From: editorial@copernicus.org
Subject: gmd-2020-311 (author) - manuscript not accepted for final publication
Date: February 9, 2021 at 5:42 AM
To: mckinley@ldeo.columbia.edu

You are receiving the following email copy due to your co-authorship of the manuscript gmd-2020-311. The original message was sent to the contact author defined upon manuscript registration. Please contact us in case of any discrepancies with regard to the manuscript.

Dear Jake Stamell,

We regret to inform you that your following submission was not accepted for final publication in GMD:

Title: Strengths and weaknesses of three Machine Learning methods forpCO2 interpolation
Author(s): Jake Stamell et al.
MS No.: gmd-2020-311
MS type: Development and technical paper
Iteration: Revised Submission

The Topical Editor's explanation for the rejection can be found in your MS overview. Please log in with your Copernicus Office user ID at: https://editor.copernicus.org/GMD/my_manuscript_overview

We thank you very much for your understanding and hope that you will consider GMD again for the publication of your future scientific papers.

In case any questions arise, please do not hesitate to contact me.

Kind regards,

The editorial support team
Copernicus Publications
editorial@copernicus.org

[My MS overview] [My MS archive] [My subjects & expertise]

MS Records

gmd-2020-311 Submitted on 16 Sep 2020

Strengths and weaknesses of three Machine Learning methods forpCO2 interpolation

Jake Stamell, Rea Rustagi, Lucas Gloege, and Galen McKinley

First Contact: Jake Stamell

Agreed licence: Creative Commons Attribution 4.0 International

Handling Topical Editor: Xiaomeng Huang

Manuscript type: Development and technical paper

Status: Rejected (GMD) **Iteration: Revised Submission**

Revised Submission

Topical Editor Decision: Reject (09 Feb 2021) by Xiaomeng Huang
Comments to the Author:
Dear authors:

Thank you for submitting the manuscript "Strengths and weaknesses of three Machine Learning methods for pCO2 interpolation" to GMD. Based on the referees' rating and your response, I am declining this manuscript for publication in GMD.

This manuscript compares the advantages and disadvantages of the three approaches, however it is not innovative enough for the data reconstruction. I am enclosing the reviews, which you may find helpful if you decide to revise the manuscript and submit to another journal. I am sorry that I cannot be more encouraging at this time.

Thank you for your interest in GMD.

Sincerely,

Xiaomeng Huang
Handling Topical Editor

Uploaded Files validated (09 Feb 2021) by Svenja Lange

File Upload (08 Feb 2021) by Jake Stamell

- Manuscript
- Author's Response
- Author's tracked changes
- Abstract

Interactive Discussion

Final Response (19 Jan 2021) waiting for final Author Comment

Discussion started (22 Oct 2020), expected end 13 Jan 2021, extended until 22 Jan 2021

- Interactive Discussion

Minimum number of Referee Reports required: 2

Nominated Referee
nominated 10 Nov 2020, missed nomination deadline

Nominated Referee
nominated 01 Nov 2020, missed nomination deadline

Nominated Referee
nominated 20 Nov 2020, missed nomination deadline

Nominated Referee
nominated 07 Dec 2020, missed nomination deadline

Nominated Referee
nominated 19 Nov 2020, missed nomination deadline

Nominated Referee
nominated 22 Oct 2020, declined 23 Oct 2020

Nominated Referee
nominated 19 Nov 2020, declined 19 Nov 2020

Nominated Referee
nominated 30 Nov 2020, declined 04 Dec 2020

Nominated Referee
nominated 22 Oct 2020, declined 23 Oct 2020

Nominated Referee
nominated 23 Oct 2020, declined 23 Oct 2020

Nominated Referee
nominated 23 Oct 2020, declined 23 Oct 2020

Nominated Referee
nominated 23 Oct 2020, nomination terminated by Topical Editor

Nominated Referee
nominated 23 Oct 2020, nomination terminated by Topical Editor

Anonymous Referee #2
nominated 07 Dec 2020, accepted 15 Dec 2020, report 19 Jan 2021 **[Report #2]**

Anonymous Referee #1
nominated 01 Nov 2020, accepted 03 Nov 2020, report 18 Nov 2020 **[Report #1]**

