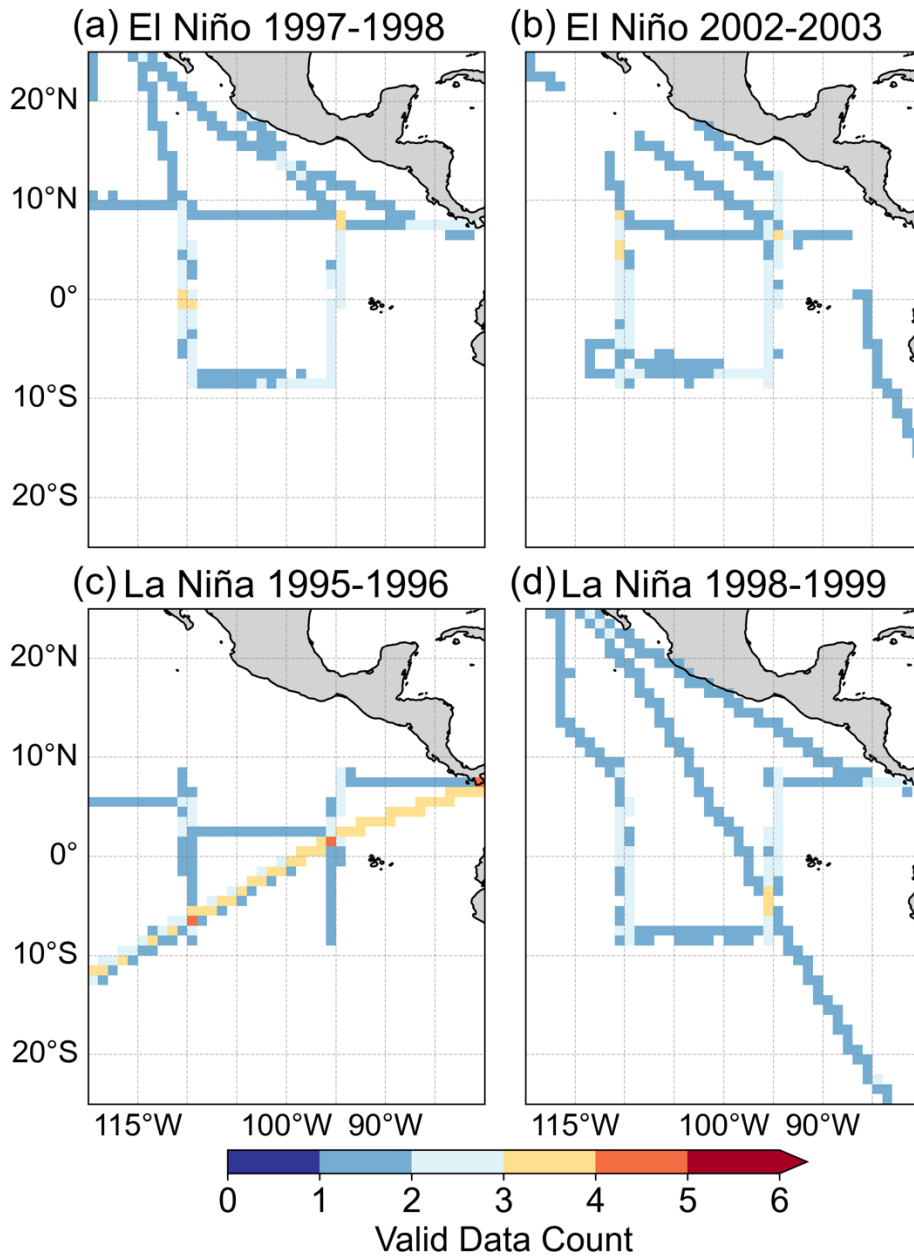
Figure S8. Spatial distribution of SOCAT spCO2 observations in the Equatorial Pacific Ocean (240°E–280°E) during selected ENSO events. The color scale indicates the valid data count per 1°×1° grid cell during four distinct ENSO events: (a) El Niño 1997–1998, (b) El Niño 2002–2003, (c) La Niña 1995–1996, and (d) La Niña 1998–1999.

Section S5 Workflow, Ablation experiments, and Validation of spCO2 reconstruction

5.1 Workflow of the spCO2 reconstruction using the ViT-based framework

The spCO2 reconstruction workflow based on the ViT framework is organized into four main stages—Data Processing, Model Architecture, Training & Validation, and Evaluation & Analysis—as illustrated in Fig. 2 (in main text). At the top, the data processing panel shows the input sources (CMIP6, MOM6, SOCAT) and the preprocessing steps: temporal harmonization to a monthly cadence, spatial regridding to a 1°×1° grid, and feature normalization. These boxes indicate that all inputs are brought to a common spatio-temporal grid and scale before being passed to the model. The model architecture panel depicts how physical variables are converted into model inputs: variable tokenization, variable aggregation, and then fed into a Transformer backbone that learns spatial and temporal dependencies. The model output block illustrates that the network predicts monthly spCO2 on the same 1° grid. The training & validation panel summarizes our multi-stage training strategy: (i) pretraining on CMIP6-derived fields, (ii) fine-tuning using MOM6 plus 80% of SOCAT, and (iii) evaluation using a withheld 20% SOCAT validation split and independent tests at long-term station sites. Finally, the evaluation & analysis panel shows the main evaluation products derived from the reconstruction: model performance metrics, climatology, seasonal cycle, interannual variability, and downstream analyses (air-sea CO2 flux calculation and uncertainty analysis).

