Author Responses Addressing Review from Referee #2 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) are also numbered and use a red font.

RC0.0: 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.

AR0.0: We want to thank Referee 2 for their helpful comments and suggestions. We have done our best to address each of the comments below.

RC1.1: **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.

AR1.1: We have updated the portion of the introduction that you reference in the updated manuscript. In particular, we more clearly state the objective of the paper earlier and mention NNEs as the method we use in the paper to help us achieve that objective.

RC1.2: **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.

AR1.2: In the updated manuscript, we have narrowed the focus in terms of the specific audience. We also state that this method is applicable to other research areas and other components of ESMs, although we only focus on marine phytoplankton in our paper.

RC1.3: **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.

AR1.3: Similar to AR1.2, our primary audience is modellers, but we also briefly discuss how these methods can be applied to other oceanographic datasets and why they might be of interest to other Earth scientists.

Although it is a big claim to state that these methods can be applied to observations, it is based on the work of another manuscript we are currently working on. In that manuscript, we demonstrate that using *climatologies* of ESM outputs and interpolated *climatologies* of observations from various data sources, we can compare the two. For example, using sensitivity analyses we can examine the general trend in the apparent relationships for various ESMs and how they compare to the trend in observations. Additionally, our preliminary results suggest that we can capture a large portion of the variance in climatological observational datasets with machine learning (60-80%). Although observations are certainly more sparse and imbalanced, using climatologies we can make a “first-pass” of the comparison between the two.

RC2.1: 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)

AR2.1: The example of weak upwelling, low nutrients, and nutrient sensitivity was only meant to serve as an example of something that an ESM **could** do to possibly compensate. The main point we were trying to describe is that the output of ESMs (such as spatiotemporal distributions) might match observations for the wrong reasons, e.g., incorrect assumptions, equations, etc. Different configurations of an ESM can arrive at the same answer for very different reasons. This means that just because the output of an ESM contour map matches a contour map of observations, the ESM might have arrived at the correct distribution for a reason other what actually happened in the real world. We have updated the wording in the updated manuscript to better reflect this.

RC2.2: Line 14-15 - Are the authors' referencing their own previous work here? If so, this should be made clear: "Recently, we demonstrated that... "

AR2.2: Yes, this is based on previous work. We implemented the suggested wording in the updated manuscript.

RC2.3: 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"

AR2.3: Changed the text in the updated manuscript to the suggested wording.

RC2.4: Line 20 - Three test cases are mentioned, but only two are elaborated in the abstract.

AR2.4: Added additional information to the abstract describing the third case.

RC2.5: Line 28 - Yes, but also by the increasingly prohibitive computational expense of adding complexity and resolution.

AR2.5: Added the additional description in the updated manuscript.

RC2.6: 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.

AR2.6: We were not trying to say that it is varied because of uncertainty, but rather that each of the eight ecosystem models that are used in Laufkötter et al. (2015) use different Q_{10} values for the various ways they try to represent it. For example, in Table 3 of Laufkötter et al. (2015), some models use a single value of Q_{10} across temperature and functional groups, while in others it is different over different temperature ranges or across phytoplankton and zooplankton functional groups.

RC2.7: 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?

AR2.7: In the updated manuscript, we introduce the concept of NNEs earlier in the introduction and move the objective of the paper into the same paragraph (originally Line 71-72). Along with the comments from Referee 1, we have also more clearly specified the objective.

RC2.8: 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?

AR2.8: It mainly depends on the context of the dataset. Intrinsic relationships in the real-world could be things like lab growth rate experiments in which one driver (such as a nutrient) is varied for a single model organism. Intrinsic relationships in the context of ESMs would be the biogeochemical equations that are programmed into them, which generally have a functional form similar to what is observed in laboratory experiments. In both cases, intrinsic relationships refer to the fundamental relationships that are driving a system forward at the smallest timescale for which data is available.

In the updated manuscript, we have revised the description of the intrinsic relationship ESM example.

RC2.9: 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?”.

AR2.9: The terms *intrinsic* and *apparent* relationships are based on the context of the dataset, similar to what we state in AR2.8.

Apparent relationships in the context of ESMs are the relationships that emerge from the output of ESMs where the intrinsic relationships programmed into the model have interacted across time and space and then had their short timescale values averaged into fields, such as monthly averages.

We have added an example of apparent relationships with respect to ESMs in the updated manuscript.

RC2.10: Line 71-72 – Perhaps ‘determining *the most significant sources of* differences in ESM outputs’?

AR2.10: Updated with the suggested wording.

RC2.11: Line 73 – Can ‘*combinations of these two*’ be considered as an independent ‘primary driver’?

AR2.11: In general, there are two primary drivers that lead to differences in how ESMs simulate phytoplankton biogeography: physical forcings and phytoplankton physiology. Insofar as both of these act to affect nutrient cycling, they can also act in combination to produce indirect impacts.

RC2.12: 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).

AR2.12: Added a clarification sentence in the updated manuscript.

RC2.13: Line 97 – A reader’s question here might be - “why not identical?”

AR2.13: We stated “similar physical circulations” because the physical circulation in our ESM can be slightly affected by the biological cycle by changing the absorption of shortwave radiation. Since we changed the intrinsic biological relationships in Case 2, this results in small differences in the circulation between the two model runs. We have added a clarification near the referenced section in the updated manuscript.

RC2.14: 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.

AR2.14: We have included additional details in the updated manuscript.

RC2.15: 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.

