Spatial reconstruction of long-term (2003-2020) sea surface pCO$_2$ in the South China Sea using a machine learning based regression method aided by empirical orthogonal function analysis.

Authors presented a machine learning approach to reconstruct ocean pCO$_2$ over the South China Sea using the new drivers based on EOFs of Remote Sensing-derived pCO$_2$. These new drivers contribute to the estimation accurate pCO$_2$ product at high spatial resolution. The final product represents a monthly 0.05°x0.05°surface ocean pCO$_2$ for the period 2003-2020. The results show a good agreement with validation data and independent observations. Authors discussed the seasonal effect on the reconstruction and mentioned seasonal processes that can affect the ocean pCO$_2$. One of the interesting points in this work is the estimation of uncertainties. Authors introduced the estimation of uncertainties from features used in pCO$_2$ reconstruction. The article is well structured, and it is easy to follow.
However, I found that the article missed the clarity and not all important details are presented or well explained. Below, I listed points that need to be improved and clarified before publication.

[Response]: We thank the reviewer for the positive comments. We have listed our point-by-point responses as of below.

Comments:
- The description and correct definition of data used. In your study you use the data from field survey that you call "observations" or "observed data". Also, you use remote sensing-derived data. However, it is not clear that the data from remote sensing is not direct measurements of pCO$_2$, and it is derived product as you mentioned in 2.3 (line 156). In you abstract you speak about the comparison between "the remote sensing and observed data" (line 23) that is ambiguous. The remote sensing data are observations too and it is not exactly what was used in the paper as it was derived product. I suggest you call the data from filed survey "in situ data", and call the data derived from remote sensing "remote sensing-derived data" everywhere in the manuscript.
[Responds]: The reviewer is right that remote sensing is also an observation tool. Revisions will be made throughout the manuscript.

- Please add more details about how and what exactly was measured during the field survey. Is it the surface fugacity of CO2? If yes, you need to mention it and precise that you estimate pCO$_2$ from fugacity.
[Responds]: We thank the reviewer for the suggestion. The details of the in situ pCO$_2$ data collections were described in Li et al. (2020). In most cruises, pCO$_2$ was measured continuously with a non-dispersive infrared spectrometer (Li-Cor® 7000) or by Cavity Ring-Down Spectroscopy (Picarro G2301) integrated in a GO-8050 system (General Oceanic Inc. USA) onboard research vessels. We will add the following information in our revision "During the cruises, sea surface pCO$_2$ was measured continuously. The measurement and data processing followed those of the SOCAT (Surface Ocean CO2 Atlas, http://www.socat.info/news.html) protocol (Li et al., 2020).".

