Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2020-262-AC2, 2020

Biogeosciences Discussions

Interactive comment on "Linking intrinsic and apparent relationships between phytoplankton and environmental forcings using machine learning – What are the challenges?" *by* Christopher Holder and Anand Gnanadesikan

Christopher Holder and Anand Gnanadesikan

cholder2@jh.edu

Received and published: 25 September 2020

Author responses to Luke Gregor (Referee 2)

In the following responses RC stands for Referee Comment and AR stands for Author Response. For sections where draft paragraphs for the revised manuscript are included, the beginning and end of the draft paragraphs are denoted with *BD* (Begin Draft) and *ED* (End Draft).

For detailed descriptions of the tables and figures included with this Author Response,

please see the Supplemental PDF included with this Author Response.

RC0: The study by Holder and Gnanadesikan tries to assess if machine learning is able to extract the intrinsic relationship between phytoplankton growth and limiting nutrients and light from observed concentrations of nutrients and light intensity. This topic was investigated with three experiments of increasing complexity asking the following questions (with my brief understanding of the outcomes):

1. Are ML methods able to extract the relationship from observations at all at instantaneous time scales? a. Yes, but NNE is better at extracting the relationship than RF despite both achieving fair results 2. If time scales are averaged, can the relationships still be extracted? a. Not very well. In most cases the estimated half-saturation is lower than it should be. I.e. even the better of the two ML methods, NNE, is not very accurate. 3. Can the approach work in a more complex model setup where biomass losses are also accounted for?

While I appreciate the question the study is asking and think that this work is important, I found that the manuscript was not very easy to follow (my summaries of the results above might illustrate this). Part of the difficulty may be that the topic is not within my immediate field of expertise, but then I feel there are stylistic changes to be made that will improve the manuscript. I have overall comments in the document below and I linked a PDF document with comments at the very end of this document (I used Adobe Reader). I hope these comments help improve the flow of the manuscript.

AR0: We would like to thank Luke Gregor as Referee 2. We have found the comments and suggestions they provided to be very helpful in restructuring this manuscript. In particular, the supplement to these comments has provided some very specific and constructive feedback.

RC1: The title can be improved. To someone who is not familiar with the "intrinsic" and "apparent" terminology, the title is not informative. Something along the lines of : Can machine learning extract the mechanisms of phytoplankton growth from large-scale

observations?

AR1: We understand how including the terminology in the title can lead to confusion. The revised manuscript will have a different title. The current draft title we have is "Can machine learning extract the mechanisms controlling phytoplankton growth from large-scale observations? – A proof of concept study".

RC2: The use of "intrinsic" and "apparent" relationships this early in the manuscript made it difficult to understand the study as I am not familiar with the terminology.

AR2: The terms "intrinsic" and "apparent" relationships are actually terms that we are defining for the first time in this manuscript. They have not been previously introduced in oceanography literature. Since these terms are used frequently throughout the paper, we find it helpful to introduce them early, including in the abstract.

RC3: I don't have major concerns with the introduction and it builds a good case for why this study is relevant.

AR3: We thank Referee 2 for this kind complement.

RC4: The questions posed (L72-75) and ideas presented (L100-102) are useful in framing the study but are not carried clearly through the manuscript. It would be very useful for the reader to have these questions and ideas as a guide for why each experiment was performed. For example, L72-75 from the basis of experiment 1, but these questions are not explicitly answered in the discussion. And lines 100-102 form the basis of the design for experiment 2.

AR4: We have changed some aspects of the introduction based on other Referee comments, which encompass similar feedback as in RC4 above. In the revised version of the manuscript we plan to remove lines 100-102, as these lines list results in the introduction. We also plan to modify the introduction in the revised manuscript. A draft form of a portion of the introduction that more clearly highlights the main points of the paper is listed below:

BD To investigate when and why the link between intrinsic and apparent relationships break, we try to answer two main questions in this paper: 1. Can ML techniques find the correct underlying intrinsic relationships and, if so, what methods are most skillful in finding them? 2. How do you interpret the apparent relationships that emerge when they diverge from the intrinsic relationships we expect? In addressing the first question, we first needed to demonstrate that we had an ML method that would correctly extract intrinsic relationships from apparent relationships. We constructed a simple model in which the intrinsic and apparent relationships operated on the same time and spatial scale and were only separated by a scaling factor, but in which the environmental drivers had realistic inter-relationships. Having a better handle on the results from the first question, we were able to move onto the second question where we look at where the link between intrinsic and apparent relationships break. We modified the first scenario to allow the intrinsic and apparent relationships to operate on different timescales – allowing us to evaluate the impact of time-averaging on the retrieval of intrinsic relationships. Finally, we conduct a proof-of-concept study with real output from an ESM. *ED*

RC5: There is no overview of the methods. I think this would be useful in addition to an accompanying diagram outlining all the experiments and the use of the machine learning approaches used. It would help the reader understand the flow of the study. BLING is used throughout the study, albeit with different outputs from the model, but it may make sense to introduce the model before the experiment configurations are described.

