Review 1 of "Anthropogenic climate change drives Non-stationary phytoplankton variance", submitted to Biogeosciences.

I appreciate the dedication that the authors have put into revising the manuscript. From my perspective, with the exception of the new ML analysis, which requires further details and clarifications, the manuscript is otherwise prepared for publication pending the resolution of the two comments below. While I acknowledge the relevance of the BRT method for driver identification, I believe that additional details are necessary to instill confidence in the results. Specifically, incorporating supplementary tests could enhance the evaluation of the zooplankton's impact on phytoplankton CoV.

We thank the reviewer for their careful reading of the revised manuscript and their constructive suggestions. In response to this feedback, we have included additional detail in the methods section on the specific hyperparameters used in our machine learning approach, as well as how the model was tuned. We have also elaborated on the machine learning model's performance by revising Figure 5 to include RMSEs for each of the four regional analyses.

Additionally, the reviewer makes a general comment regarding the effect of zooplankton grazing controls on the predictive skill of the model. To address this comment, we withheld all zooplankton grazing terms (zooplankton carbon, diatom grazing, and small phytoplankton grazing) when performing the predictor importance analysis in the Equatorial Pacific (the only region with a strong zooplankton dependence). When zooplankton grazing terms were withheld in this region, the RMSE increased by 7%, indicating that the predictive model performs slightly worse without zooplankton grazing included.