Minor Revision

Preprint posted (22 Oct 2020) by Caren Zimara
▸ Preprint

Production File Validation completed (22 Oct 2020) by Lorena Grabowski

File Upload (22 Oct 2020) by Jake Stamell

Short Summary: Using simulated surface ocean pCO2 from Earth System Models, we test three Machine Learning methods [more]

[Assets]

Topical Editor Initial Decision: Start review and discussion (20 Oct 2020) by Xiaomeng Huang

Uploaded Files validated (20 Oct 2020) by Svenja Lange

File Upload (19 Oct 2020) by Jake Stamell
▸ Manuscript

▸ Author's Response

Initial Submission

Topical Editor Initial Decision: Start review and discussion after revisions (review by editor) (13 Oct 2020) by Xiaomeng Huang
Comments to the Author:
Dear authors,

Thank you for submitting your manuscript to GMD. I consider that the manuscript is within the scope of the journal. It will be proceeded to the review and discussion phase, once you have addressed the following questions:

1) Please take a loot the Similarity Report and try your best to reduce the similarity index. The report shows that most of the similar words are from your own github project, and a few words are similar to two another papers. It is recommended to modify the abstract on the website or the abstract of this manuscript.

2) Please explain why not use some kinds of deep learning method to study your pCO2 problem? Deep learning may achieve better results than the three traditional machine learning methods you used. However, I clarify that this is only a suggestion.

Please feel free to contact me if you have any questions regarding these, or any other, issues.

Kind regards,

Xiaomeng Huang
Handling Topical Editor

Topical Editor found (08 Oct 2020) Xiaomeng Huang agreed to serve as Topical Editor

Topical Editor Call all Topical Editors (08 Oct 2020)

Topical Editor Call Second Choice - Reminder (04 Oct 2020) a reminder was sent to all corresponding Topical Editors

Topical Editor Call Second Choice (01 Oct 2020)

Topical Editor Call First Choice - Reminder (27 Sep 2020) a reminder was sent to all corresponding Topical Editors

Topical Editor Call First Choice (24 Sep 2020)

Uploaded Files validated (24 Sep 2020) by Anna Wenzel

iThenticate.com Similarity Report completed (24 Sep 2020)
▸ Report
(README)

File Upload (16 Sep 2020) by Jake Stamell
▸ Manuscript

Registered (16 Sep 2020)

Competing interests: The contact author has declared that neither they nor their co-authors have any competing interests.

Cover Letter (Information for the Topical Editor):
Dear Geophysical Model Development Editors –

My co-authors and I submit this manuscript, Strengths and Weaknesses of Three Machine Learning Methods for pCO2 Interpolation, for your consideration. This paper represents a significant advance in our understanding of computational tools to model global ocean carbon levels.

Direct estimates of ocean pCO2 from volunteer observing ships and moored platforms are sparse in space and time. Various machine learning based approaches have been proposed to upscale these observations to create a gap-free data product; however, it is difficult to evaluate the performance of each approach. Therefore, it is thus far unclear what the merits of each approach are.

Here, we use a recently developed Large Ensemble testbed that aggregates data from four independent climate models to train three leading machine learning approaches (XGBoost, Neural Network, and Random Forest) to reconstruct the full pCO2 field. With this synthetic data, we train each machine learning algorithm as if we had sparse real-world data and then holistically evaluate the performance on various time scales using a suite of metrics. Our three main findings are:

1. that XGBoost reconstructs the full pCO2 field with the lowest average RMSE;
2. that the Neural Network captures the decadal variability with high correlation; and,
3. that RF reconstruction performance is comparatively weak.

Our work is of great interest to the scientific community because the ocean is integral in sequestering fossil fuel emissions (Friedlingstein et al. 2019). Understanding this uptake is an essential component of predicting the future state of Earth's climate, because while we have a good understanding of the mean global uptake, regional variations, especially in the Southern Ocean, are not well understood. This work can also help researchers in all Earth science disciplines choose which machine learning algorithm best suits their needs.

Two colleagues have reviewed a draft of this manuscript: Drs. Luke Gregor and Pierre Gentine. The paper is not under consideration or published elsewhere. We thank you in advance for your attention to this manuscript.