5.2 Ablation experiments on the contribution of CMIP6 pre-training

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 pre-training): 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 pre-training): 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 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.
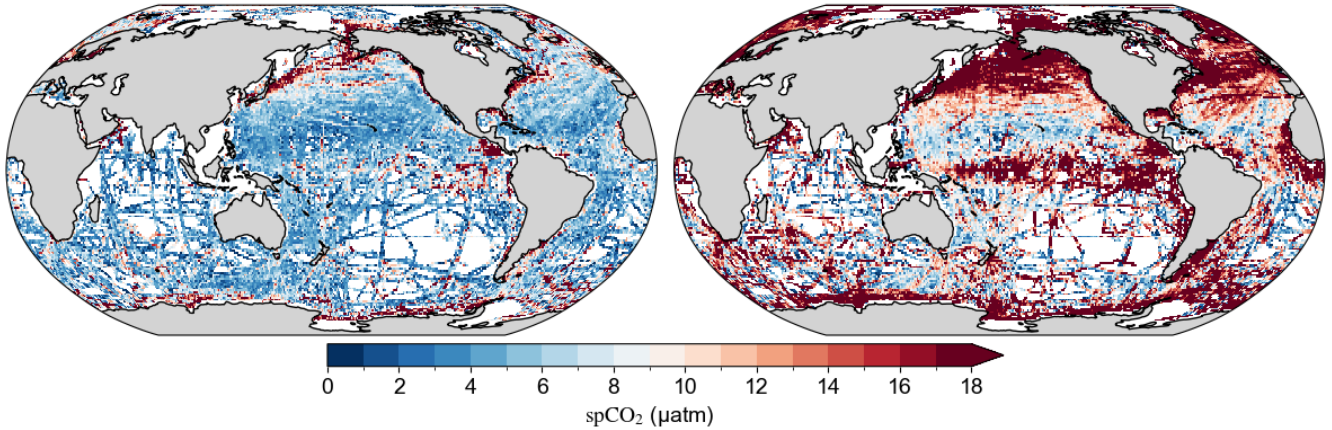

Figure S9. Impact of CMIP6 pre-training on reconstructed spCO$_2$ fields. (a) Test 1 (with CMIP6 pre-training): CMIP6 pre-training followed by MOM6 & SOCAT fine-tuning; (b) Test 2 (without CMIP6 pre-training): no CMIP6 pre-training, trained only on MOM6 & SOCAT. Inclusion of CMIP6 pre-training reduces validation RMSE by 9.69 µatm (~56.57% relative reduction), justifying the two-stage training strategy.

5.3 Ablation experiments on the contribution of MOM6 fine-tuning

To assess the role of MOM6 fine-tuning in our reconstruction framework, we designed two comparative experiments while keeping all other model settings identical:
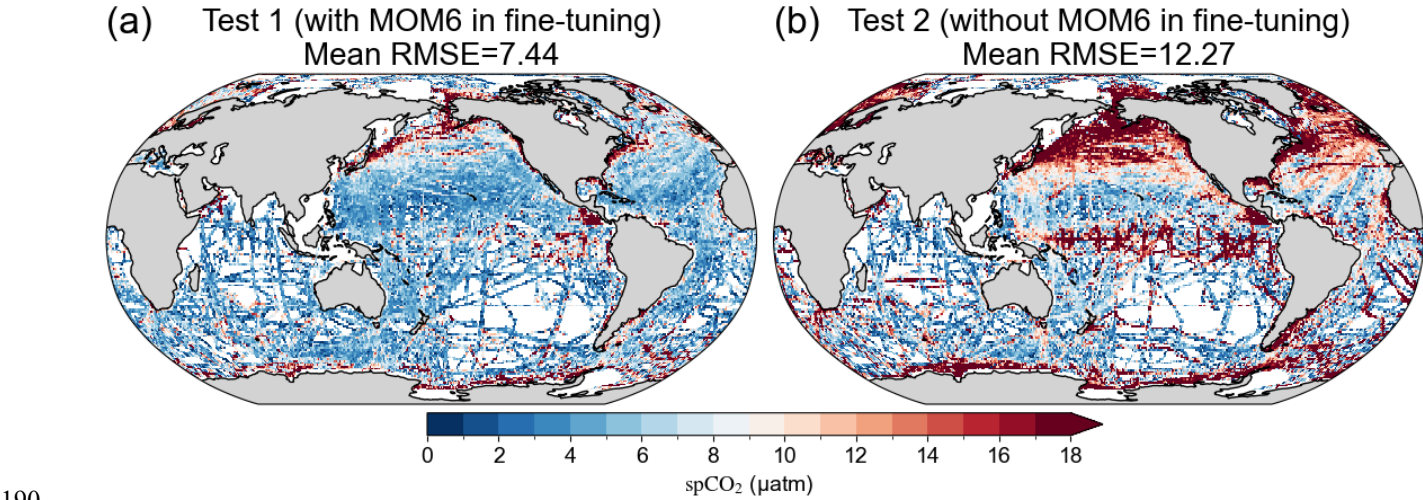
(a) Test 1 (with MOM6 in fine-tuning): The model was first pretrained on CMIP6 outputs and then fine-tuned using both MOM6 simulation outputs and SOCAT observations. MOM6 provides continuous, physically consistent global fields, while SOCAT supplies essential observational constraints.

