Review for "*Using Neural Network Ensembles to Separate Biogeochemical and Physical Components in Earth System Models*" by Holder et al. (gmd-2021-167)

14th Nov. 2021

SUMMARY

The authors' present an efficient and innovative approach for identifying the major sources of difference between Earth System Model (ESM) predictions. Specifically, they use neural network ensembles (NNEs) to identify whether the phytoplankton biomass predictions of different ESMs are more affected by changes in ocean circulation, or by differences in biogeochemical formulation. They conclude that, in the context of their test cases, the NNEs were able to accurately identify the relationships between variables in the ESMs – when they have access to all of the variables that affect phytoplankton biomass.

On the whole, the authors' have developed a robust, well-designed and meticulously implemented framework for examining variability in the outputs of ESMs. Their NNEs appear to perform exceptionally well, and serve as a powerful demonstration of the predictive capabilities of such models. The manuscript is generally well-written, and, in the context of their initially stated aims, the authors' have done an excellent job.

That said, there are areas where I feel this paper could be improved. I recommend minor revisions before publication, primarily in terms of ensuring that the motivation of the study and the broader importance of the results are both clearly communicated.

GENERAL COMMENTS

1.1 ***Main motivations for study unclear from abstract and intro***

I feel that the Abstract and Introduction could do a better job of presenting the work in terms of the specific problems that the authors' methods are addressing, and why it matters. They touch on several factors that might lead different ESMs to produce different outputs, but it is not necessarily clear, in my opinion, how the work being introduced addresses these problems.

- Different ESMs yield different predictions because of variations and uncertainties in input parameters (line 36-37)
- Uncertainty as to whether ESMs are using the "correct relationships" (line 45-46)
- Traditional methods for estimating ESM sensitivity are inadequate (line 40-44)

Thereafter, it’s concluded that these factors indicate a need for a method that can identify whether different ESM predictions of phytoplankton biomass result from differences in biogeochemical formulation, or in physical circulations (line 47-49). It’s not immediately clear how the proposed method will help alleviate the previously raised issues.

1.2 ***Clarify intended audience***

- Who is the intended audience?
- Researchers who develop ESMs? If so, how will your methods/results help improve their models?
- Researchers who work with observational data? How might this work help them better utilise their data?
- Researchers who build ML models? How might this work inform theirs? etc

The conclusion does elaborate on some of the reasons why other researchers (ostensibly those who wish to compare different ESMs) might find value in this work, and reads more clearly than the abstract and introduction, if a little unfocused.

1.3 ***Clarify broader importance***

The conclusion does briefly elaborate on some of the reasons why other researchers might find value in this work. Is the primary target audience those who wish to compare the outputs of different ESMs? Those who wish to improve their existing ESM? Those who wish to more efficiently utilize observational data?

In NNEs, the authors’ demonstrate a very powerful tool, that can (and is being) applied to all of the above use cases across multiple fields. That said, the application of this methodology in the current study is quite specific to extracting relationships from ESMs. However, the authors’ suggest that directly applying their methods to observational data will help calibrate and improve ESMs, and thus yield better predictions of e.g. changes we might expect under climate change.

This is a big claim to make, and I question whether it is meaningful in the context of this manuscript. In the current work, the authors’ have access to a complete, perfect knowledge of all of the variables that affect e.g. plankton biomass, within each ESM (itself a highly simplified representation of the real Earth system). In addition, their NNEs have access to ALL the depth-integrated data in every part of the simulated global ocean for each ESM, across arbitrary time.

Real world data clearly represents a very different set of challenges and constraints. Observational datasets are orders of magnitude more sparse and imbalanced. Even if we were able to sample the entire ocean, our knowledge of the important physical and chemical fluxes driving growth and distribution is incomplete, even without including the significant added complexities of biotic interactions and adaptive evolution.

The results of the present work do not appear to be sufficient to make claims as to the direct applicability of these methods, as presently described, to generating more accurate representations of the natural world.

SPECIFIC COMMENTS

Line 11-14 – Is this really a "compensating error"? In the real ocean, oligotrophic regions are dominated by small phytoplankton species (greater surface-area-to-volume ratio, greater uptake of, and thus "sensitivity to", nutrients)

Line 14-15 - Are the authors' referencing their own previous work here? If so, this should be made clear: "Recently, we demonstrated that... "

Line 17-18 - Suggest being more specific about the types of results being examined e.g. "...why different ESMs produce different spatiotemporal distributions of phytoplankton biomass"

Line 20 - Three test cases are mentioned, but only two are elaborated in the abstract.

Line 28 - Yes, but also by the increasingly prohibitive computational expense of adding complexity and resolution.

Line 33-35 - This seems like a reasonable metric to vary, particularly when modelling different plankton community structures. It is not necessarily variable as a result of uncertainty.