Sincerely,

Jake Stamell

917-992-5575
jake.stamell@columbia.edu

Suggested Referees:	Elizabeth Barnes, abarnes@atmos.colostate.edu, Colorado State University, USA, https://sites.google.com/rams.colostate.edu/barnesresearchgroup/prof-barnes C. Deser, cdeser@ucar.edu, National Center for Atmospheric Research, USA, http://www.cgd.ucar.edu/staff/cdeser/ Marion Gehlen, marion.gehlen@lsce.ipsl.fr, LSCE, France, http://sobums.lsce.ipsl.fr/index.php/whos-involved/participants/12-marion-gehlen Andrew Jacobson, andy.jacobson@noaa.gov, National Oceanic and Atmospheric Administration, USA, https://cce-signin.gsfc.nasa.gov/cgi-bin/participants/getrec.pl?name_id=12222&amp;wid=12 Pedro Montiero, pmonteir@csir.co.za, University of Cape Town, South Africa, https://socco.org.za/team/dr-pedro-monteiro/

First Choice Index Terms: Climate and Earth system modeling

Second Choice Index Terms: Oceanography

From: **Galen McKinley** mckinley@ldeo.columbia.edu
Subject: gmd-2020-311
Date: February 11, 2021 at 9:21 AM
To: hxm@tsinghua.edu.cn
Cc: Jake Frank Stamell jfs2167@columbia.edu, Rea Radhika Rustagi rrr2151@columbia.edu, Lucas Gloege ljg2157@columbia.edu
Bcc: Galen McKinley mckinley@ldeo.columbia.edu

Dear Editor Xiaomeng Huang -

I am writing to ask that you please re-consider your decision on our manuscript.

You state that the approaches we use are not innovative. But we are not proposing approaches to perform pCO2 reconstruction here. We are evaluating the ***uncertainties*** in these approaches in as consistent a manner as possible.

We are the first to directly illustrate the strengths and weaknesses with respect to extrapolation to unseen points by the NN, XGB and RF approaches using the Large Ensemble Testbed. We use these approaches that have been used by others for pCO2 extrapolation precisely because these are commonly-used and we want to explore what the strengths and weaknesses in them. The previous authors were not able to understand the degradation of performance from test data to extrapolated points in the way that we can using the Large Ensemble Testbed. This work is, thus, a useful complement to these previous works.

I would like also to note that our work with the Large Ensemble Testbed has been enthusiastically received by funding agencies and the ocean carbon cycle community and the subset of this community working on developing pCO2 products. We are in frequent contact with Luke Gregor, Marion Gehlen, Peter Landschutzer, Christian Rodenbeck, who are the lead authors on these products. We have used our Large Ensemble Testbed in collaboration with Peter Landschutzer and others to evaluate his NN approach, the SOM-FFN (Gloege et al 2021 https://www.essoar.org/doi/abs/10.1002/essoar.10502036.1 , presently in review with Global Biogeochemical Cycles). We are just starting to work with Christian Rodenbeck to evaluate his MLS method (a process based approach, not ML) with the Large Ensemble Testbed. Our efforts are supported with ample finding by NOAA, and we are collaborating on it with the groups at NOAA (AOML, PMEL) that collect about 50% of all the pCO2 data in the SOCAT database.

We have made a great effort in terms of changes with the revision to explain that our work is not to reconstruct pCO2 for the global ocean, but to ***understand the uncertainties*** in these types reconstructions. We are certainly glad to do more in terms of this explanation if you see that we have not adequately described the work. We could also perhaps change the title to “Strengths and Weaknesses of three ***commonly-used*** Machine Learning methods for pCO2 interpolation”.

We respectfully ask that you please re-consider your decision on our manuscript.

Thank you,

Galen A. McKinley, Professor
Department of Earth and Environmental Sciences
Columbia University, New York, NY
Lamont Doherty Earth Observatory, Palisades, NY
mckinley@ldeo.columbia.edu 845.365.8585
url: mckinley.ldeo.columbia.edu

From: **Tsinghua** hxm@tsinghua.edu.cn
Subject: Re: gmd-2020-311
Date: February 11, 2021 at 2:06 PM
To: Galen McKinley mckinley@ldeo.columbia.edu
Cc: Jake Frank Stamell jfs2167@columbia.edu, Rea Radhika Rustagi rrr2151@columbia.edu, Lucas Gloege ljg2157@columbia.edu

Dear Galen Mckinley:

I regret to make the rejection decision. It is based on the reviewers' opinion of the scientific significance and scientific quality of your work. I believe that you should be able to see their reports.