(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 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 S10. Impact of MOM6 fine-tuning on reconstructed spCO₂ 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.

5.4 Ablation experiments on the contribution of different CMIP6 pre-training data combination

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.

(a) Test 1 (3-model CMIP6 pre-training)
Mean RMSE=10.48

(b) Test 2 (4-model CMIP6 pre-training)
Mean RMSE=9.54

spCO$_2$ (μatm)

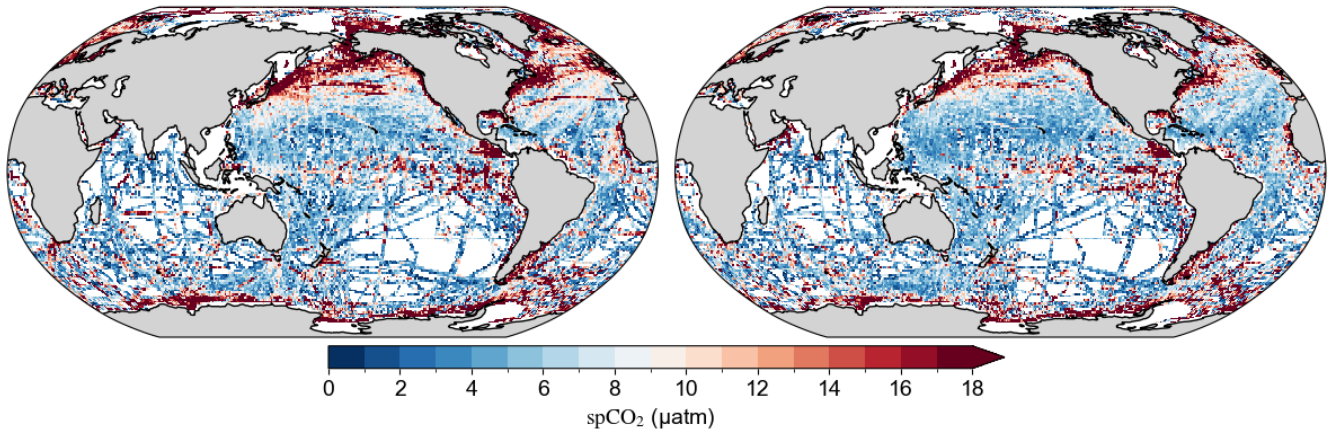**Figure S11. The sensitivity of reconstructed spCO$_2$ fields to the choice of CMIP6 models. (a) Test 1 (3-model CMIP6 pre-training): three CMIP6 simulations (GFDL-ESM4, NorESM2-LM, NorESM2-MM) pre-training followed by MOM6 & SOCAT fine-tuning; (b) Test 2 (4-model CMIP6 pre-training): four CMIP6 simulations (CESM2, CESM2-FV2, CESM2-WACCM, CESM2-WACCM-FV2) pre-training followed by MOM6 & SOCAT fine-tuning.**

5.5 Ablation experiments on the contribution of SOCAT fine-tuning

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.

(a) Test 1 (with SOCAT in fine-tuning)
Mean RMSE=7.44

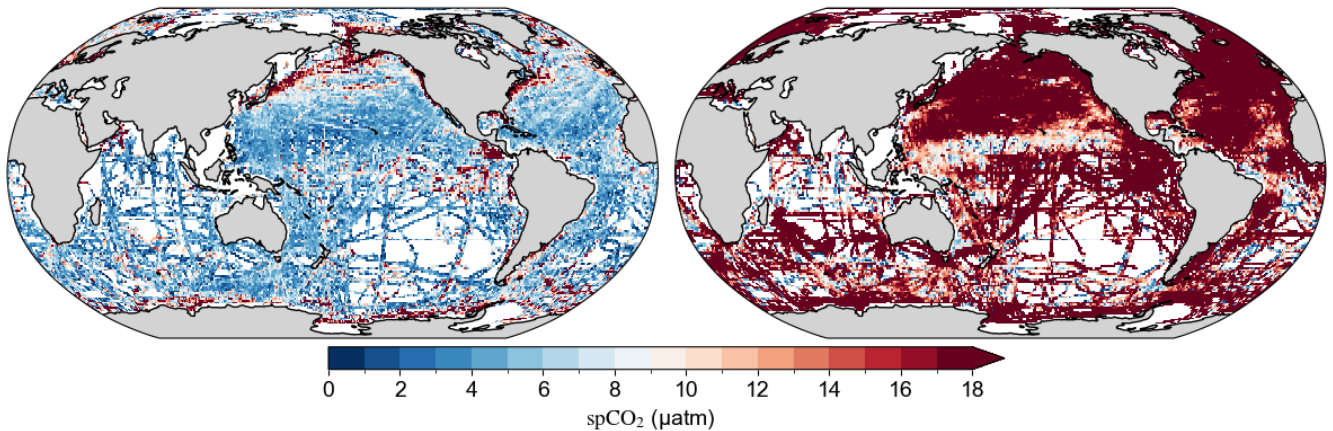**(b)** Test 2 (without SOCAT in fine-tuning)
Mean RMSE=26.87

