Author responses to Luke Gregor's (Referee 2) comments in the Supplemental PDF

The Supplemental PDF attached to Referee 2's comments was a PDF of the original submitted manuscript with specific comments by Referee 2 as highlighted PDF comments.

Any minor grammatical errors (commas, periods, and other punctuation) that were noted in Referee 2's comments will be corrected in the revised manuscript. So the discussion can be focused on the comments, we have not included those grammatical errors in this Author Response. However, we do want to thank Referee 2 for finding grammatical errors that we missed.

For ease of reading, in responses addressing the specific comments we have included the referenced paragraph from the original manuscript, along with their associated line numbers in black-colored font. The text sections in the original manuscript that were highlighted by Referee 2 are in orange-colored font. Any line numbers that are referenced refer to the line numbers of the original submitted manuscript. Referee 2's comments then follow the paragraph in green-colored font and our Author Responses follow this in red-colored font.

Acronyms used in this Author Response include OMT (Original Manuscript Text), RHS (Referee Highlighted Section), Referee Comment (RC), and Author Response (AR).

OMT:

Lines 10-30

Abstract. Controls on phytoplankton growth are typically determined in two ways: by varying one driver of growth at a time such as nutrient or light in a controlled laboratory setting (intrinsic relationships) or by observing the emergence of relationships in the environment (apparent relationships). However, challenges remain when trying to take the intrinsic relationships found in a lab and scaling them up to the size of ecosystems (i.e., linking intrinsic relationships in the lab to apparent relationships in large ecosystems). We investigated whether machine learning (ML) techniques could help bridge this gap. ML methods have many benefits, including the ability to accurately predict outcomes in complex systems without prior knowledge. Although previous studies have found that ML can find apparent relationships, there has yet to be a systematic study that has examined when and why these apparent relationships will diverge from the underlying intrinsic relationships. To investigate this question, we created three scenarios: one where the intrinsic and apparent relationships operate on the same time and spatial scale, another model where the intrinsic and apparent relationships have different timescales but the same spatial scale, and finally one in which we apply ML to actual ESM output. Our results demonstrated that when intrinsic and apparent relationships are closely related and operate on the same spatial and temporal timescale, ML is able to extract the intrinsic relationships when only provided information about the apparent relationships. However, when the intrinsic and apparent relationships operated on different timescales (as little separation as hourly to daily), the ML methods underestimated the biomass in the intrinsic relationships. This was largely attributable to the decline in the variation of the measurements; the hourly time series had higher variability

than the daily, weekly, and monthly-averaged time series. Although the limitations found by ML were overestimated, they were able to produce more realistic shapes of the actual relationships compared to MLR. Future research may use this type of information to investigate which nutrients affect the biomass most when values of the other nutrients change. From our study, it appears that ML can extract useful information from ESM output and could likely do so for observational datasets as well.

RHS1: Abstract (Line 10)

RC1: General comment: I find that the language used in the abstract might complicate the message.

AR1: Some of the confusion appears to be in the language and terminology used in the beginning of the abstract, which includes the lines mentioned in RC2, RC3, and RC4. The revised manuscript will clarify the language in the abstract to reflect the main points of the paper more accurately. A draft version of the revised manuscript includes rewriting the first five sentences of the abstract from the original manuscript and replacing them with:

> "A key challenge for biological oceanography is relating the physiological limitations controlling phytoplankton growth to the spatial distribution of those plankton. Physiological mechanisms are often isolated by varying one driver of growth such as nutrient or light in a controlled laboratory setting producing what we call "intrinsic relationships". We contrast these with the "apparent relationships" which emerge in the environment in climatological data. Although previous studies have found that machine learning (ML) can find apparent relationships, there has yet to be a systematic study examining when and why these apparent relationships diverge from the underlying intrinsic relationships found in the lab, and how and why this may depend on the method applied."

RHS2: emergence of relationships in the environment (apparent relationships). (Line 12)

RC2: This could be much more clear and explicit - i.e. observed nutrient concentrations and light intensity.

AR2: Please see our response in AR1.

RHS3: We investigated whether machine learning (ML) techniques could help bridge this gap. (Lines 14-15)

RC3: Be much more specific here. See my general comments what the title should be.

AR3: Please see our response in AR1.

RHS4: prior knowledge. (Line 16)

RC4: prior knowledge needs to be qualified - ML uses data to predict. Perhaps the authors mean "knowledge of the system" as stated later

AR4: The revised manuscript will clarify this with the term “knowledge of the system”. Please also see our response in AR1.

RHS5: apply ML to actual ESM output. (Line 21)

RC5: To do what?

AR5: The objective was to apply ML to actual ESM output as a proof-of-concept to the kinds of relationships and information one can find. Please also see our response in AR1.

RHS6: intrinsic relationships. (Line 25)

RC6: Why are you estimating intrinsic relationships? Is this predicted with the intrinsic relationship?

AR6: This sentence has been reworded. The new sentence reads:

> “When the intrinsic and apparent relationships operated on different timescales (as little separation as hourly to daily), NNEs fed with apparent relationships in time-averaged data produced responses with the right shape but underestimated the biomass.”

A key point here is that the intrinsic relationships are what gets coded into the models. If these are incorrect (i.e. the model gets the right answer in a given location due to compensating errors in the intrinsic relationships) it will have the wrong sensitivity to climate change.

RHS7: limitations found by ML were overestimated, (Line 27)

RC7: First mention of limitations - does this refer to limitations of growth?

AR7: The revised manuscript will make it clearer what is meant by the term “limitations”. Please see our draft version in AR1.

RHS8: MLR (Line 28)

RC8: write this out in full

AR8: The revised manuscript will have this written out as “Multiple Linear Regression”.

