Supplement of

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

Xueying Zhang[1], Enhui Liao[1]*, Wenfang Lu[2,3], Zelun Wu[4], Guansuo Wang[5,6], Xueming Zhu[3], Shiyu Liang[7]*

[1]School of Oceanography, Shanghai Jiao Tong University, Shanghai, 200240, China
[2]School of Marine Sciences, State Key Laboratory of Environmental Adaptability for Industrial Products, Sun Yat-sen University, Zhuhai, Guangdong, 519082, China
[3]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, Guangdong, 519082, China
[4]School of Marine Science and Policy, University of Delaware, Newark, Delaware, 19716, USA
[5]Observation and Research Station of Huaniaoshan East China Sea Ocean-Atmosphere Integrated Ecosystem, Ministry of Natural Resources, Shanghai, 200137, China
[6]East China Sea Forecasting and Hazard Mitigation Center, Ministry of Natural Resources, Shanghai, 200137, China
[7]John Hopcroft Center for Computer Science, Shanghai Jiao Tong University, Shanghai, 200240, China

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

Contents of this file

Section S1 Monte Carlo method for uncertainty estimation

Section S2 The data description and climate mode selection

Section S3 Table S1-S2

Section S4 Fig. S1-S8

Reference

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:

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

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 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 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), 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.

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 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 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

Section S4 Fig. S1-S8

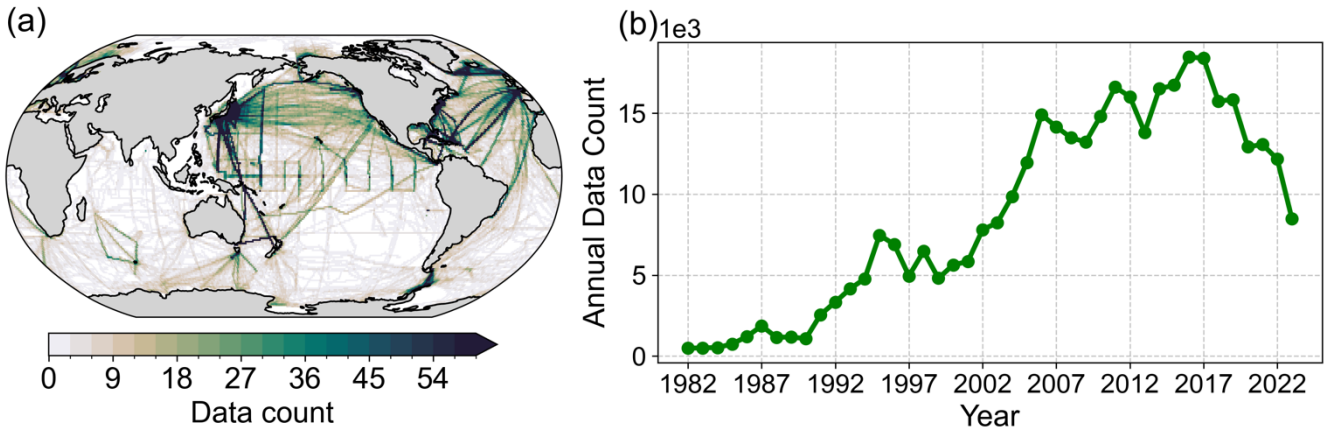

Figure S1. Data availability for $spCO_2$ reconstruction. (a) Spatial distribution of the number of $spCO_2$ data points. (b) Annual data count over the period from 1982 to 2023.