Line 36-49 – This paragraph first mentions the uncertainty associated with ESM input parameters, then the coupled nature of a given input to multiple outputs, and then the difficulty in knowing whether ESMs are modelling the “correct relationships”. These are all valid – if separate – points. But I’m struggling to link these points to the proposed ‘solution’ in lines 47-49. Will the NNE help to identify which relationships are ‘most correct’, or extract new ‘more correct’ relationships? Or is its primary function to more clearly identify the reasons why ESM predictions of biological variables diverge?

Line 56-59 - This definition was a little confusing to read. Are the “intrinsic relationships” those which are known as true drivers of a target variable? E.g. those captured by lab growth rate experiments, or, as in the current context, the biogeochemical equations underlying ESMs?

Line 61-64 - Similar to the previous point, are your “apparent relationships” a reference to data-derived correlations?

For the record, I really like the terms “intrinsic” and “apparent”, but I think your description of these terms is much more clear in your previous work “Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations?”.

Line 71-72 – Perhaps ‘determining *the most significant sources of* differences in ESM outputs’?

Line 73 – Can ‘*combinations of these two*’ be considered as an independent ‘primary driver’?

Line 78-79 - Possibly worth clarifying that you’re only referring to the link between circulation changes and patterns of co-limitation in the ESM (to avoid readers’ potentially drawing parallels to real ocean dynamics, to which such findings may not apply).

Line 97 – A reader’s question here might be - “why not identical?”

Line 162 - For all case descriptions, perhaps include details on e.g. how long each ESM was run for (in model years), output formats (e.g. daily/monthly averaged values) and the model resolution.

Line 166-167 - Perhaps expand on this, as it seems like an important point. We know from your (very clear and helpful!) Fig. 1 that nutrient distribution is coupled to circulation, but biomass itself is not, and that changes in biomass are a function only of nutrient distribution. With this in mind, the reader might be wondering whether it is even possible – given the constraints of BLING - to “push the biology into fundamentally new states” by varying circulation alone.

Line 236-237 - Why were these activation functions chosen?

Line 257-258 - Perhaps mention these previous sensitivity tests earlier (e.g. line 235+) - “we previously determined that {x,y} were not sensitive to {p,q} (ref) so our individual NN’s were constructed using…”

Line 274-275 - I think it’s worth including more detail on the actual data that the NNEs are using. How many datapoints do the training and test sets contain? Are the training and test sets randomly sampled in both time and space from the ESM outputs? Are they drawn from different temporal periods? Or from different spatial regions?

Were any resampling techniques employed to address potential imbalance in the randomly-sampled data? Or were the datasets large enough to effectively capture variance? Did you use all of the depth-integrated output data from the model runs for training/testing?

Line 281-282 – I'd suggest being more explicit here on how r-squared was calculated - this is a notoriously tricky, often misused metric. E.g. NNE predictions of mean annual biomass for each point are plotted against the 'true' ESM values… standard or adjusted, etc.

Line 290-294 – What were your criteria for what constituted a significant increase or decrease in RMSE?

Line 320-321 - I suspect that some readers will have questions about the extremely high performance seen here, across both metrics. It would be helpful again to provide more detail on the nature of the training and test datasets, how they were sampled, etc. Is the "mean value of the total biomass" calculated as a total global mean, or the mean for a given point? Across what time period?

Line 324-328 - Is this unexpected? The fact that "physical circulation would simply act to change the location of where combinations of light and nutrients were found" seems like a given, considering that "biomass is not directly affected by changes in the physical circulation" in BLING.I think that an explicit clarification of the importance of this result, in this context, would be helpful to place here.

Line 483-485 - This is a really interesting result in terms of the importance of including the correct variables in predictive models. In this case, we happen to know all of the variables that affect our target within the system. When applying such models to real-world data, we don't, and it can have significant consequences for predictive accuracy.

TECHNICAL COMMENTS

Line 1-2 - The use of "components" in the title is a little broad. Perhaps substitute with "Drivers of Plankton Biogeography"?

Line 92-94 - Advise being more specific here - what is meant by "push the biology into fundamentally new states"? I liked "produce new patterns of colimitation", as given in line 79.

Line 110 – Introduce symbol for phytoplankton biomass (B) in this line

Line 138 - Typo, should read "than the"

Line 142 - Typo "nutrient and temperature"

Line 180 – Is this meant to read ‘Section 3.4’?

Line 311 - Again, I think your original phrase “new patterns of colimitation” is more descriptive and appropriate than “fundamentally new states”

Fig. 3 and 4 - Agreed, the inclusion of both large and small phytoplankton in Fig 3 makes it difficult to read. Suggest splitting them up