OMT:

Lines 44-56

Limitations on phytoplankton growth are usually characterized in two ways – which we term intrinsic and apparent. Intrinsic relationships are those where the effect of one driver (nutrient/light) at a time is observed, while all others are held constant (often at levels where they are not limiting). An example of such intrinsic relationships is the Michaels-Menten growth rate curves that emerge from laboratory experiments (Eppley and Thomas, 1969). Apparent relationships are those which emerge in the observed environment. An example of apparent relationships is those that emerge from satellite observations, which provide spatial distributions

of phytoplankton on timescales (say a month) much longer than the phytoplankton doubling time, which can be compared against monthly distributions of nutrients. A significant challenge that remains is determining how intrinsic relationships found in the laboratory scale up to the apparent relationships observed at the ecosystem scale (i.e., scaling the small to the large). Differences may arise between the two because apparent relationships reflect both intrinsic growth and loss rates, which are near balance over the long monthly timescales usually considered in climatological analyses. Biomass concentrations may thus not reflect growth rates. Differences may also arise because different limitation factors may not vary independently.

RHS9: Limitations on phytoplankton growth are usually characterized in two ways (Line 44)

RC9: This paragraph makes the terminology used in the abstract much clearer

AR9: The terms “intrinsic” and “apparent” relationships are terms that we define for the first time. To the best of our knowledge, these terms have not previously appeared in oceanography literature. Because we use these terms frequently throughout the manuscript, we included them in the abstract.

OMT:

Lines 58-74

Earth System Models (ESMs) have proved valuable in linking intrinsic and apparent relationships. The intrinsic relationships are programmed into ESMs as equations that are run forward in time and the output is typically provided as monthly-averaged fields. The output of these ESMs is then compared against observed fields such as chlorophyll and nutrients and can be analyzed to find apparent relationships between the two. If the ESM output is close to the observations we find in nature, we say that the ESM is performing well. However, as recently pointed out by Löptien and Dietze (2019), ESMs can trade-off biases in physical parameters with biases in biogeochemical parameters (i.e., they can arrive at the same answer for different reasons). Using two versions of the UVic 2.9 ESM, they showed that they could increase mixing (thus bringing more nutrients to the surface) while simultaneously allowing for this nutrient to be more efficiently cycled – producing similar distributions of surface properties. However, the carbon uptake and oxygen concentrations predicted by the two models diverged under climate change. Similarly, Sarmiento et al. (2004) showed that physical climate models would be expected to produce different spatial distributions of physical biomes due to differences in patterns of upwelling and downwelling, as well as the annual cycle of sea ice. These differences would then be expected to be reflected in differences in biogeochemical cycling, independent of differences in the biological models. These studies highlight the importance of constraining not just individual biogeochemical fields, but also their relationships with each other. What is less clear is: 1. Can robust relationships be found? 2. If so, what methods are most skillful in finding them? 3. How do you interpret the apparent relationships that emerge when they diverge from the intrinsic relationships we expect?

RHS10: relationships (Line 73)

RC10: between what and what?

AR10: Relationships in this sentence refers to relationships in very broad terms between biogeochemical fields and the target variable. The target variable could be phytoplankton biomass, chlorophyll concentrations, or other biogeochemical variables.

We have plans to revise this section in the revised manuscript and our current draft removes these questions from this paragraph and combines them with other sentences in the last paragraph of the introduction to more clearly highlight the main points of the manuscript.

OMT:

Lines 76-81

Recently, researchers have turned to machine learning (ML) to help in uncovering the dynamics of ESMs. ML is capable of fitting a model to a dataset without any prior knowledge of the system and without any of the biases that may come from researchers about what processes are most important. As applied to ESMs, ML has mostly been used to constrain physics parameterizations, such as longwave radiation (Belochitski et al., 2011; Chevallier et al., 1998) and atmospheric convection (Brenowitz and Bretherton, 2018; Gentine et al., 2018; Krasnopolsky et al., 2010, 2013; O'Gorman and Dwyer, 2018; Rasp et al., 2018).

RHS11: As applied to ESMs, ML has mostly been used to constrain physics parameterizations, such as longwave radiation (Belochitski et al., 2011; Chevallier et al., 1998) and atmospheric convection (Brenowitz and Bretherton, 2018; Gentine et al., 2018; Krasnopolsky et al., 2010, 2013; O'Gorman and Dwyer, 2018; Rasp et al., 2018). (Lines 78-81)

RC11: There is also now the study by Kasim et al (preprint). https://www.researchgate.net/publication/338762727_Up_to_two_billion_times_acceleration_of_scientific_simulations_with_deep_neural_architecture_search

AR11: Thank you for bringing this publication to our attention. In the revised manuscript we will conduct an additional search for the information in this paragraph of the manuscript to ensure we include recent literature.

OMT:

Lines 83-94

With regards to phytoplankton, ML has not been explicitly applied within ESMs but has been used on phytoplankton observations (Bourel et al., 2017; Flombaum et al., 2020; Kruk and Segura, 2012; Mattei et al., 2018; Olden, 2000; Rivero-Calle et al., 2015; Scardi, 1996, 2001; Scardi and Harding, 1999) and has used ESM output as input for an ML model trained on phytoplankton observations (Flombaum et al., 2020). Rivero-Calle et al. (2015) used random forest (RF) to identify the drivers of coccolithophore abundance in the North Atlantic through feature importance measures and partial dependence plots. The authors were able to find an

apparent relationship between coccolithophore abundance and environmental levels of CO_2, which was consistent with intrinsic relationships between coccolithophore growth rates and ambient CO_2 reported from 41 laboratory studies. They also found consistency between the apparent and intrinsic relationships between coccolithophores and temperature. While they were able to find links between particular apparent relationships found with the RFs and intrinsic relationships between laboratory studies, it remains unclear when and why this link breaks.

RHS12: it remains unclear when and why this link breaks. (Line 93)

RC12: The agreement between these two variables might be due to the scales of variability? Or a consistent reponse between the response of coccolithophores to temperature at a low and resolutions

AR12: Thank you for these suggestions. As we show in this paper, scales of variability can affect the link between intrinsic and apparent relationships.

OMT:

Lines 95-102

ML has been used to examine apparent relationships of phytoplankton in the environment (Flombaum et al., 2020; Rivero-Calle et al., 2015; Scardi, 1996, 2001) and it is reasonable to assume that ML could find intrinsic relationships when provided a new independent dataset from laboratory growth experiments. However, it has yet to be determined under what circumstances the apparent relationships captured by ML are no longer equal to the intrinsic relationships that actually control phytoplankton growth. In this paper, we identify two drivers of such divergence. The first is colimitation that limits the biological responses actually found in the ocean, which causes non-parametric ML methods to produce apparently non-physical results. The second is climatological averaging of the input and output variables, which can distort these relationships in the presence of non-linearity.