spCO$_2$ (μatm)

Figure S12. 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.

5.6 Ablation experiments on the contribution of physical-biogeochemical Constraints on seasonal cycle

245 To assess the impact of MOM6-derived physical-biogeochemical constraints on the seasonal cycle of spCO₂, 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.

250 (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, 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

255 ~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 spCO₂ 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

260 demonstrate that including physical-biogeochemical constraints play a substantial and reliable role in improving the seasonal cycle representation of spCO₂, rather than merely introducing stochastic or localized enhancements.
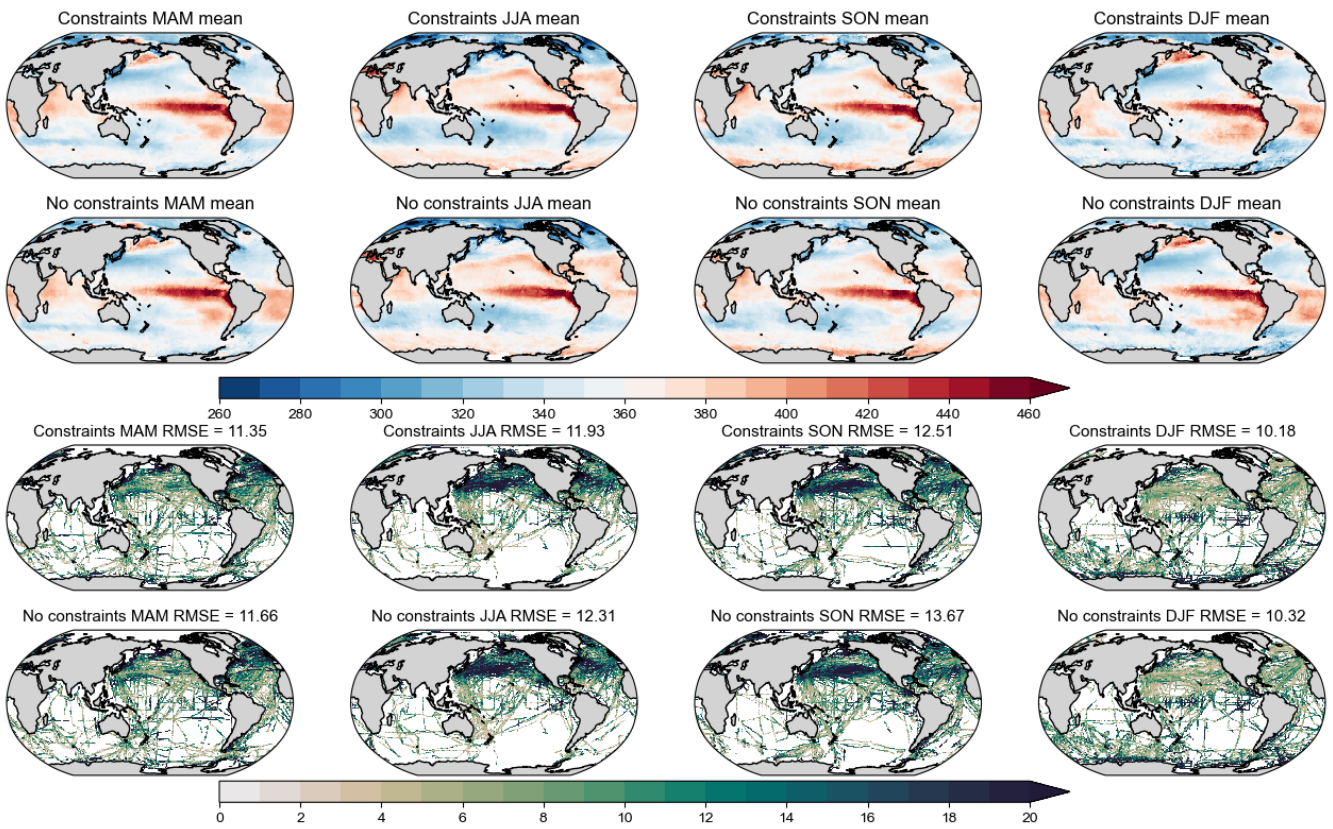
Figure S13. Seasonal comparison of SJTU-AViT spCO₂ means and RMSE with and without physical-biogeochemical constraints.
(a-d) Test 1 (with physical-biogeochemical constraints): seasonal mean spCO₂ from SJTU-AViT with physical-biogeochemical
265 **constraints for MAM (March-May), JJA (June-August), SON (September-November), and DJF (December-February). (e-h) Test 2**
(without physical-biogeochemical constraints): seasonal mean spCO₂ from SJTU-AViT without constraints. (i-l) Test 1 (with
physical-biogeochemical constraints): seasonal RMSE of spCO₂ between SJTU-AViT and SOCAT with constraints. (m-p) Test 2
(without physical-biogeochemical constraints): seasonal RMSE of spCO₂ between SJTU-AViT and SOCAT without constraints. For
RMSE calculations, SJTU-AViT spCO₂ was interpolated to SOCAT observation locations and times.