In terms of the current revised version, I agree with that the highlight of the work is the use of LET data to generate pCO2 data, and you did perform a bias analysis and comparison using the three ML methods to present the strength and weakness of these methods .

Although you emphasize that the focus of the paper is about uncertainty, but the sources of uncertainty are really not analyzed and presented in the text. In fact, the number of layers or decision trees in the ML models, all the hyper-parameters, random initialization, as well as the selection of different features (driver data) can have a huge impact on the pCO2 data product. If your work is focused on uncertainty, then more detailed experiments and analysis about model structural uncertainty, model parametric uncertainty, uncertainty bound, and sources of uncertainty maybe necessary. The experimental results of bias are not the whole story of uncertainty.

Thus it is recommended to resubmit the manuscript to GMD or the other journals after major revisions. Thank you for your understanding and support.

Best regards,
Xiaomeng Huang

From: **Galen McKinley** mckinley@ldeo.columbia.edu
Subject: Re: gmd-2020-311
Date: February 11, 2021 at 2:47 PM
To: Tsinghua hxm@tsinghua.edu.cn
Cc: Jake Frank Stamell jfs2167@columbia.edu, Rea Radhika Rustagi rrr2151@columbia.edu, Lucas Gloege ljg2157@columbia.edu

Dear Editor Huang -

I am quite confused by your editorial process. We, the author team, are not getting clear information upon which to act.

The reviews you presented to us did not ask us to change our uncertainty analysis to address components such as selection of drier data or hyper-parameters, as does your third paragraph here. Is this what you expected us to do in our revision? How could we know what you were expecting from us if you did not tell us so?

The reviews that you sent to use suggest that the reviewers did not understand the purpose of the Large Ensemble Testbed. The reviewers seem to have thought we were trying to reconstruct real-world pCO2. We made great effort in our revision to correct this mis-understanding by modifying the text of the manuscript. But now you are giving us new reasons for rejection that appear to be inconsistent with the original reviewers comments. This is unfair and we would like to have the issue clarified.

Can you please cite the parts of the reviews to which we have not adequately responded and that are the reason for your rejection?

Thank you
Professor Galen McKinley

From: **Tsinghua** hxm@tsinghua.edu.cn
Subject: Re: gmd-2020-311
Date: February 11, 2021 at 4:02 PM
To: Galen McKinley mckinley@ldeo.columbia.edu
Cc: Jake Frank Stamell jfs2167@columbia.edu, Rea Radhika Rustagi rrr2151@columbia.edu, Lucas Gloege ljg2157@columbia.edu

Dear Galen Mckinley:

Sorry for not explaining the referee's rating and my decision clearly. You can see the referees' attitude from the following two figures.

