Spatial reconstruction of long-term (2003-2020) sea surface pCO2 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 pCO2 over the South China Sea using the new drivers based on EOFs of Remote Sensing-derived pCO2. These new drivers contribute to the estimation accurate pCO2 product at high spatial resolution. The final product represents a monthly 0.05°x0.05° surface ocean pCO2 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 pCO2. One of the interesting points in this work is the estimation of uncertainties. Authors introduced the estimation of uncertainties from features used in pCO2 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.

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 pCO2, 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.

  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 pCO2 from fugacity.

  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.

  Please make corresponding changes in Figure 5: observed data to in situ data; RS pCO2 data to RS-derived pCO2 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".

  Please add more information on the datasets that you introduced in lines 150-152.
- Figures' captions. Please add more information in figures' captions. Each subplot needs to be introduced in the caption.

- 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.
- Abbreviations. Please define abbreviations when you use it for the first time: for example, SSS in line 184.
- 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 pCO2 distribution. How can you be sure that algorithms will provide the same accuracy during different seasons when other features can become more important?
- 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 pCO2_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 pCO2 uncertainties and identify the feature that brings larger bias.
- Conclusion. Line 424, please specify which machine learning method. Line 426, please specify that you used remote sensing-derived data.
- Data pre-processing. Are the data used in ML method pre-processed: interpolated on the same grid, normalized?

- Line 148: “relatively low pCO2”, what does it mean, how low is it?
- Line 164: “current algorithm”, please precise, what algorithm are you talking about.
- Line 187: “our observed data”, please precise which data.
- 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 pCO2 product.
- 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.”
- 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.
- Line 322: “The greatest bias occurs in the Pearl River plume area in summer”. Could you please indicate how large is this bias?
- Line 323: what is tpCO2?
- Line 376: you say here that “the lowest value occurs in January”, in the next sentence you say “pCO2 first decreases in December and then increases in January”. It means that the lowest value is in December. Please clarify it.
- Line 417: “…a source or sink of atmospheric CO2 is influenced by seasonal changes and physical processes”. Please specify seasonal changes and physical processes.

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.

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

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.

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

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

Line 145: "Spatially, the pCO2 distribution in the basin is relatively homogeneous, but shows large variability in the northern region". I suppose you meant "Spatially, the pCO2 distribution in the basin is relatively homogeneous with large variability in the northern region".

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

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

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

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

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