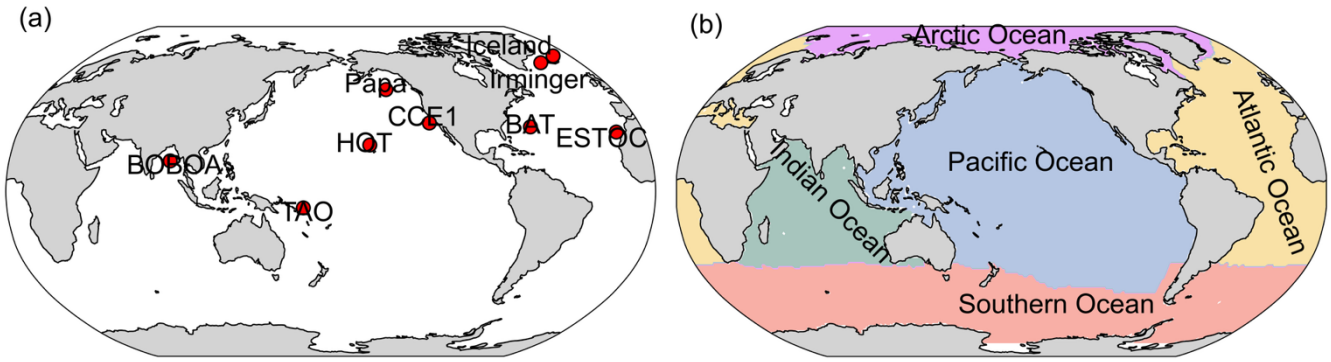

Figure S2. Spatial distribution of independent in situ observations and the definition of ocean basins used in this study.