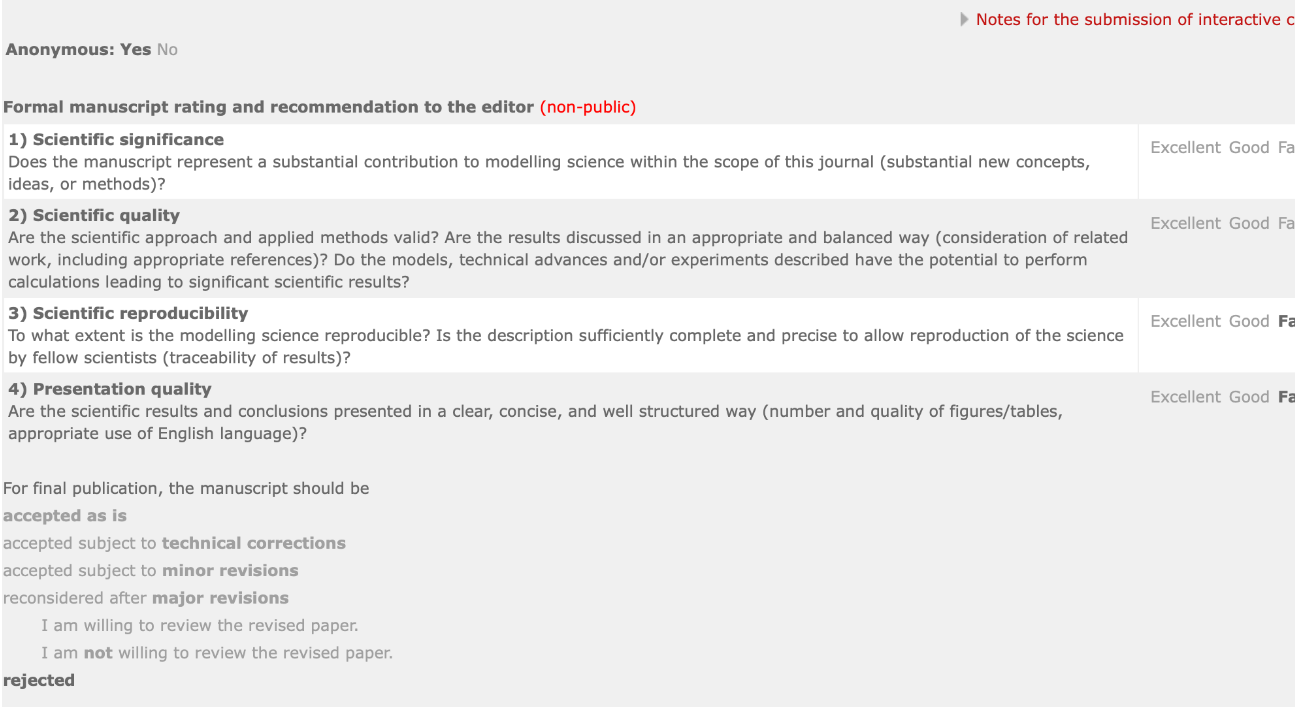
Moreover, I do not give any new reasons for rejection. I just notice the referee #1 comment about "the three method ML methods presented in this study were already compared in Gregor et al. (2019), ……". The novelty of exploring the strengths and weakness of individual methods have some value but not enough.

The major comment of referee #1 is "What is the overall novelty and significance of the resent study?". In your revised version, your answer is "Quantify extrapolation uncertainties". However, after reading your revision version, I think the uncertainty about model structural, model parametric, uncertainty bound of three ML methods are not analyzed and discussed, especially the source of uncertainty among NN, RF and XGB. I think the goal of quantify extrapolation uncertainties do not achieve.

Best regards,

Xiaomeng Huang

Report #1 rating:



Notes for the submission of interactive c

Anonymous: Yes No

Formal manuscript rating and recommendation to the editor (non-public)

1) Scientific significance Does the manuscript represent a substantial contribution to modelling science within the scope of this journal (substantial new concepts, ideas, or methods)?	Excellent Good Fa
2) Scientific quality Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)? Do the models, technical advances and/or experiments described have the potential to perform calculations leading to significant scientific results?	Excellent Good Fa
3) Scientific reproducibility To what extent is the modelling science reproducible? Is the description sufficiently complete and precise to allow reproduction of the science by fellow scientists (traceability of results)?	Excellent Good **Fa**
4) Presentation quality Are the scientific results and conclusions presented in a clear, concise, and well structured way (number and quality of figures/tables, appropriate use of English language)?	Excellent Good **Fa**

For final publication, the manuscript should be

accepted as is

accepted subject to **technical corrections**

accepted subject to **minor revisions**

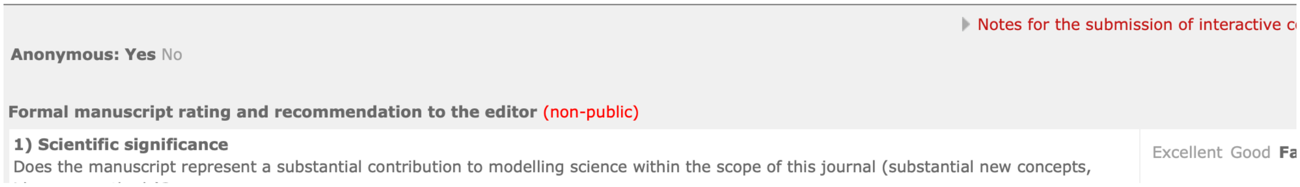
reconsidered after **major revisions**

I am willing to review the revised paper.