AR5: We have included a diagram (Table 1) outlining the details of each scenario which include: the predictor variables, the target variable, the equations used to calculate biomss, a description of the source file, and a short description of each scenario.

Because the machine learning approaches are the same for each Scenario, we didn't think it would be necessary to include a table or diagram showing this. However, in the revised manuscript, we will state more clearly in the methods that the same machine

learning approaches are used for each scenario.

The main reason for including the description of BLING in the third scenario was so readers would not get confused as to which equations and model are being used for each scenario if it was introduced before the explanation of the Scenarios. However, we will consider whether to move the BLING description before the scenario explanations in the revised manuscript now that we have included a diagram (Table 1) outlining the details of each scenario.

RC6: It would make sense to formalise the following structure for each experiment:

âĂć A brief introduction to the experiment

âĂć HEADING for data

âĂć HEADING for Machine learning parameterisation / application

AR6: We agree with the idea about formalizing the structure of each experiment. The revised manuscript will include a structure for each experiment similar to that described in RC6 above.

RC7: In experiment 2, the authors create hourly data by simulating variability of light conditions. The data are then averaged again to create daily, weekly and monthly data. If I understand correctly, the hourly data is analogous to the data used in experiment 1 - i.e. there is no temporal averaging in the "apparent data". It would be much more methodologically consistent to use the hourly data in experiment 1 and easier for the reader to follow. Either, the authors should implement this, or should make this explicit and state the reason that a separate experiment is needed.

AR7: Yes, the Referee is correct in their understanding that the hourly data is analogous to the data used in Scenario 1 where there was no temporal averaging. We agree that it would be easier for the reader to follow, and we spent several days testing this strategy.

The main issue we ran into was with the size of the hourly dataset. Across all longitudes, latitudes, and hours for a single year, this results in a dataset with 56,214,560 observations. We attempted to randomly sample the dataset with up to 500,000 points to train the machine learning algorithms. Quantities of observations higher than 500,000 were leading to computer crashes because of the computational power required for training the ML algorithms. While it is technically feasible to train random forests and neural networks on this number of observations, this would still require very long spans of time for training each ML method. Since we would like this paper and the methods to be accessible to everyone, we would like our Matlab code to be able to run on a standard laptop. With this in mind, we chose the first BLING scenario since it was already at monthly timescales and the number of observations was significantly less than the amount in an hourly dataset over the course of a year. The number of samples in the monthly dataset of Scenario 1 is only 77,328 compared to the 56 million of the hourly dataset.

Additionally, as we now show (please see our response AR11 to Referee 1) that adding length of day as a variable or going to very high percentiles of other variables does appear to allow the NNEs to correctly extrapolate the correct relationships even in the time-averaged datasets.

RC8: Another question is regarding the model: what is the variability of the nutrients at a daily resolution (native model resolution), and the averaged resolutions (weekly, monthly). Show some violin/box plots for the normalised data.

AR8: A figure including boxplots for the time-averaged datasets of Scenario 2 will be included in the revised manuscript. A draft version of that figure is included below in Fig. 1.

RC9: I still don't fully understand what the predictors and target variables are for each experiment and what is the role of the intrinsic? From what I understand, predictors are always the "apparent" data and biomass is the target. The intrinsic is what describes

the relationship between the biomass and the “apparent data”. Please make this more clear. Addressing the points above in the structure section will help with this.

AR9: With the inclusion of Table 1 below, the predictors and target variables for each Scenario are included there. However, the revised version of the manuscript will also clarify this in the methods section as well.

