



# Spatial reconstruction of long-term (2003–2020) sea surface $p\text{CO}_2$ in the South China Sea using a machine-learning-based regression method aided by empirical orthogonal function analysis

Zhixuan Wang<sup>1,2</sup>, Guizhi Wang<sup>1,2,3</sup>, Xianghui Guo<sup>3</sup>, Yan Bai<sup>4</sup>, Yi Xu<sup>1,2</sup>, and Minhan Dai<sup>1,2</sup>

<sup>1</sup>State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen 361102, China

<sup>2</sup>College of Ocean and Earth Sciences, Xiamen University, Xiamen 361102, China

<sup>3</sup>Fujian Provincial Key Laboratory for Coastal Ecology and Environmental Studies, Xiamen University, Xiamen 361102, China

<sup>4</sup>State Key Laboratory of Satellite Ocean Environment Dynamics, Second Institute of Oceanography, State Oceanic Administration, Hangzhou 310012, China

**Correspondence:** Minhan Dai (mdai@xmu.edu.cn)

Received: 19 September 2022 – Discussion started: 4 October 2022

Revised: 8 March 2023 – Accepted: 13 March 2023 – Published: 17 April 2023

**Abstract.** The South China Sea (SCS) is the largest marginal sea of the North Pacific Ocean, where intensive field observations, including mappings of the sea surface partial pressure of  $\text{CO}_2$  ( $p\text{CO}_2$ ), have been conducted over the last 2 decades. It is one of the most studied marginal seas in terms of carbon cycling and could thus be a model system for marginal sea carbon research. However, the cruise-based sea surface  $p\text{CO}_2$  datasets are still temporally and spatially sparse. Using a machine-learning-based method facilitated by empirical orthogonal function (EOF) analysis, this study provides a reconstructed dataset of the monthly sea surface  $p\text{CO}_2$  in the SCS with a reasonably high spatial resolution ( $0.05^\circ \times 0.05^\circ$ ) and temporal coverage between 2003 and 2020. The data input to our model includes remote-sensing-derived sea surface salinity, sea surface temperature, and chlorophyll, the spatial pattern of  $p\text{CO}_2$  constrained by EOF, atmospheric  $p\text{CO}_2$ , and time labels (month). We validated our reconstruction with three independent testing datasets that are not involved in the model training. Among them, Test 1 includes 10 % of our in situ data, Test 2 contains four independent in situ datasets corresponding to the four seasons, and Test 3 is an in situ monthly dataset available from 2003–2019 at the South East Asia Time-series Study (SEATs) station located in the northern basin of the SCS. Our Test 1 validation demonstrated that the reconstructed  $p\text{CO}_2$  field successfully simulated the spatial and temporal patterns of sea surface  $p\text{CO}_2$  observations. The root mean square error (RMSE) between our reconstructed data and in situ data in Test 1 averaged  $\sim 10 \mu\text{atm}$ , which is much smaller (by  $\sim 50 \%$ ) than that between the remote-sensing-derived data and in situ data. Test 2 verified the accuracy of our retrieval algorithm in months lacking observations, showing a relatively small bias (RMSE of  $\sim 8 \mu\text{atm}$ ). Test 3 evaluated the accuracy of the reconstructed long-term trend, showing that, at the SEATs station, the difference between the reconstructed  $p\text{CO}_2$  and in situ data ranged from  $-10$  to  $4 \mu\text{atm}$  ( $-2.5 \%$  to  $1 \%$ ). In addition to the typical machine learning performance metrics, we assessed the uncertainty resulting from reconstruction bias and its feature sensitivity. These validations and uncertainty analyses strongly suggest that our reconstruction effectively captures the main spatial and temporal features of sea surface  $p\text{CO}_2$  distributions in the SCS. Using the reconstructed dataset, we show the long-term trends of sea surface  $p\text{CO}_2$  in five subregions of the SCS with differing physicobiogeochemical characteristics. We show that mesoscale processes such as the Pearl River plume and China coastal currents significantly impact sea surface  $p\text{CO}_2$  in the SCS during different seasons. While the SCS is overall a weak source of atmospheric

$\text{CO}_2$ , the northern SCS acts as a sink, showing a trend of increasing strength over the past 2 decades. The data used in this article are available at <https://doi.org/10.57760/sciencedb.02050> (Wang and Dai, 2022).

## 1 Introduction

The ocean possesses a large portion of the global capacity for atmospheric carbon dioxide ( $\text{CO}_2$ ) sequestration, annually mitigating 22 %–26 % of the anthropogenic  $\text{CO}_2$  emissions associated with fossil fuel burning and land use changes over the period from 2012–2021 (Friedlingstein et al., 2022). Ocean margins are an essential part of the land–ocean continuum, representing a particularly challenging regime to study (e.g., Chen and Borges, 2009; Dai et al., 2022; Laruelle et al., 2015), as they are often characterized by large spatial and temporal variations in air–sea  $\text{CO}_2$  fluxes that lead to larger uncertainties in their overall estimation and predictions than those made in the open ocean (Dai et al., 2013, 2022; Cao et al., 2020; Laruelle et al., 2015; Chen and Borges, 2009, and references therein). Limited spatiotemporal coverage of in situ observations is a large source of these uncertainties.

In recent years, many studies have used numerical models or data-based approaches to improve estimates of the partial pressure of carbon dioxide ( $p\text{CO}_2$ ) at the sea surface and the accuracy of the global carbon budget for periods and regions with poor coverage of in situ data (e.g., Rödenbeck et al., 2015; Wanninkhof et al., 2013). Numerical models can successfully quantify the generally increasing trend in oceanic  $p\text{CO}_2$  and simulate some critical carbon cycling processes (e.g., net ecosystem production) but still suffer from regional and seasonal differences in their estimates of ocean carbonate parameters (e.g., Luo et al., 2015; Mongwe et al., 2016; Tahata et al., 2015; Wanninkhof et al., 2013). Thus, data-based approaches, which typically apply statistical interpolation and regression methods, have become an important complement to numerical models (e.g., Jones et al., 2014; Lefèvre et al., 2005; Landschützer et al., 2014, 2017; Telszewski et al., 2009). Statistical interpolation improves the spatial coverage of in situ data but does not work for periods in which in situ data are unavailable. Regression methods allow the mapping of the relationships between in situ  $p\text{CO}_2$  data and other parameters that may drive changes in surface ocean  $p\text{CO}_2$ , and then the extrapolation of this relationship to improve estimates of the spatiotemporal distribution of  $p\text{CO}_2$ . Machine learning methods, and remote-sensing-derived products (as proxy variables in regression methods) have aided the development of data-based methods (Rödenbeck et al., 2015; Bakker et al., 2016) and can improve the model results for the oceanic carbonate system by numerical assimilation methods. Consequently, machine learning has increasingly become a routine approach for reconstructing sea surface  $p\text{CO}_2$  in open-ocean regimes (e.g., Zeng et al., 2017; Li et al., 2019); however, it remains challenging

to extend this method to ocean margins, which are more dynamic in both time and space.

The South China Sea (SCS) is the largest marginal sea of the North Pacific Ocean, with a surface area of  $3.5 \times 10^6 \text{ km}^2$ . Although extensive field observations of sea surface  $p\text{CO}_2$  have been conducted in the SCS over the past 2 decades, their spatial and temporal coverage is still limited with respect to coverage of different physicobiogeochemical domains and subseasonal timescales (e.g., Guo and Wong, 2015; Li et al., 2020; Zhai et al., 2005, 2013). Therefore, there is a strong need for improved surface water  $p\text{CO}_2$  coverage in the SCS to constrain air–sea  $\text{CO}_2$  fluxes and improve initial conditions of numerical models. Moreover, the reasonably high spatiotemporal resolution of  $p\text{CO}_2$  data can help identify the controlling factors of  $p\text{CO}_2$  changes in the SCS and reliably resolve long-term changes.

Zhu et al. (2009) presented an empirical approach to estimate sea surface  $p\text{CO}_2$  in the northern SCS using remote-sensing-derived (RS-derived) data, including sea surface temperature (SST) and chlorophyll *a* (Chl *a*). Their reconstructed  $p\text{CO}_2$  data were generally consistent with the in situ data. However, uncertainties remained large, primarily caused by limited in situ data from only two summer cruises in their study. Jo et al. (2012) developed a neural-network-based algorithm using SST and Chl *a* to estimate sea surface  $p\text{CO}_2$  in the northern SCS. In their study, in situ sea surface  $p\text{CO}_2$  data were collected from three cruises during May 2001 and February and July 2004. The reconstruction also suffered a relatively large bias (Wang et al., 2021). Bai et al. (2015) employed a mechanic semi-analytical algorithm (MeSAA) to estimate satellite remote-sensing-derived sea surface  $p\text{CO}_2$  in the East China Sea from 2000–2014 and then expanded the application of this algorithm to estimate sea surface  $p\text{CO}_2$  for the whole China seas region, including the South China Sea. These authors explained that their MeSAA did not fully account for some localized processes, which resulted in a RMSE of about  $45 \mu\text{atm}$  for the SCS (Wang et al., 2021). Yu et al. (2022) subsequently used a non-linear regression method to develop a retrieval algorithm for seawater  $p\text{CO}_2$  in the China seas, and the RS-derived  $p\text{CO}_2$  data from 2003–2018 were provided by the Sat $\text{CO}_2$  platform (<http://www.SatCO2.com>, last access: 8 October 2022). In this retrieval algorithm, the input parameters included sea surface temperature, Chl *a* concentrations, remote sensing reflectance at three bands (Rrs412, 443, and 488 nm), the temperature anomaly in the longitudinal direction, and the theoretical thermodynamic background  $p\text{CO}_2$  under the corresponding SST. Although the RMSE associated with the RS-derived  $p\text{CO}_2$  product was relatively large ( $21.1 \mu\text{atm}$ ),

it successfully showed the major spatial patterns of sea surface  $p\text{CO}_2$  in the China seas (Yu et al., 2022).

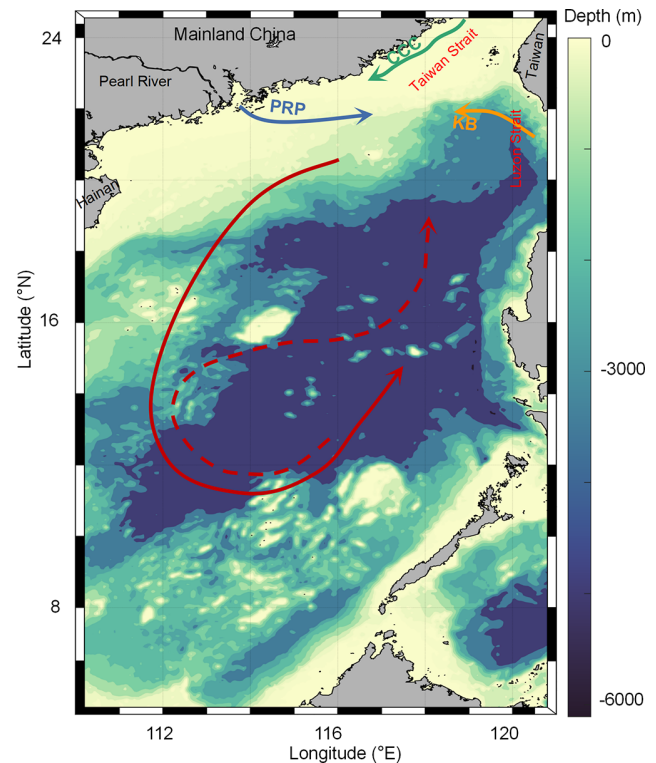
To take advantage of both the high spatiotemporal resolution of the RS-derived  $p\text{CO}_2$  data and the accuracy of the in situ data, Wang et al. (2021) reconstructed a basin-scale sea surface  $p\text{CO}_2$  dataset in the SCS during summer using an empirical orthogonal function (EOF) based on a multilinear regression method. They demonstrated that the spatial modes of RS-derived data calculated using the EOF can effectively provide spatial constraints on the data reconstruction, and thus, this approach is adopted in this study. However, the reconstructed results may still be subject to bias when the standard deviation of spatial in situ data is relatively large because of the influence of outliers (Wang et al., 2021). Therefore, many studies have used machine-learning-based regression methods to reduce the influence of outliers in open-ocean areas and have achieved a RMSE of  $< 17 \mu\text{atm}$  in most cases (e.g., Zeng et al., 2017; Li et al., 2019).

Building on the ability of the EOF method to significantly improve reconstructions in terms of spatial patterns and accuracy (Wang et al., 2021), we developed a machine-learning-based regression method facilitated by the EOF to fully resolve the long-term spatial distribution of sea surface  $p\text{CO}_2$  at a resolution of  $0.05^\circ \times 0.05^\circ$  in the SCS. Our reconstructed model uses input data that include remote-sensing-derived sea surface salinity, sea surface temperature, and Chl  $a$ , the spatial pattern of  $p\text{CO}_2$  constrained by the EOF, atmospheric  $p\text{CO}_2$ , and time labels (month). In addition to assessing typical machine learning performance metrics, we evaluated the uncertainty resulting from the bias of the reconstruction and its sensitivity to the features.

## 2 Study site and data sources

### 2.1 Study area

The SCS, located in the northwestern Pacific, is a semi-enclosed marginal sea with a maximum water depth of ca. 4700 m (e.g., Gan et al., 2006, 2010). The rhombus-shaped deep-water basin, with a southwesterly–northeasterly direction, accounts for about half of the total area of the SCS (Fig. 1). Largely modulated by the Asian monsoon and topography, the SCS exhibits seasonally varying surface circulation, river inputs, and upwelling. The circulation of the upper layer shows a large cyclonic circulation structure in winter (Fig. 1), while in summer it exhibits an anticyclonic circulation structure (Fig. 1; Hu et al., 2010). In the northern SCS, the Pearl River discharges into the SCS with an annual freshwater input of  $3.26 \times 10^{11} \text{ m}^3$  (e.g., Dong et al., 2004; Dai et al., 2014). The area influenced by the Pearl River plume may extend southeastward to a few hundred kilometers from the estuary in summer because of the monsoonal wind stress (Dai et al., 2014). The northern and western coastal regions of the SCS feature summer coastal upwelling, such as the eastern Guangdong and Qiongdong upwelling systems in



**Figure 1.** Topographic map of the South China Sea (SCS) showing the basin-wide cyclonic circulation in winter (solid line) and anticyclonic circulation over the southern half of the SCS in summer (dashed line). Also shown are the Kuroshio branch (KB; orange line), the China coastal current (CCC; green line), and the Pearl River plume (PRP; blue line).

the northern SCS and the Vietnam upwelling systems in the western SCS (e.g., Cao et al., 2011; Chen et al., 2012; Gan et al., 2006, 2010; Li et al., 2020). These seasonal changes in sea surface circulation lead to strong seasonal characteristics of sea surface  $p\text{CO}_2$  in the SCS.