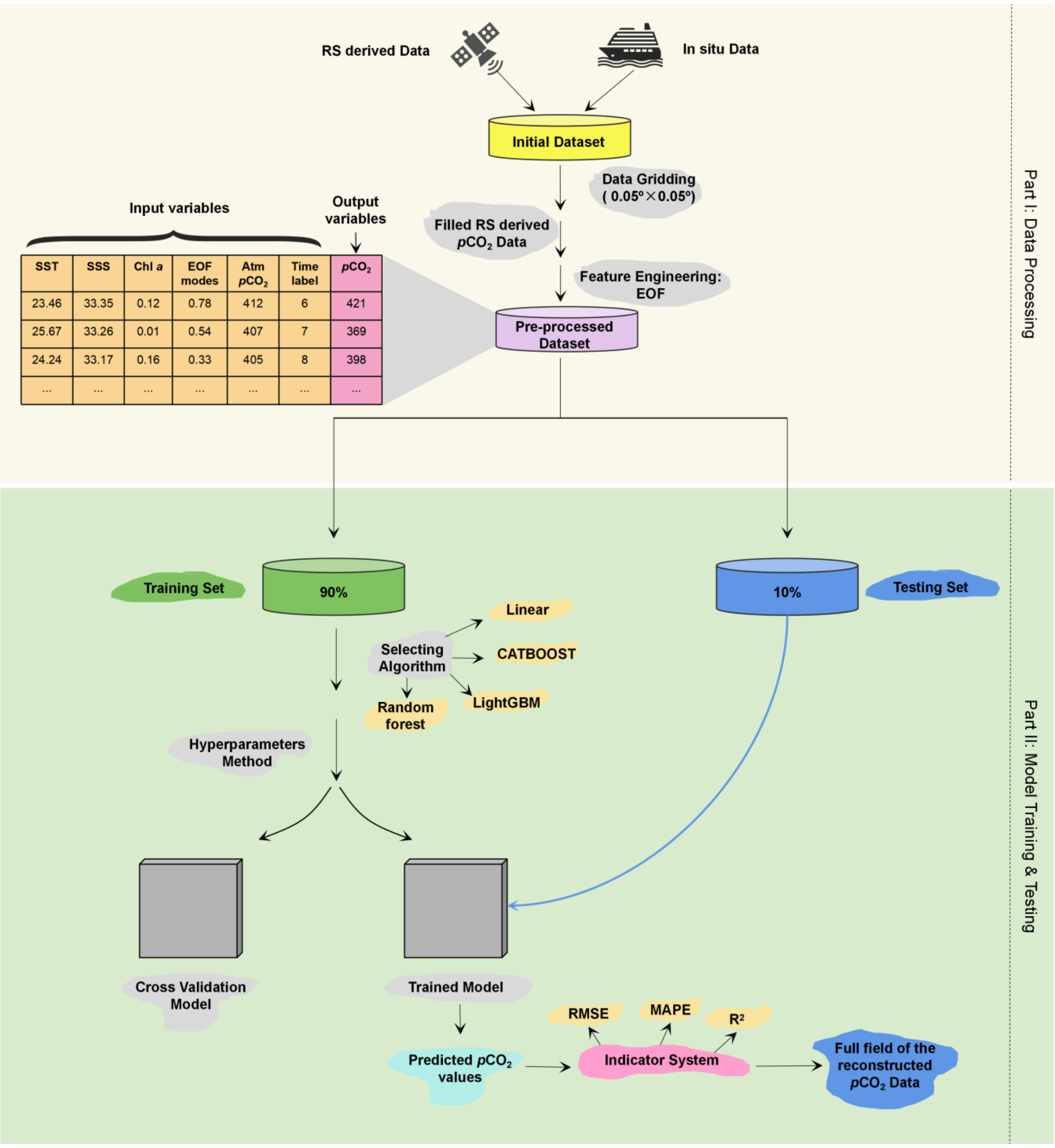
- Please add more details on how remote sensing-derived data were produced. The website you cite in your paper www.SatCO2.com shows only homepage and it is impossible to navigate as all other webpages where we could find details about the product is forbidden. There is a little description of the product in introduction (lines 80-86), however, there is no indication that this product will be used further in the article.
[Responds]: We will add the following information to show how remote sensing (RS)-derived data were produced: "The remote sensing-derived sea surface pCO$_2$ is produced following Yu et al. (2022). The input parameters include sea surface temperature, chlorophyll-a concentration, remote sensing reflectance of three bands (Rrs412, 443, 488 nm), the temperature anomaly in longitude direction, and the theoretical thermodynamic background pCO$_2$ under corresponding SST. Although

the root mean squared errors (RMSE) associated with the RS-derived $p\mathrm{CO}_2$ product were relatively large (21.1 μatm), it successfully showed the spatial distribution of the $p\mathrm{CO}_2$ in China Seas (Yu et al., 2022).”

In the revision we will also add the following information “Wang et al. (2021) demonstrate that the spatial modes of remote sensing-derived data calculated using EOF are effective in providing spatial constraints on the data reconstruction and are thus adopted in this study.” to explain how the RS-derived $p\mathrm{CO}_2$ data were used in this study.