RHS13: identify (Line 99)

RC13: The use of "identify" here makes me think that this is a result from the study. Perhaps "propose"?

AR13: Other reviewers have also pointed to this portion of the introduction as possibly containing results from the paper. To avoid confusion, we will remove the following sentences from the introduction of the revised manuscript:

> “In this paper, we identify two drivers of such divergence. The first is colimitation that limits the biological responses actually found in the ocean, which causes non-parametric ML methods to produce apparently non-physical results. The second is climatological averaging of the input and output variables, which can distort these relationships in the presence of non-linearity.”

OMT:

Lines 115-118

In the first scenario, we wanted to determine how well different ML methods could extract intrinsic relationships when only provided information on the apparent relationships and when the intrinsic and apparent relationships were operating on the same timescale. In this scenario, the apparent relationships were simply the result of multiplying the intrinsic relationships between predictors and biomass by a scaling constant.

RHS14: In this scenario, the apparent relationships were simply the result of multiplying the intrinsic relationships between predictors and biomass by a scaling constant. (Lines 117-118)

RC14: I would add that three machine learning methods are used after this sentence.

AR14: In the revised manuscript, we will introduce earlier in the methods that we are using three machine learning methods and what they are (MLR, RF, and NNE).

OMT:

Lines 148-152

The final dataset consisted of three input/predictor variables and one response term with a total of 77,328 "observations." The input variables given to each of three ML methods (Multiple Linear Regression, Random Forests, and Neural Network Ensembles, described in more detail below) were the concentrations (not the limitation terms) for the micronutrient, macronutrient, and light. The response variable was the biomass we calculated from Eq. 1 and 2.

RHS15: (Multiple Linear Regression, Random Forests, and Neural Network Ensembles, described in more detail below) (Lines 149-150)

RC15: I would introduce the use of these methods earlier. i.e. the general concept explained briefly in the opening paragraph.

AR15: Please see our response in AR14.

OMT:

Lines 154-162

The dataset was then randomly split into training and testing subsets, with 60% of the observations going to the training subset and the remainder going to the testing subset. This provided a convenient way to test the generalizability of each ML method by presenting them with "new" observations from the test subset and ensuring the models did not overfit the data. The input and output values for the training subset were then used to train a model for each ML method. Once each method was trained, we provided the trained models with the input values of

the testing subset to acquire their respective predictions. These predictions were then compared to the actual output values of the test subset. To assess model performance, we calculated the coefficient of determination (R^2), the mean squared error (MSE), and the root mean squared error (RMSE) between the ML predictions and the actual output values for the training and testing subsets.

RHS16: randomly split into training and testing subsets, (Line 154)

RC16: COMMENT: I'm usually not a fan of random splits (particularly in a simulated environment), as training and testing data would have very similar distributions. But since the goal here is to see if it is possible to do exactly that (can ML capture the relationships), it makes sense.

AR16: Our main purpose for splitting the data into training and testing subsets is to ensure the machine learning methods are not overfitting the data. As you (Referee 2) state in RC16, we want to ensure that the machine learning models capture the relationships.

RHS17: convenient (Line 155)

RC17: Not sure I'd use convenient here. This is standard machine learning practice.

AR17: Because it is a standard machine learning practice, the revised manuscript will replace the term “convenient” with “standard” in the referenced sentence.

RHS18: “new” observations from the test subset and ensuring the models did not overfit the data. (Line 156-157)

RC18: Would be great to see a plot of the distribution of the training and test data (box plot). as well as the spatial distribution.

AR18: Since the observations in the training and testing subsets were randomly sampled from a large dataset, we felt it was apparent that the subsets contained observations of equal magnitude and spatiotemporal distribution. This was further reinforced in the similar performances of the training/testing subsets for the machine learning methods. However, we would like to remind the Referee and other readers that source files and code are freely available on the Zenodo data repository (https://doi.org/10.5281/zenodo.3932388, Holder and Gnanadesikan, 2020).

Holder, C. D. and Gnanadesikan, A.: Linking intrinsic and apparent relationships between phytoplankton and environmental forcings using machine learning - What are the challenges?, doi:10.5281/zenodo.3932388, 2020.

OMT:

Line 195

$$\overline{L_{Irr}} = \frac{\overline{Irr}}{K_{Irr}+Irr} \neq \frac{\overline{Irr}}{K_{Irr}+\overline{Irr}} \quad (4)$$

RHS19:

$$\overline{L_{Irr}} = \frac{\overline{Irr}}{K_{Irr}+Irr} \neq \frac{\overline{Irr}}{K_{Irr}+\overline{Irr}} \quad (4)$$

(Line 195)

RC19: Maybe add brackets around the fraction with the overbar applied to the entire equation

AR19: To clarify that the overbar applies to the entire fraction term, the revised manuscript will change the highlighted term to include brackets.

OMT:

Lines 196-198

(Eq. 4 appears before this in the original manuscript)

where the overbar denotes a time-average, and Irr stands for irradiance (light). We wanted to investigate how such time averaging biased our estimation of the intrinsic relationships from the apparent ones; i.e., how does the link between the intrinsic and apparent relationships change with different amounts of averaging over time?

RHS20: how does the link between the intrinsic and apparent relationships change with different amounts of averaging over time? (Lines 197-198)

RC20: Having this idea in the title would make the manuscript so much clearer!

AR20: We plan on changing the title of the manuscript in the revised version. Similar to what you (Referee 2) proposed in the general comments, our new draft title currently reads as: "Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations? – A proof of concept study".

OMT:

Lines 204-210

(Eq. 5 appear before this in the original manuscript)

where Irr_{Int} is the hourly interpolated value of irradiance, Irr_{daily} is the **daily-mean** value of irradiance, t is the hour of the day being interpolated, $t_{Sunrise}$ is the hour of sunrise, and T_{Day} is the total length of the day. The resulting curve preserves the day to day variation in the daily mean irradiance due to clouds but allows a realistic variation over the course of the day. The hourly values for the micronutrient and macronutrient were assigned using a standard interpolation between each of the daily values. These hourly interpolated values were then used to calculate the hourly biomass from Eq. 1 and 2. Note that we are not claiming the biomass itself would be

zero at night but assume that on a long enough timescale, it should approach the average of the hourly biomass.