The SCS is subject to dynamic water exchanges with the East China Sea via the Taiwan Strait and the western Pacific via the Luzon Strait (Fig. 1). In winter, driven by the winter monsoon, the China coastal current (CCC; green line in Fig. 1; Han et al., 2013; Yang et al., 2021) flows south along the Chinese mainland through the Taiwan Strait, and occupies the northern SCS with cold, fresh, nutrient-rich waters. The strong northeasterly winds in winter also slow down the western boundary ocean current, forcing the intrusion of Kuroshio water, featuring high surface salinity and high total alkalinity, into the SCS via the Luzon Strait (orange line in Fig. 1; Du et al., 2013; Park, 2013; Yang et al., 2021). These water exchange processes increase the complexity of the spatial distribution of sea surface  $p\text{CO}_2$  in the SCS, which, as a result, has strong seasonal characteristics and spatial variability.

## 2.2 Observational $p\text{CO}_2$ data

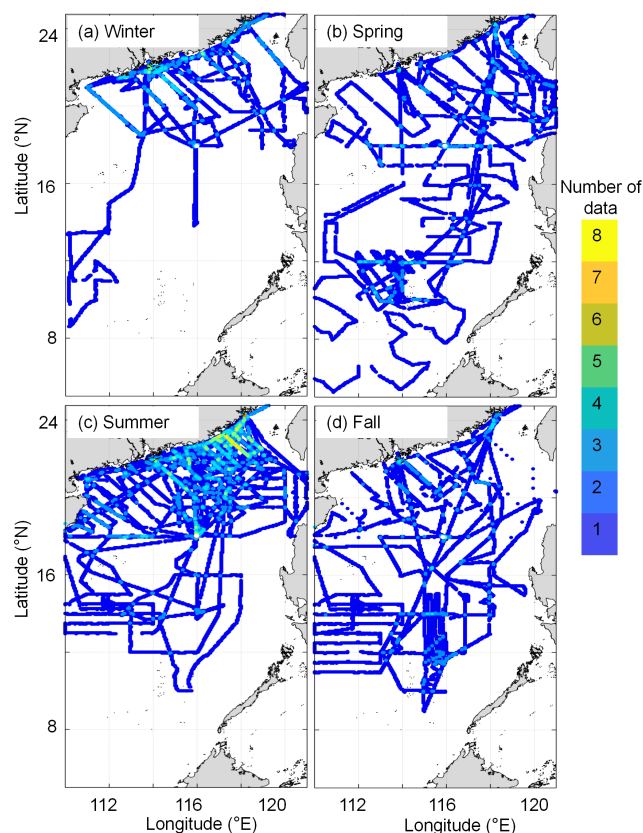
Data collected from field surveys during the study period 2003–2020 are summarized in Table 1. Most observations were made in July, with fewer observations made in March and December of each year. The rough sea state in the SCS in winter and early spring limited the field surveys during these seasons. Data collected from July 2000 to January 2018 were originally published in Li et al. (2020). The in situ  $p\text{CO}_2$  were collected from R/Vs *Dongfanghong-2* and *Tan Kah Kee* (TKK) (shown in Table 1). During the cruises, sea surface  $p\text{CO}_2$  was measured during the cruise. The measurements and data processing followed the SOCAT (Surface Ocean  $\text{CO}_2$  Atlas) protocol (Li et al., 2020). More details of the data collection methods are provided in Li et al. (2020). The spatial coverage and frequency of the observations are shown in Fig. 2, revealing pronounced seasonal changes across a large spatial area. For example, the spatial coverage of the in situ data in spring and fall are relatively uniformly distributed, and the south end of the spatial coverage reaches  $5^\circ\text{N}$  in spring, whereas during other seasons the data are concentrated in the northern and central regions of the SCS. In addition, only one observation was made in the basin area in winter, while the northern coastal area was more frequently surveyed, especially in summer.

Figure 3 shows the spatial and temporal distributions of in situ sea surface  $p\text{CO}_2$ . Seasonally, the lowest  $p\text{CO}_2$  occurs in January, and the highest concentrations occur in May and June. Spatially, the  $p\text{CO}_2$  distribution in the basin is relatively homogeneous, although is highly variable in the northern region. In the northern coastal area in summer, the  $p\text{CO}_2$  distribution is affected by the Pearl River plume (yielding low values) and coastal upwelling (yielding high values), which last into early fall. In winter and early spring, relatively low  $p\text{CO}_2$  values ( $\sim 350\ \mu\text{atm}$ ) were found in the near-shore area. In addition, the high  $p\text{CO}_2$  values recorded on the western side of the Luzon Strait in December demonstrate the influence of winter upwelling during some of the surveys.

In addition to the above in situ sea surface  $p\text{CO}_2$  data, we selected in situ sea surface  $p\text{CO}_2$  data collected during four independent surveys across the four seasons in September 2018 (fall), December 2018 (winter), August 2019 (summer), and April 2020 (spring) to verify the accuracy of our reconstruction model in extrapolating periods lacking training datasets. Furthermore, we used an additional dataset of sea surface  $p\text{CO}_2$  calculated from observed dissolved inorganic carbon and total alkalinity during 2003–2019 at the Southeast Asia Time-series Study (SEATs) station (data from Dai et al., 2022) to test the long-term consistency of the reconstruction.

## 2.3 Remote-sensing-derived sea surface $p\text{CO}_2$ data

The gridded ( $0.05^\circ \times 0.05^\circ$ ) RS-derived  $p\text{CO}_2$  data cover almost the entire SCS ( $5\text{--}25^\circ\text{N}$ ,  $109\text{--}122^\circ\text{E}$ ) and show ma-



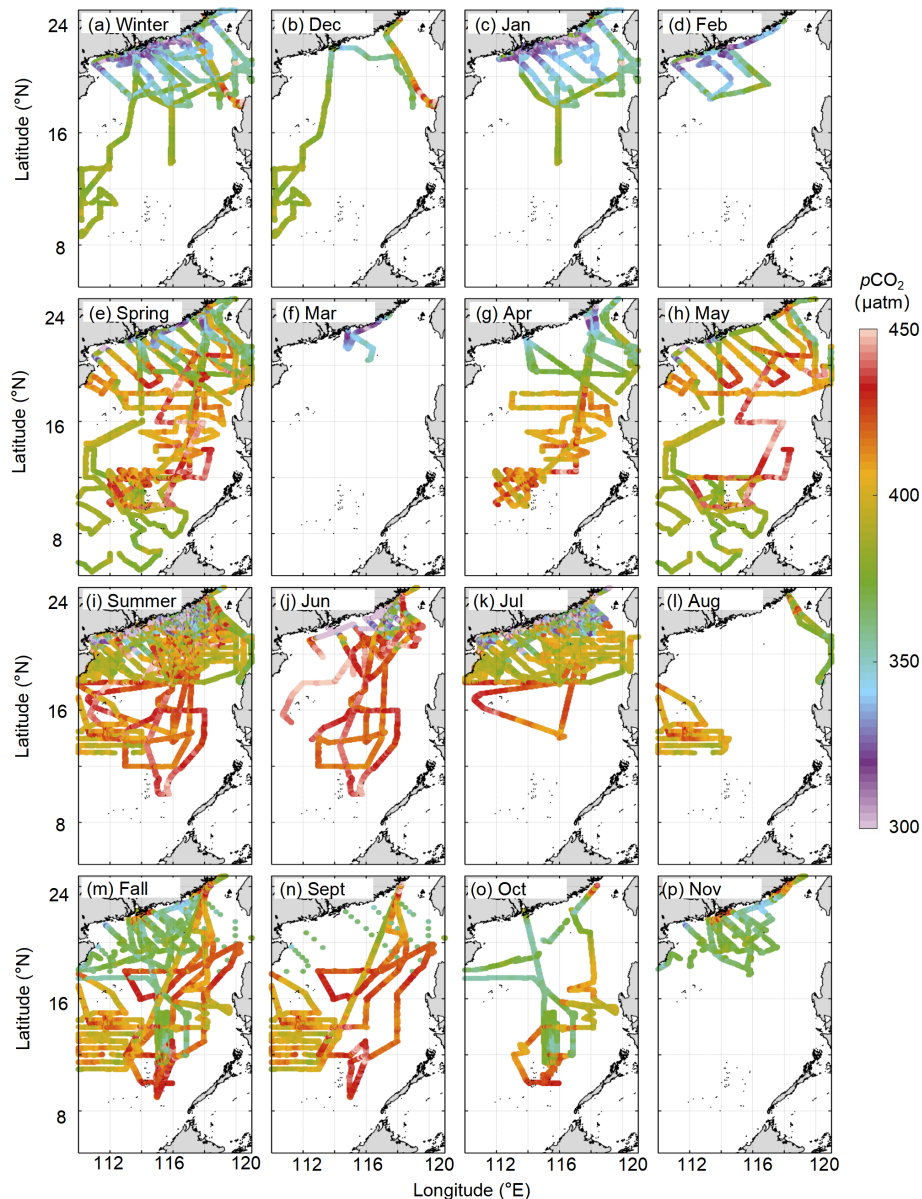
**Figure 2.** Cruise tracks of the observations conducted in the South China Sea in each season from 2000 to 2020. (a) Winter, (b) spring, (c) summer, and (d) fall are shown. The data collected before February 2018 are from Li et al. (2020), except for those collected in July 2015 and June 2017.

jor variations in sea surface  $p\text{CO}_2$  at the basin scale (Wang et al., 2021; Yu et al., 2022). Further details of the RS-derived  $p\text{CO}_2$  data can be found on the Sat $\text{CO}_2$  platform (<http://www.SatCO2.com>).

A grid-to-grid comparison was undertaken between the RS-derived  $p\text{CO}_2$  and the in situ  $p\text{CO}_2$  data (Table 2). The differences in between range from 35 to  $120\ \mu\text{atm}$  in the near-shore area. The largest biases occur in summer when the RMSE is up to  $29.95\ \mu\text{atm}$  (Table 2). Relatively large discrepancies may reflect the limitations of the current algorithm (MeSAA and nonlinear regression), which only considers biological processes and the turbidity induced by the Pearl River discharge (characterized by  $\text{Chl } a$  and the remote sensing reflectance at  $555\ \text{nm}$  ( $R_{rs555}$ ) and does not take into account the riverine dissolved inorganic carbon and the input of other substances that may affect  $p\text{CO}_2$  (Bai et al., 2015; Yu et al., 2022; Wang et al., 2021).

To remove the influence of the bias in RS-derived  $p\text{CO}_2$  data on our reconstructed results, this study used the EOF method to compute the spatial patterns of the RS-derived  $p\text{CO}_2$  data as input data instead of directly using the RS-





**Figure 3.** Seasonal and monthly sea surface  $p\text{CO}_2$  fields in the South China Sea. (a) Winter. (b) December, (c) January. (d) February. (e) Spring. (f) March. (g) April. (h) May. (i) Summer. (j) June. (k) July. (l) August. (m) Fall. (n) September. (o) October. (p) November. The data sources are given in Table 1.

derived  $p\text{CO}_2$  data. Moreover, using EOF modes of the RS-derived  $p\text{CO}_2$  as input data in the reconstructed model can provide spatial constraints on the  $p\text{CO}_2$  reconstruction.

#### 2.4 Other data

The RS-derived SST data produced by MODIS (Moderate Resolution Imaging Spectroradiometer; <https://oceancolor.gsfc.nasa.gov/>, last access: 8 October 2022) are adopted in our reconstruction. The uncertainty in this dataset in the SCS is  $\sim 0.27^\circ$  (Qin et al., 2014). For sea surface salinity (SSS) data, Wang et al. (2022) found rela-

tively large differences between different open-source SSS databases (i.e., multi-satellite fusion data from <https://podaac.jpl.nasa.gov/>, last access: 8 October 2022; model data from <https://climatedataguide.ucar.edu/>, last access: 8 October 2022; multidimensional covariance model data from <https://resources.marine.copernicus.eu/>, last access: 8 October 2022) and the in situ SSS data. Thus, Wang et al. (2022) produced an RS-derived SSS database using machine learning methods based on the MODIS Aqua remote sensing data. The bias between the RS-derived SSS (Wang et al., 2022) and in situ data was near zero (mean absolute error, MAE,

**Table 1.** Summary of seasonal in situ data of sea surface  $p\text{CO}_2$  in the South China Sea for the period 2003–2020 used in this study.

Season	Spring			Summer		
	March	April	May	June	July	August
Cruise time	Mar 2004	Apr 2005	May 2004	Jun 2006	Jul 2004	Aug 2007
		Apr 2008	May 2011	Jun 2016	Jul 2005	Aug 2008
		Apr 2009	May 2014	Jun 2017*	Jul 2007	Aug 2019*
		Apr 2012	May 2020*	Jun 2019*	Jul 2008	
		Apr 2020*		Jun 2020*	Jul 2009	
					Jul 2012	
Cruise time					Jul 2015*	
					Jul 2019*	
Season	Fall			Winter		
	September	October	November	December	January	February
Cruise time	Sep 2004	Oct 2003	Nov 2006	Dec 2006	Jan 2009	Feb 2004
	Sep 2007	Oct 2006	Nov 2010		Jan 2010	Feb 2006
	Sep 2008				Jan 2018	
	Sep 2020*					
Data sources	Li et al. (2020) * This study					

**Table 2.** Biases between the seasonal remote-sensing-derived  $p\text{CO}_2$  data and in situ  $p\text{CO}_2$  data and between the reconstructed and the in situ  $p\text{CO}_2$  data ( $\mu\text{atm}$ ). The remote-sensing-derived  $p\text{CO}_2$  data during 2003–2019 are from <http://www.SatCO2.com>, and the source of the in situ data can be found in Table 1. The reconstructed  $p\text{CO}_2$  data are from Sect. 3; all data were gridded into  $0.05^\circ \times 0.05^\circ$ ; the slash (/) means no data). MAE is the mean absolute error. RMSE is the root mean square error.  $R^2$  is the coefficient of determination. MAPE is the mean absolute percentage error.