- Please make corresponding changes in Figure 5: observed data to in situ data; RS $p\mathrm{CO}_2$ data to RS-derived $p\mathrm{CO}_2$ data. As you use SSS data reconstructed using ML it is incorrect to put it together with observed SST and Chl-a, or you should precise it in your figure like “ML SSS”.

[Responds]: Accepted and we will modify Figure 5 accordingly (Figure R1). We note that the SSS data over 2003-2020 in the South China Sea used in the present study were reconstructed based on the MODIS-Aqua remote sensing data (Wang et al., 2022). We will add this information in our revision.



SST	SSS	Chl *a*	EOF modes	Atm $p\mathrm{CO}_2$	Time label	$p\mathrm{CO}_2$
23.46	33.35	0.12	0.78	412	6	421
25.67	33.26	0.01	0.54	407	7	369
24.24	33.17	0.16	0.33	405	8	398
...	...	...	...	...	...	...

Figure R1. Procedure for the reconstruction of sea surface $p\mathrm{CO}_2$ using machine learning. RS derived data = remote

sensing-derived data, RMSE = root mean square error, MAPE= mean absolute percentage error, and R^2 = coefficient of determination, and MAE = average absolute error.

- Please add more information on the datasets that you introduced in lines 150-152.

[Responds]: Accepted. We will add more details as follows "In addition to the above in situ sea surface pCO$_2$ data, we selected four independent surveys corresponding to four seasons, September 2018 (fall), December 2018 (winter), August 2019 (summer), and April 2020 (spring), and the in situ sea surface pCO$_2$ data collected in these surveys are used to verify the accuracy of our reconstruction model in extrapolation to periods lacking training datasets. Furthermore, we used another dataset of sea surface pCO$_2$ calculated from observed dissolved inorganic carbon and total alkalinity, during 2003–2019 at the Southeast Asia Time-Series (SEATs) station (data from Dai et al., 2022) to test the long-term consistency of the reconstruction.".

- Figures' captions. Please add more information in figures' captions. Each subplot needs to be introduced in the caption.

[Responds]: Accepted. We will introduce each subplot accordingly.

- Tables. Please keep same number of digits in fractional part for your results in tables: Table 2, Summer RMSE has 3 digits while all other values limited by 2 digits in fractional part. Also, please use the same numbers in the text and in tables, line 163.

[Responds]: Accepted, and we will retain 2 significant digits after the decimal point.

- Abbreviations. Please define abbreviations when you use it for the first time: for example, SSS in line 184.

[Responds]: Accepted. SSS stands for the sea surface salinity. We will define all abbreviations at their first appearances.

- Verification of different regression algorithms. Lines 255-261. To test the capacity of different algorithm you choose the summer season due to its "greatest temporal sampling coverage". However, we can see in your article that there is a strong seasonality in pCO$_2$ distribution. How can you be sure that algorithms will provide the same accuracy during different seasons when other features can become more important?

[Responds]: The reviewer is correct that we performed complementary experiments for other three seasons, showing that the difference resulted from different algorithms for other seasons was minor (<2 μatm in RMSE, Table R1).

Table R1. RMSE associated with between different algorithms in different seasons.

Season	Random Forest	LightGBM	CATBOOST	Multi-linear regression (Wang et al., 2021)
Spring	10.65 μatm	9.52 μatm	8.17 μatm	NaN*
Summer	26.53 μatm	27.83 μatm	16.15 μatm	20.13 μatm
Fall	10.34 μatm	11.56 μatm	10.35 μatm	NaN
Winter	12.48 μatm	12.75 μatm	11.52 μatm	NaN

***NaN stands for the missing value**

In the revision, we will add Table R1 into the MS along with the following information: "From the above options, we chose three ensemble learning algorithms as the machine learning-based regression portion, and multi-linear 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 the different regression algorithms in the pCO$_2$ reconstruction for seasonal training data. We show that in summer, the CATBOOST algorithm yields the best degree of accuracy, with an RMSE of 16 μatm (Table R1). For comparison, the RMSE of LightGBM was 27 μatm, and that of Random Forest was 26 μatm. The RMSE was nearly 20 μatm using the linear regression algorithm employed by Wang et al. (2021). Thus, CATBOOST appears to provide a reliable algorithm for reconstructing $p\mathrm{CO}_2$. Note that different algorithms for other three seasons only resulted in minor difference (~2 µatm in RMSE).".