I am **not** willing to review the revised paper.

rejected

Report #2 rating:



Notes for the submission of interactive c

Anonymous: Yes No

Formal manuscript rating and recommendation to the editor (non-public)

1) Scientific significance Does the manuscript represent a substantial contribution to modelling science within the scope of this journal (substantial new concepts,	Excellent Good **Fa**

ideas, or methods)?	
2) Scientific quality Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)? Do the models, technical advances and/or experiments described have the potential to perform calculations leading to significant scientific results?	Excellent [✓ **Good**] Fa
3) Scientific reproducibility To what extent is the modelling science reproducible? Is the description sufficiently complete and precise to allow reproduction of the science by fellow scientists (traceability of results)?	Excellent [✓ **Good**] Fa
4) Presentation quality Are the scientific results and conclusions presented in a clear, concise, and well structured way (number and quality of figures/tables, appropriate use of English language)?	Excellent Good [✓ **Fa**]

For final publication, the manuscript should be

accepted as is

accepted subject to **technical corrections**

accepted subject to **minor revisions**

[✓ **reconsidered after major revisions**]

[✓ **I am willing to review the revised paper.**]

I am **not** willing to review the revised paper.

rejected

Reviewer 1 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?

Thank you for this comment.

The reviewer is correct that these types of approaches have been applied to pCO2 reconstruction in past studies such as Gregor et al. (2019). However, these past studies have not be able to quantify extrapolation uncertainty. Gregor et al. (2019) attempted to resolve this issue using independent data sets; however, these data sets still suffered from data sparsity issues. Because we work on simulated pCO2 from the Large Ensemble Testbed, we add the ability to assess uncertainties in extrapolation beyond the test data, or “unobserved data” in our updated terminology.

Yes, Gloege et al. (2020) also show overestimation of decadal variability with a specific NN implementation, the SOM-FFN (Landschutzer et al. 2016). For the NN in this paper, we make different choices, such as not breaking the ocean up into biomes. There are other NNs, such as in Gregor et al. (2019) and in CMEMS (Denvil-Sommier et al. 2019) that make their own independent choices. Our goal is to provide a reasonable basis for comparison between the NN, RF and XGB methodologies. Our implementations all use the same input driver data and do not use spatial clustering into biomes so that they are as simple as possible. Our focus is on understanding the extrapolation skill across different methodologies - something that has not been presented in the literature before. For example, Gregor et al. (2019) combine different methodologies to create an ensemble method, but do not explore in detail the strengths and weakness of the individual methods.

We have added text to the abstract and text to clarify that the goal of this work is to quantify extrapolation uncertainty for NN, RF and XGB applied to surface ocean pCO2. This is our novel contribution.

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?

Thank you.

We have added text to the introduction manuscript to clarify that the LET is composed of 4 initial condition large ensemble climate model simulations. These are plausible evolutions of the Earth System from 1982-2016. These are the models used for future climate projection by the IPCC and are thus carefully vetted for many variables, including the ocean carbon cycle (e.g. Long et al. 2013 for CESM).

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 add text to make this clear.

We have added to the text additional description of the testbed and its purpose. We have clarified our terminology by replacing "features" with "driver data" and explaining specifically what we mean by "unobserved data".

We also add discussion of shortcomings of the LET, such as an inability to assess the impact of sub-gridscale variability, in the last section of the Discussion.

More generally, we have attempted to improve readability by reducing repetition and adding additional sections to the Discussion.

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?

We have added to the text to clarify that only the spatial pattern of SOCATv5 data is used, not these data themselves.

We have worked to reduce repetition about details, but do feel it is useful to our likely readers if we provide a more general overview of the ML approaches in the introduction, in addition to specific details of our implementation in Methods and Appendix.

In the introduction, we state that the reason these approaches are selected is because they are currently very common for a variety of industry and scientific applications, and because they have been used for pCO2 extrapolation using real data by authors such as Gregor et al. 2019.

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

We have carefully checked to be sure that the MAE statistics cited in the text on line 244-246 are the same as in Table 1. No change to the text is required.

Reviewer 2 Comments

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?

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 pCO_2 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.