RHS21: The hourly values for the micronutrient and macronutrient were assigned using a standard interpolation between each of the daily values. (Lines 207-208)

RC21: i.e. light is the only input/predictor that varies hourly.

AR21: The revised manuscript will include the clarification that light is the only predictor variable that varies with a daily cycle.

OMT:

Lines 221-224

As a demonstration of their capabilities, the ML methods were also applied directly to monthly averaged output from the BLING model itself using the same predictors in Scenarios 1 and 2, but using the biomass calculated from the actual BLING model. As described in Galbraith et al. (2010), BLING is a biogeochemical model where biomass is diagnosed as a non-linear function of the growth rate smoothed in time. The growth rates, in turn, have the form (continues on to Eq. 6 in the original manuscript)

RHS22: biomass is diagnosed as a non-linear function of the growth rate smoothed in time. (Lines 223-224)

RC22: Aha! This gives ligitimacy to the assumption made in EQ1 and Scenario 2, where hourly data is averaged to give biomass. I use bring this as a justification for the assumption.

AR22: The reviewer rightly points out that this assumption could have been pointed out earlier, and that this might have motivated the manuscript better. We now say at line 40 that:

> "As we will show, under certain formulations of ecosystem dynamics the phytoplankton biomass has a direct relationship to this growth rate."

OMT:

Line 224

$$B = \left(\frac{\tilde{\mu}}{\lambda} + \frac{\tilde{\mu}^s}{\lambda^s}\right) S_* \qquad (7)$$

RHS23:

$$B = \left(\frac{\tilde{\mu}}{\lambda} + \frac{\tilde{\mu}^s}{\lambda^s}\right) S_* \qquad (7)$$

(Line 234)

RC23: What is superscript a?

AR23: That is actually a “3,” not a lowercase a. The μ and λ in the fraction term after the “+” are both cubed. We will make the fonts in the equation larger in the revised manuscript.

OMT:

Lines 270-273

It should be noted that we are not trying to suggest that MLR is always ineffective for studying ecological systems. MLR is a very useful and informative approach for studying linear relationships within marine ecological systems (Chase et al., 2007; Harding et al., 2015; Kruk et al., 2011). However, we highly encourage our readers to try ML as it can provide insight into the non-linear portions of a dataset.

RHS24: It should be noted that we are not trying to suggest that MLR is always ineffective for studying ecological systems. MLR is a very useful and informative approach for studying linear relationships within marine ecological systems (Chase et al., 2007; Harding et al., 2015; Kruk et al., 2011). However, we highly encourage our readers to try ML as it can provide insight into the non-linear portions of a dataset.

RC24: Good! I was going to comment on this if it was not mentioned in the text.

AR24: Yes, we wanted to make it clear that we were not invalidating multiple linear regression. It is a very useful method!

OMT:

Lines 276-285

RFs are an ensemble ML method utilizing a large number of decision trees to turn “weak learners” into a single “strong learner” by averaging multiple outputs (Breiman, 2001). In general, RFs work by sampling (with replacement) about two-thirds of a dataset and constructing a decision tree. At each split, the random forest takes a random subset of the predictors and examines which variable can be used to split a given set of points into two maximally distinct groups. This use of random predictor subsets helps to ensure the model is not overfitting the data. The process of splitting the data is repeated until an optimal tree is constructed or until the stopping criteria are met, such as a set number of observations in every branch (then called a leaf / final node). The process of constructing a tree is then repeated a specified number of times, which results in a group (i.e., “forest”) of decision trees. Random forests can also be used to construct regression trees in which a new set of observations traverse each decision tree with its associated predictor values and the result from each tree is aggregated into an averaged value.

RHS25: two-thirds of a dataset and constructing a decision tree. (Line 278)

RC25: might be good to add that this is commonly referred to as bootstrap aggregation, or bagging in the machine learning world

AR25: We will include this term in the revised manuscript.

RHS26: Random forests can also be used to construct regression trees in which a new set of observations traverse each decision tree with its associated predictor values and the result from each tree is aggregated into an averaged value. (Lines 283-285)

RC26: This sentence is not completely clear. Are you trying to say that the predicted value is the average of all tree's prediction values.

AR26: Yes, that is what we are trying to say. When using random forest for regression (instead of classification), the predicted value is the average of all the individual trees' predictions.

OMT:

Lines 287-293

Here, we used the same parameters for RF in the three scenarios to allow for a direct comparison between the scenarios and to minimize the possible avenues for errors. Each RF scenario was implemented using the TreeBagger function in MATLAB 2019b, where 500 decision trees were constructed with each terminal node resulting in a minimum of five observations per node. An optimization was performed to decide the number of decision trees that minimized the error while still having a relatively short runtime of only several minutes. For reproducible results, the random number generator was set to "twister" with an integer of "123". Any remaining options were left to their default values in the TreeBagger function.

RHS27: minimum of five observations per node. (Line 290)

RC27: How did you decide on 5? This might result in overfitting given the nature and size of the training data set. Is there some sort of hyper-parameter selection process?

AR27: Five observations per node was the default number in the Matlab function used to construct the random forests. This means that at the end of each leaf of a decision tree within the random forest, five observations were being averaged for that single leaf. However, we still would not expect random forests to overfit the data. By construction, random forests generally do not overfit datasets for a couple reasons:

1. The random way in which data is selected to build individual decision trees.
    a. When any one decision tree is being constructed, the data is randomly selected with replacement from the available data until it reaches the same number of observations in the dataset. In general, this type of random sampling means that about 2/3 of the observations in a dataset will be captured by this type of sampling. This is then used to construct a decision tree. The process is repeated until the specified number of trees is reached. In our case, this would mean that our training subset is being randomly sampled for the construction of the decision

trees. This type of decision tree construction for random forests also means that no decision tree will be trained with every sample, which further decreases the likelihood of overfitting.