- Uncertainties. The method to estimate uncertainties should be presented in section 3.4 and not in the section where you discuss your results. In part 1 of equation 6 the function MAX does not do anything as you apply it to a scalar. What is $p\mathrm{CO}_2$_recon in this equation? Does the part 2 of equation 6 represent the sum over the features? Do you base your estimation on the error propagation method (absolute/relative error of a function)? It would be interesting to see the effect of individual features on $p\mathrm{CO}_2$ uncertainties and identify the feature that brings larger bias.

[Responds]: Following suggestions, we will move the method to estimate uncertainties to section 3.5 and modify Equation 6 as follows (Equation R1). And Figure 11 will be modified to Figure R2 to identify the uncertainty caused by each feature.

$$Uncertainty = MAX([\frac{\sum_{i=1,j=1,k=1}^{n}\frac{|OR\_Monthly\_Data(i,j,k)-Obs\_Monthly\_Data(i,j,k)|}{Obs\_Monthly\_Data(i,j,k)}}{num(i)+num(j)},\ldots,\frac{\sum_{i=1,j=1,k=n}^{n}\frac{|OR\_Monthly\_Data(i,j,k)-Obs\_Monthly\_Data(i,j,k)|}{Obs\_Monthly\_Data(i,j,k)}}{num(i)+num(j)}]) * 100\% * pCO2\_recon$$

$$+ \quad \sum_{i=1}^{n}(\frac{\partial pCO2}{\partial Feature_i})dFeature_i \quad . \qquad (R1)$$