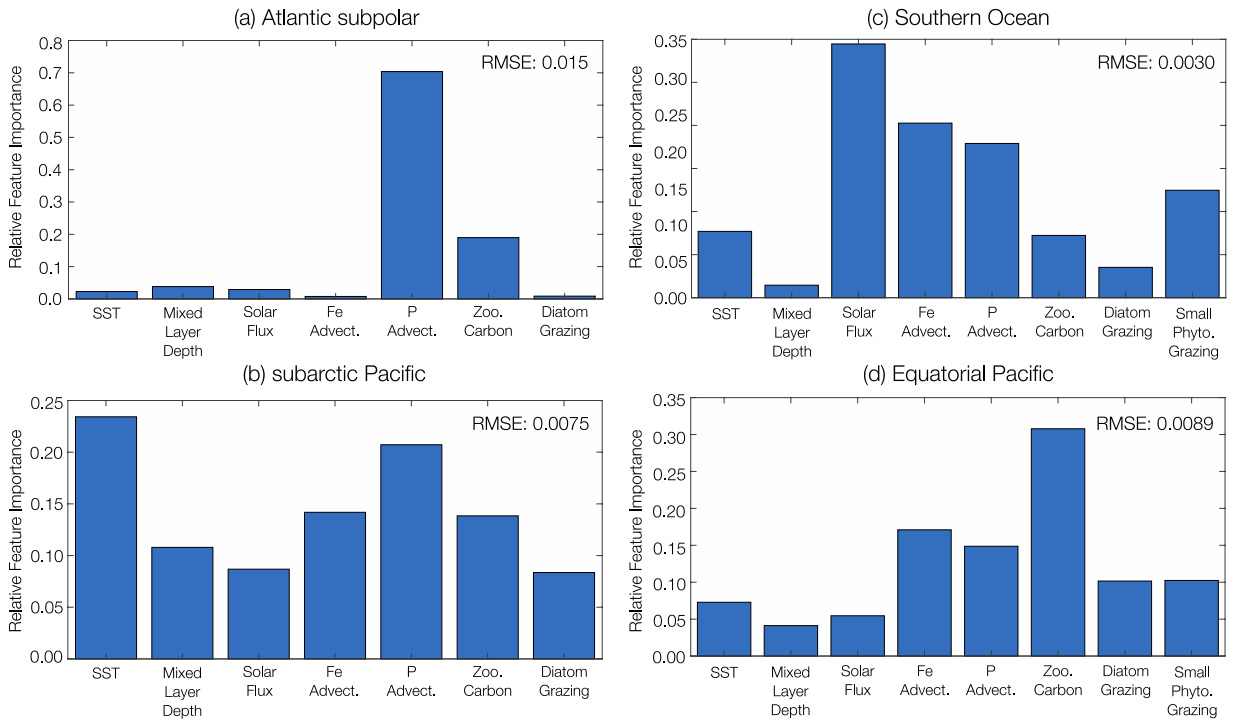
Comment #1: ML Methods

Though I lack expertise in ML, it seems that insufficient information is provided to comprehensively interpret the new findings. Recognizing that the paper's main focus might not be on this aspect, it's essential to provide a rational explanation for the utilization of this method rather than treating it as a "black box."

Specifically:

- Could you elaborate on the model's performance? Visualizing a time series would help confirm whether BRT effectively reconstructs CoV time series. Metrics like RMSE and r2 for both training and testing sets would also provide clarity.

Thank you for this suggestion. We have elaborated on the model's performance by including the RMSE for the testing dataset in Figure 5 for each region analyzed using the machine learning method. We have noted the addition of the RMSE in the figure caption.



"Figure 5: Relative importance of predictor variables on phytoplankton biomass coefficient of variation across the RCP8.5 forcing scenario (2006 to 2100). Marine ecological regions are defined in Tagliabue et al. (2021). Regions were selected which aligned with the highest fisheries catch in the (a) Atlantic and (b) Pacific basins and the biogeochemically important (c) Southern Ocean and (d) Equatorial Pacific regions. The dominant phytoplankton functional type is considered in each region. In regions with a mixed ecological assemblage, total phytoplankton carbon is considered. The RMSE (mmol C m^{-2}) for the testing dataset of each machine learning analysis is included in the upper right corner of each panel."

- Additionally, showing partial dependency plots (examples in Dannouf et al. (2022) or Lamb et al. (2021)), could help elucidate each variable's contribution.

Thank you for bringing this to our attention. Partial dependency plots are key to understanding the contribution of each predictor variable when using other 'black box' statistical/machine learning approaches (e.g. Gaussian Process Regression Models or Neural Networks). However, since we apply a Boosted Regression Tree which implicitly provides predictor importance, we can effectively reconstruct the relative importance of all predictor variables on phytoplankton carbon without the use of a partial dependency analysis. However, we now include the RMSE of our testing dataset from our machine learning approach in Figure 5 to elaborate on the model's performance.

- You would also need to specify how you tuned the model (i.e. how you choose the hyperparameters: learning rate, depth, number of trees).

We agree that the manuscript would benefit from more detail on how the machine learning model was tuned. We have added text to the methods to describe how the machine model was tuned and to include the specific hyperparameter values used.

L32: "The machine learning model was tuned to a learning rate of 1 and a tree depth of 10, generating 100 trees. We tuned several hyperparameters to generate the highest quality predictive results with the least computational expense. While learning rate can affect the quality of the solution, we experimented with a range of learning rates (0.1-1) with no change in the predictive results. Similarly, we tuned the tree depth using a range of 1 to 10 splits, and tree depths less than 10 produced a higher RMSE of the testing dataset."

- Clarification is needed on whether your predictors are regional time series spanning 2006 to 2100, as hinted at in L138-141.

Thank you for this feedback. We have modified the text to clarify the temporal extent of our regional analyses.

L26: " Our predictor variables are the regional mean, ensemble mean temperature, mixed layer depth, incoming shortwave radiation, physically mediated iron, physically mediated phosphate, zooplankton carbon, and zooplankton grazing (diatom, small phytoplankton, or their sum) annually resolved from 2006 to 2100, while our response variable is CoV of phytoplankton carbon (diatom, small phytoplankton, or their sum) annually resolved from 2006 to 2100."

