The reader has made valuable comments. We have recalculated fluxes of all products used in the comparison, compared fluxes of NIES-ML2 obtained by CCMP and ERA5 wind, and revised the manuscript to include the changes. The followings are out point-to-point response to the reader's comments.

1. Our validations were discussed in the "Model Performance" section. Maybe because of the title, the reader thought we did set data aside for validation. Now we changed the title to "Model Validation". There are different ways to validate a machine learning model. A common method is to randomly divide a data set into two parts, one for training and another for validation. We believe this is not a good practice as random sampling would make the two parts having the same distribution and a validation won't fail unless no relationship exists between the target and the predicters. We believe it would better that the sampling domain of the testing data differs from that of the training data. We used a so-called leave-one-year-out method, that is that for each year from 1980 to 2020, one year's data was set aside for validation and others for training. Thus 41 validations were done. The results were summarized in Table 2 of the manuscript.

2. Regarding model configuration. It is a complicated issue and there is no universal method to obtain the "best configuration" for a given problem. We selected the configuration parameters for RF based on our experience with using RF for global forest GPP reconstruction and the lesson we learnt from ocean CO2 reconstruction in the past. Few data are available in the southern oceans in some months (we added an example the supplement material). An overfitting would result in hot spots in the areas. So we raised the default number of trees and end-node leaves to prevent the problem. As GBM is also a tree-based model, we opted to use the same parameters. As for the FNN, we wrote the model code used in Zeng et al. (2014) and did extensive test at that time. The full-batch method of the code is very slow with the data size of this study. It would take a few months if we used the code to do the same calculations (a lot of iterations are involved for a rate extraction of many years). We switch to use python's MLPRegressor. Its mini-batch method training is much faster. We did a few tests by comparing its results with our old program and figuration seems working well.

3. We did water vapor correction. The expressions on the matter in our manuscript are misleading. (We revised them.) The CO2 from NOAA's Marine Boundary Layer

Reference is mole fraction (xCO2). We converted xCO2 to pCO2 by pCO2=xCO2*(P_s-P_{h2o}). The vapor pressure of seawater P_{h2o} was calculated by the method of Weiss and Price (1980).

4. We recalculated out fluxes with α=0.271 for ERA5 wind and α=0.257 for CCMP wind. We also recalculated fluxes of the products used for comparison with the same method and adjusted the fluxes as if the products have the same spatial coverage of NIES-ML3 so that they can be put together in one figure for comparison.

5. As the reader pointed out, pCO2 of HOT, BATS, and GLODAP are not directly measured. We don't think it is logical to use them for validation. It would be difficult to judge that a disagreement is resulted from the method or from the systematic difference between SOCAT and GLODAP. It would be better to merge the two datasets for CO2 reconstruction. Even doing so won't solve the data scarcity program in southern oceans in boreal summer (refer to supplement Fig. 1d and https://essd.copernicus.org/articles/12/3653/2020/#section5&gid=1&pid=1). Also, using data from a few sites cannot draw any statistically sound conclusion. In our validations, when a year's data was set aside for testing, the bias is far from zero even though thousands of data point were included. But the overall bias of 41 validations is negligible.

6. The reader also suggests comparing our results with those of Denman et al. (2007) and Gruber et al. (2019). Unfortunate their time series are short. Gruber's publication is quite new, but the data used is more than 10 years older. Our study emphasizes how the annual increase rates of CO2 used or embedded in different methods could affect the long-term variation of CO2 flux. That's the reason we didn't use SeaFlux for comparison in the first place because the product only include estimate after 1990.