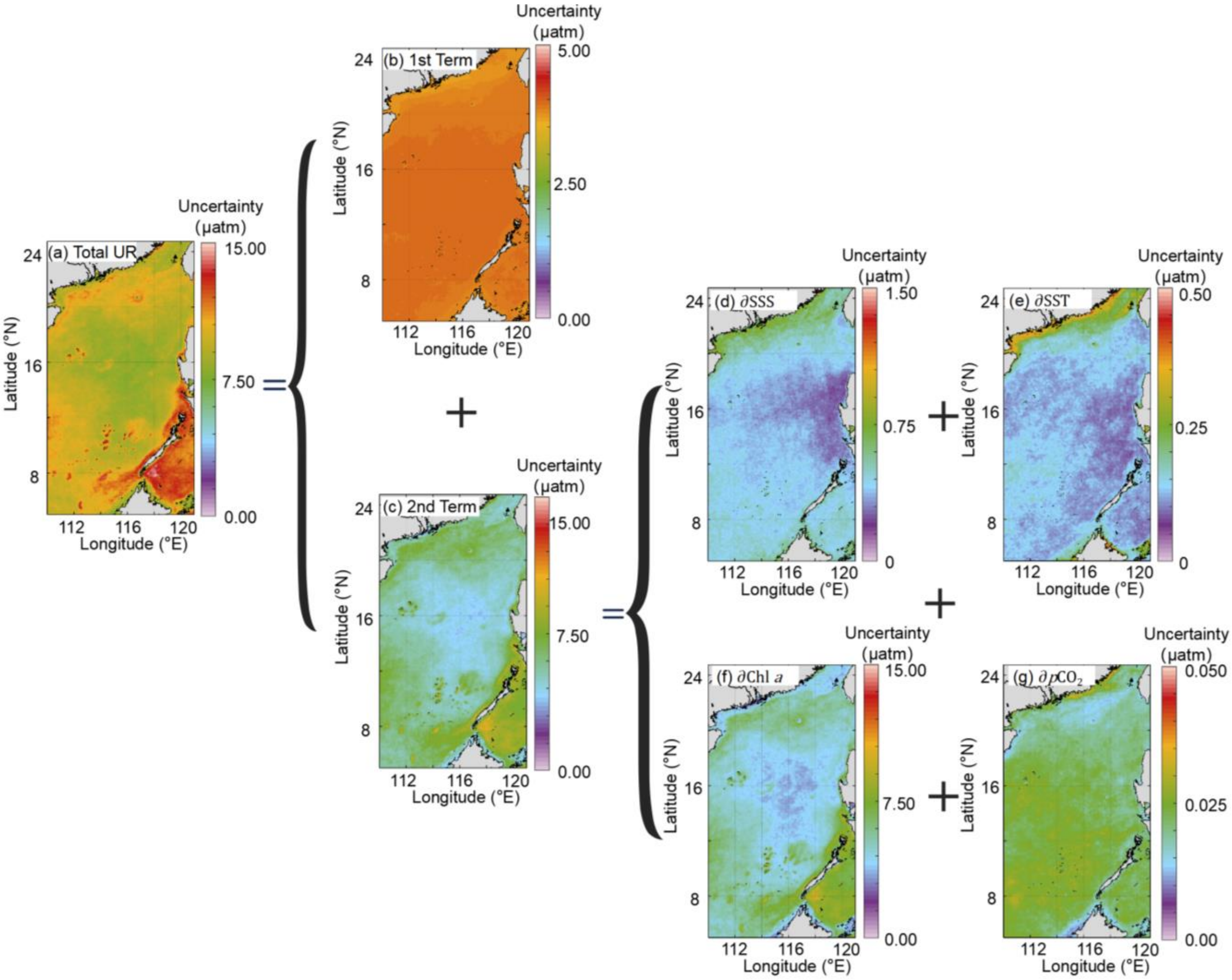


Figure R2. Uncertainties of the reconstructed sea surface $p\mathrm{CO}_2$ fields (a, Total uncertainty in Equation 6; b. the first term of Equation 6; c. the second term of Equation 6; d stands for the $(\frac{\partial pCO2}{\partial SSS})dSSS$ in the the second term of Equation 6; e stands for the $(\frac{\partial pCO2}{\partial SST})dSST$ in the the second term of Equation 6; f stands for the $(\frac{\partial pCO2}{\partial Chl\ a})dChl\ a$ in

the the second term of Equation 6; g stands for the $(\frac{\partial pCO2}{\partial RS\_derived\_pCO2})dRS\_derived\_pCO2$ in the the second term of Equation 6.

For the first term in Equation R1, k stands for kth month, $OR\_Monthly\_Data(i,j,k)$ stands for the kth monthly reconstructed data at longitude(i) and latitude(j), and $Obs\_Monthly\_Data(i,j,k)$ stands for the kth monthly in situ data at longitude (i) and latitude (j). Therefore, the *MAX* in first term stands for that the maximum value between the k monthly bias ratios. And 'pCO$_2$*_recon*' stands for the reconstructed CO$_2$ data.
The second term in Equation R1 represents the sum over the features. According to Equation R1, the bias of RS derived pCO$_2$ used in the second term of Equation R1 is ~21 µatm (Table 2), the bias of SST is ~ 0.27° (Qin et al., 2014), the bias of SSS is ~0.33 (Wang et al., 2022), and the bias of Chl-a is ~115% (Zhang et al., 2006), and the results can be found in Fig. R1. of the overall uncertainty (Fig. R1 a) is greater in the coastal area (~13 µatm) than in the basin (~10 µatm). And this spatial pattern is mainly determined by the second term. The spatial distribution of the first term in Equation R1 (Fig. R1 b) calculated from a "max bias ratio" is consistent with that of pCO$_2$. The second term in Equation R1 (Fig. R1 c) is calculated from the propagation of bias of each variable. The bias of Chl *a* (Fig. R1 f) shows the greatest effect on the reconstruction between features. Although the bias of RS derived pCO$_2$ has relatively large bias, the final influence of its bias on the reconstruction model results is negligible due to the EOF method (Fig. R1 g).
We will include this description of uncertainty in Section 4.3 of the paper.