2. The random way in which the predictors are used in the construction of individual decision trees.
    a. When the decision trees are being constructed, only some of the predictors are available each time a split is determined. These predictors are randomly selected. Given those predictors, an error metric is used to determine the best split that will minimize the error. This random way in which predictors are selected decreases the chances of overfitting the data.

Additionally, we only allowed the random forest to be trained on the training subset. The other 40% of the data was in the testing subset and had never been "seen" by the random forests. Since the performance metrics of the training and testing subsets for random forest were very similar, this suggested to us that the relationships had been captured by the random forest in the training subset and those same relationships were present in the testing subset.

OMT:

Lines 304-310

Feed-forward NNs consist of nodes connected by synapses (or weights) and biases with one input layer, (usually) at least one hidden layer, and one output layer. The nodes of the input layer correspond to the input values of the predictor variables, and the hidden and output layer nodes each contain an "activation function." Each node from one layer is connected to all other nodes before and after it. The values from the input layer are transformed by the weights and biases connecting the input layer to the hidden layer, put through the activation function of the hidden layer, modified by the weights and biases connecting the hidden layer to the output layer, and finally entered into the final activation function of the output node.

RHS28: synapses (Line 304)

RC28: I would stick with weights rather than synapses. There has been a move away from comparing NNs with the brain as this gives far too much credit to the capabilities of NNs

AR28: The revised manuscript will use the term "weights" in place of "synapses".

RHS29: Each node from one layer is connected to all other nodes before and after it. (Line 306-307)

RC29: By design FFNNs are fully connected, meaning that each node from one layer is connected to all other nodes in preceding and succeding layers

AR29: Yes, that is correct. We included this line so that readers who may not be experienced in machine learning or neural networks understood the details of FFNNs.

OMT:

Lines 322-330

To minimize the differences between scenarios, we used the same framework for the NNs in each scenario. Each NN consisted of three input nodes (one for each of the predictor variables), 25 nodes in the hidden layer, and one output node. The activation function within the hidden nodes was a hyperbolic tangent sigmoid function and the activation function within the output node used a linear function. The stopping criteria for each NN was set as a validation check such that the training stopped when the error between the predictions and observations increased for six consecutive epochs. An optimization was performed to decide the number of nodes in the hidden layer that 11 minimized the error while maintaining a short training time. Additionally, sensitivity analyses were performed using different activation functions to ensure the choice of activation function had minimal effect on the outcome and apparent relationships found by the NNEs.

RHS30: validation check (Line 326)

RC30: What portion of the data was used for validation?

AR30: We used the default values in the Matlab function for training feedforward neural networks. These default values are 70% in the training set, 15% in the validation set, and 15% in the testing set. The function used in Matlab randomly partitions the data into these categories.

It should be noted that those partition values were only applied to our training subset, not the testing subset. The Matlab function used to train the NNs partitions our training subset into its own training, validation, and test sets. For example, this means that out of the 46,397 observations in our training subset of Scenario 1, 32,477 observations (70%) went to the training set of the NN function, 6,960 observations (15%) went to the validation set of the NN function, and 6,960 observations (15%) went to the test set of the NN function.

RHS31: An optimization was performed to decide the number of nodes in the hidden layer that (Line 327)

RC31: What kind of optimisation? Grid search? What ranges of nodes were explored?

AR31: Please see our responses to Referee 1 (AR8) for these details.

RHS32: ensure the choice of activation function had minimal effect on the outcome and apparent relationships found by the NNEs. (Lines 329-330)

RC32: I would phrase this differently. Perhaps the activation function does make a difference. There are tens of activation functions (https://stats.stackexchange.com/questions/115258/comprehensive-list-of-activation-functions-in-neural-networks-with-pros-cons). It would be more accurate to list the activation functions tested and state that it did not make a difference within these options.

AR32: The revised manuscript will list the activation functions that we tested.

OMT:

Lines 332-335

Each NNE scenario used the feedforwardnet function in MATLAB 2019b. Any options not previously specified remained at their default values in the feedforwardnet function. The NNEs contained ten individual NNs for each scenario. For reproducibility, the random number generator was set to “twister,” and the random number seed was set to the respective number of its NN (i.e., 1, 2, 3, up to 10).

RHS33: For reproducibility, the random number generator was set to “twister,” and the random number seed was set to the respective number of its NN (i.e., 1, 2, 3, up to 10). (Lines 334-335)

RC33: Details like this can be either in supplementary material and/or in the code.

AR33: Since this information is already in the code, we will remove this sentence from the revised manuscript.

OMT:

Lines 337-341

Each variable was scaled between -1 and 1 based on its respective maximum and minimum. This step ensures that no values are too close to the limits of the hyperbolic tangent sigmoid activation function, which would significantly increase the training time of each NN. These scalings were also applied to the RF and MLR methods for consistency between methods and the scaling did not affect the results of either method (results not shown). The results presented in this paper were then transformed back to their original scales to avoid confusion from scaling.

RHS34: scaled between -1 and 1 based on its respective maximum and minimum. (Line 337)

RC34: I'm interested to know why data wasn't scaled with MEAN and STDEV rather? This approach is usually a bit more robust to outliers. Perhaps this is not a problem with model data? i.e. model averages will not have outliers?

AR34: Please see our response in AR35.

RHS35: This step ensures that no values are too close to the limits of the hyperbolic tangent sigmoid activation, (Line 337-338)

RC35: Scaling the data ensures that the gradient of each variable has the same "steepness". https://stats.stackexchange.com/questions/322822/how-normalizing-helps-to-increase-the-speed-of-the-learning

AR35: Yes, it makes sure they have the same “steepness,” but it also ensures that the output of the activation function we are using (tangent sigmoid; tanh) are concentrated in a narrow range.

For example, if the input to the tanh activation function is between -1 and 1 (inside the red bars of Fig. 1), the range of the output is between about -0.76 and 0.76. In contrast, if the input is outside the range of -1 to 1 (outside the red bars of Fig. 1), the output quickly approaches the extremes of -1 and 1 on the y-axis. If the outputs are toward the extreme ends, this can cause the NNs to get "stuck" in those extremes during training which affects how much the weights can be adjusted during each epoch (ie. more epochs are needed for training which leads to longer training times).

