Author Responses Addressing Review from Referee #1 for "Using Neural Network Ensembles to Separate Biogeochemical and Physical Components in Earth System Models" *by* Holder et al.

For these responses, we address each Referee comment individually and include our response below it. The Referee Comments (RC) are numbered and use a black font, while the Author Responses (AR) use a red font.

RC0.0: In their article, Holder et al. use the approach of neural network ensembles (NNE) to extract relationships between predictor (nutrients, irradiance, temperature) and target (small and large phytoplankton biomass) variables within ocean biogeochemical models. Specifically, they investigate whether the NNE approach is capable of determining why different models produce different results. They study three test cases, where they either alter the physical formulation controlling the circulation or biological equations. Thereby, they focus on the two different types of relationship, i.e., intrinsic vs. apparent relationships. They conclude that the NNE approach is capable of characterizing these relationships and can thus be considered as a parsimonious representation of the system, including extrapolative power.

Overall, this study provides a valuable contribution of how one can leverage "Machine Learning" approaches to better understand the dynamics of a complex model, such as ocean biogeochemical or Earth system models. Also, not being an expert in ocean biogeochemical modeling, I consider the presented methods and analyses to be robust. The manuscript is well written, however some passages need restructuring and the manuscript needs to be tuned for its target audience — please see my comments.

I recommend minor revisions of the manuscript before publication.

AR0.0: We want to thank Referee 1 for their helpful comments and suggestions. We have done our best to address each of the comments below.

RC1.1: Overall, the manuscript based on the title and the bigger part of the abstract aims at a larger readership working with Earth system models (ESM), the main part of the manuscript is, however, very focused on ocean biogeochemistry modeling. For example, the abstract is very ESM-general until line 12, but then jumps into a very specific problem on phytoplankton. Earth system modelers, who are not so familiar with ocean biogeochemistry, might be a bit lost here and in general throughout the article. I suggest to either sharpen the focus of the manuscript to only aim for the ocean biogeochemistry community, or to be more inclusive for Earth system modeler in general. The latter solution would require that you clearly state that the ocean biogeochemistry problem investigated in this study is used as a case study to demonstrate your approach, introduce the reader more to the problem of small vs. large phytoplankton prediction, and how one could adapt your approach/case study to other aspects of the Earth system.

AR1.1: In the updated manuscript, we specify that although we focus on phytoplankton and ocean components that our results are applicable to other components of ESMs as well.

RC2.1: Please stick to the tenses, i.e. do not switch between present and past tense when describing your results. I recommend that you always use present tense when describing your study at hand, i.e. when describing your methods, your results etc., and only use past tense when referring to already published studies.

AR2.1: This was a very helpful comment! In the updated manuscript, we replaced past tenses with present tenses in the following sections: Methods, Case Descriptions, Results and Discussion. We replaced present tenses in the Conclusions section with past tenses.

RC2.2: L1: The title is too general. There are also biogeochemical and physical components in the land-surface models. Better to add "Ocean" in the title.

AR2.2: Added "Ocean" into the title for clarity.

RC2.3: L22: The abstract misses a concluding sentence. Please add a sentence that gives a general outlook of your study and highlights its significance for the discipline of Earth system modeling.

AR2.3: Added two short concluding sentences to the end of the abstract in the updated manuscript.

RC2.4: L27: It is limited not only by imperfect knowledge, but also by the fact that we cannot resolve the processes in current models and current HPC facilities.

AR2.4: Updated in the revised manuscript.

RC2.5: L46: Maybe better "are indeed being modelled".

AR2.5: Updated with the suggested sentence fragment.

RC2.6: L50: Include a sentence here that shortly explains the concept behind the NNE.

AR2.6: Added a transition sentence at the bottom of the paragraph ending in "… differences in physical circulations and climate sensitivities." Also added a sentence to the paragraph following the one previously mentioned which briefly introduces and explains the concept behind NNEs.

RC2.7: L51: Again, the use of tenses in this manuscript is a bit misleading. It is better to write: "… (NNEs) **are** able to extract …" instead of "… were able …". It's not that they lost the capability to do so in the meantime.

AR2.7: We address this particular comment as part of the RC2.1 comment above.

RC2.8: L64: Better "high irradiance" instead of "high light".

AR2.8: Updated with the suggested wording.

RC2.9: L71: The paper is, on the one hand, specific about ocean biogeochemical modeling and, on the other hand, it tries to be more general about Earth system modeling. One could add a statement here that you look into phytoplankton physiology as a case study, but the approach is also applicable to other problems in the Earth system.

AR2.9: We added a sentence specifying that this approach is applicable to other components of ESMs and specified that we focus on marine phytoplankton physiology in our study.