		RS-derived $p\text{CO}_2$ data	Training data	Testing data I	Testing data II	Testing data III
Spring	MAE	9.00	2.44	4.76	1.68	/
	RMSE	12.70	3.47	7.43	2.26	/
	$R^2$	/	0.98	0.92	/	/
	MAPE	/	0.01	0.01	/	/
Summer	MAE	16.75	2.48	8.46	5.73	/
	RMSE	29.95	3.54	14.69	15.18	/
	$R^2$	/	0.99	0.89	/	/
	MAPE	/	0.01	0.02	/	/
Fall	MAE	9.93	2.41	4.90	7.133	/
	RMSE	13.08	3.39	6.85	8.94	/
	$R^2$	/	0.98	0.92	/	/
	MAPE	/	0.01	0.01	/	/
Winter	MAE	9.25	2.18	5.61	11.41	/
	RMSE	14.26	3.14	8.82	12.63	/
	$R^2$	/	0.98	0.89	/	/
	MAPE	/	0.01	0.01	/	/
Annual	MAE	11.95	2.41	6.30	5.27	6.19
	RMSE	20.66	3.43	10.79	11.18	8.26
	$R^2$	/	0.99	0.91	/	/
	MAPE	/	0.01	0.01	/	/

of  $\sim 0.25$ ). Next, we used Chl  $a$  (from <https://oceancolor.gsfc.nasa.gov/>, last access: 8 October 2022) as an indicator of biological influence, which has a bias of  $\sim 0.35$  on a log scale and  $\sim 115\%$  in the SCS (Zhang et al., 2006). Atmospheric  $p\text{CO}_2$  also influences sea surface  $p\text{CO}_2$  through air–sea  $\text{CO}_2$  exchange. We chose the atmospheric  $\text{CO}_2$  mole fraction ( $x\text{CO}_2$ ) data from the monthly mean  $\text{CO}_2$  concentrations measured at the Mauna Loa Observatory, Hawaii (<https://gml.noaa.gov/>, last access: 8 October 2022), and then calculated the atmospheric  $p\text{CO}_2$  values from  $x\text{CO}_2$  using the method in Li et al. (2020).

### 3 Methods

The  $p\text{CO}_2$  reconstruction procedure is shown in Fig. 4. It includes (1) data processing and (2) model training and testing. For the former, we first gridded the in situ data and RS-derived  $p\text{CO}_2$  data into  $0.05^\circ \times 0.05^\circ$  boxes with a monthly temporal resolution. Second, we filled missing  $p\text{CO}_2$  measurements with the RS-derived  $p\text{CO}_2$  data, according to Fay et al. (2021; see more details in Sect. 3.1). We then used EOF to ignore any biases in the RS-derived  $p\text{CO}_2$  dataset itself or from the  $p\text{CO}_2$  filling method. Third, the gridded in situ  $p\text{CO}_2$  data and their corresponding RS-derived data were divided into a training set (90 %) and a testing set (10 %) to calculate the  $p\text{CO}_2$  retrieval model. To ensure that the model had sufficient training samples in the coastal area, we divided the entire SCS into two regions along the 200 m isobath (as shown in Fig. 5). The data from these two regions were divided into training and testing sets with the same ratios listed above (9 : 1) and then combined to obtain the final training and testing sets. Note that all the data used in the machine learning have been interpolated on the same grid.

For model training and testing, we chose a relatively reliable algorithm to undertake the  $p\text{CO}_2$  reconstruction. Next, we determined the optimal range of the parameters using hyperparameter methods (code from <https://github.com/optuna/>, last access: 8 October 2022) for the training set. The final optimal parameter values were then determined using the  $K$ -fold and cross-validation methods (code from <https://github.com/suryanktiwari/Linear-Regression-and-K-fold-cross-validation>, last access: 8 October 2022) for the training set. These optimal parameters were applied to the chosen algorithm. Finally, the testing set was used to verify the accuracy of the  $p\text{CO}_2$  retrieval algorithm produced by the training set, and some indicators of the model's accuracy were calculated. More detailed methods employed in the present study are described below.

#### 3.1 Remote sensing data filling

As mentioned in the Sat $\text{CO}_2$  platform (<http://www.SatCO2.com>), RS-derived  $p\text{CO}_2$  datasets have some missing values. Thus, we used the  $p\text{CO}_2$  data-filling method, suggested by Fay et al. (2021), to obtain the missing data points. First, a

scaling factor for a filled month was calculated according to Eq. (1):

$$\text{sf}_{p\text{CO}_2} = \text{mean}_{x,y} \left( \frac{p\text{CO}_2^{\text{ens}}}{p\text{CO}_2^{\text{clim}}} \right), \quad (1)$$

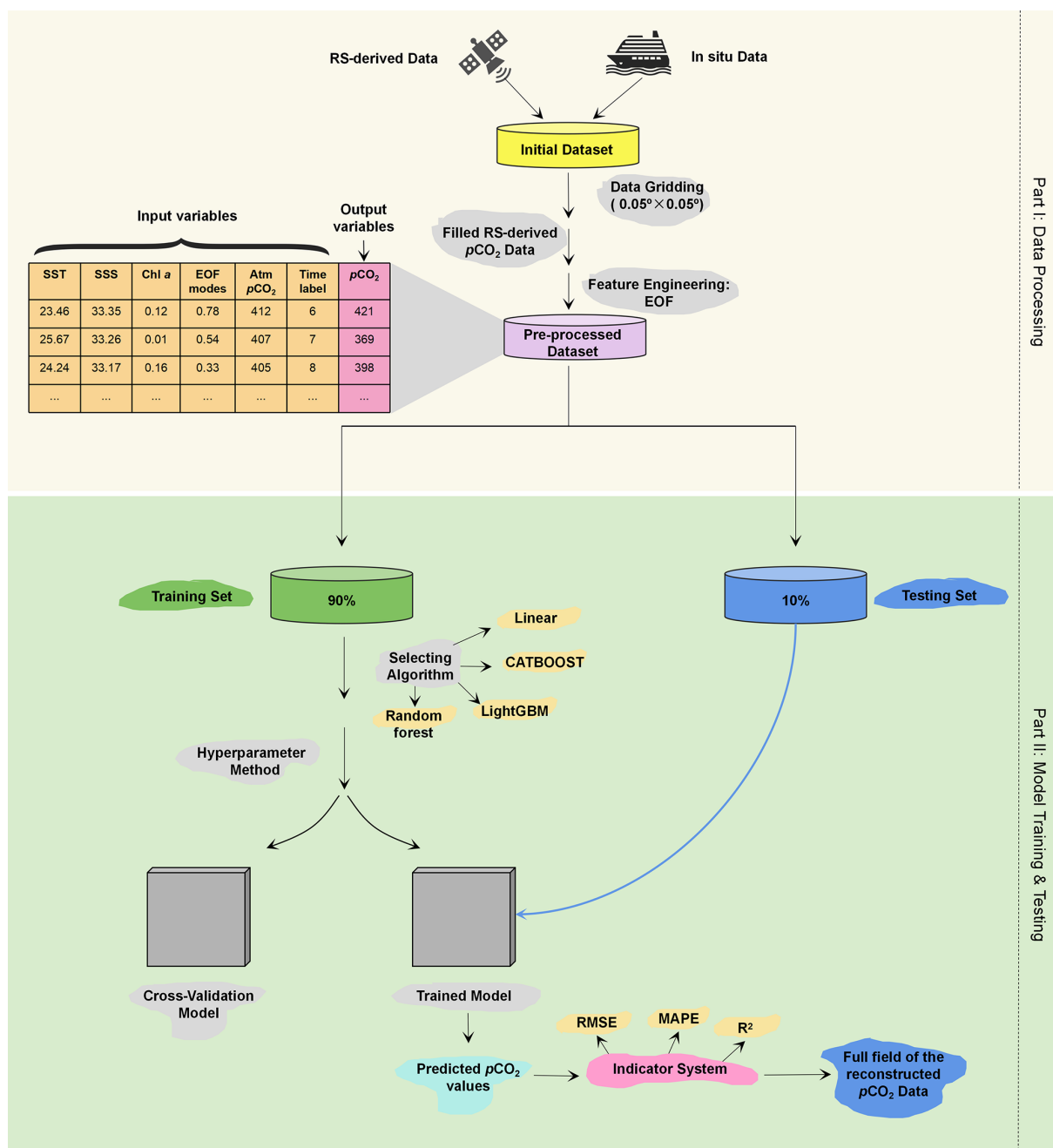
where  $\text{sf}_{p\text{CO}_2}$  is the scaling factor,  $p\text{CO}_2^{\text{ens}}$  is the monthly RS-derived  $p\text{CO}_2$  data, and  $p\text{CO}_2^{\text{clim}}$  is the monthly climatology RS-derived  $p\text{CO}_2$  data.  $x$  and  $y$  indicate that we took the area-weighted average over longitude ( $x$ ) and latitude ( $y$ ) to produce the monthly  $\text{sf}_{p\text{CO}_2}$  value. Then, the filled portion of the data can be calculated from the  $p\text{CO}_2^{\text{clim}}$  data multiplied by the  $\text{sf}_{p\text{CO}_2}$  value (see Fay et al., 2021, for details of this method).

Briefly, this filling method scales the climatological monthly  $p\text{CO}_2$  field values to fill in the missing measurements. Therefore, although specific values may be biased, the interpolated measurements still retain the main spatial distribution pattern of the filled months.

#### 3.2 Feature engineering and selection

As mentioned above, the  $p\text{CO}_2$  data-filling method may bias some of the actual values. To avoid the influence of such biases on the reconstructed results, instead of directly using the RS-derived  $p\text{CO}_2$  data as features in our reconstructed model, we used the EOF method to obtain the main spatiotemporal distribution patterns of the RS-derived  $p\text{CO}_2$  data as features in our reconstructed model. The EOF reflects the spatial commonality of variables shown in the time series, and thus it is widely used to calculate spatial patterns of climate variability (e.g., Levitus et al., 2005; Dye et al., 2020; McMonigal and Larson, 2022). Typically, the spatial commonality of variables (EOF modes) is found by computing the eigenvalues and eigenvectors of a spatially weighted anomaly covariance matrix of a field. Each EOF mode's corresponding variance represents its degree of interpretation of the spatial pattern of a variable. For each of the 12 months, the cumulative variance contribution of the first eight EOF values was consistently  $> 90\%$ , indicating that it could explain the main  $p\text{CO}_2$  spatial characteristics during each month; we therefore selected them as features.

The features selected in our reconstructed model can be divided into two main categories. In the first category, the features are related to the underlying physicochemical mechanisms controlling the  $p\text{CO}_2$  distribution; for example, SST exerts a primary control on the seasonal variations in surface water  $p\text{CO}_2$  in the northern SCS (Zhai et al., 2005; Chen et al., 2007; Li et al., 2020). In the second category, they provide spatiotemporal information for the  $p\text{CO}_2$  reconstruction. Previous studies (Landschützer et al., 2014; Laruelle et al., 2017; Denvil-Sommer et al., 2019) have shown that Chl  $a$  plays a critical role in fitting the influence of biological activity to  $p\text{CO}_2$ , especially in the northern SCS (Landschützer et al., 2014; Laruelle et al., 2017; Denvil-Sommer



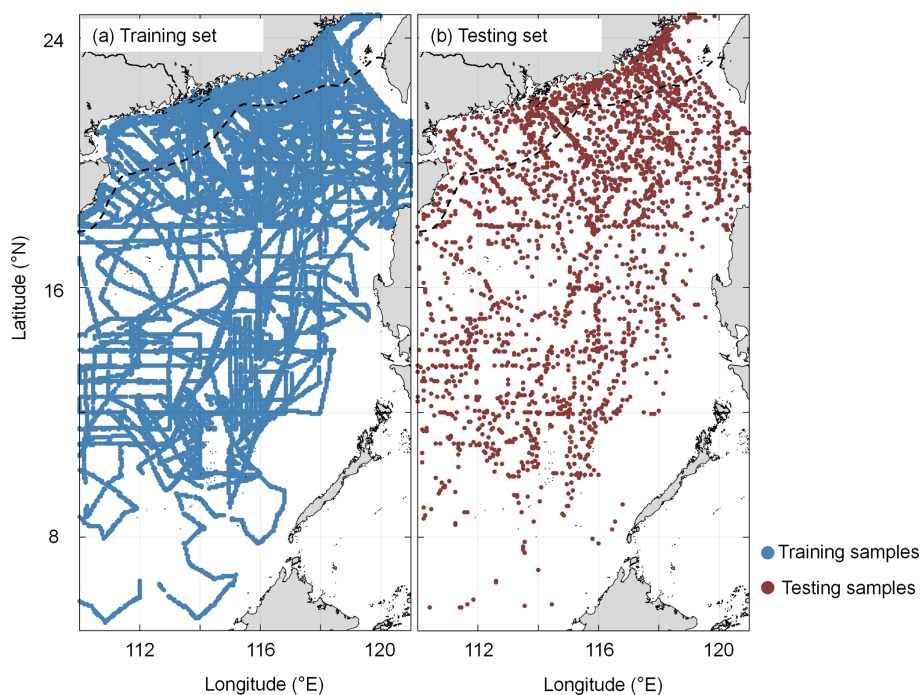
**Figure 4.** Procedure for the reconstruction of surface water  $p\text{CO}_2$  using machine learning. RS-derived data are remote-sensing-derived data. RMSE is the root mean square error. MAPE is the mean absolute percentage error.  $R^2$  is the coefficient of determination. MAE is the mean absolute error.

et al., 2019). Sutton et al. (2017) suggest that increasing atmospheric  $p\text{CO}_2$  controls the overall increase in seawater  $p\text{CO}_2$ . For the features that provide spatiotemporal information for the  $p\text{CO}_2$  reconstruction, in the present study we selected the first eight EOF values of  $p\text{CO}_2$  as the main spatial distribution feature and the monthly information of the in situ datasets as the temporal feature.