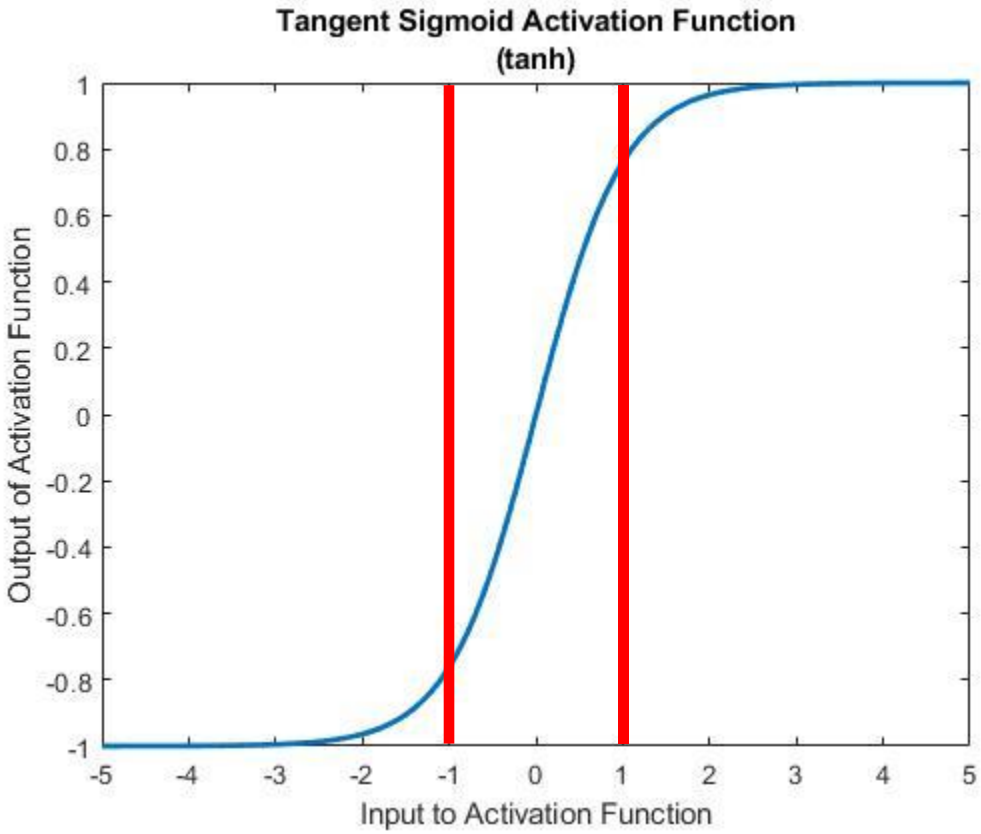

Figure 1: Tangent sigmoid activation function.

RHS36: These scalings were also applied to the RF and MLR methods for consistency between methods and the scaling did not affect the results of either method (results not shown). (Lines 339-340)

RC36: Rephrase: Scaling of the input variables is not necessary for RF and MLR, but was still applied for each of the methods for the sake of consistency with NNE.

AR36: The revised manuscript will rephrase the highlighted sentence.

RHS37: paper were then transformed back to their original scales to avoid confusion from scaling. (Line 341)

RC37: Scaling the output/target variable is not usually done. Though this would make no difference. I think you can thus leave this sentence out.

AR37: We received questions in past submissions when we did not specify that the values were transformed back to their original values. For clarity, we would like to keep this sentence (or something similar) in the revised manuscript.

OMT:

Lines 345-349

In Scenario 1, the RF and NNE both outperformed the MLR as demonstrated by higher R_2 values, lower MSE, and lower RMSE (Table 1). The decreased performance of the MLR is not inherently surprising, given the non-linearity of the underlying model, but it does demonstrate that the range of nutrients and light produced as inputs by ESM2Mc is capable of producing a non-linear response. Additionally, each method showed similar performances between the training and testing subsets suggesting adequate capture of the model dynamics in both subsets.

RHS38: Additionally, each method showed similar performances between the training and testing subsets suggesting adequate capture of the model dynamics in both subsets. (Lines 348-349)

RC38: I have a feeling that this might be due to the random shuffling of testing and training data; i.e. the training dataset is almost perfectly representative of the test dataset

AR38: It was our intention that the training and testing subsets be representative of one another and of the complete dataset. The highlighted sentence provides support to that intention.

OMT:

Lines 367-373

When we computed an "effective" half-saturation for the nutrient curves in the top row of Fig. 2, we got values for K_N that were far lower than the actual ones specified in the model (Table 4). The "effective" half-saturation of when other predictors are held at their 25_{th} percentile for the micro- and macronutrient were underestimated by one and two orders of magnitude, respectively. It was only at the higher percentiles that the micronutrient "effective" half-saturation was adequately captured when the macronutrient was not limiting. Furthermore, the "effective" half-saturation of the macronutrient was not captured even when the other variables were held at their 75_{th} percentiles because the 75_{th} percentile of the micronutrient still limited growth.

RHS39: "effective" (Line 367)

RC39: why is this in quotations - maybe pseudo/quasi is a better word here?

AR39: These quotations will be removed in the revised manuscript. Additionally, we will remove other unnecessary quotation marks in the revised manuscript.

OMT:

Lines 424-436

Despite the fact that it agreed well with the observations, the RF prediction deviated from the true response to a given variable when other variables are held at higher percentiles (Fig. 2). This can likely be explained by the range of the training subset and how RFs acquire their predictions. When presented with predictor information, RFs rely on the information contained within their training data. If they are presented with predictor information that goes outside the range of the dataspace of the training set, RFs will provide a prediction based on the range of the training set. When performing the sensitivity analysis, the values of the predictors in the higher percentiles were probably outside the range of the training subset. For example, the bottom left plot of Fig. 2 shows how RF deviates from the true response as the concentration of the macronutrient increases – actually decreasing as nutrient increases despite the fact that such a result is not programmed into the underlying model. Although there may be observations in the training subset where the light and micronutrient are at their 75th percentile values when the macronutrient is low, there likely are not any observations where high levels of the macronutrient, micronutrient, and light are co-occurring. Without any observations meeting that criteria, the RF provided the highest prediction it could based on the training information. We discuss this point in more detail below.