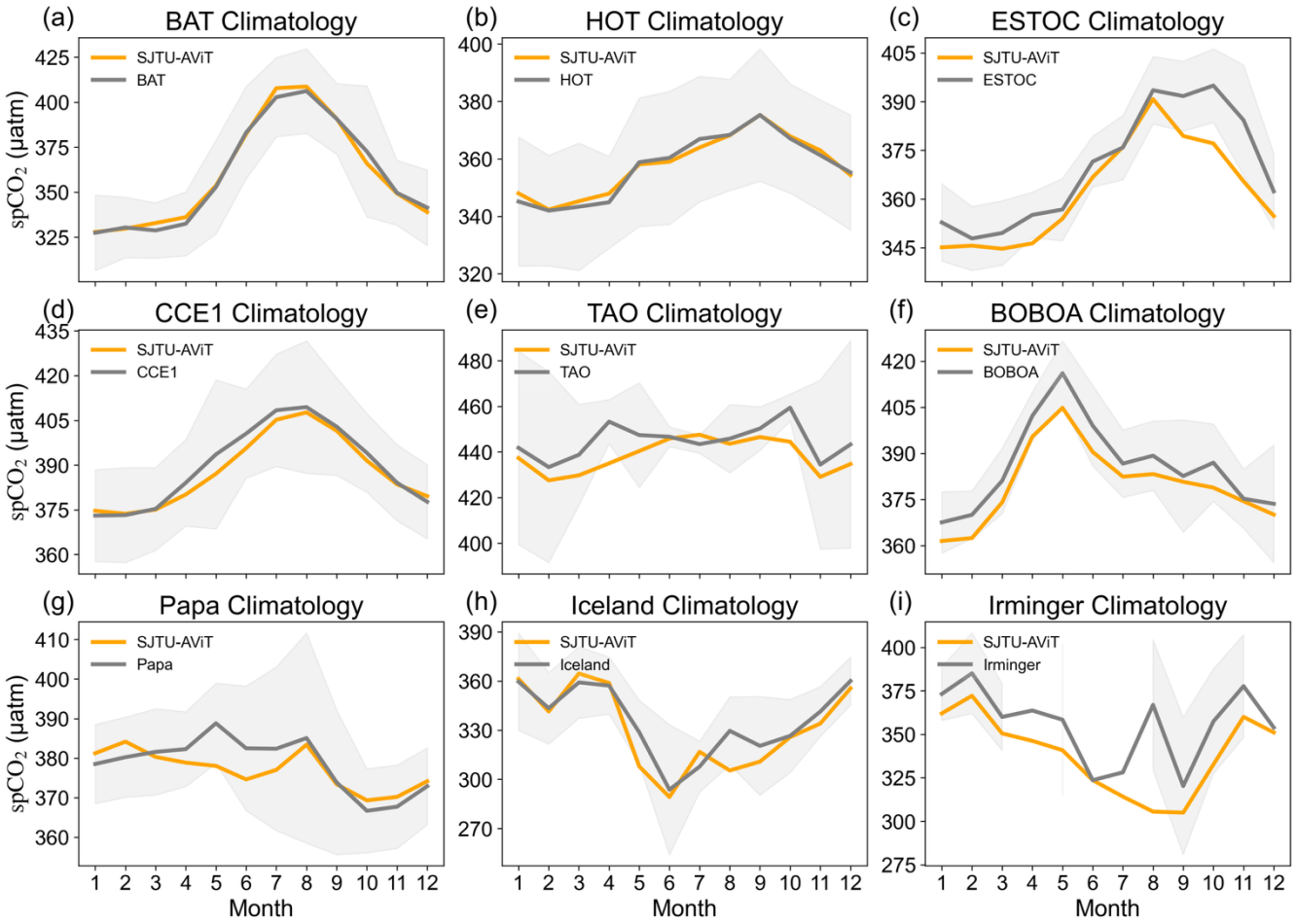

Figure S3. Independent test of seasonal cycles of $spCO_2$ climatology between SJTU-AViT and in situ observations. These in situ data are independent data and are not used to train the model. The station description and location refer to Supplement Section S2 and Fig. S2. The $spCO_2$ in SJTU-AViT is interpolated to match the station locations and times in the comparison. The lines represent the monthly mean $spCO_2$ 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 $spCO_2$ patterns at each station.

