This work by Stamell et al. compares the performance of three machine learning approaches, i.e., feed forward neural network (NN), XGBoost (XGB) and random forest (RF), based on the Large Ensemble Testbed. The authors did a lot of work, however, there are many unclear parts in the manuscript.

Major comments:

1. The literature review in this manuscript only mentioned previous studies using SOM-FFN to interpolate the pCO_2 field. What about the other methods, especially the three methods tested in this study? Have they been used in estimating pCO_2 field before? What are the major improvements of this study?
2. The SOM-FNN performed well in interpolating the pCO_2 field, but is likely to overestimate in the Southern Ocean. Is this issue improved in the three methods from this study?
3. I am a little confused about the data used to test the three ML methods. What are the target data or ground truth data when training the model? The data from the Large Ensemble Testbed (LET) are the ensemble of Earth system models, which are not observational data. While the SOCATv5 data product, which are actual measurements data, seems not to be included in the model training. Please clarify.
4. How are the train (60%), validate (20%) and test data (20%) split? Are they spatial-temporal randomly divided, or according to the locations or times? Different split methods lead to the evaluation of different model abilities. Split according to locations indicates the model's ability in spatial interpolation, while split according to times indicates the model's ability in temporal prediction. Please clarify.

Minor comments:

What is the sample size of the data?

Line 196: "fianlly" should be "finally"