Regarding the intrinsic and apparent relationships, the intrinsic relationships are those in which the effects of other variables affecting the target variable can be accounted. For example, if one is measuring the effect of macronutrient concentrations on phytoplankton in the lab, it is possible for them to hold concentrations of other variables (light, micronutrient, water temperature, salinity, etc.) at some particular value. Apparent relationships are those for which the effects of other variables affecting the target variable cannot be accounted (ex. taking measurements in the field). Another way of saying this is that intrinsic relationships are the underlying relationships governing a system where you can adjust one variable at a time (such as a lab). Apparent relationships are determined by how the intrinsic relationships combine in the environment when variables cannot be adjusted one variable at a time. We will try to clarify this distinction in the revised manuscript.

RC10: The authors should only NNE results for experiment 2 (figure 4). Is there a reason for this? My presumption is that the intrinsic relationship estimated by RF for micronutrients is poor, thus only NNE is shown. This should be cleared up (unless I missed this).

AR10: The presumption of the Referee is correct. Because the RF performs poorly and is incapable of extrapolating outside the range of the training dataset, we chose to limit further analyses of Scenario 2 to NNEs. We will clarify this in the revised manuscript.

RC11: From what I understand, the half-saturation constants are the metric for whether the method is able to capture the intrinsic from the apparent. Make this much more clear - also in the abstract

AR11: That is correct. We are using the calculated half-saturation constants as a metric to help identify if the methods are capturing the true relationships. We will clarify this in the revised manuscript.

RC12: The subheadings could be the questions posed in the introduction (see my previous comments on this section). This would help guide the reader

AR12: Yes, we agree that subheadings in the discussion could aid in guiding the reader. As the Referee suggests, we will consider using the questions posed in the introduction as subheadings. The revised manuscript may include subheadings in the discussion section.

RC13: I think the authors should make the point that given the simplicity of the definition of biomass, one would expect the ML methods to perfectly represent the Michaelis-Menten curves. The authors do correctly state that RF is less likely to estimate accurately as the method is not able to extrapolate. This then increases the importance of showing the distributions of the training and test data set distributions. A further comment: what is the envelope around the estimated curves and why is there a large variability for the NNE at larger values?

AR13: To keep the number of figures in the manuscript to a minimum, we had not included boxplots of each variable in each Scenario. However, we see the use that information can provide. The revised manuscript may include the distributions of the training and test subsets for each Scenario in the Supplementary Materials section.

The gray regions around the dashed lines for the random forest (RF) and neural network ensemble (NNE) predictions show the standard deviation in the predictions. For example, the NNEs are composed of 10 individual neural networks and each one produces its own predictions. For the sensitivity analysis figures, the dashed lines for NNE show the average prediction of those 10 individual neural networks. Similarly, the gray regions show the range of one standard deviation for those predictions. We will clarify this in the revised manuscript.

The large variability for the NNEs at the larger values is likely because those particular conditions are outside the range of the dataset on which the NNEs were trained. For example, it is rare that any of the observations would have high macronutrient, high micronutrient, and high irradiance occurring at the exact same time and location. Without any observations in the training subset meeting those types of criteria and the NNE never having seen what those conditions actually produce, the NNE predictions become less certain.

RC14: The discussion around scenario/experiment 3 is not clear and I don't feel that there is a take-home message after reading this section.

AR14: The purpose of Scenario 3 is largely to provide a proof-of-concept to how the techniques we demonstrate in Scenarios 1 and 2 can be applied to Earth System Model output. The revised manuscript will expand on this and better highlight the main goal of Scenario 3.

RC15: The captions are not standalone for both figures and tables.

AR15: The revised manuscript will include more detailed descriptions of the tables and figures. A draft version for Figure 2 of the original manuscript currently reads:

BD Figure 2: Sensitivity analysis for Scenario 1 showing the true and predicted relationships for how each predictor affects the biomass when the other predictors are set at specific percentiles. The columns correspond to the predictors and the rows correspond with the percentile value at which the other predictors were set. The black line shows the true intrinsic relationship calculated from Eq. 1 and 2. The dashed lines show the predicted apparent relationships for each method (MLR – red; RF – blue; NNE – green). The gray region around the RF and NNE dashed lines shows the standard deviation of the predictions. *ED*

RC16: The reader needs to know what the target variable in each table is and there are no units.

AR16: The revised manuscript will include the target variable and its units in the description of each table.

RC17: What is the envelope around the dashed lines.

AR17: The gray regions around the dashed lines for the random forest (RF) and neural network ensemble (NNE) predictions show the standard deviation in the predictions. For example, the NNEs are composed of 10 individual neural networks and each one produces its own predictions. For the sensitivity analysis figures, the dashed lines for NNE show the average prediction of those 10 individual neural networks. Similarly, the gray regions show the range of one standard deviation for those predictions.

