Guan et al. utilize an ensemble machine learning method to estimate a long-term global atmospheric column carbon dioxide dataset based on multi-source data. The comprehensive validation of predicted XCO_2 confirmed the generalization of the model and the reliability of the dataset. In particular, the dynamic normalization strategy significantly improves the performance of the model. The application of the dataset reveals significant seasonal distribution and long-term changing trends in global XCO2. However, there are some issues that the authors need to address before the manuscript can be considered for publication. The specific comments are listed below.

1. L77-78: The data on ocean area is the highlight of the study in the Introduction, but in the text, in addition to considering the characteristics of CHL-a as a variable, there is no verification and application analysis for the ocean. OCO-2, carbon tracker, and CAMS are also covered in marine areas. It is suggested that the authors reorganize this part to better highlight the innovation aspect.
2. L175, 177: What does “each period” refer to? It is for each year?
3. L184: In Step 3, were the variables of CAMS and CHLEVI removed to train another model based on the training dataset between 2015 to 2018 for the mapping from 2000 to 2002?
4. L312: why the validation results of the supplementary model (SR0002), which removed the CAMS and CHLEVI variables, perform better (R^2 is 0.898 in Fig. S3c, and slope is 1.27 in Fig. S3d) in the temporal extension validation compared to the main model (R^2 is 0.886 in Figure 6c, and slope is 1.79 in Figure 6d). It seems that model SR0002 can better solve the problems of low-valued and high-value overvaluation of machine learning models. It is also confirmed by the comparison between Figure 7c, j and Fig. S4c, j (with lower RMSE values). Perhaps model SR0002 has a better performance for the prediction with temporal extension. I suggest that the ranges of the color bar between Figure 5 and Fig. S2 be consistent for the comprehensive comparison.
5. Fig. S5, 6: What is the time range in the validation for each TCCON station of models SR0002 and SR0320?
6. L237: Why are there only 150000 data that were randomly selected from other regions instead of all data to train the model for the spatial expansibility validation?

7. L276-L279: It is suggested that the spatial expansibility verification can provide evaluation results on a grid scale, rather than on the shape of the satellite observations for supporting the results of the continental regions with high RMSE and lower R^2.
8. L174: The dynamic normalization strategy has a surprising improvement for the modeling. Is this the author's original contribution? If so, please provide the relevant background in the introduction. If not, please add a reference. And why is the performance of predicted XCO_2 without dynamic normalization worse than that of the input variables (CT XCO_2 and CAMS XCO_2) between 2019 to 2020 (Figure 6)? What is the result of the training set (2015-2018)? Are the hyperparameters of the model with dynamic normalization the best? Is there any overfitting in this model?
9. Section 3.4.3 and L252, I'm curious about the reason for the higher R^2 in the extrapolation periods compared to the training periods. The annual verification accuracy between TCCON station observations and OCO-2 satellite estimates in matching grids may explain this.
10. L369: Please add the calculation method of trends, and the statistically significant (P values).
11. Section 4.1: The evaluation indicators and spatial distribution between Stacking and ETR models are very similar. Can authors provide additional evidence that the stacking model yields better prediction, such as the ability to predict extreme values, the spatial difference between the stacking model and the ETR model, and so on?
12. L468: I suggest that authors add a comparison table between the predicted dataset and previous studies in the Supplement, summarizing spatiotemporal resolution, time span, whether it contains the ocean, verification method and accuracy, etc. Especially, the comparison of the ocean and resolution can support your Introduction.
13. L475-476: The expression "we developed a novel validation method to evaluate the spatiotemporal extensibility …" may not be accurate enough. As far as I know, many atmospheric studies have used similar temporal and spatial evaluation methods.