Comments:
Stamell et al. compared three machine learning methods in interpolating surface pCO2 in the global ocean. The manuscript is lack of novelty and suffers many technical difficulties.

Major comments:

1) In essence, the authors developed three pCO2 models using three machine learning (ML) approaches (NN, RF, and XGB), and evaluated their overall performance in monitoring the seasonal, sub-decadal, decadal variabilities of surface pCO2. There are too many such studies in the published literature. In fact, the three machine learning (ML) methods presented in this study were already compared in Gregor et al. (2019), in which 6 supervised ML methods (including the three used in this study) were applied to reconstruct global surface pCO2. The authors concluded that all methods had overestimation in the reconstructed pCO2, yet Gloege et al (2020) also found overestimation using NN. So what is the overall novelty and significance of the present study?

2) There are lots of technical difficulties. The authors stated that Large Ensemble Testbed (LET) consists of 100 members across four initial-condition ensemble models, and each member is a representation of the real ocean climate system. In my understanding, data from these members are actually from the Earth system models. How accurate are these modeled data particularly those (SST, SSS, MLD, Chl-a, .etc) used to train the pCO2 model? How accurate are the pCO2 from the LET comparing to the *in situ* observations (SOCAT v5)? Without any evaluation, it is questionable to say these modeled data represent the real ocean system. I am not an expert on Earth system models, but why the authors say '100-member LET consists of 25 randomly selected member …' (L114-115)? What is difference between these members? The authors argued that the use of many members was to test the reconstruction capabilities of the ML across different ocean states, however, what is the impact of ocean state differences on the reconstructed pCO2? Also, there are some technical words that are quite difficult to follow without clear explanation (e.g., full field driver data, unseen data, LET). The ML was trained based on grid data at 1 by 1 degree, what is the impact of real spatial variability within the grid on the uncertainties of the reconstructed pCO2?

3) As to the overall structure of the manuscript, the authors presented details of the three ML methods in both Introduction and Methods. The earth system modeled data and SOCATv5 data are not well described, for example, the data coverage both spatially and temporally, and why they are used. In the ML approaches, again, why these three approaches were selected?

Specific comment:
L244: Statistics to the 'unseen' data is different from those listed in Table 1.