RC18: Please also note the supplement to this comment: https://bg.copernicus.org/preprints/bg-2020-262/bg-2020-262-RC2-supplement.pdf

AR18: The additional Referee comments in the supplement are very helpful. We will address these in a separate Author Response and/or implement the suggestions in the revised manuscript.

Please also note the supplement to this comment:
https://bg.copernicus.org/preprints/bg-2020-262/bg-2020-262-AC2-supplement.pdf

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2020-262, 2020.

Scenario	Predictors	Target	Equations Used	Source File Description	Scenario Description
1	Macronutrient (mol kg^{-1}); Micronutrient (mol kg^{-1}); Irradiance (W m^{-2})	Biomass (mol kg^{-1})	1, 2	Monthly Output from BLING	Nutrient distributions (predictors) from BLING were fed to Eq. 1 and 2 to calculate the biomass (target)
2	Macronutrient (mol kg^{-1}); Micronutrient (mol kg^{-1}); Irradiance (W m^{-2})	Biomass (mol kg^{-1})	1, 2, 5	Daily Output from BLING	1) Hourly values for the predictors were interpolated using the Daily Output of BLING 1a) The macronutrient and micronutrient hourly values were calculated using a standard interpolation between the daily points. 1b) The irradiance hourly values were calculated from Eq. 5 using the value of the BLING daily input, hour of day, time of year, and location. 2) Hourly values of the predictors were fed to Eq. 1 and 2 to calculate hourly values for the biomass (target) 3) Daily-averaged values were calculated by averaging 24 hours for each location through one year 4) Weekly-averaged values were calculated by averaging 168 hour blocks of time for each location through the year 5) Monthly-averaged values were calculated by averaging the number of hours in each month (days per month * 24) for each location through the year 6) The true relationships were calculated by using the range of the hourly values for the predictors and calculating the biomass based on Eq. 1 and 2.
3	Macronutrient (mol kg^{-1}); Micronutrient (mol kg^{-1}); Irradiance (W m^{-2})	Biomass (mol kg^{-1})	6, 7 (Equations within BLING used to determine the biomass)	Monthly Output from BLING	Nutrient distributions from the BLING Output were used as the predictors; Biomass from the BLING Output itself was used as the target

Fig. 1. Table 1: Details for each Scenario

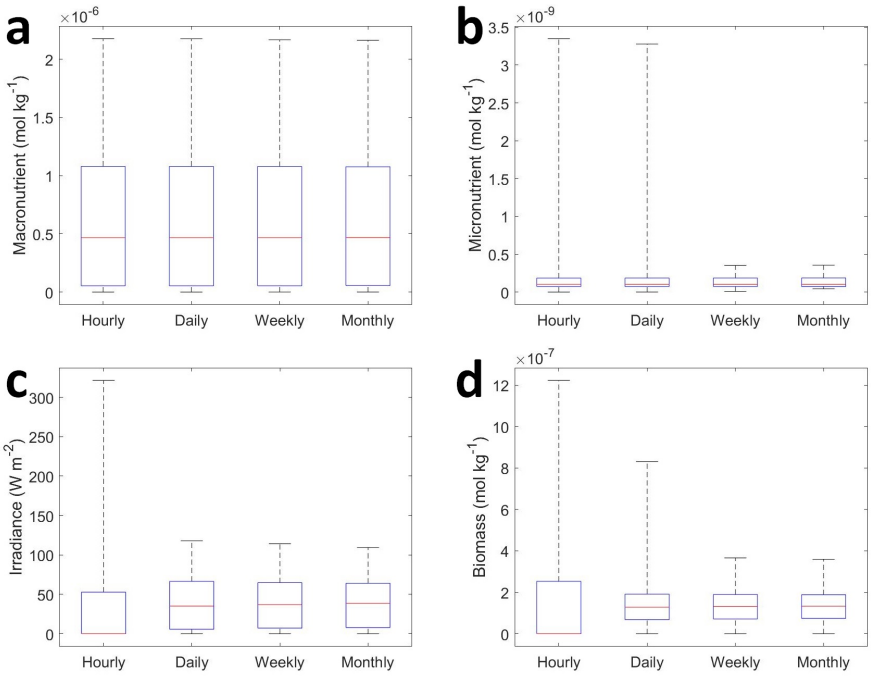

Fig. 2. Figure 1: Boxplots showing the variability in each of the predictor and target variables for each time-averaged dataset of Scenario 2.