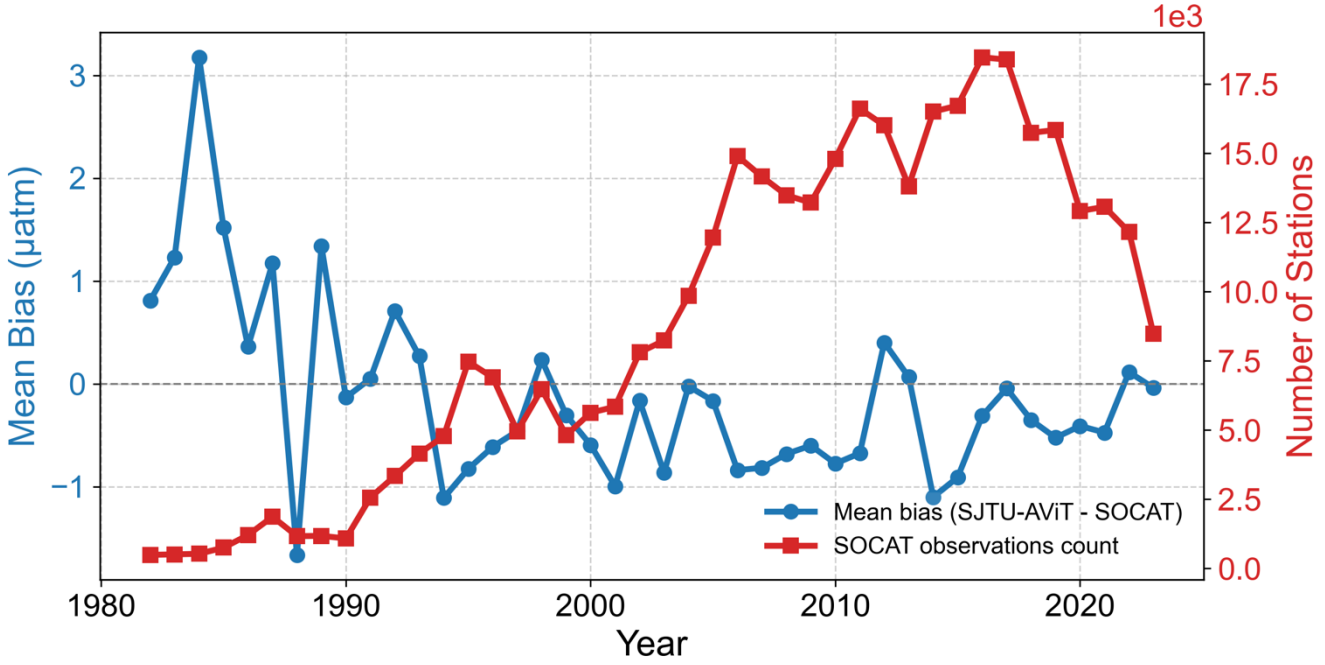

Figure S4. Temporal evolution of bias and SOCAT observation count from 1982 to 2023. The blue line represents the bias in long-term mean $spCO_2$ (SJTU-AViT minus SOCAT), while the red bars show the annual number of SOCAT observations contributing to the data. The increasing observation count over time correlates with a decrease in the mean bias, suggesting improvements in model performance as more observational data became available.