### 3.3 Algorithm selection

Ensemble learning, which is the process of training multiple machine learning models and combining their output to improve the reliability and accuracy of predictions, is one of the most powerful machine learning techniques (e.g., Zhan et al., 2022; Cheng et al., 2020). In other words, several different models are used as the basis to develop an op-





**Figure 5.** Spatial distributions of training samples (a) and testing samples (b). The dashed black line shows the 200 m isobath.

timal predictive model. There are two main ways to employ ensemble learning, namely bagging (to decrease the model's variance) or boosting (to decrease the model's bias). The random forest algorithm (code from <https://scikit-learn.org/stable/>, last access: 6 May 2022) is an extension of the bagging method, as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. The light gradient-boosting machine (LightGBM; code from <https://github.com/microsoft/LightGBM/>, last access: 6 May 2022) is a gradient-boosting framework that uses tree-based learning algorithms. LightGBM can be used for regression, classification, and other machine learning tasks; it exhibits rapid, high-performance as a machine learning algorithm. CatBoost (code from <https://github.com/catboost/>, last access: 6 May 2022) is a gradient-boosting algorithm which improves prediction accuracy by adjusting weights according to the data distribution and by incorporating prior knowledge about the dataset. This can help to reduce overfitting and improve general performance.

From the above options, we chose three ensemble learning algorithms as the machine-learning-based regression portion and multilinear regression methods (Wang et al., 2021) as the linear regression portion. We then used the  $K$ -fold and cross-validation methods to verify the applicability of different regression algorithms in the  $p\text{CO}_2$  reconstruction for seasonal training data. The results show that, in summer, the CatBoost algorithm yields the best degree of accuracy, with an RMSE of  $16 \mu\text{atm}$  (Table 3). In contrast, the RMSE of LightGBM was  $27 \mu\text{atm}$  and that of random forest (RF) was

$26 \mu\text{atm}$ . The RMSE was nearly  $20 \mu\text{atm}$ , using the linear regression algorithm employed by Wang et al. (2021). Thus, CatBoost appears to provide a reliable algorithm for reconstructing  $p\text{CO}_2$ . In the other three seasons, however, using different algorithms resulted in minor differences ( $\sim 2 \mu\text{atm}$  in RMSE).

### 3.4 Evaluation metrics

It is necessary to evaluate the accuracy of any model based on certain error metrics before applying it to specific scenarios. Common model evaluation metrics include RMSE, mean absolute percentage error (MAPE),  $R^2$  (coefficient of determination), and MAE.

The mean squared error (MSE) is the standard deviation of the residuals (prediction error), and the residuals are the distances between the fitted line and the data points (i.e., the residuals show the degree of concentration of the reconstructed data around the regression line). In regression analysis, RMSE is commonly used to verify experimental results. To assess bias, the RMSE needs to combine the magnitude of the model data and is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_{ri})^2}, \quad (2)$$

where  $y$  stands for the in situ data,  $y_r$  represents the reconstructed data, and  $n$  is the number of data points.

The MAPE is a statistical measure used to define the accuracy of a machine learning algorithm on a particular dataset.

**Table 3.** RMSEs associated with different algorithms in the four seasons.

Season	Random forest	LightGBM	CatBoost	Multilinear regression (Wang et al., 2021)
Spring	10.65 $\mu\text{atm}$	9.52 $\mu\text{atm}$	8.17 $\mu\text{atm}$	NaN*
Summer	26.53 $\mu\text{atm}$	27.83 $\mu\text{atm}$	16.15 $\mu\text{atm}$	20.13 $\mu\text{atm}$
Fall	10.34 $\mu\text{atm}$	11.56 $\mu\text{atm}$	10.35 $\mu\text{atm}$	NaN
Winter	12.48 $\mu\text{atm}$	12.75 $\mu\text{atm}$	11.52 $\mu\text{atm}$	NaN

\* NaN stands for missing values.

It is commonly used because, compared to other metrics, it uses a percentage to measure the magnitude of the bias and is easy to understand and interpret; the lower the value of the MAPE, the better a model is at forecasting. MAPE is calculated as follows:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - y_{ri}|}{|y_i|}. \quad (3)$$

The regression error metric, the coefficient of determination ( $R^2$ ), can describe the performance of a model by evaluating the accuracy and efficiency of the modeled results; i.e., it indicates the magnitude of the dependent variable, as calculated by the regression model, that can be explained by the independent variable. It is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - y_{ri})^2}. \quad (4)$$

MAE is the average absolute difference between the in situ data (true values) and the model output (predicted values). The sign of these differences is ignored so that cancelations between positive and negative values do not occur. It is calculated as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - y_{ri}|. \quad (5)$$

### 3.5 Uncertainty

In previous studies, RMSE and MAE have primarily been used to represent the uncertainties in reconstructed datasets. However, this expression of uncertainty ignores the sensitivity of the reconstructed model to the features; i.e., the biases that the features themselves pass to the reconstructed model are ignored. Moreover, it is clearly unreasonable to use a single RMSE or MAE value to represent the entire region because the spatial bias pattern in the coastal region clearly differs from that in the basin.

Thus, here we present a novel method for calculating uncertainty, as shown below:

Uncertainty

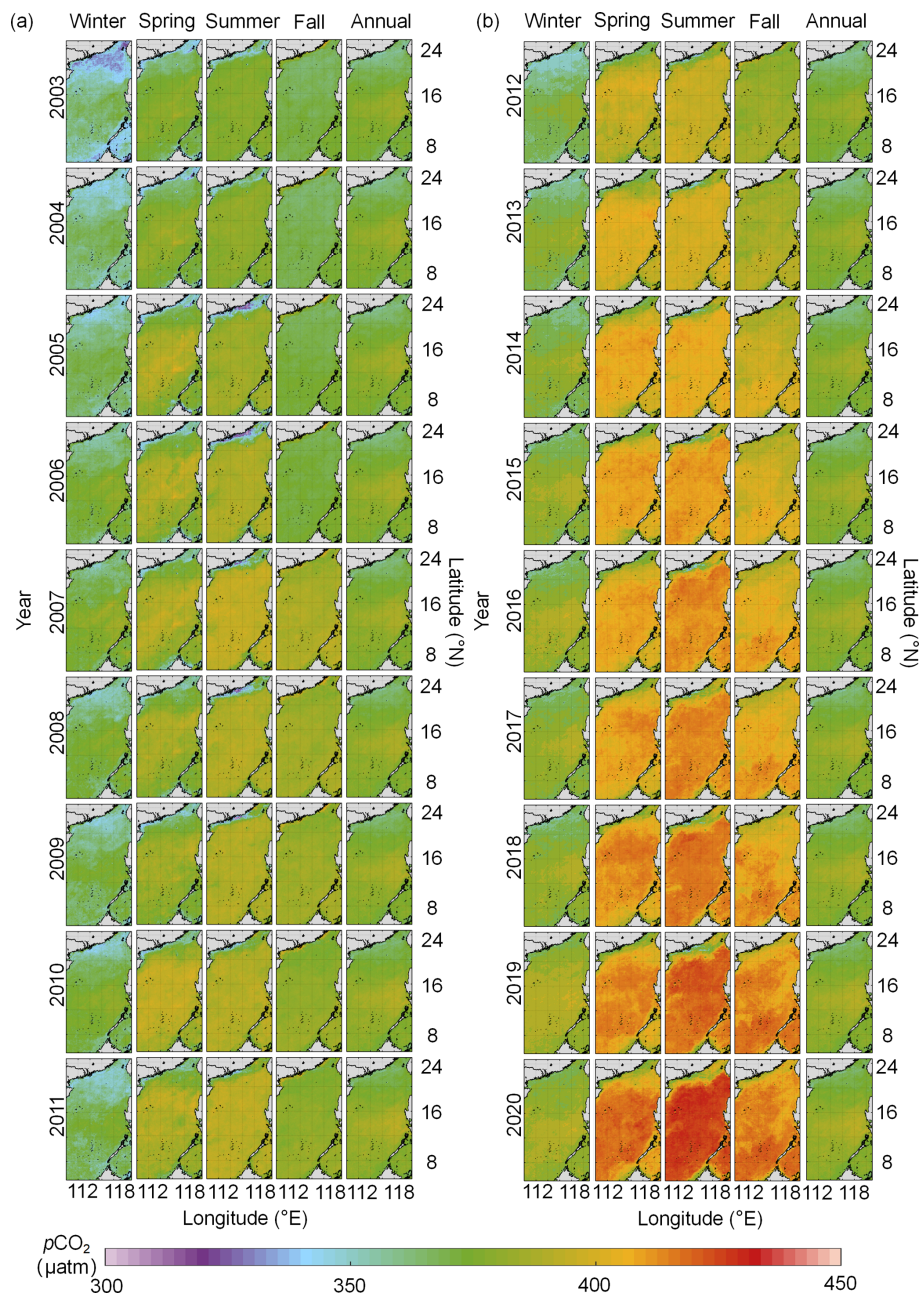
$$= \text{MAX} \left( \left[ \frac{\sum_{i=1, j=1, k=1}^n \frac{|\text{OR\_Monthly\_Data}(i, j, k) - \text{Obs\_Monthly\_Data}(i, j, k)|}{\text{Obs\_Monthly\_Data}(i, j, k)}}{\text{num}(i) + \text{num}(j)}, \dots, \frac{\sum_{i=1, j=1, k=n}^n \frac{|\text{OR\_Monthly\_Data}(i, j, k) - \text{Obs\_Monthly\_Data}(i, j, k)|}{\text{Obs\_Monthly\_Data}(i, j, k)}}{\text{num}(i) + \text{num}(j)} \right] \right) \cdot 100\% \cdot p\text{CO}_2\text{-recon} + \left( \frac{\partial p\text{CO}_2}{\partial \text{Feature}} \right) \text{dFeature}. \quad (6)$$

Equation (6) includes two terms. The first term is the conservative bias between the reconstructed  $p\text{CO}_2$  fields and the in situ data, and the second is the sum over sensitivity of the reconstructed model to the features. For the first term in Eq. (6),  $k$  stands for the  $k$ th month,  $\text{OR\_Monthly\_Data}(ijk)$  stands for the  $k$ th monthly reconstructed data at longitude ( $i$ ) and latitude ( $j$ ), and  $\text{Obs\_Monthly\_Data}(ijk)$  stands for the  $k$ th monthly in situ data at longitude ( $i$ ) and latitude ( $j$ ). Therefore, MAX in the first term stands for the maximum of the  $k$  monthly bias ratios. And  $p\text{CO}_2\text{-recon}$  stands for the reconstructed  $p\text{CO}_2$  data. In the second term, dFeature stands for the bias of the features. We conducted a sensitivity analysis using a chain rule to evaluate the influence of these biases in the features on  $p\text{CO}_2$ . Then we estimated  $p\text{CO}_2$  changes due to the variabilities in these features by constraining these features based on our model and computed  $\frac{\partial p\text{CO}_2}{\partial \text{Feature}}$ . For example, for  $\frac{\partial p\text{CO}_2}{\partial \text{SST}}$ , we only changed the value of SST and kept the values of the other features constant to calculate the effect of each additional unit of SST on the simulated  $p\text{CO}_2$ .

## 4 Results and discussion

### 4.1 Results

The reconstructed  $p\text{CO}_2$  fields show relatively low values in the northern coastal region of the study area and generally high values in the middle and southern basins (Fig. 6). The continuous changes in the spatiotemporal distribution can be found in the reconstruction results (Fig. 6). The reconstructed  $p\text{CO}_2$  fields show a trend of slow but sustained increases from 2003 to 2020. Spatial patterns of  $p\text{CO}_2$  change between 2003 and 2020, such that the coastal portion of the northern SCS shows relatively complex variability from

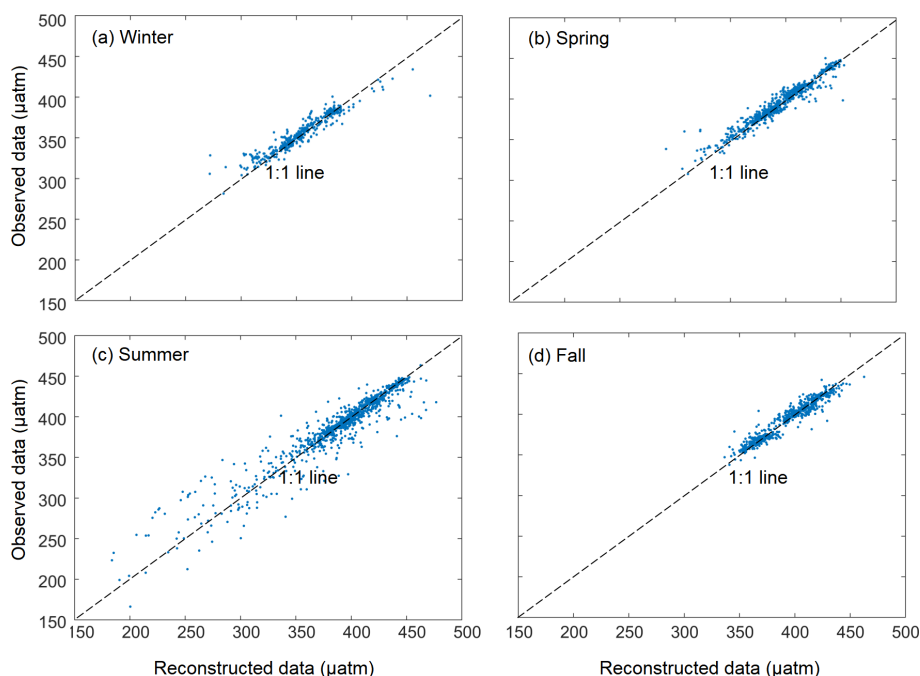


**Figure 6.** Reconstructed seasonal and annual  $p\text{CO}_2$  fields in the South China Sea from 2003 to 2020 (a, 2003–2011; b, 2012–2020).

multiple controlling factors, such as coastal upwelling, river plumes, biological activity, etc. However,  $p\text{CO}_2$  values in the middle and southern basins are relatively homogeneous, as they are mainly controlled by atmospheric  $p\text{CO}_2$  forcing and SST. Temporal changes in  $p\text{CO}_2$  between 2003 and 2020 are relatively large ( $\sim 44 \mu\text{atm}$ ) in summer and relatively small ( $\sim 33 \mu\text{atm}$ ) in winter.