- Conclusion. Line 424, please specify which machine learning method. Line 426, please specify that you used remote sensing-derived data.
[Responds]: Accepted. The machine learning method is CATBOOST, and the input data we used in machine learning includes remote sensing derived data (sea surface salinity, sea surface temperature, chlorophyll), the spatial patterns of pCO$_2$ calculated by Empirical Orthogonal Function, atmospheric CO$_2$, and time labels (month). We will specify these information in the revision.

- Data pre-processing. Are the data used in ML method pre-processed: interpolated on the same grid, normalized?
[Responds]: Yes, all the data used in ML were interpolated on the same grid. In the revision, we will add this information "All these data used in machine learning have been interpolated on the same grid".

- Line 148: "relatively low pCO$_2$", what does it mean, how low is it?
[Responds]: "relatively low pCO$_2$" means ~350 µatm. We will add this info in the revision.

- Line 164: "current algorithm", please precise, what algorithm are you talking about.
[Responds]: It refers to mechanic semi-analytical algorithm (MeSAA) and non-linear regression. In the revision, we will add this information.

- Line 187: "our observed data", please precise which data.
[Responds]: "our observed data" stands for the in situ data. In the revision, we will make modifications accordingly.

- You should mention in section 2.3 that there is a section 3.1 where you explain how you fill missing points in RS-derived pCO$_2$ product.
[Responds]: Accepted. We will mention this information in the end of section 2.3.

- Could you please provide a figure to show the distribution of training samples you mentioned in lines 201-202: “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 depth contour.”

[Responds]: Accepted, and such a figure will be added (Figure R3).

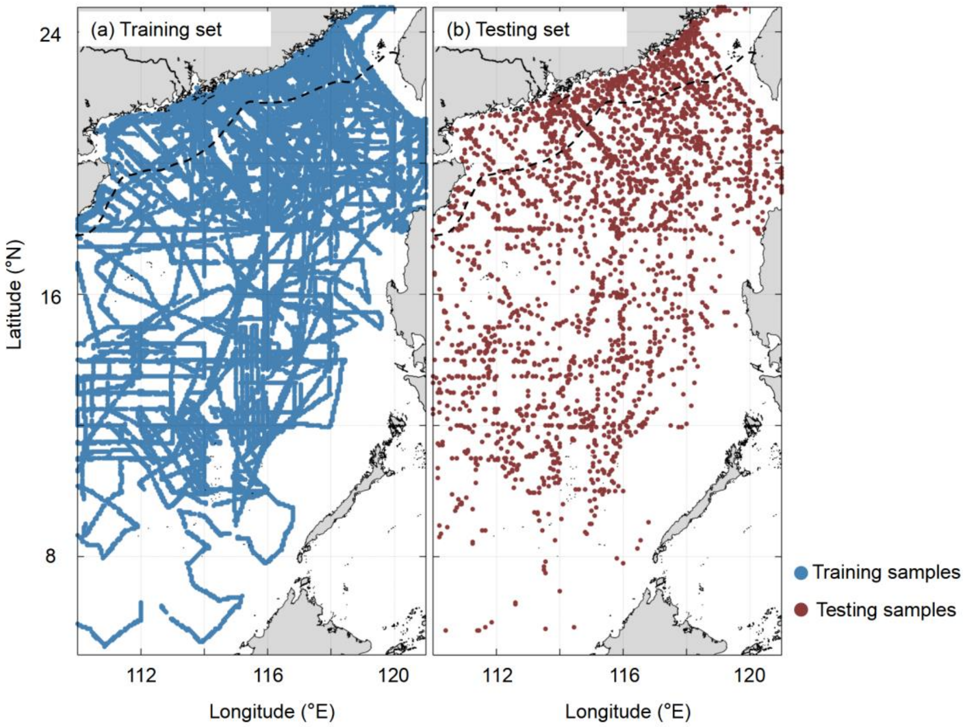


Figure R3. Spatial distribution of training samples (a) and testing samples (b). The black dash line stands for the 200m depth contour.