270 ## 5.7 Algorithm uncertainty assessment using synthetic data

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

275 learning model had never seen before. The SD data were divided into two subsets: (i) SD_SOCAT: SD outputs sampled at the spatiotemporal locations of SOCAT observations. (ii) SD_nonSOCAT: the remaining SD outputs.

Following our standard workflow (CMIP6 pretraining, MOM6 fine-tuning, and SD_SOCAT fine-tuning), we reconstructed spCO₂ 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 error 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.
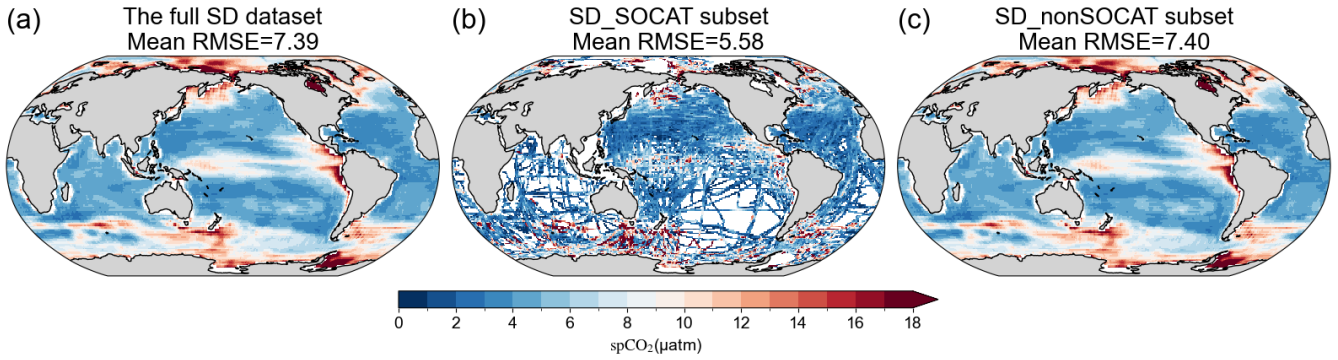
Figure S14. Spatial distribution of RMSE (µatm) between the reconstructed spCO₂ field and the Scott Doney RECCAP2 simulation (SD data). (a) RMSE for the full SD dataset. (b) RMSE for the SD_SOCAT subset, i.e., SD data sampled at SOCAT observation locations and used in training. (c) RMSE for the SD_nonSOCAT 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).