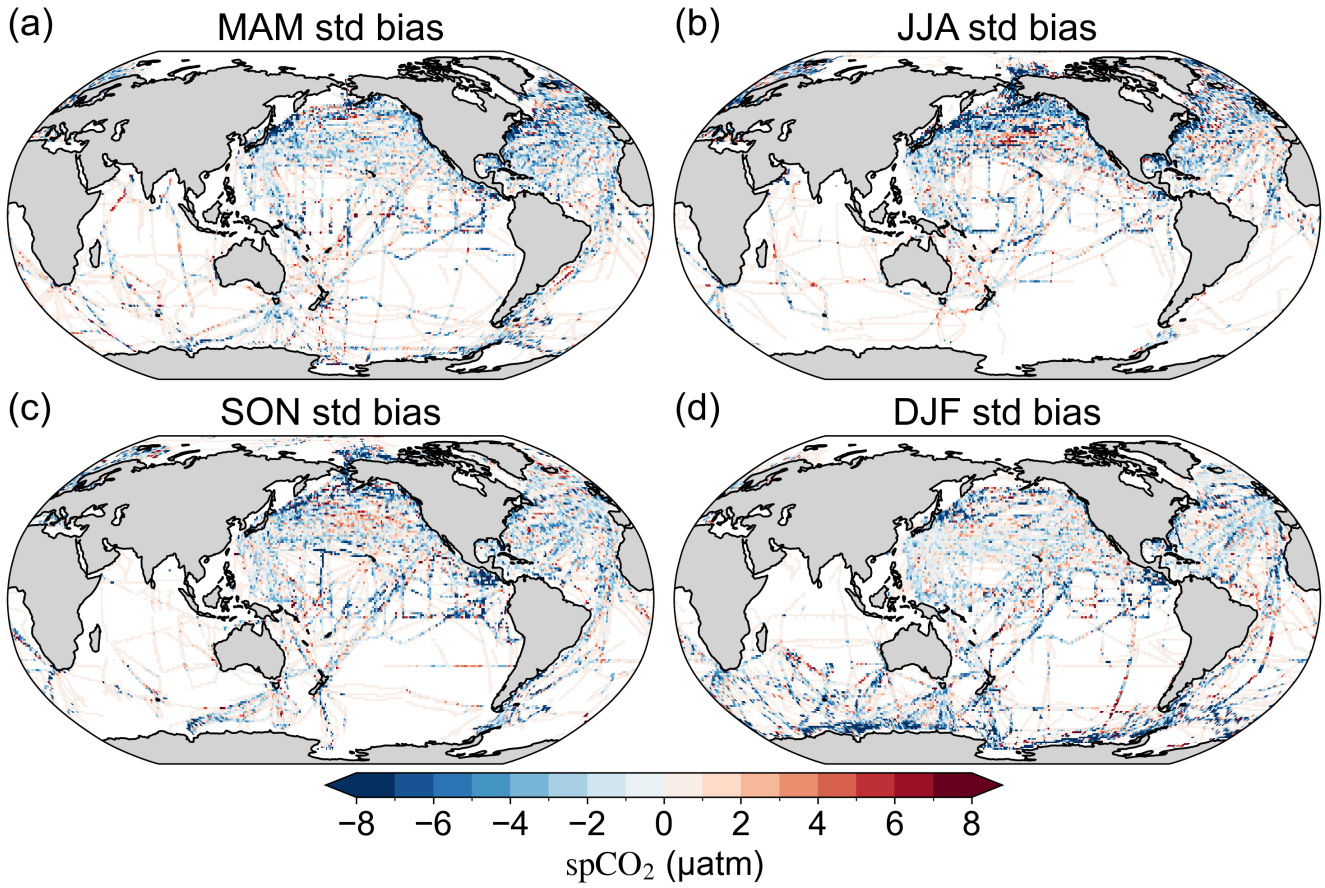

Figure S5. Bias in the standard deviation of $spCO_2$ between SJTU-AViT and SOCAT at each season from 1982 to 2023. (a) MAM (March–May), (b) JJA (June–August), (c) SON (September–November), and (d) DJF (December–February). The bias is calculated as the difference between SJTU-AViT and SOCAT standard 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 spCO2 in SJTU-AViT is interpolated to match the SOCAT observation locations and times in the comparison (see detailed computation in section 2.3).

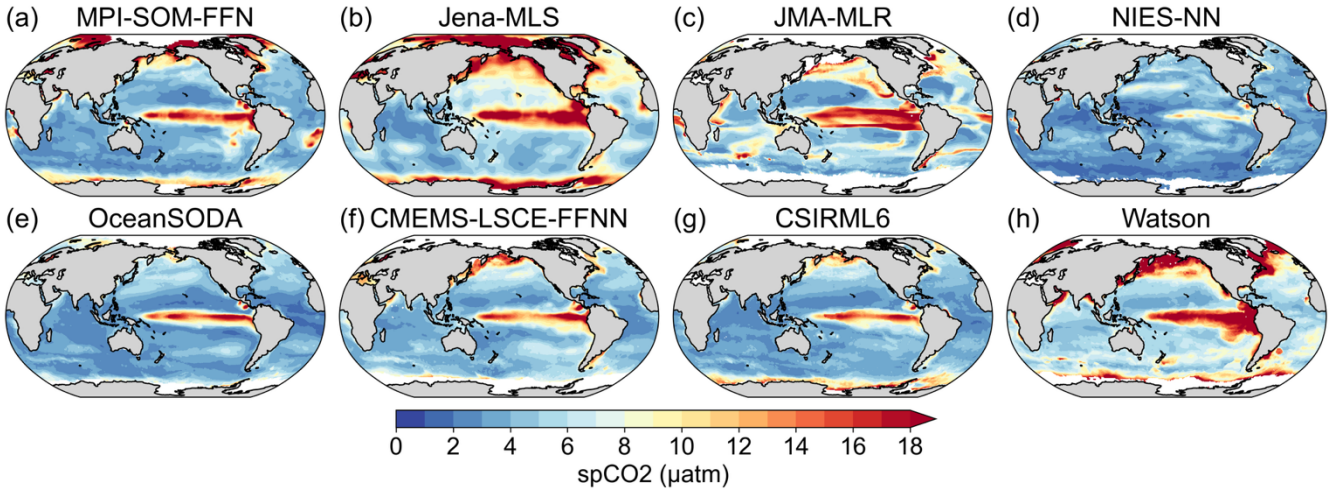

Figure S6. Spatial distribution of standard deviation in interannual time scale of reconstructed $spCO_2$ at multiple data products from 1985 to 2018. All the panels show the standard deviation of residuals after removing long-term trends and seasonal cycles. The color scale represents the magnitude of variability in $spCO_2$, with higher values (red) indicating greater variability.