RHS40: it (Line 424)

RC40: It would be better to put "the RF prediction" first and then "it"

AR40: The revised manuscript will restructure this sentence to be: “Despite the fact that the RF prediction agreed well with the observations, it deviated from the true response to a given variable when other variables were held at higher percentiles.”

OMT:

Lines 462-474

When comparing the apparent relationships of the time-averaged datasets with those of the hourly intrinsic relationships, the methods almost always underestimated the true response to light and nutrient (Fig. 3 and 4). This result is not entirely unexpected. The averaging of the hourly values into daily, weekly, and monthly timescales quickly leads to a loss of variability, especially for light (Fig. 5). In fact, the variability was lost in the daily time averaging with the longer timescales showing only small differences in the possible range of values (Fig. 5). The loss of variability means that the light limitation computed from the averaged light is systematically higher than the averaged light limitation. To match the observed biomass, the asymptotic biomass at high light has to be systematically lower (see Appendix A for the mathematical proof). Differences were much smaller for nutrients as they varied much less over the course of a month in our dataset. Our results emphasize that when comparing apparent

relationships in the environment to intrinsic relationships from the laboratory, it is essential to take into account which timescales of variability averaging has removed. Insofar as most variability is at hourly time scales, daily-, weekly-, and monthly-averaged data will produce very similar apparent relationships (Fig. 4). But if there was a strong week-to-week variability in some predictor, this may not be the case.

RHS41: has (Line 472)

RC41: was / has been

AR41: This correction will be made in the revised manuscript.

OMT:

Lines 495-501

The large increases in biomass in the micronutrient plots and hindrance of biomass in the light and macronutrient plots suggest that the system is limited by the concentration of micronutrient (Fig. 7). The biomass remained low even when macronutrient and light were at favorable levels because even when at the 75_{th} percentile value, the micronutrient was still limiting (Fig. 8). Conceptually this makes sense since the micronutrient limitation in the BLING model hinders growth, but also limits the efficiency of light-harvesting (Galbraith et al., 2010). Additionally, the computation of the "effective" half-saturation constants demonstrates that the half-saturation constant for light drops sharply as nutrients drop (Table 4).

RHS42: computation of the "effective" (Line 500)

RC42: Could be replaced by "estimated"

AR42: Yes, "estimated" seems to be the word more in line with our intention. The term "effective" will be replaced in the revised manuscript.

OMT:

Lines 504-507

Our main objective in this manuscript was to use ML to determine under what conditions intrinsic and apparent relationships between phytoplankton are no longer equal, to identify whether such divergence depends on the ML method or how the input data is handled, and to understand how such divergence is related to underlying biological dynamics.

RHS43: how the input data is handled, (Line 506)

RC43: Be more specific here, bring in the time aspect.

AR43: The revised manuscript will include more specific information, such as temporal averaging.

OMT:

Lines 538-548

Both RFs and NNEs performed well when the predictions they were asked to make were within the range of the training data. However, the sensitivity analyses illustrated the impact of RFs inability to extrapolate outside that range and that RF's suggested systematic decreases in biomass at high values of a limiting variable. Nonetheless, RFs were able to capture the same relationships as the NNEs when the sensitivity analysis was querying environments within the range of the training data. It seems that as long as RFs are presented with information across the range of the dataset, RFs will perform just as well as NNEs in a sensitivity analysis. This strengthens the conclusions of Rivero-Calle et al. (2015) in that physiologically reasonable relationships between forcing variables and biomass found using RF are reliable so long as the forcing variables (in this case pCO_2 and temperature) vary over their entire range independently of other variables (nutrients and light). However, when variation in pCO_2 is related to variation in nutrients and light (i.e., in the seasonal climatology where pCO_2 is high in the winter, light is low, and nutrients are high) RFs are unable to extract a clear signal of pCO_2 limitation.

RHS44: This strengthens the conclusions of Rivero-Calle et al. (2015) in that physiologically reasonable relationships between forcing variables and biomass found using RF are reliable so long as the forcing variables (in this case pCO2 and temperature) vary over their entire range independently of other variables (nutrients and light). However, when variation in pCO2 is related to variation in nutrients and light (i.e., in the seasonal climatology where pCO2 is high in the winter, light is low, and nutrients are high) RFs are unable to extract a clear signal of pCO2 limitation. (Lines 543-548)

RC44: This is new literature in the conclusion. I would try to bring this into the discussion rather.

AR44: We will move the highlighted sentences to the discussion in the revised manuscript.

OMT:

Lines 562-573

ML techniques have several benefits that could make them useful for biological oceanographers and ecosystem modelers. Many ML methods (including the two presented here) do not require any prior knowledge of a system to construct a model. Additionally, new methods are continually being developed for viewing the dynamics of the ML models. Given these advantages, ML could provide a compact form for representing relationships between ecosystem parameters such as biomass and primary productivity and their environmental drivers (nutrients and light) in observational data and complex models. Preliminary work indicates that we can use NNEs in particular to: 1. Compare model relationships with those derived from observational datasets, rather than simply using spatial patterns of errors. 2. Evaluate whether differences between models reflect important differences in biological parameters or whether they are due to

differences in the physical circulation. We would expect that two different physical models run with the same biological scheme would produce the same relationships. 3. Evaluating whether global warming really would be expected to drive ecosystems outside their historical parameter range. We will report on these results in a future manuscript.

RHS45: dynamics (Line 564)

RC45: What do you mean by dynamics? Do you mean feature importances as an example?

AR45: In this sentence, we were referring broadly to new ways of viewing the relationships that are found by machine learning methods.

RHS46: compact form (Line 565)

RC46: unclear what is meant by this?

AR46: In this sentence, we were trying to state that machine learning could provide a way of comparing relationships found in observational datasets to relationships found in output of Earth System Models.

RHS47: relationships (Line 568)

RC47: relationships of what?

AR47: In this sentence, we were referring broadly to biogeochemical relationships in Earth System Models and observational datasets.