5.8 Independence validation of spCO₂ reconstruction

To address the concern regarding the potential lack of independence between the model and the validation data, we conducted an additional analysis using an independent reconstructed data product from MPI-SOM-FFN (Landschützer et al. 2016). Specifically, when calculating the detrended and deseasonalized SOCAT STD, we applied the long-term trends and seasonal cycles derived from the MPI-SOM-FFN data product instead of the SJTU-AViT estimates. The results, shown in the Fig. S15, demonstrate that the overall spatial distribution of SOCAT STD remains highly consistent, with only minimal deviations (1.68 µatm). This indicates that the small deviations observed between SJTU-AViT and SOCAT are not artifacts of model-data dependence. Therefore, the analysis confirms the robustness of our methodology and supports the credibility of the interannual variability assessment.
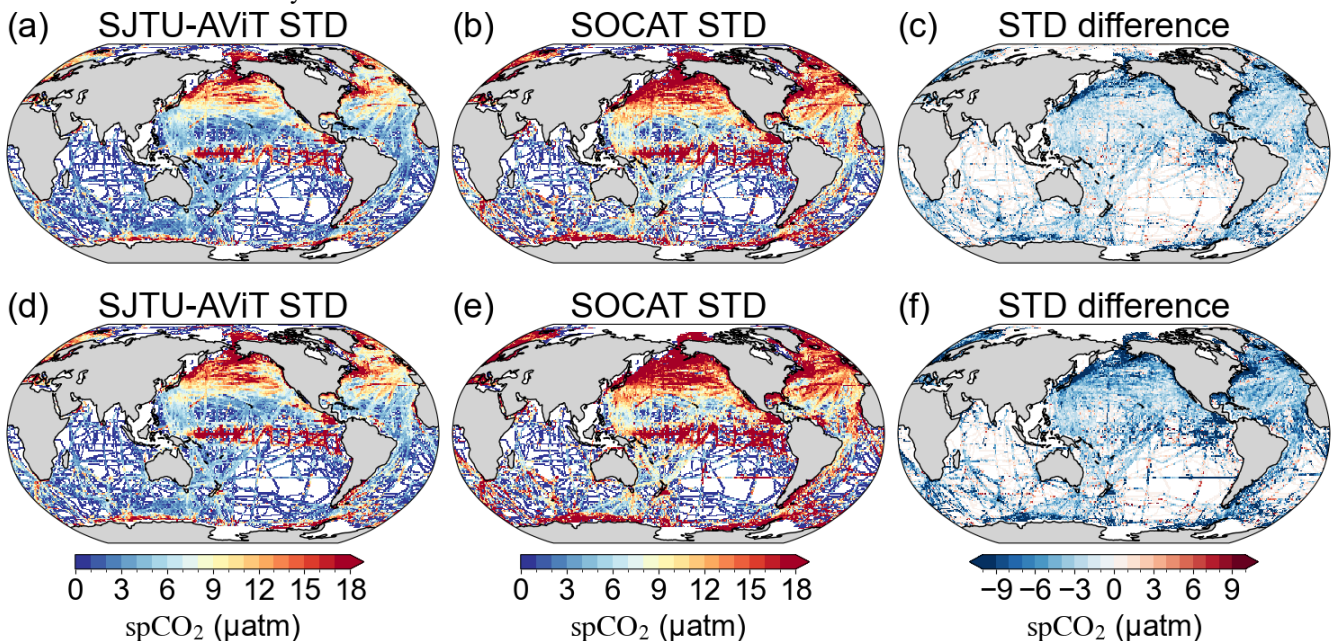
Figure S15. Comparison of spCO₂ standard deviations on timescales longer than one year between SJTU-AViT, SOCAT, and MPI-SOM-FFN data product. (a) Standard deviation of spCO₂ from the SJTU-AViT at SOCAT observation grid points. (b) Standard deviation of spCO₂ from SOCAT data (the long-term trends and seasonal cycles derived from the SJTU-AViT). (c) Standard deviation bias between SJTU-AViT and SOCAT (panel a minus panel b). (d) Standard deviation of spCO₂ from the

SJTU-AViT at SOCAT observation grid points. (e) Standard deviation of $spCO_2$ from SOCAT data (the long-term trends and seasonal cycles derived from the MPI-SOM-FFN). (f) Standard deviation bias between SJTU-AViT and SOCAT (panel d minus panel e).

310 **5.9 Validation of seasonal phase consistency with SOCAT observations**

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. 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
315 model well reproduces the seasonal cycle of $spCO_2$, with peak and minimum months largely consistent with SOCAT observations.

(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.
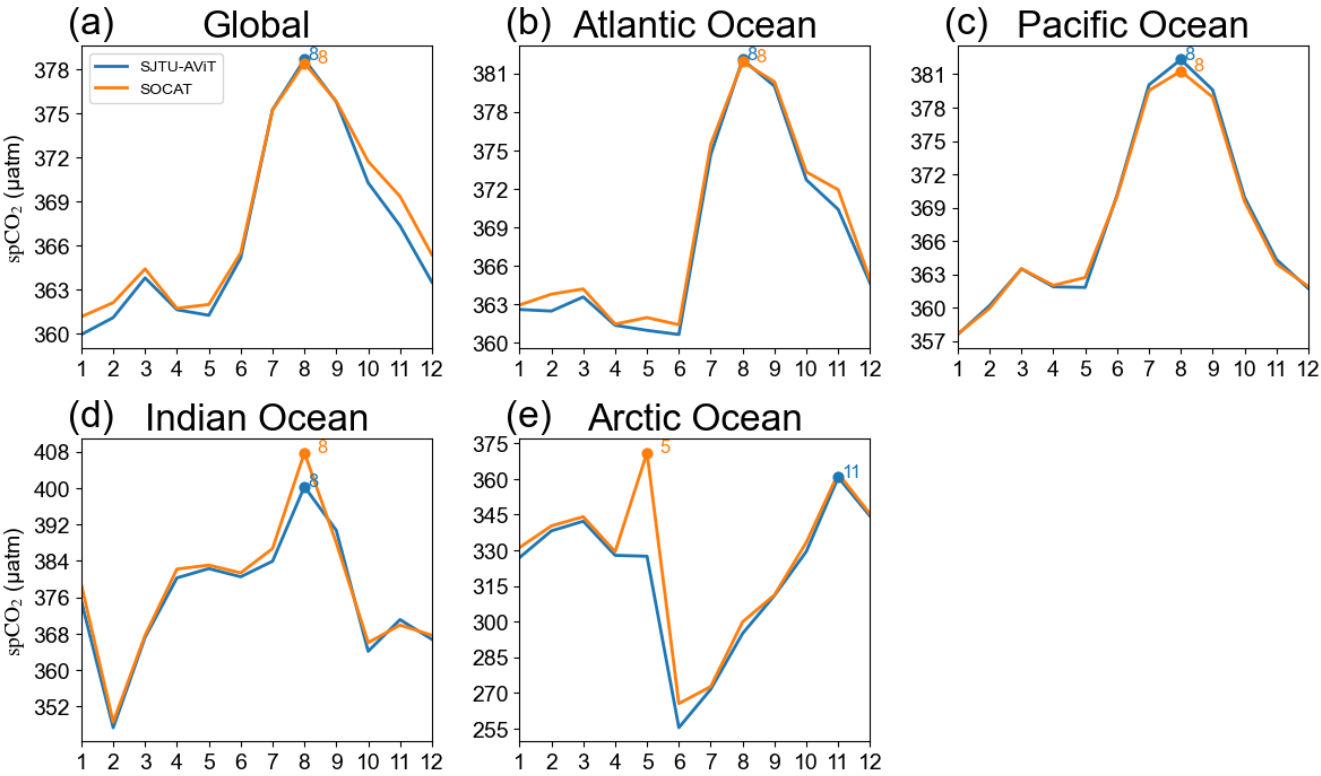