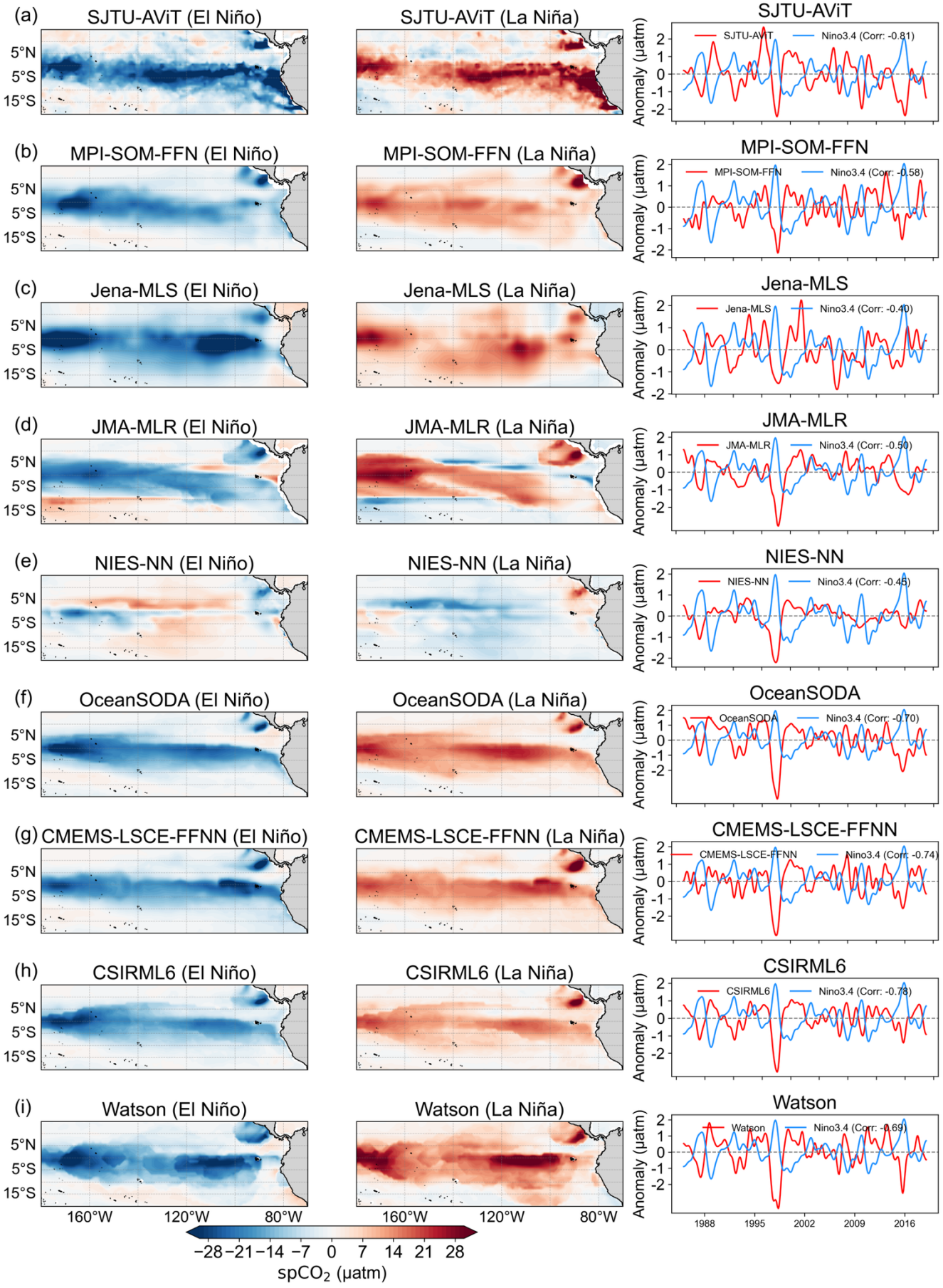

Figure S7. Spatial and temporal patterns of $spCO_2$ anomalies during ENSO events in the equatorial Pacific Ocean: comparison between SJTU-AViT and multiple data products. The left column shows composite spatial distribution of $spCO_2$ anomalies during eight El Niño events. The middle column shows composite spatial distribution of $spCO_2$ anomalies during seven La Niña events. The right column shows the time series of $spCO_2$ 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.