- Figure 8: It is difficult to see the results for test set. The results for training set look very similar and homogeneous, I would suggest keep only test set here.

[Responds]: Accepted, and we will only keep the results of test sets in Figure 8 as shown in Figure R4.

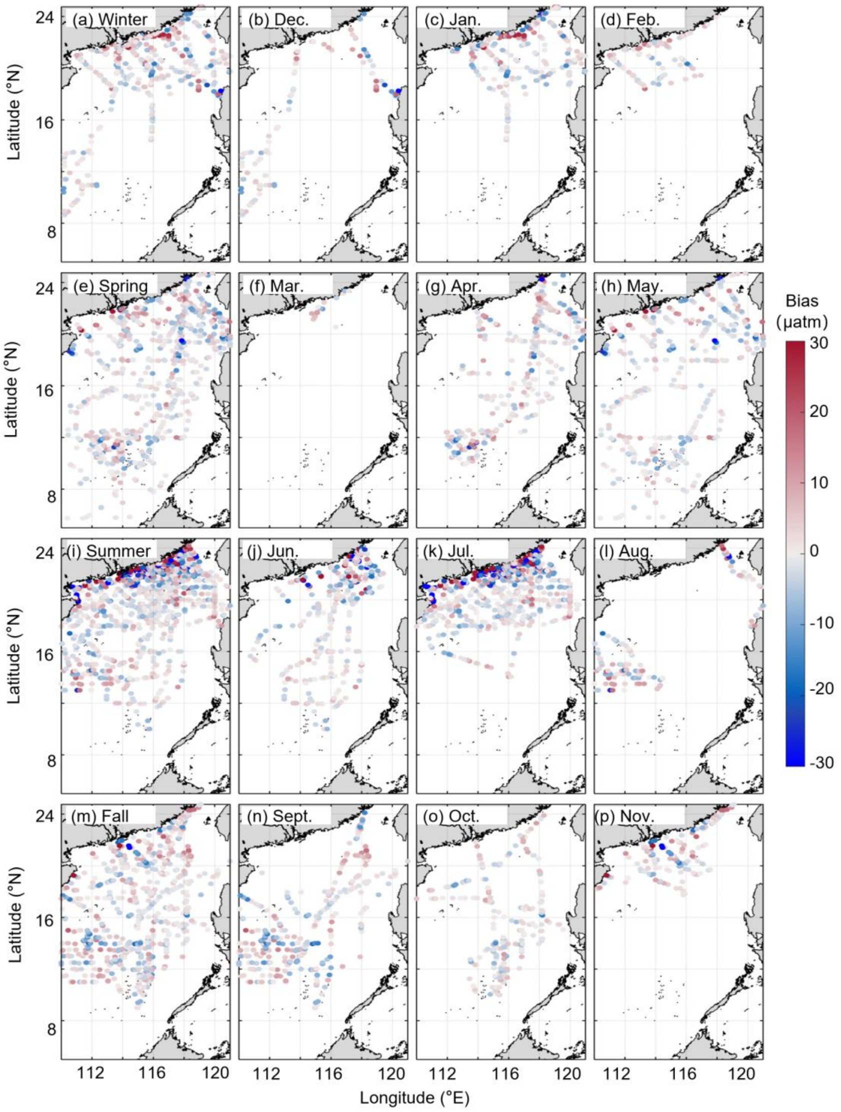


Figure R4. Differences between reconstructed seasonal and monthly pCO$_2$ and the in situ pCO$_2$ for the test set (a. winter; b. December; c. January; d. February; e. Spring; f. March; g. April; h. May; i. Summer; j. June; k. July; l. August; m. Fall; n. September; o. October; p. November).