320

Figure S16. Monthly spCO₂ 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.
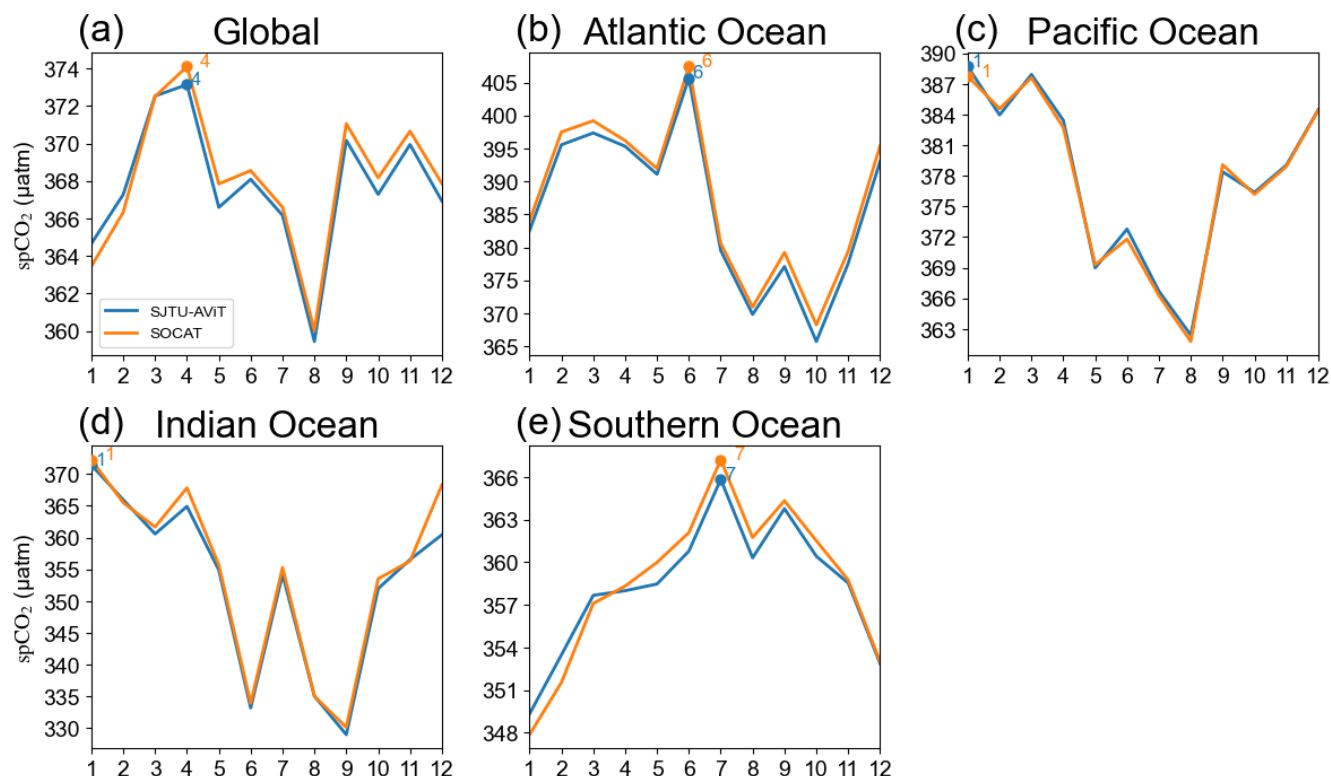
SH monthly climatology