RC2.10: L73–83: This section reads a bit like you already discuss your results. It would work better if you used present tense and explain the different approaches which are applied in this research, and why.

AR2.10: We changed the wording of these sentences to use the present tense, instead of past tense, so that we are not discussing the results of our study in the introduction.

RC2.11: L110: "**ocean** biogeochemical components of ESMs"

AR2.11: Updated with the suggested wording.

RC2.12: L112; Equation 1: Please replace “Light” with **I** for irradiance.

AR2.12: Replaced “light” with irradiance in Equation 1 and in the sentence describing the variables of Equation 1. Also replaced the term “light” with “irradiance” throughout the rest of the text as well, including the abbreviations in the equations.

RC2.13: L131: “computationally cheap”.

AR2.13: Updated with the suggested wording.

RC2.14: L132: either “in” or “within” the model.

AR2.14: Removed “in” and kept “within.”

RC2.15: Case Descriptions: Could you include for each case an equation describing how the NN is set up? E.g. something like Biomass = NN(Irradiance, Nutrients, Temperature) with proper variable names?

AR2.15: We tried to implement this suggestion where we gave each Case (3.1, 3.2, and 3.3) their own equation, but this led to a lot of equations repeating themselves. Additionally, we wanted to keep most details pertaining to the framework of the NNs and NNEs in Section 3.4 for clarity.

To include the type of information requested in this comment, at the end of Section 3.1 we added a sentence stating that the details of the NNE and NN training/frameworks can be found Section 3.4. Additionally, in Section 3.4 we added an updated description for the structure of the individual NNs.

RC2.16: L190: Can you more clearly explain what the “LgSm” acronym is referring to?

AR2.16: The LgSm acronym was chosen because there are variables that specifically state the concentration of the small and large phytoplankton biomass. This differs from the PI Control where the small and large phytoplankton biomass are calculated as a fraction of the total biomass. In both instances, you still get small and large phytoplankton biomass values, but just arrive at it slightly differently. Specifying the acronym as LgSm is shorthand for stating that the small/large phytoplankton variables are specifically stated in that model run.

RC2.17: L231: “NNEs possess some capability of extrapolating outside the range of the data on which they are trained.” Very important point - you should provide a citation here!

AR2.17: Included another mention of Holder and Gnanadesikan (2021) since that is something that particular study found.

RC2.18: L232: With RF you mean Random Forests, I assume. Can you make it clear?

AR2.18: Replaced the RF acronym with the full spelling of random forests.

RC2.19: L248–250: Why did you not set up your NN system with training, test and validation datasets? So, validation dataset to prevent overtraining, and test dataset to test generalizability?

AR2.19: The Matlab function that we used for training the individual NNs separates the data into training, validation, and test datasets. We were trying to keep the specific details in the manuscript to a minimum, but we understand the need for this clarification. For clarity, we have included details about this in the updated manuscript.

We can also provide a brief explanation here and please note the specific distinction we make between *dataset* and *subset*. In the original manuscript, we mention that we split the data into training and testing datasets. Only the training dataset is provided to the Matlab function used for training the NNs. The Matlab function then takes the training dataset and splits it further into training, validation, and testing *subsets*, with 70% of the data from the training dataset going into the training *subset*, 15% to the validation subset, and 15% to the testing subset. The remaining observations in the testing *dataset* are therefore observations that none of the trained NNs have ever seen before, which makes the performance metrics even more rigorous. This provides a convenient way to test the unique performance of the NNE (collection of the trained NNs).

RC2.20: L260: Maybe you can write that hyperparameters tuning showed that the setup is not very sensitive to the selection of different hyperparameters.

AR2.20: Depending on the hyperparameters, the performance of the NNEs could be affected. For example, if the neural networks used hidden layers that had only one node or that used a linear activation function, the performance would decrease. For clarity, we have changed this paragraph to include more specific information.

RC2.21: L262: Did you also use a different scheme for normalization, e.g. normalization to zero mean and unit standard deviation.

AR2.21: We considered normalizing with the zero mean and unit standard deviation, but the predictors are either heavily right-skewed (nutrients) or bimodal (temperature). Even with a different normalization scheme, we still get values greater than 3 standard deviations from a zero mean.

We could include this normalization before scaling the variables between -1 and 1, but we already get relatively short training times for the NNs with the current parameters.

RC2.22: L340: For me, the extrapolative power of your NNE approach is a very encouraging result. You show that a NN can learn the dynamics of the system from the PI run and is able to extrapolate to extreme forcing like 4xCO2 - maybe one should make a bigger deal out of this and highlight in the abstract.