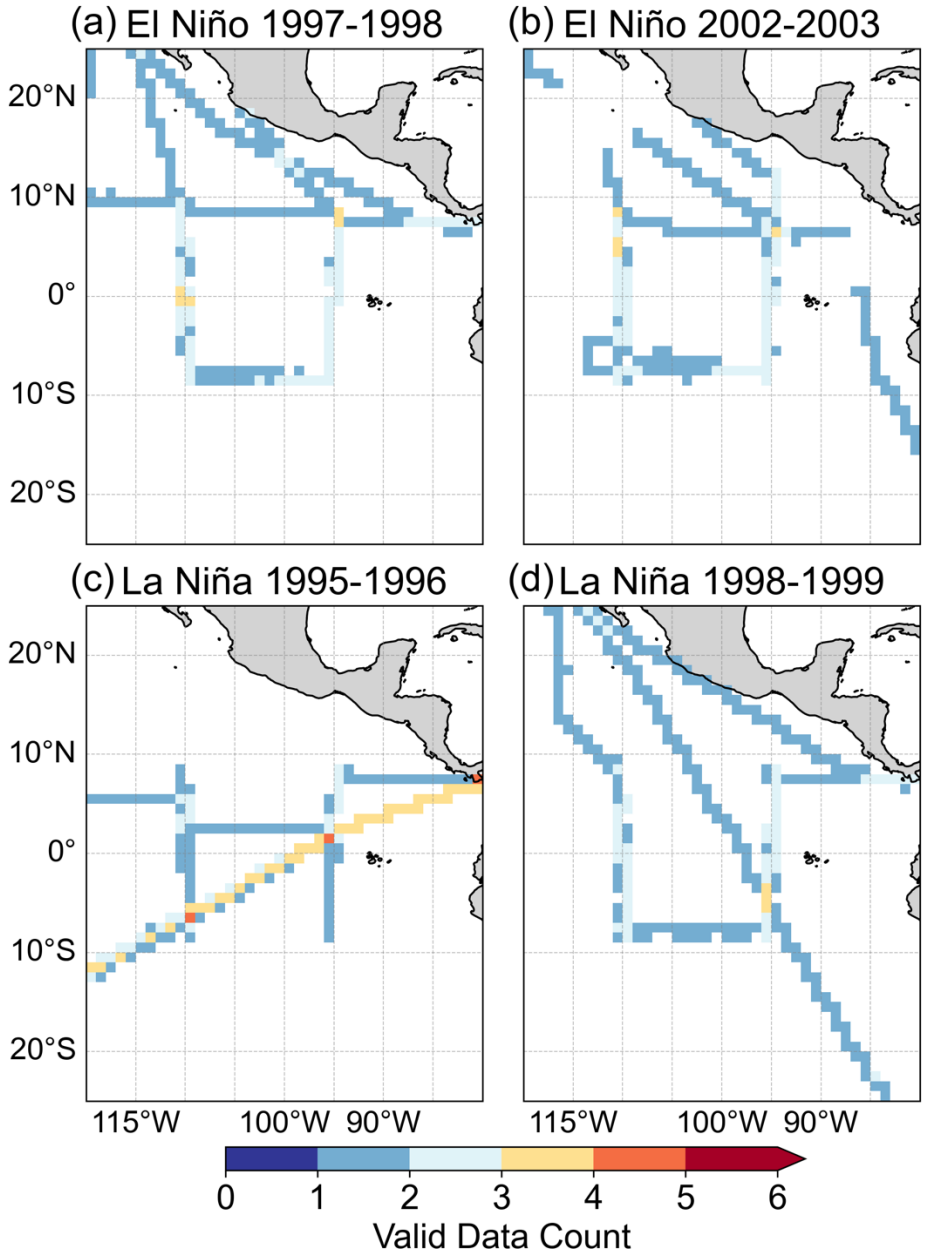

Figure S8. Spatial distribution of SOCAT $spCO_2$ observations in the Equatorial Pacific Ocean (240°E–280°E) during selected ENSO events. The color scale indicates the valid data count per 1°×1° grid cell during four distinct ENSO events: (a) El Niño 1997–1998, (b) El Niño 2002–2003, (c) La Niña 1995–1996, and (d) La Niña 1998–1999.

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