#### 4.2 Model validation

Figure 7 compares the monthly reconstructed and in situ data. For the training dataset, the reconstructed  $p\text{CO}_2$  fields of the four seasons fit the in situ data well (Fig. 7), with an average RMSE of  $3.43 \mu\text{atm}$  and an average MAE of  $2.14 \mu\text{atm}$  (Table 2). For the testing sets, although there are some outliers, most of the reconstructed  $p\text{CO}_2$  data are consistent with the in situ data, with RMSE averaging  $10.79 \mu\text{atm}$  and MAE averaging  $6.30 \mu\text{atm}$ . The  $R^2$  of the testing set is ca. 0.91. In terms of MAPE, the accuracies of the four seasonal mod-



**Figure 7.** Comparisons between the monthly reconstructed and in situ  $p\text{CO}_2$  values for the testing set. The monthly results are grouped into the four seasons, including (a) winter, in December, January, and February, (b) spring, in March, April, and May, (c) summer, in June, July, and August, and (d) fall, in September, October, and November.

els are all around 99 % (Table 2), with the highest value for spring data and the lowest value for summer data. The relatively large bias (14.67  $\mu\text{atm}$ ) in the summer may be the influence of relatively complex regional processes, such as river plumes and upwelling. The four evaluation metrics indicate that our reconstructed  $p\text{CO}_2$  field is highly accurate in simulating both the training and testing sets.

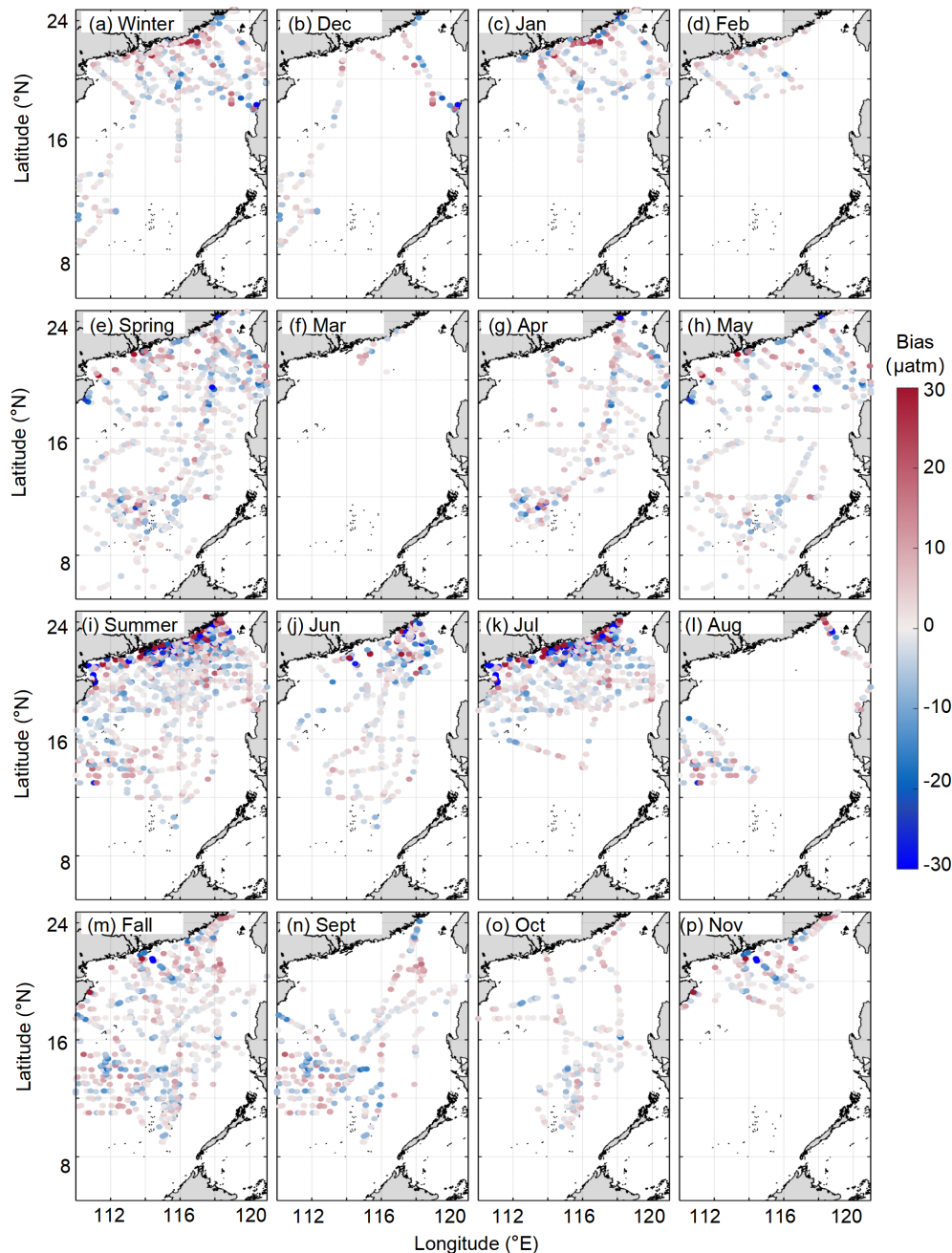
The distributions of the biases between the reconstructed fields and the in situ data for both the training and testing datasets can be found in Fig. 8. In terms of the temporal pattern, the larger biases were more concentrated in the summer. For the spatial pattern, the biases in the northern coastal area are much greater than those in the basin. However, 95 % of the biases are  $< \pm 10 \mu\text{atm}$ ; therefore, our reconstructed dataset exhibits relatively high accuracy.

Figure 9 shows the bias between our reconstructed fields and the four independent in situ datasets corresponding to the four seasons. This validation can verify the accuracy of the retrieval algorithm for months without observations, namely the applicability of the retrieval algorithm extrapolation. This comparison shows that the retrieval algorithm is relatively accurate in the basin, with a near-zero bias (MAE of  $\sim 8 \mu\text{atm}$ ; Fig. 9a). The largest bias occurs in the Pearl River plume area in summer ( $\sim 35 \mu\text{atm}$ ). The retrieval algorithm also has a high accuracy for  $p\text{CO}_2$  spatial variability, except in the Pearl River plume area in summer (22–20° N; Fig. 9b–e). The effect of the Pearl River plume on the  $p\text{CO}_2$  spatial distribution in our retrieval algorithm is

smaller than that shown by the in situ data. This is because, at around the survey time (24–28 August 2019), a large amount of precipitation ( $\sim 30 \text{ mm d}^{-1}$ ; <https://psl.noaa.gov/data/gridded/data.ncep.reanalysis2.surface.html>, last access: 8 October 2022) occurred around the Pearl River estuary region (24–20° N), which led to the intensification of the Pearl River plume. The plume has relatively low  $p\text{CO}_2$  values that eventually decreased the observed values along the coast. However, the monthly average runoff of the Pearl River during that month (August 2019; <http://www.pearlwater.gov.cn/>, last access: 8 October 2022; see the Pearl River plume index in Wang et al., 2022) was low, indicating that our retrieval algorithm is still highly reliable from the perspective of monthly averages. Thus, the inconsistencies between the reconstructed (monthly average) and the in situ datasets are mainly due to the differences in the timescales of the remote sensing and the in situ data. The reconstructed data in this study were determined on a monthly scale, while the temporal resolution of the in situ data were on the order of hours. It is clear that relatively pronounced short-term changes in  $p\text{CO}_2$ , such as the diurnal variability caused by short-term heavy precipitation, cannot be reflected in the reconstructed data.

Dai et al. (2022) produced a time series of in situ data from 2003 to 2019 at the SEATs station, which we used here to validate the accuracy of the long-term trends of our model data (results shown in Fig. 10). The long-term trend of reconstructed  $p\text{CO}_2$  data at the SEATs station is largely con-





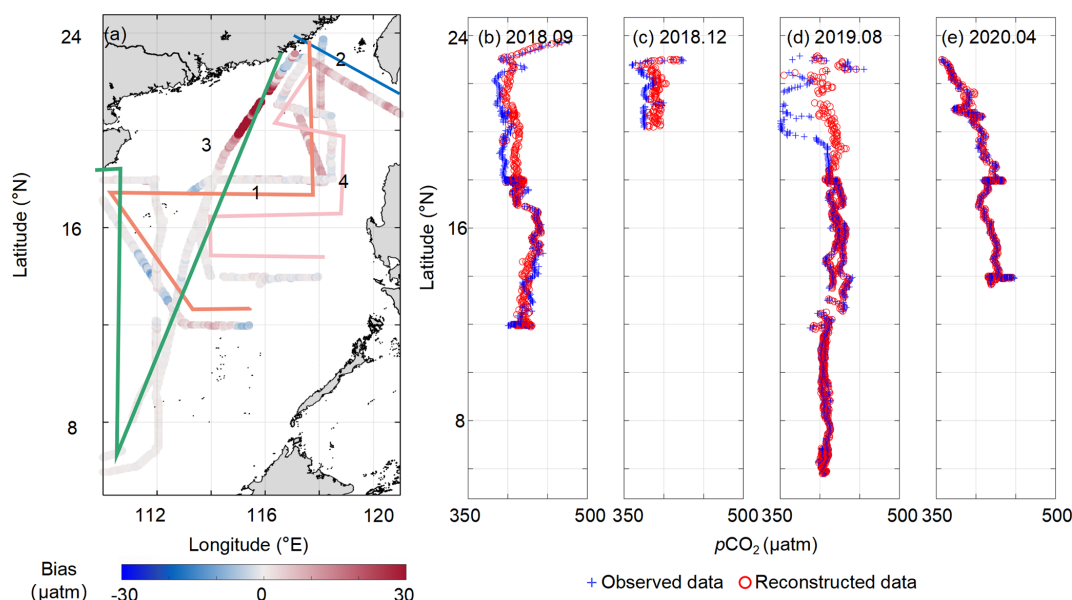
**Figure 8.** Differences between the reconstructed and in situ  $p\text{CO}_2$  data, both seasonally and monthly, for the testing set, including (a) winter, (b) December, (c) January, (d) February, (e) spring, (f) March, (g) April, (h) May, (i) summer, (j) June, (k) July, (l) August, (m) fall, (n) September, (o) October, and (p) November).

sistent with the in situ data, with differences mainly found before 2005. Thus, the long-term trend produced in our reconstructed model is also highly reliable.

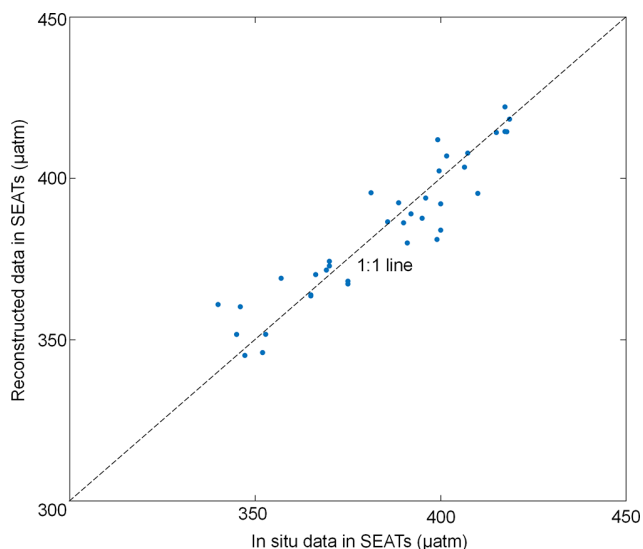
### 4.3 Uncertainties

As shown in Table 2, our reconstructed data have a high degree of accuracy, with an RMSE of  $\sim 10 \mu\text{atm}$  and MAE of

$\sim 6 \mu\text{atm}$ . According to Eq. (6), the bias of RS-derived  $p\text{CO}_2$  data used in the second term of Eq. (6) is  $\sim 21 \mu\text{atm}$  (Table 2), the bias of SST is  $\sim 0.27^\circ\text{C}$  (Qin et al., 2014), the bias of SSS is  $\sim 0.33$  (Wang et al., 2022), and the bias of Chl  $a$  is  $\sim 115\%$  (Zhang et al., 2006). We then estimated the  $p\text{CO}_2$  changes due to the variations in these features by constraining these features based on our model and computed  $\frac{\partial p\text{CO}_2}{\partial \text{Feature}}$ .



**Figure 9.** Difference between the reconstructed  $p\text{CO}_2$  data and four independently tested in situ datasets during the four seasons. In panel (a), the numbers 1–4 represent September 2018 (b), December 2018 (c), August 2019 (d), and April 2020 (e), respectively.



**Figure 10.** Comparison of the reconstructed  $p\text{CO}_2$  with in situ data at the Southeast Asia Time-series Study (SEATs) station ( $116^\circ \text{E}$ ,  $18^\circ \text{N}$ ). The in situ data are from Dai et al. (2022), which were calculated from dissolved inorganic carbon and total alkalinity values.