- Line 322: "The greatest bias occurs in the Pearl River plume area in summer". Could you please indicate how large is this bias?

[Responds]: It is about 35 µatm. We will add this information in the revision.

- Line 323: what is tpCO2?

[Responds]: It should be 'the pCO$_2$', and we will remove this typo in the revision.

- Line 376: you say here that "the lowest value occurs in January", in the next sentence you say "pCO$_2$ first decreases in December and then increases in January". It means that the lowest value is in December. Please clarify it.

[Responds]: It is a typo, "pCO$_2$ decreases in December and then increases in January" should be "then increases after January". The lowest value is in January. In the revision, we will make these corrections.

- Line 417: "…a source or sink of atmospheric CO2 is influenced by seasonal changes and physical processes". Please specify seasonal changes and physical processes.

[Responds]: Accepted. We will add more details as follows "Subregion_B can be a significant sink zone of atmospheric CO_2 as demonstrated by its low sea surface $p$$CO_2$ when the Pearl River plume is spreading into a wider spatial coverage in summer. In contrast in winter when the Kuroshio intrusion is strong, both subregions B and D have high sea surface $p$$CO_2$, indicating that both subregions are sources of atmospheric CO_2."

Typo and style:

Line 15: I would suggest using word "sparse" instead of "incomplete". Line 37: Please change "…annually mitigates…" to "…annually mitigated…" as you refer to the concrete period of 1960-2019; or change the sentence completely.

[Responds]: Accepted. We will use "sparse" instead of "incomplete", and change "…annually mitigates…" to "…annually mitigated…".

Line 50: "Numerical ocean models of performance.." Please remove "of performance".

[Responds]: Accepted. We will remove "of performance" in this sentence.

Line 53: I would not use the word "alternative". The data-based approaches are different methods to study ocean biogeochemistry that can be complementary to biogeochemical models.

[Responds]: Accepted. We will rewrite this sentence as follows "data-based approaches have become an important complementary to numerical models"

Line 119: "CCC, yellow line in Fig. 1". There is no yellow line in Fig. 1. CCC corresponds to the green line.

[Responds]: The reviewer is correct, and we will make the corrections.

Fig. 2: Please change the name of your colorbar to "number of data".

[Responds]: Accepted. We will change the name of our colorbar in Fig.2 to "number of data"

Line 145: "Spatially, the $p$$CO_2$ distribution in the basin is relatively homogeneous, but shows large variability in the northern region". I suppose you meant "Spatially, the $p$$CO_2$ distribution in the basin is relatively homogeneous with large variability in the northern region".

[Responds]: Accepted. We will rewrite this sentence as your suggestion.

Line 288: Please change "the continuity changes.." to "the continuous changes".

[Responds]: Accepted. We will change "the continuity changes" to "the continuous changes".

Line 300: Please add that these estimations are over the seasons.

[Responds]: Accepted. We will add this information in the revision.

Line 322: Please change "The greatest bias" to "The largest bias".

[Responds]: Accepted. We will change "The greatest bias" to "The largest bias".

Line 358: Please change "Equation (7)" to "Equation (6)".

[Responds]: Accepted. We will make these corrections in the revision.

Line 408: Missing space between "uncertainty" and "is".

[Responds]: Accepted. We will remove this typo in the revision.

References

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, 2022. https://doi.org/10.3390/rs14236147.

Yu, S., Song, Z., Bai, Y., and He, X.: Remote Sensing based Sea Surface partial pressure of CO2 ($p$$CO_2$) in China Seas (2003-2019) (2.0). Zenodo, 2022. https://doi.org/10.5281/zenodo.7372479.