AR2.22: Yes, it is an encouraging result, but we were trying to be careful about how we stated this result. We did not want to state that this method is great for extrapolating. Using any method for extrapolation comes with higher uncertainty in the regions of the dataspace where the model was not trained. For example, NNEs will have higher uncertainty in the regions where all the predictor variables are very high, because there are not any observations from that region of the dataspace in the training subset. Any predictions the NNEs make in that unexplored region will be less certain than regions of the dataspace that were included in the training subset.

One way to explain the predictability of the 4xCO2 from the NNE trained on the PI Control run is that the PI Control run and the 4xCO2 are being governed by the same equations. Although they have different inputs, the models are still run with the same internal equations and constants. If one of the constants (e.g., different half-saturation constant for one of the nutrients) between the two runs differed, the apparent relationships would be different, and the accuracy of the predictions would decrease when using one NNE to predict the outcome of the other. In the original version of the manuscript, we state this in Lines 378-379, “When the biological equations remain the same, changing the physical parameters simply change where combinations of nutrients and light occur.” To make this point more apparent, we have added an additional sentence clarifying this in the final paragraph of Section 4.1 in the updated manuscript.

RC2.23: L437: You have not introduced the abbreviations Chl:C. I know, it is clear for the reader with ocean biogeochemistry background, but your title addresses a larger readership. So, please introduce all abbreviations.

AR2.23: Defined the acronym.

RC2.24: L498: Better rename to “Summary & Conclusions”.

AR2.24: Updated the section name to the suggested wording.

RC2.25: L518: Rephrase “we can be relatively confident” to something like “their predictions can be considered reliable.”

AR2.25: We agree the current wording could be improved. We have revised this in the updated manuscript.

RC2.26: Conclusions: Overall, I find them too long and not to-the-point. Can you boil it down to a few concise statements?

AR2.26: In the updated manuscript, we have shortened the conclusions to what we consider to be the essential points that we wanted to highlight. We kept the summary of each case (L499-L515) to remind readers of the main objectives for each one. We condensed the next three paragraphs (L517-L533) into a single paragraph to summarize the main conclusions. We kept the next two paragraphs (L535-L551) which discuss the implications of the research and how the research can be utilized by oceanographers and climate scientists.

RC2.27: Figure 1 & 2: I cannot comment on the specifics of the ocean biogeochemical models. Ideally, another referee with the needed expertise should comment on these aspects.

AR2.27: Understood and noted.

RC2.28: Sensitivity Analysis Figures: The colored lines are the actual model run output, right? Or is it the mean NNE? The grey shading is the NNE, right? Could you put this in the legend? If the actual model output is not included in the figure, where do you show the performance for NNE versus actual model output except R^2 and RMSE values in the tables.

AR2.28: The colored lines are the average of the predictions from the NNs that make up the respective NNE. They grey shading is equivalent to plus/minus one standard deviation relative to the predictions of those NNs. In the updated manuscript, we have kept the legend the same in order to minimize the space required for the legend. However, we have updated the description of each sensitivity analysis figure to make it clear that the lines are the average prediction of the NNEs.

The actual relationship of the model is not included since the model output does not have that capability. One purpose of the apparent relationships is to allow for the visualization of those relationships. The proof-of-concept for using the apparent relationships in this way is discussed in Holder and Gnanadesikan (2021). Within that manuscript the actual model output is shown, along with the predictions from several machine learning methods.

RC2.29: Sensitivity Analysis Figures: Why do you show in e.g. Figure 3 small and large phytoplankton biomass together and Figure 4 only small phytoplankton biomass. Can you not remove small phytoplankton biomass from Figure 3 and corresponding subsequent figures?

AR2.29: We included Figure 4 with only the small phytoplankton biomass since the small phytoplankton lines are overshadowed by the responses of the large phytoplankton in the higher percentiles of Figure 3, such as the 75th percentile nutrient and temperature subplots. The benefit of including them both on the same original plot is that it allows for the comparison of the apparent relationships between small and large phytoplankton. Since the large phytoplankton relationships are clear in all the subplots of Figure 3, we did not think it was necessary to create a separate figure for large phytoplankton like we did for small phytoplankton in Figure 4.

RC2.30: Sensitivity Analysis Figures: I find the black arrows at the axis to be a bit misleading - do you need them?

AR2.30: We understand how they can be misleading. The only black arrows we kept were the ones labeling the biomass, so that it is obvious that the y-axis on each plot is for biomass. The rest of the black arrows are not necessary for communicating the purpose of the figures and we have removed them in the updated manuscript.

RC2.31: Sensitivity Analysis Figures: What does “ex.” in the captions mean? Example? Better use e.g. then.

AR2.31: Yes, “ex” was being used as shorthand for “for example.” We have updated all instances of “ex” with “e.g.” in the updated manuscript.

RC2.32: Figure 5: I’d prefer if you added the unit next to the colorbar.

AR2.32: Label and units added to the colorbars of the contour plots.