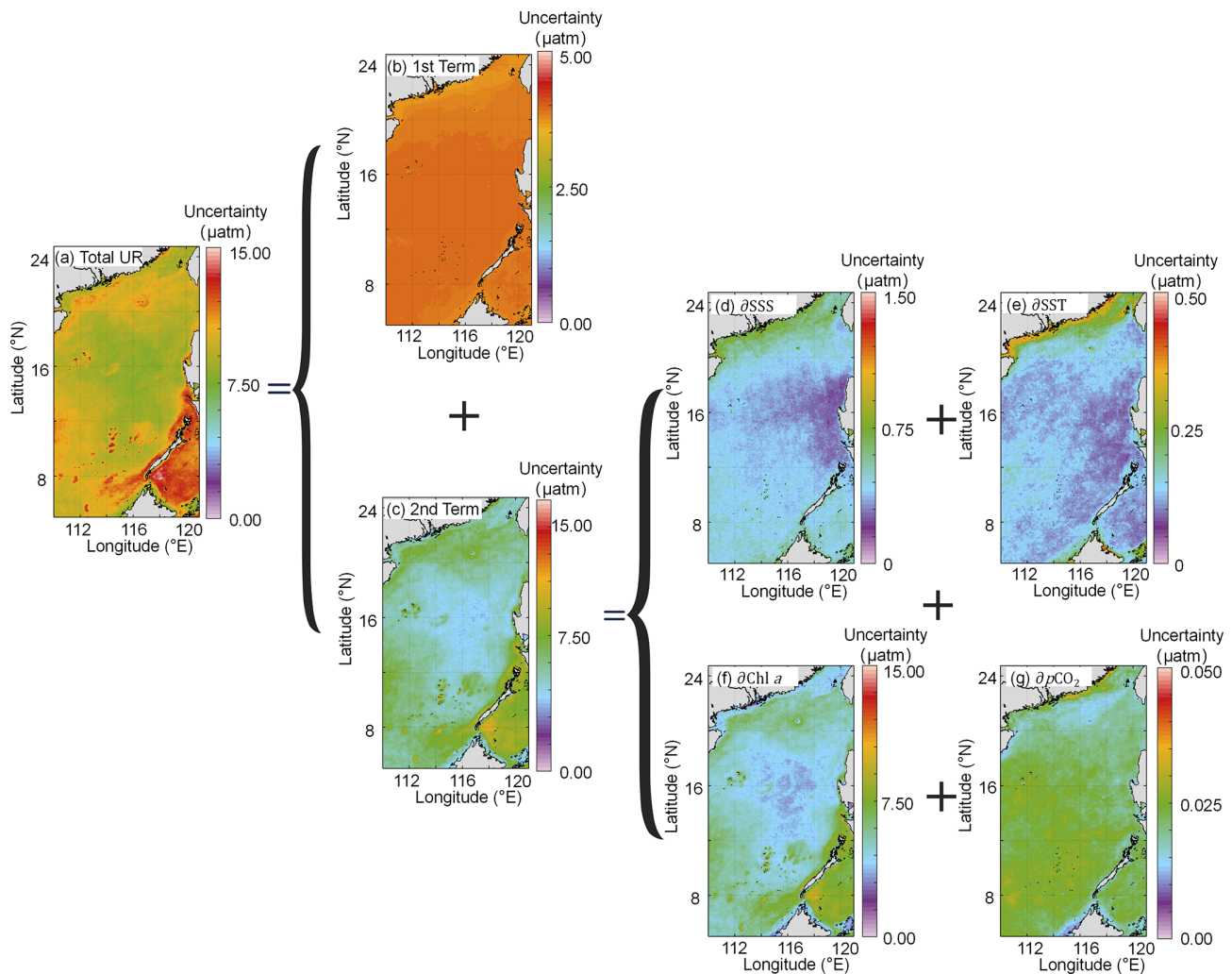
The overall uncertainty in the reconstructed dataset is greater in the coastal area ( $\sim 13 \mu\text{atm}$ ) than in the basin ( $\sim 10 \mu\text{atm}$ ; Fig. 11a), and this spatial pattern is mainly determined by the second term in Eq. (6). The spatial distribution of the first term in Eq. (6) (Fig. 11b), calculated from a max bias ratio, is consistent with that of  $p\text{CO}_2$  (Fig. 11b). The second term in Eq. (6) (Fig. 11c) is calculated from the propagation of the bias from each variable (Fig. 11c). The Chl  $a$

bias (Fig. 11f) shows that it has the greatest effect on the reconstruction, among all the features (Fig. 11f). Although the bias of the RS-derived  $p\text{CO}_2$  data is relatively large, the final influence that it has on the results from the retrieval algorithm is negligible due to the use of the EOF method (Fig. 11g).

#### 4.4 Spatial and temporal $p\text{CO}_2$ features

The climatological monthly reconstructed  $p\text{CO}_2$  fields are shown in Fig. 12. The highest values occur in May and June, and the lowest values occur in January. In winter,  $p\text{CO}_2$  first decreases in December and then increases after January; the  $p\text{CO}_2$  value is ca.  $325 \mu\text{atm}$  in the northern coastal area and ca.  $350 \mu\text{atm}$  in the basin. In spring,  $p\text{CO}_2$  gradually increases from the basin to the northern coastal area, and the high  $p\text{CO}_2$  values in the central basin gradually expand outward starting in April. In summer,  $p\text{CO}_2$  gradually declines, starting in June. In fall,  $p\text{CO}_2$  increases from north to south, and the southern region shows consistently high values.

To better show specific regions in the northern coastal area, we magnified the reconstructed  $p\text{CO}_2$  fields at locations north of  $18^\circ \text{N}$  (Fig. 13). The reconstructed  $p\text{CO}_2$  fields successfully reflect the influence of the meso–microscale processes on  $p\text{CO}_2$  in this northern coastal area of the SCS. For example, in winter, the relatively low  $p\text{CO}_2$  values, which last into early spring, are mainly controlled by the low SST and the high  $p\text{CO}_2$  around Luzon Strait affected by winter upwelling. In summer, the reconstructed  $p\text{CO}_2$  field shows that the influence of the Pearl River plume on  $p\text{CO}_2$  is the strongest in July and lasts until September; it also effectively shows the influence of coastal upwelling in the northeastern shelf ( $\sim 23^\circ \text{N}$ ,  $117^\circ \text{E}$ ). Thus, our reconstructed  $p\text{CO}_2$  fields



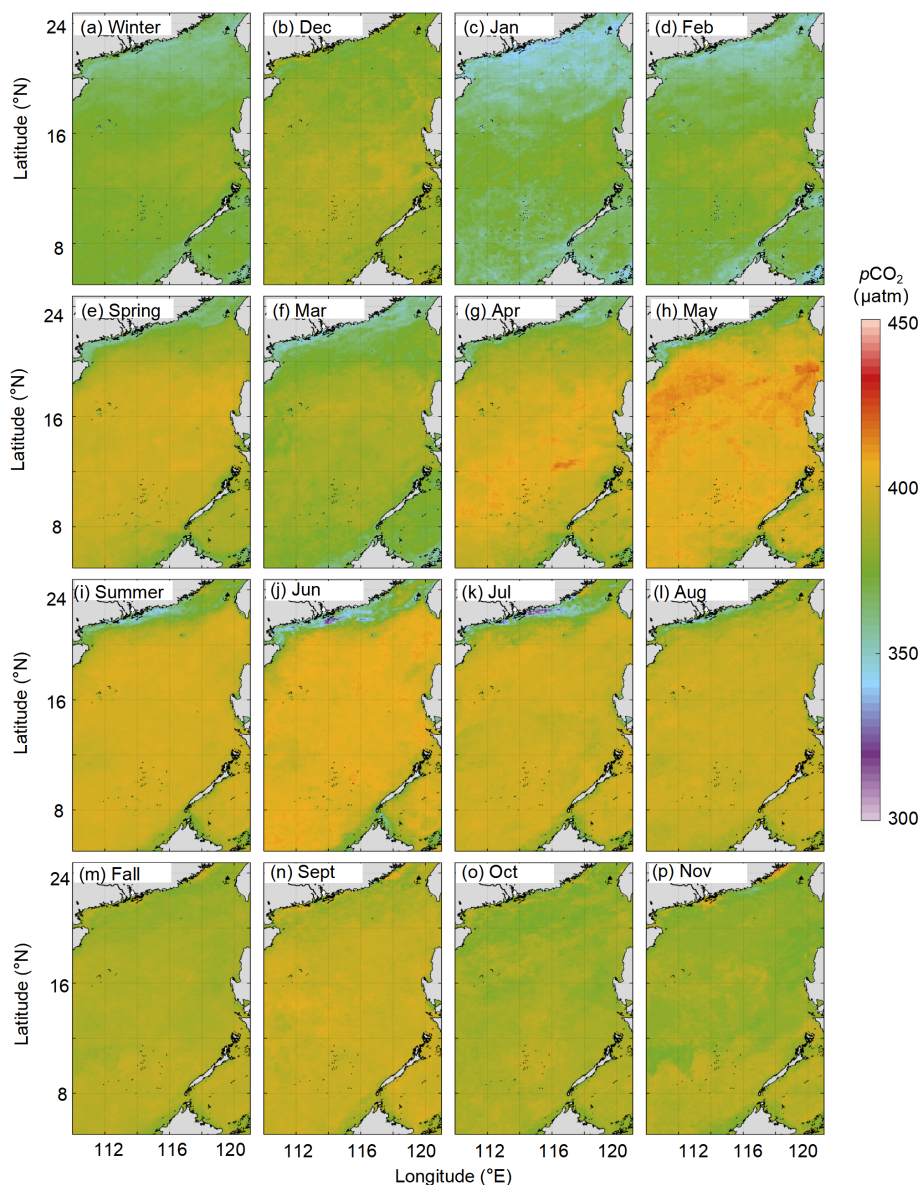
**Figure 11.** Uncertainties in the reconstructed  $p\text{CO}_2$  fields. **(a)** Total uncertainty in Eq. (6). **(b)** The first term of Eq. (6). **(c)** The second term of Eq. (6). **(d)**  $(\frac{\partial p\text{CO}_2}{\partial \text{SSS}})d\text{SSS}$  in the second term of Eq. (6). **(e)**  $(\frac{\partial p\text{CO}_2}{\partial \text{SST}})d\text{SST}$  in the second term of Eq. (6). **(f)**  $(\frac{\partial p\text{CO}_2}{\partial \text{Chl } a})d\text{Chl } a$  in the second term of Eq. (6). **(g)**  $(\frac{\partial p\text{CO}_2}{\partial \text{RS\_derived\_}p\text{CO}_2})d\text{RS\_derived\_}p\text{CO}_2$  in the second term of Eq. (6).

clearly reflect the spatial pattern of the in situ  $p\text{CO}_2$  (Fig. 3), which are generally consistent with previously reported patterns (Li et al., 2020; Zhai et al., 2013; Gan et al., 2010).

We divided SCS into five subregions, according to Li et al. (2020). In Fig. 14, Subregion\_A stands for the northern coastal area of the SCS, Subregion\_B stands for the slope area of the northern SCS, Subregion\_C stands for the SCS basin, Subregion\_D stands for the region west of the Luzon Strait, and Subregion\_E stands for the slope and basin area of the western SCS. All\_region indicates the whole region containing the five subregions described above. We then calculated the deseasonalized long-term trend of spatially averaged monthly data for each subregion, and the results are shown in Fig. 14 and Table 3. This deseasonalized trend is consistent with that of the in situ data, and its uncertainty is on the 95 % confidence interval (much lower than that shown

by the in situ data). We can thus also infer that the long-term trend of our reconstructed data shows high reliability in all subregions and that our data can serve as an important basis for predicting future changes in  $p\text{CO}_2$  in the SCS.

In Fig. 14a–e, we found that the sea surface  $p\text{CO}_2$  of the entire SCS is slightly higher than the atmospheric  $p\text{CO}_2$ , indicating that the SCS is a weak source of atmospheric  $\text{CO}_2$ . This conclusion is consistent with previous studies (e.g., Li et al., 2020). Moreover, compared to the rate of atmospheric  $\text{CO}_2$  increase ( $\sim 2.2 \mu\text{atm yr}^{-1}$ ), for Subregion\_A, the  $p\text{CO}_2$  trend is much slower than that of atmospheric  $p\text{CO}_2$ , and the spatially averaged monthly mean  $p\text{CO}_2$  is lower than the atmospheric  $p\text{CO}_2$ . Thus, carbon accumulation in this region is expected to increase in the future. For Subregion\_C and Subregion\_E, the spatially averaged monthly mean  $p\text{CO}_2$  is higher than the atmospheric  $p\text{CO}_2$ ; thus, these two regions



**Figure 12.** Long-term (2003–2020) seasonal and monthly averaged  $p\text{CO}_2$  field ( $\mu\text{atm}$ ). (a) Winter. (b) December. (c) January. (d) February. (e) Spring. (f) March. (g) April. (h) May. (i) Summer. (j) June. (k) July. (l) August. (m) Fall. (n) September. (o) October. (p) November.

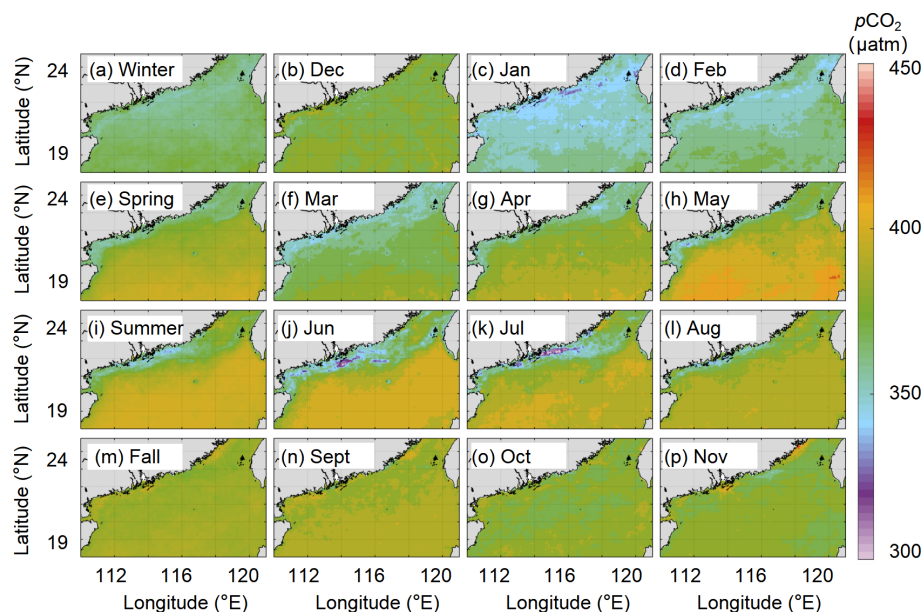
**Table 4.** Deseasonalized long-term trend of the spatially averaged monthly  $p\text{CO}_2$  data for each subregion of the South China Sea ( $\mu\text{atm yr}^{-1}$ ).