AR2.15: This is true on short timescales. In short timescales, such as the size of the timestep of BLING (intrinsic relationships), the biology is not being pushed into fundamentally new biological states from changing circulations. However, the apparent relationships arise both from changing combinations of light, macronutrients, micronutrients, and temperature, as well as time averaging of these relationships. There is no guarantee that, for example under climate change, that the primary drivers will combine and average in the same way.

We have updated the definitions and examples of intrinsic and apparent relationships in the updated manuscript.

RC2.16: Line 236-237 - Why were these activation functions chosen?

AR2.16: We chose to use the hyperbolic tangent sigmoid function for the hidden layer because we showed in previous work (Holder and Gnanadesikan 2021; their Appendix B) that the choice of activation function for the hidden layer did not really affect the performance of the NNEs as long as the activation function was non-linear. Specifically, we tested the following activation functions: hyperbolic tangent sigmoid, logarithmic sigmoid, inverse, positive linear, linear, soft maximum, and radial basis. The settings we chose for this current manuscript allowed us to have reasonably fast training times while keeping high performance metrics.

We have reworded the text in the updated manuscript and moved it closer to the beginning of Section 3.4.

RC2.17: 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…”

AR2.17: We have moved the text closer to the beginning of Section 3.4 in the updated manuscript.

RC2.18: 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?

AR2.18: We understand the request for more information on the dataset. This kind of information is a bit of a paradox and largely depends on the preferences of the reviewers. In manuscripts where we put this information in the first submission, some reviewers request that we put it in a supplement/appendix. In contrast, when we do not include this information in the first submission, some reviewers request this information. It can be difficult to strike a balance between providing too much information and too little information. We have tried to address this here.

In the updated manuscript, we have added Appendix A which contains more specific information on the datasets for each model run in each of the three cases. Appendix A includes information on the size of each dataset, the sizes of the training and testing subsets, and additional information on how the data was partitioned into training and testing subsets. We have also included more information on how the individual NNs were trained in the main body of the text.

With regards to whether the datasets were large enough to capture the variance, the datasets do appear to be large enough to effectively capture variance.

We did not use the depth-integrated data. We only used the surface values since this is where we also have information from remote sensing products in observational datasets. In a forthcoming manuscript, we demonstrate that the analysis developed here can be extended to such products, providing a useful constraint for ESMs.

RC2.19: 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.

AR2.19: Yes, R^2 can be a tricky metric when used by itself, especially on non-linear models. That is primarily why we also included RMSE as an additional metric, so they could be considered together.

The R^2 calculated when we compare the predictions of the NNEs to the "true" values of the ESM is the square of the Pearson correlation coefficient (i.e., standard R^2). The NNEs are predicting the monthly value of biomass (not mean annual biomass) and these are compared against the "true" monthly biomass values from the ESM. We have included this extra information in the updated manuscript.

RC2.20: Line 290-294 – What were your criteria for what constituted a significant increase or decrease in RMSE?

AR2.20: We understand the confusion from our use of the word "significant." We changed this to "substantial" in the updated manuscript so that we are not using a statistical term out of context.

RC2.21: 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?

AR2.21: We have included more information about the training and testing subsets and the sampling procedure in the updated manuscript. For more specific information, please see our response (AR2.18) where we address this in more detail.

RC2.22: 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.

AR2.22: It was not necessarily unexpected, but we were also not certain. We wanted to ensure that this result was verifiable, rather than just assuming, in case there were indirect effects we forgot to consider. This result reinforces what we find in the other cases.

RC2.23: 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 realworld data, we don't, and it can have significant consequences for predictive accuracy.

AR2.23: The application to real world data is something that we are currently trying to address in a separate manuscript. However, our preliminary results suggest that the inclusion of about 10 to 11 biogeochemical variables do well at predicting climatological phytoplankton biomass values (R^2 values between 0.6 to 0.85).

RC3.1: Line 1-2 - The use of "components" in the title is a little broad. Perhaps substitute with "Drivers of Plankton Biogeography"?

AR3.1: Changed "Components" to "Drivers of Plankton Biogeography."

RC3.2: 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.

AR3.2: Changed to "new patterns of co-limitation."

RC3.3: Line 110 – Introduce symbol for phytoplankton biomass (B) in this line

AR3.3: Added symbol *B* for phytoplankton biomass.

RC3.4: Line 138 - Typo, should read "than the"

AR3.4: Corrected.

RC3.5: Line 142 - Typo "nutrient and temperature"

AR3.5: Corrected.

RC3.6: Line 180 – Is this meant to read 'Section 3.4'?

AR3.6: Yes, it was supposed to be Section 3.4. This has been corrected in the updated manuscript.

RC3.7: Line 311 - Again, I think your original phrase "new patterns of colimitation" is more descriptive and appropriate than "fundamentally new states"

AR3.7: Changed "push the biology into fundamentally new states," to "lead to new patterns of co-limitation."

RC3.8: 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

AR3.8: Could you please clarify? Are you suggesting separate figures for small and large phytoplankton, such that we keep Figure 4 with **only** small phytoplankton and change Figure 3 to have **only** large phytoplankton? Or are you agreeing with the current layout with Figure 3 having **both** large and small phytoplankton and Figure 4 having **only** small phytoplankton?

The reason we chose to include both large and small phytoplankton in Figure 3 was so we could visualize the differences between them on the same plot. We did not give large phytoplankton its own figure since the apparent relationships of the large phytoplankton are already easily visible in Figure 3.