Figure S17. 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.
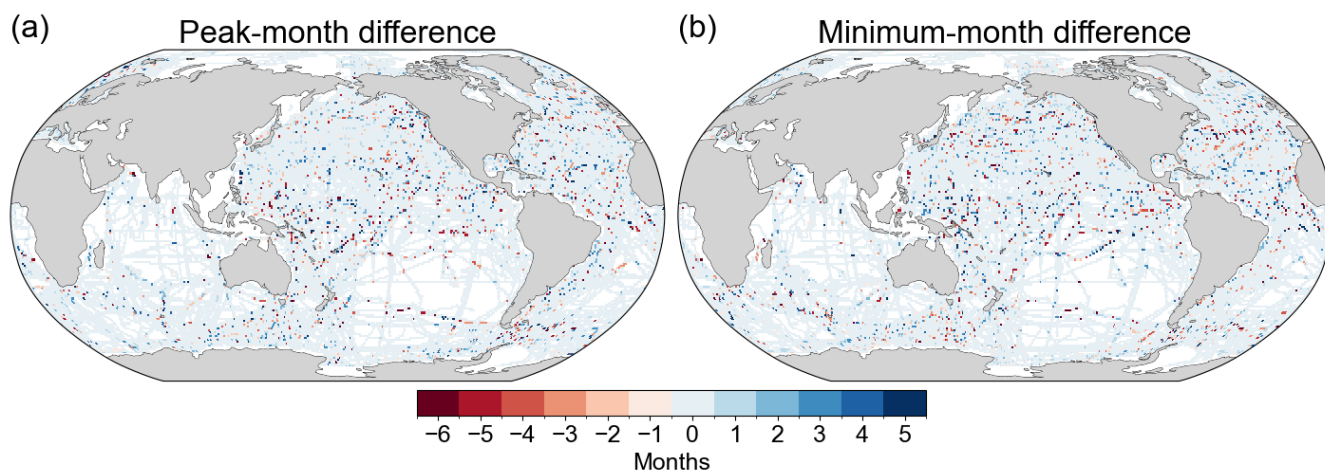

Figure S18. Grid-scale maps of spCO$_2$ peak- and minimum-month differences (SJTU-AViT − SOCAT, in months, range ±6). For the peak-month difference map, positive values indicate that SJTU-AViT peaks later than SOCAT; for the minimum-month difference map, positive values indicate that SJTU-AViT minimums 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.

Reference

Bentsen, M., et al.: NCC NorESM2-MM model output prepared for CMIP6 CMIP historical, edited, Earth System Grid Federation, 2019a.

Bentsen, M., et al.: NCC NorESM2-MM model output prepared for CMIP6 ScenarioMIP ssp245, edited, Earth System Grid Federation, 2019b.

Bentsen, M., et al.: NCC NorESM2-MM model output prepared for CMIP6 ScenarioMIP ssp585, edited, Earth System Grid Federation, 2019c.

Danabasoglu, G.: NCAR CESM2 model output prepared for CMIP6 CMIP historical, edited, Earth System Grid Federation, 2019a.

Danabasoglu, G.: NCAR CESM2 model output prepared for CMIP6 ScenarioMIP ssp245, edited, Earth System Grid Federation, 2019b.

Danabasoglu, G.: NCAR CESM2 model output prepared for CMIP6 ScenarioMIP ssp585, edited, Earth System Grid
Federation, 2019c

Danabasoglu, G.: NCAR CESM2-FV2 model output prepared for CMIP6 CMIP historical, edited, Earth System Grid
Federation, 2019d.

Danabasoglu, G.: NCAR CESM2-WACCM model output prepared for CMIP6 CMIP historical, edited, Earth System Grid
Federation, 2019e.

Danabasoglu, G.: NCAR CESM2-WACCM model output prepared for CMIP6 ScenarioMIP ssp245, edited, Earth System
Grid Federation, 2019f.

Danabasoglu, G.: NCAR CESM2-WACCM model output prepared for CMIP6 ScenarioMIP ssp585, edited, Earth System
Grid Federation, 2019g.

Danabasoglu, G.: NCAR CESM2-WACCM-FV2 model output prepared for CMIP6 ScenarioMIP ssp585, edited, Earth
System Grid Federation, 2019h.

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.

John, J. G., et al.: NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 ScenarioMIP ssp585, edited, Earth System
Grid Federation, 2018.

Landschützer, P., Gruber, N., and Bakker, D. C. E.: Decadal variations and trends of the global ocean carbon sink, Global
Biogeochem. Cycles, 30, 1396–1417, https://doi.org/10.1002/2015GB005359, 2016.

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.

Krasting, J. P., et al.: NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP historical, edited, Earth System
Grid Federation, 2018.

Seland, Ø., et al.: NCC NorESM2-LM model output prepared for CMIP6 CMIP historical, edited, Earth System Grid
Federation, 2019a.

Seland, Ø., et al.: NCC NorESM2-LM model output prepared for CMIP6 ScenarioMIP ssp245, edited, Earth System Grid
Federation, 2019b.

Seland, Ø., et al.: NCC NorESM2-LM model output prepared for CMIP6 ScenarioMIP ssp585, edited, Earth System Grid
Federation, 2019c.