	All_region	Subregion_A	Subregion_B	Subregion_C	Subregion_D	Subregion_E
Reconstructed $p\text{CO}_2$	$2.12 \pm 0.17$	$1.82 \pm 0.14$	$2.23 \pm 0.12$	$2.17 \pm 0.12$	$2.20 \pm 0.13$	$2.16 \pm 0.13$
In situ $p\text{CO}_2$	$2.10 \pm 0.79$	$1.80 \pm 0.86$	$1.73 \pm 0.84$	$1.81 \pm 0.85$	$1.41 \pm 1.16$	$2.13 \pm 1.10$

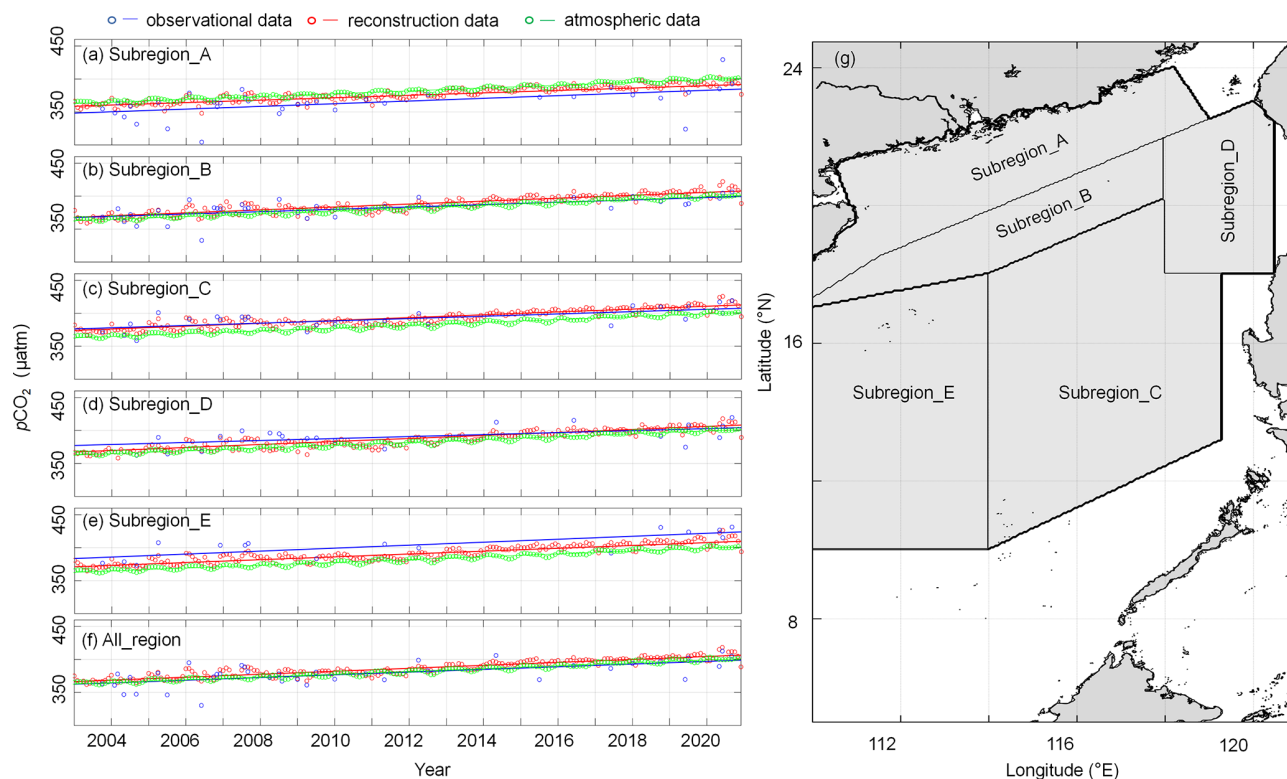
will still provide a weak source of atmospheric  $\text{CO}_2$  in the future. Finally, whether Subregion\_B and Subregion\_D act as a source or sink of the atmospheric  $\text{CO}_2$  is influenced by seasonal changes and physical processes. Subregion\_B can be a zone of significant sink of atmospheric  $\text{CO}_2$ , as demonstrated by its low sea surface  $p\text{CO}_2$  when the Pearl River

plume spreads more widely in summer. In contrast, in winter, when the Kuroshio intrusion is strong, both Subregion\_B and Subregion\_D have high sea surface  $p\text{CO}_2$ , indicating both subregions are sources of atmospheric  $\text{CO}_2$ .





**Figure 13.** Long-term (2003–2020) seasonal and monthly averaged  $p\text{CO}_2$  field in the region north of  $18^\circ\text{N}$  ( $\mu\text{atm}$ ). (a) Winter. (b) December. (c) January. (d) February. (e) Spring. (f) March. (g) April. (h) May. (i) Summer. (j) June. (k) July. (l) August. (m) Fall. (n) September. (o) October. (p) November.



**Figure 14.** Time series of spatially averaged monthly  $p\text{CO}_2$  data in five subregions (a–e) and the entire South China Sea (f) under study. The subregions are shown in panel (g). The lines indicate the deseasonalized long-term trend of the spatially averaged monthly  $p\text{CO}_2$  data for each subregion, with the slopes shown in Table 3. The deseasonalized method can be found in Landschützer et al. (2016).

## 5 Data availability

The data (the reconstructed  $p\text{CO}_2$  data, the in situ  $p\text{CO}_2$  data before 2018 ( $0.5^\circ \times 0.5^\circ$ ), and the remote-sensing-derived  $\text{CO}_2$  data) for this paper are available at <https://doi.org/10.57760/sciencedb.02050> (Wang and Dai, 2022).

## 6 Conclusions

Based on the machine learning method, we reconstructed the sea surface  $p\text{CO}_2$  fields in the SCS with an  $0.05^\circ \times 0.05^\circ$  spatial resolution over the last 2 decades (2003–2020) by calculating the statistical relationship between the in situ  $p\text{CO}_2$  data and RS-derived data. The input data we used in machine learning include RS-derived data (sea surface salinity, sea surface temperature, and chlorophyll), the spatial patterns of  $p\text{CO}_2$  calculated by EOF, atmospheric  $\text{CO}_2$ , and time labels (month). The machine learning method (CatBoost) used in this study was facilitated by the EOF method, which provides spatial constraints for the data reconstruction. In addition to the typical machine learning performance metrics, we present a novel method for uncertainty calculation that incorporates the bias of both the reconstruction and the sensitivity of reconstructed models to its features. This method effectively shows the spatiotemporal patterns of bias and makes up for the spatial representation of the typical performance metrics.

We validate our reconstruction with three independent testing datasets, and the results show that the bias between our reconstruction and in situ  $p\text{CO}_2$  data in the SCS is relatively small (about  $10 \mu\text{atm}$ ). Our reconstruction successfully captures the main features of the spatial and temporal patterns of  $p\text{CO}_2$  in the SCS, indicating that we can use these reconstructed data to further analyze the effect of meso–microscale processes (e.g., the Pearl River plume and CCC) on sea surface  $p\text{CO}_2$  in the SCS.

We divided the SCS into five subregions, separately calculated the deseasonalized long-term trend of  $p\text{CO}_2$  in each subregion, and compared them with the long-term trend of atmospheric  $p\text{CO}_2$ . Our results show that the reconstructed data are consistent with those of in situ data. Moreover, the strength of the  $\text{CO}_2$  sink in the northern SCS shows an increasing trend, whereas  $p\text{CO}_2$  trends in other subregions are essentially the same as that of atmospheric  $p\text{CO}_2$ .

This high spatiotemporal resolution of sea surface  $p\text{CO}_2$  data is helpful to clarify the controlling factors of  $p\text{CO}_2$  change in the SCS and may be useful to predict changes in  $\text{CO}_2$  source or sink patterns in this system.

**Author contributions.** MD conceptualized and directed the field program of in situ observations. XG and YX participated in the in situ data collection. YB provided the remote-sensing-derived  $p\text{CO}_2$  data. MD, GW, and ZW developed the reconstruction method, wrote

the codes, analyzed the data, and plotted the figures. ZW wrote the paper. MD, XG, and GW contributed to the writing, editing, and revising of the draft of this paper.

**Competing interests.** The contact author has declared that none of the authors has any competing interests.

**Disclaimer.** Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Acknowledgements.** We thank the National Natural Science Foundation of China (grant nos. 42188102, 42141001, and 41890800) and the National Basic Research Program of China (973 Program; grant no. 2015CB954000) for their support.

**Financial support.** This research has been supported by the National Natural Science Foundation of China (grant nos. 42188102, 42141001, and 41890800) and the Dream Project of the Ministry of Science and Technology of the People's Republic of China (grant no. 2015CB954000).

**Review statement.** This paper was edited by Giuseppe M. R. Manzella and reviewed by two anonymous referees.

## References

- Bai, Y., Cai, W., He, X., Zhai, W., Pan, D., Dai, M., and Yu, P.: A mechanistic semi-analytical method for remotely sensing sea surface  $p\text{CO}_2$  in river-dominated coastal oceans: A case study from the East China Sea, *J. Geophys. Res.-Oceans*, 120, 2331–2349, 2015.
- Bakker, D. C. E., Pfeil, B., Landa, C. S., Metzl, N., O'Brien, K. M., Olsen, A., Smith, K., Cosca, C., Harasawa, S., Jones, S. D., Nakaoka, S., Nojiri, Y., Schuster, U., Steinhoff, T., Sweeney, C., Takahashi, T., Tilbrook, B., Wada, C., Wanninkhof, R., Alin, S. R., Balestrini, C. F., Barbero, L., Bates, N. R., Bianchi, A. A., Bonou, F., Boutin, J., Bozec, Y., Burger, E. F., Cai, W.-J., Castle, R. D., Chen, L., Chierici, M., Currie, K., Evans, W., Featherstone, C., Feely, R. A., Fransson, A., Goyet, C., Greenwood, N., Gregor, L., Hankin, S., Hardman-Mountford, N. J., Harlay, J., Hauck, J., Hoppema, M., Humphreys, M. P., Hunt, C. W., Huss, B., Ibáñez, J. S. P., Johannessen, T., Keeling, R., Kitidis, V., Körtzinger, A., Kozyr, A., Krasakopoulou, E., Kuwata, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lo Monaco, C., Manke, A., Mathis, J. T., Merlivat, L., Millero, F. J., Monteiro, P. M. S., Munro, D. R., Murata, A., Newberger, T., Omar, A. M., Ono, T., Paterson, K., Pearce, D., Pierrot, D., Robbins, L. L., Saito, S., Salisbury, J., Schlitzer, R., Schneider, B., Schweitzer, R., Sieger, R., Skjelvan, I., Sullivan, K. F., Sutherland, S. C., Sutton, A. J., Tadokoro, K., Telszewski, M., Tuma, M., van Heuven, S. M. A.

- C., Vandemark, D., Ward, B., Watson, A. J., and Xu, S.: A multi-decade record of high-quality  $f\text{CO}_2$  data in version 3 of the Surface Ocean  $\text{CO}_2$  Atlas (SOCAT), *Earth Syst. Sci. Data*, 8, 383–413, <https://doi.org/10.5194/essd-8-383-2016>, 2016.
- Cao, Z., Dai, M., Zheng, N., Wang, D., Li, Q., Zhai, W., Meng, F., and Gan, J.: Dynamics of the carbonate system in a large continental shelf system under the influence of both a river plume and coastal upwelling, *J. Geophys. Res.-Biogeo.*, 116, G02010, <https://doi.org/10.1029/2010JG001596>, 2011.
- Cao, Z., Yang, W., Zhao, Y., Guo, X., Yin, Z., Du, C., Zhao, H., and Dai, M.: Diagnosis of  $\text{CO}_2$  dynamics and fluxes in global coastal oceans, *Natl. Sci. Rev.*, 7, 786–797, 2020.
- Chen, C. and Borges, A. V.: Reconciling opposing views on carbon cycling in the coastal ocean: Continental shelves as sinks and near-shore ecosystems as sources of atmospheric  $\text{CO}_2$ , *Deep-Sea Res. Pt. I*, 56, 578–590, 2009.
- Chen, C., Lai, Z., Beardsley, R. C., Xu, Q., Lin, H., and Viet, N. T.: Current separation and upwelling over the southeast shelf of Vietnam in the South China Sea, *J. Geophys. Res.-Oceans*, 117, C03033, <https://doi.org/10.1029/2011JC007150>, 2012.
- Chen, F., Cai, W. J., Benitez-Nelson, C., and Wang, Y.: Sea surface  $p\text{CO}_2$ –SST relationships across a cold-core cyclonic eddy: Implications for understanding regional variability and air-sea gas exchange, *Geophys. Res. Lett.*, 34, 265–278, 2007.
- Cheng, C., Xu, P. F., Cheng, H., Ding, Y., Zheng, J., Ge, T., and Xu, J.: Ensemble learning approach based on stacking for unmanned surface vehicle's dynamics, *Ocean Eng.*, 207, 107388, <https://doi.org/10.1016/j.oceaneng.2020.107388>, 2020.
- Dai, M., Gan, J., Han, A., Kung, H., and Yin, Z.: Physical Dynamics and Biogeochemistry of the Pearl River Plume, in: *Biogeochemical Dynamics at Large River-Coastal Interfaces*, edited by: Bianchi, T., Allison, M. and Cai, W. J., Cambridge University Press, Cambridge, 321–352, 2014.
- Dai, M., J. Su, Zhao, Y., Hofmann, E. E., Cao, Z., Cai, W., Gan, J., Lacroix, F., Laruelle, G., Meng, F., Müller, J., Regnier, P., Wang, G., and Wang, Z.: Carbon fluxes in the coastal ocean: Synthesis, boundary processes and future trends, *Annu. Rev. Earth Pl. Sc.*, 50, 593–626, 2022.
- Dai, M. H., Cao, Z., Guo, X., Zhai, W., Liu, Z., Yin, Q., Xu, Y., Gan, J., Hu, J., and Du, C.: Why are some marginal seas sources of atmospheric  $\text{CO}_2$ ?, *Geophys. Res. Lett.*, 40, 2154–2158, 2013.
- Denvil-Sommer, A., Gehlen, M., Vrac, M., and Mejia, C.: LSCE-FFNN-v1: a two-step neural network model for the reconstruction of surface ocean  $p\text{CO}_2$  over the global ocean, *Geosci. Model Dev.*, 12, 2091–2105, <https://doi.org/10.5194/gmd-12-2091-2019>, 2019.
- Dong, L., Su, J., Wong, L. Cao, Z. and Chen, J.: Seasonal variation and dynamics of the Pearl River plume, *Cont. Shelf Res.*, 24, 1761–1777, 2004.
- Du, C., Liu, Z., Dai, M., Kao, S.-J., Cao, Z., Zhang, Y., Huang, T., Wang, L., and Li, Y.: Impact of the Kuroshio intrusion on the nutrient inventory in the upper northern South China Sea: insights from an isopycnal mixing model, *Biogeosciences*, 10, 6419–6432, <https://doi.org/10.5194/bg-10-6419-2013>, 2013.
- Dye, A. W., Rastogi, B., Clemesha, R. E. S., Kim, J. B., Samelson, R. M., Still, C. J., and Williams, A. P.: Spatial patterns and trends of summertime low cloudiness for the Pacific Northwest, 1996–2017, *Geophys. Res. Lett.*, 47, e2020GL088121, <https://doi.org/10.1029/2020GL088121>, 2020.
- Fay, A. R., Gregor, L., Landschützer, P., McKinley, G. A., Gruber, N., Gehlen, M., Iida, Y., Laruelle, G. G., Rödenbeck, C., Roobaert, A., and Zeng, J.: SeaFlux: harmonization of air-sea  $\text{CO}_2$  fluxes from surface  $p\text{CO}_2$  data products using a standardized approach, *Earth Syst. Sci. Data*, 13, 4693–4710, <https://doi.org/10.5194/essd-13-4693-2021>, 2021.
- Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Gregor, L., Hauck, J., Le Quéré, C., Luijkx, I. T., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Alkama, R., Arneeth, A., Arora, V. K., Bates, N. R., Becker, M., Bellouin, N., Bittig, H. C., Bopp, L., Chevallier, F., Chini, L. P., Cronin, M., Evans, W., Falk, S., Feely, R. A., Gasser, T., Gehlen, M., Gkritzalis, T., Gloege, L., Grassi, G., Gruber, N., Gürses, Ö., Harris, I., Hefner, M., Houghton, R. A., Hurtt, G. C., Iida, Y., Ilyina, T., Jain, A. K., Jersild, A., Kadono, K., Kato, E., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Landschützer, P., Lefèvre, N., Lindsay, K., Liu, J., Liu, Z., Marland, G., Mayot, N., McGrath, M. J., Metzl, N., Monacchi, N. M., Munro, D. R., Nakaoka, S.-I., Niwa, Y., O'Brien, K., Ono, T., Palmer, P. I., Pan, N., Pierrot, D., Pocock, K., Poulter, B., Resplandy, L., Robertson, E., Rödenbeck, C., Rodriguez, C., Rosan, T. M., Schwinger, J., Séférian, R., Shutler, J. D., Skjelvan, I., Steinhoff, T., Sun, Q., Sutton, A. J., Sweeney, C., Takao, S., Tanhua, T., Tans, P. P., Tian, X., Tian, H., Tilbrook, B., Tsujino, H., Tubiello, F., van der Werf, G. R., Walker, A. P., Wanninkhof, R., Whitehead, C., Willstrand Wranne, A., Wright, R., Yuan, W., Yue, C., Yue, X., Zaehle, S., Zeng, J., and Zheng, B.: Global Carbon Budget 2022, *Earth Syst. Sci. Data*, 14, 4811–4900, <https://doi.org/10.5194/essd-14-4811-2022>, 2022.
- Gan, J., Li, H., Curchitser, E. N., and Haidvogel, D. B.: Modeling South China sea circulation: Response to seasonal forcing regimes, *J. Geophys. Res.-Oceans*, 111, C06034, <https://doi.org/10.1029/2005JC003298>, 2006.
- Gan, J., Lu, Z., Dai, M., Cheung, A. Y. Y., Liu, H., and Harrison, P.: Biological response to intensified upwelling and to a river plume in the northeastern South China Sea: A modeling study, *J. Geophys. Res.-Oceans*, 115, C09001, <https://doi.org/10.1029/2009JC005569>, 2010.
- Guo, X. and Wong, G.: Carbonate chemistry in the Northern South China Sea shelf-sea in June 2010, *Deep-Sea Res. Pt. II*, 117, 119–130, 2015.
- Han, A. Q., Dai, M. H., Gan, J. P., Kao, S.-J., Zhao, X. Z., Jan, S., Li, Q., Lin, H., Chen, C.-T. A., Wang, L., Hu, J. Y., Wang, L. F., and Gong, F.: Inter-shelf nutrient transport from the East China Sea as a major nutrient source supporting winter primary production on the northeast South China Sea shelf, *Biogeosciences*, 10, 8159–8170, <https://doi.org/10.5194/bg-10-8159-2013>, 2013.
- Hu, J., Kawamura, H., Li, C., Hong, H., and Jiang, Y.: Review on current and seawater volume transport through the Taiwan Strait, *J. Oceanogr.*, 66, 591–610, 2010.
- Jo, Y., Dai, M., Zhai, W., Yan, X., and Shang, S.: On the Variations of Sea Surface  $p\text{CO}_2$  in the Northern South China Sea - A Remote Sensing Based Neural Network Approach, *J. Geophys. Res.-Oceans*, 117, C08022, <https://doi.org/10.1029/2011JC007745>, 2012.
- Jones, S. D., Quéré, C., and Rödenbeck, C.: Spatial decorrelation lengths of surface ocean  $f\text{CO}_2$  results in

