Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-311-AC2, 2021

Geoscientific Model Development Discussions

Interactive comment on "Strengths and weaknesses of three Machine Learning methods for pCO_2 interpolation" *by* Jake Stamell et al.

Jake Stamell et al.

jake.stamell@columbia.edu

Received and published: 4 February 2021

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 pCO2 field. What about the other methods, especially the three methods tested in this study? Have they been used in estimating pCO2 field before? What are the major improvements of this study?

Thank you for this comment. In the introduction, we also discuss Gregor et al. 2019 in which methods in the class of XGB and RF were applied. Also, in response to the comment from Reviewer 1, we clarify that the goal of this work is to quantify extrapolation uncertainty for NN, RF and XGB applied to surface ocean pCO2.

2. The SOM-FNN performed well in interpolating the pCO2 field, but is likely to overestimate in the Southern Ocean. Is this issue improved in the three methods from this study?

Thank you for this comment. In the Conclusion, we have expanded the third paragraph to explicitly address this point.

"Decadal variability is of particular interest to the ocean carbon cycle community (Landschützer et al., 2015; Gruber et al., 2019). We have previously shown that the commonly-used SOM-FFN observation-based pCO2product (Landschützer et al., 2016) likely overestimates the amplitude of Southern Ocean decadal variability due to data sparsity (Gloege et al., 2021). Here, we also find overestimation of the amplitude of decadal variability for all approaches. Nonetheless, we do find that the NN performs slightly better than XGB (Figure 4). The creation of a non-linear mapping, without creating distinct regions in driver data space, appears to lead the NN to better extrapolate to the poorly-sampled decadal timescale."

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.

Thank you for this comment. We now state specifically in the introduction that SOCATv5 data are not directly used. We use only the pattern of sampling of pCO2 that occurred in SOCATv5. All data are drawn from the LET ensemble members. We have clarified in the text that the goal of this work is to ask the question – if we only have

pCO2 as sampled in the real world, how likely is it that we could reconstruct global coverage pCO2? What is the relative skill of one method vs. another? If test data comparisons indicate a certain level of skill, does this represent the skill in extrapolation? Clearly, there are not enough data from the real ocean to evaluate extrapolation. So, we use a proxy – Earth System Models – that represent the physical and biogeochemical processes responsible for pCO2 evolution in the real world.

We do not attempt to predict real-world pCO2 in this effort. Our goal is to understand the skill of methodologies currently in use to make such predictions. We have endeavored to clarify this throughout.

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.

The split is random in both space and time. We add clarification to the text of section 2.2. and A2 to clarify this.

Minor comments:

What is the sample size of the data? Line 196: "fianlly" should be "finally"

Thank you, we have corrected this typo. We have also added the dataset size. It is ~ 14.2M total, with ~ 220K for training, validation, and testing.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-311, 2020.