RHS48: Evaluate whether differences between models reflect important differences in biological parameters or whether they are due to differences in the physical circulation. (Lines 569-570)

RC48: Also, as pointed out by the author in the introduction, it may also differentiate whether models are similar for the right reason.

Would one not have to be careful of the implementation of the "intrinsic" equations and half-saturation constants of the model?

AR48: Yes, one would need to be careful of the implementation of the intrinsic equations and half-saturation constants. We are currently working on a manuscript that discusses such questions.

OMT:

Line 758

Table 3: Scenario 3 comparison of MLR, RF, and NNE method performance for the training and testing sets.

RHS49: Scenario 3 comparison of MLR, RF, and NNE method performance for the training and testing sets. (Line 758)

RC49: Be specific about what the target variable is (biomass). Further, there are no units

AR49: The revised manuscript will include the target biomass variable, along with its units, in the captions for the Tables.

OMT:

Line 763

(References Table 4 in the original manuscript)

RHS50: -2.11 x 10^4 (This is the value in Table 4 associated with the 25th percentile of the Micronutrient for Scenario 3; Line 763)

RC50: Typo?

AR50: That is not a typo. One result we found was limitations in the variables can affect the estimate of the half-saturation. The issue is that there are not any observations where the macronutrient is at the 25th percentile or below when the micronutrient is limiting. The micronutrient is more or less uncorrelated with biology in this range and any half-saturations derived from it will be poor estimates. This clarification will be made in the revised manuscript.

RHS51: 1.85 (This is the value in Table 4 associated with the 25th percentile of the Light variable for Scenario 3; Line 763)

RC51: Should these light half-saturation estimates be the same as those above? If so, why is this so far off? If not, then make this clear!

AR51: In BLING, we use the Geider et al. model for light limitation. In this model, the chlorophyll to carbon ratio θ adjusts with the light; it becomes lower as light gets higher and higher as light gets lower. Since $Irr_k \propto \frac{Lim_{nut}}{\theta}$, this means that the ratio Irr/Irr_k ends up being a lot more constant than one might expect – essentially it only drops to zero when the plankton cannot make any more chlorophyll. Please note that in Scenarios 1 and 2 we assumed that $Irr_k$was independent of *Irr*. At very low values when nutrients are highly limiting, Irr_k is very small; while at higher values of nutrients, it is larger.

OMT:

Lines 791-792

Figure 6: Scatter plots from the BLING model (a: surface biomass vs. temperature-normalized growth rate; b: mean nutrient limitation vs. monthly-averaged nutrients; c: mean light limitation vs. monthly-averaged Irr, Irr_k).

RHS52: Figure 6: Scatter plots from the BLING model (a: surface biomass vs. temperature-normalized growth rate; b: mean nutrient limitation vs. monthly-averaged nutrients; c: mean light limitation vs. monthly-averaged Irr, Irr_k). (Lines 791-792)

RC52: should the dependent (predicted) variables not be on the y-axes?

AR52: As it is currently constructed, the horizontal axis represents an "input" constructed from monthly mean variables (light, nutrient), while the vertical axis represents a target computed by the model. We will clarify this in the revised manuscript.

OMT:

Lines 795-798

Figure 7: Sensitivity analysis for Scenario 3 with the columns corresponding to the predictors and the rows corresponding with the percentile value at which the other predictors were set. The gray circles show the observations from the BLING model and the dashed lines show the predicted apparent relationships for each method.

RHS53: Figure 7: Sensitivity analysis for Scenario 3 with the columns corresponding to the predictors and the rows corresponding with the percentile value at which the other predictors were set. The gray circles show the observations from the BLING model and the dashed lines show the predicted apparent relationships for each method.

RC53: I have two comments here:

> 1) It would be more useful to have a 2D histogram in this data representation. Represent this data as a 2D contour where the colormap is scaled logarithmically. Using a grey colormap will then allow you to still plot the dashed lines in colour.
>
> 2) What are the actual curves for the intrinsic relationship?

AR53: Addressing comment 1: Yes, that is a good suggestion. We will try that implementation to see if it improves the layout of the plots.

Addressing comment 2: Since Scenario 3 was being used as a proof-of-concept, we did not include the intrinsic relationships. The reason for this was we were attempting to demonstrate the type of information one could gain from having access to Earth System Model output only, but not necessarily access to the computational resources to run the Earth System Model code itself.

RHS54: (The Referee comment is posted next to the Legend in Figure 7; Line 794)

RC54: Are the different models necessary here? If RF is susceptible to "missing data" why should it be used in an even more complex scenario? The same applies for MLR - it is clearly not complex enough to capture this signal.

AR54: The revised manuscript will only include sensitivity curves for the NNE in this figure.

RHS55: (The Referee comment highlights the label "75th percentile" in Figure 7; Line 794)

RC55: Is it still useful to show the different percentiles at this point? Is this point not proven in experiment 2.

AR55: The different percentiles still affect the result. For example, in Figure 4 the predicted biomass values increase with higher percentiles. The values on the y-axis change between the subplots.

RHS56: (The Referee comment highlights the "P" in the units for Biomass in Figure 7; Line 794)

RC56: What is P?

AR56: Here, P stands for phosphorus. This is analogous to the macronutrient term.

OMT:

Lines 801-802

Figure 8: A 3-D scatter plot showing the concentrations from Scenario 3 for the macronutrient, micronutrient, and light with the color of the data points corresponding to the biomass concentrations.

RHS57: Figure 8: A 3-D scatter plot showing the concentrations from Scenario 3 for the macronutrient, micronutrient, and light with the color of the data points corresponding to the biomass concentrations. (Lines 801-802)

RC57: I don't think this is a very informative plot. It is difficult to see what is really going on with the majority of the data being obscured by other data. In addition to this, it is quite strange to have the x-axis (macro) going from large to small, rather than small to large

AR57: The main purpose of this figure was to show that some of the relationships found by the sensitivity analysis were reflected in the data. Specifically, that increases in the micronutrient led to large increases in biomass.

It was necessary to plot it with the macronutrient going from large to small, otherwise the area we were wanting to focus on would be obscured by the other data.

We will consider removing this plot in the revised manuscript.