- NetCDF format, *Global Biogeochem. Cy.*, 26, GB2042, <https://doi.org/10.1029/2010GB004017>, 2014.
- Landschützer, P., Bakker, D. C. E., Gruber, N., and Schuster, U.: Recent variability of the global ocean carbon sink, *Global Biogeochem. Cy.*, 28, 927–949, 2014.
- Landschützer, P., Gruber, N., and Bakker, D.: Decadal variations and trends of the global ocean carbon sink, *Global Biogeochem. Cy.*, 30, 1396–1417, 2016.
- Landschützer, P., Gruber, N., and Bakker, D. C. E.: An updated observation-based global monthly gridded sea surface  $p\text{CO}_2$  and air-sea  $\text{CO}_2$  flux product from 1982 through 2015 and its monthly climatology, Dataset, [https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/SPCO2\\_1982\\_2015\\_ETH\\_SOM\\_FFN.html](https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/SPCO2_1982_2015_ETH_SOM_FFN.html) (last access: 8 October 2022), 2017.
- Laruelle, G., Lauerwald, R., Pfeil, B., and Regnier, P.: Regionalized global budget of the  $\text{CO}_2$  exchange at the air-water interface in continental shelf seas, *Global Biogeochem. Cy.*, 28, 1199–1214, 2015.
- Laruelle, G. G., Landschützer, P., Gruber, N., Tison, J.-L., Delille, B., and Regnier, P.: Global high-resolution monthly  $p\text{CO}_2$  climatology for the coastal ocean derived from neural network interpolation, *Biogeosciences*, 14, 4545–4561, <https://doi.org/10.5194/bg-14-4545-2017>, 2017.
- Lefèvre, N., Watson, A., and Waston, A.: A comparison of multiple regression and neural network techniques for mapping in situ  $p\text{CO}_2$  data, *Tellus B*, 57, 375–384, 2005.
- Levitus, S., Antonov, J. I., Boyer, T. P., Garcia, H. E., and Locarnini, R. A.: EOF analysis of upper ocean heat content, 1956–2003, *Geophys. Res. Lett.*, 32, L18607, <https://doi.org/10.1029/2005GL023606>, 2005.
- Li, Q., Guo, X., Zhai, W., Xu, Y., and Dai, M.: Partial pressure of  $\text{CO}_2$  and air-sea  $\text{CO}_2$  fluxes in the South China Sea: Synthesis of an 18-year dataset, *Prog. Oceanogr.*, 182, 102272, <https://doi.org/10.1016/j.pocean.2020.102272>, 2020.
- Li, Y., Xie, P., Tang, Z., Jiang, T., and Qi, P.: SVM-Based Sea-Surface Small Target Detection: A False-Alarm-Rate-Controllable Approach, *IEEE Geosci. Remote*, 16, 1225–1229, 2019.
- Luo, X., Hao, W., Zhe, L., and Liang, Z.: Seasonal variability of air-sea  $\text{CO}_2$  fluxes in the Yellow and East China Seas: A case study of continental shelf sea carbon cycle model, *Cont. Shelf Res.*, 107, 69–78, 2015.
- McMonigal, K. and Larson, S. M.: ENSO explains the link between Indian Ocean dipole and Meridional Ocean heat transport, *Geophys. Res. Lett.*, 49, e2021GL095796, <https://doi.org/10.1029/2021GL095796>, 2022.
- Mongwe, N. P., Chang, N., and Monteiro, P.: The seasonal cycle as a mode to diagnose biases in modelled  $\text{CO}_2$  fluxes in the Southern Ocean, *Ocean Model.*, 106, 90–103, 2016.
- Park, J. H.: Effects of Kuroshio intrusions on nonlinear internal waves in the South China Sea during winter, *J. Geophys. Res.-Oceans*, 118, 7081–7094, 2013.
- Qin, H., Chen, G., Wang, W., Wang, D., and Zeng, L.: Validation and application of MODIS-derived SST in the South China Sea, *Int. J. Remote Sens.*, 35, 4315–4328, 2014.
- Rödenbeck, C., Bakker, D. C. E., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S., Landschützer, P., Metzl, N., Nakaoka, S., Olsen, A., Park, G.-H., Peylin, P., Rodgers, K. B., Sasse, T. P., Schuster, U., Shutler, J. D., Valsala, V., Wanninkhof, R., and Zeng, J.: Data-based estimates of the ocean carbon sink variability – first results of the Surface Ocean  $p\text{CO}_2$  Mapping intercomparison (SOCOM), *Biogeosciences*, 12, 7251–7278, <https://doi.org/10.5194/bg-12-7251-2015>, 2015.
- Sutton, A. J., Wanninkhof, R., Sabine, C. L., Feely, R. A., Cronin, M. F., and Weller, R. A.: Variability and trends in surface seawater  $p\text{CO}_2$  and  $\text{CO}_2$  flux in the Pacific Ocean, *Geophys. Res. Lett.*, 44, 5627–5636, <https://doi.org/10.1002/2017GL073814>, 2017.
- Tahata, M., Sawaki, Y., Ueno, Y., Nishizawa, M., Yoshida, N., Ebisuzaki, T., Komiya, T., and Maruyama, S.: Three-step modernization of the ocean: Modeling of carbon cycles and the revolution of ecological systems in the Ediacaran/Cambrian periods, *Geosci. Front.*, 6, 121–136, 2015.
- Telszewski, M., Chazottes, A., Schuster, U., Watson, A. J., Moulin, C., Bakker, D. C. E., González-Dávila, M., Johannessen, T., Körtzinger, A., Lüger, H., Olsen, A., Omar, A., Padin, X. A., Ríos, A. F., Steinhoff, T., Santana-Casiano, M., Wallace, D. W. R., and Wanninkhof, R.: Estimating the monthly  $p\text{CO}_2$  distribution in the North Atlantic using a self-organizing neural network, *Biogeosciences*, 6, 1405–1421, <https://doi.org/10.5194/bg-6-1405-2009>, 2009.
- Wang, G., Shen, S. S. P., Chen, Y., Bai, Y., Qin, H., Wang, Z., Chen, B., Guo, X., and Dai, M.: Feasibility of reconstructing the summer basin-scale sea surface partial pressure of carbon dioxide from sparse in situ observations over the South China Sea, *Earth Syst. Sci. Data*, 13, 1403–1417, <https://doi.org/10.5194/essd-13-1403-2021>, 2021.
- Wang, Z. and Dai, M.: Datasets of reconstructed sea surface  $p\text{CO}_2$  in the South China Sea, *Science Data Bank* [data set], <https://doi.org/10.57760/sciencedb.02050>, 2022.
- Wang, Z., Wang, G., Guo, X., Hu, J., and Dai, M.: Reconstruction of High-Resolution Sea Surface Salinity over 2003–2020 in the South China Sea Using the Machine Learning Algorithm LightGBM Model, *Remote. Sens.*, 14, 6147, <https://doi.org/10.3390/rs14236147>, 2022.
- Wanninkhof, R., Park, G.-H., Takahashi, T., Sweeney, C., Feely, R., Nojiri, Y., Gruber, N., Doney, S. C., McKinley, G. A., Lenton, A., Le Quéré, C., Heinze, C., Schwinger, J., Graven, H., and Khatriwala, S.: Global ocean carbon uptake: magnitude, variability and trends, *Biogeosciences*, 10, 1983–2000, <https://doi.org/10.5194/bg-10-1983-2013>, 2013.
- Yang, W., Guo, X., Cao, Z., Wang, L., Guo, L., Huang, T., Li, Y., Xu, Y., Gan, J., and Dai, M.: Seasonal dynamics of the carbonate system under complex circulation schemes on a large continental shelf: The northern South China Sea, *Prog. Oceanogr.*, 197, 1026–1045, 2021.
- Yu, S., Song, Z., Bai, Y., and He, X.: Remote Sensing based Sea Surface partial pressure of  $\text{CO}_2$  ( $p\text{CO}_2$ ) in China Seas (2003–2019) (2.0), Zenodo [code], <https://doi.org/10.5281/zenodo.7372479>, 2022.
- Zeng, J., Matsunaga, T., Saigusa, N., Shirai, T., Nakaoka, S., and Tan, Z.-H.: Technical note: Evaluation of three machine learning models for surface ocean  $\text{CO}_2$  mapping, *Ocean Sci.*, 13, 303–313, <https://doi.org/10.5194/os-13-303-2017>, 2017.
- Zhai, W., Dai, M., Cai, W. J., Wang, Y., and Hong, H.: The partial pressure of carbon dioxide and air-sea fluxes in the northern South China Sea in spring, summer and fall, *Mar. Chem.*, 96, 87–97, 2005.



- Zhai, W.-D., Dai, M.-H., Chen, B.-S., Guo, X.-H., Li, Q., Shang, S.-L., Zhang, C.-Y., Cai, W.-J., and Wang, D.-X.: Seasonal variations of sea–air  $\text{CO}_2$  fluxes in the largest tropical marginal sea (South China Sea) based on multiple-year underway measurements, *Biogeosciences*, 10, 7775–7791, <https://doi.org/10.5194/bg-10-7775-2013>, 2013.
- Zhan, Y., Zhang, H., Li, J., and Li, G.: Prediction Method for Ocean Wave Height Based on Stacking Ensemble Learning Model, *J. Mar. Sci. Eng.*, 10, 1150, <https://doi.org/10.3390/jmse10081150>, 2022.
- Zhang, C., Hu, C., Shang, S., Müller-Karger, F., Yan, L., Dai, M., Huang, B., Ning, X., and Hong, H.: Bridging between SeaWiFS and MODIS for continuity of chlorophyll-*a* concentration assessments off Southeastern China, *Remote Sens. Environ.*, 102, 250–263, 2006.
- Zhu, Y., Shang, S., Zhai, W., and Dai, M.: Satellite-derived surface water  $p\text{CO}_2$  and air-sea  $\text{CO}_2$  fluxes in the northern South China Sea in summer, *Prog. Nat. Sci.*, 19, 775–779, 2009.