I believe some refinement of the references is necessary to better justify the application of this method. There are a few reference suggestions that could help in establishing a more precise methodology (although there could be more relevant references). I think that integrating the BRT analysis into this paper could be better supported without significantly increasing its length, by fairly utilizing the supplemental information as a means of support. Have a look at Elith et al. (2008) for an introduction on the ecological application of BRT. For insights into BRT applied to time series, Dannouf et al. (2022) and Lamb et al. (2021) could be valuable references. While Denvil-Sommer (2023) focuses on the application of ML to ESM simulated spatial data (rather than temporal), there's potential inspiration for method structure. Additionally, consider referring to Robert et al. (2017) for insights into cross-validation.

Elith, J., Leathwick, J.R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of animal ecology, 77(4), 802-813.

Lamb, S.E., Haacker, E.M.K., & Smidt, S.J. (2021). Influence of irrigation drivers using boosted regression trees: Kansas High Plains. Water Resources Research, 57, e2020WR028867. https://doi.org/10.1029/2020WR028867

Denvil-Sommer, A., Buitenhuis, E.T., Kiko, R., Lombard, F., Guidi, L., & Le Quéré, C. (2023). Testing the reconstruction of modelled particulate organic carbon from surface ecosystem components using PlankTOM12 and machine learning. Geoscientific Model Development, 16(10), 2995-3012.

Dannouf, R., Yong, B., Ndehedehe, C.E., Correa, F.M., & Ferrerira, V. (2022). Boosted Regression Tree Algorithm for the Reconstruction of GRACE-Based Terrestrial Water Storage Anomalies in the Yangtze River Basin. Frontiers in Environmental Science, 10, 917545.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., … & Dormann, C.F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913-929.

Thank you for providing these very helpful resources. We have included them in the text to highlight machine learning methodologies in more detail.

L21: "Unlike linear models, boosted trees are able to capture non-linear interaction between the predictors and the response, and have been used in a number of ecological applications (Elith et al., 2008; Roberts et al., 2016; Lamb et al., 2021; Dannouf et al., 2022; Denvil-Sommer et al., 2023)."

Comment #2: Predictor choice (in particular zooplankton)

However, even using BRT, it's important to clarify that having zooplankton as a predictor doesn't necessarily mean it's the cause. The only way to really test the nature of the relation would be to run the model without zooplankton, which I know would require substantial extra work, so that I don't think it is necessary. Despite this, my reservations regarding the utilization of zooplankton grazing and zooplankton biomass predictors persist. If it's more comfortable to designate light and nutrients as "bottom-up control," it might be less accurate to term grazing and zooplankton biomass as "top-down control" (as mentioned in L10, L259, L272). This is because their changes during the focus period could also reflect variations in phytoplankton. Here are some suggestions that could assist in identifying the role of zooplankton as a top-down driver of phyto CoV in different regions and reinforcing your assumptions:

We thank the reviewer for this comment. To test the effect of the zooplankton grazing controls, we withheld all zooplankton terms (zooplankton carbon, diatom grazing, and small phytoplankton grazing) when performing the predictor importance analysis in the Equatorial Pacific (the only region with a strong zooplankton dependence). When zooplankton grazing terms were withheld in this region, the RMSE increased by 7%, indicating that the predictive model performs slightly worse without zooplankton grazing included. We have opted to maintain "top-down" and "bottom-up" to describe the controls on phytoplankton biomass, as these phrases are regularly used in the literature. This terminology is commonly used in both the ocean biogeochemistry literature (e.g., Bopp et al., 2001; Hashioka et al., 2013; Prowe et al., 2011; Behrenfeld et al., 2010, 2013; Laufkötter et al., 2015) and in its seminal textbook (Sarmiento and Gruber, 2006).

- What does the correlation matrix between predictors look like, particularly for phytoplankton, zooplankton, and grazing?

As mentioned above, we withheld the zooplankton terms and recalculated the RMSE of the testing dataset when developing a BRT model in the Equatorial Pacific, where zooplankton plays a key

role. When zooplankton grazing terms were withheld in this region, the RMSE increased by 7%, indicating that the predictive model performs slightly worse without zooplankton grazing included.

- When you apply BRT using the same set of predictors but replace zooplankton biomass with phytoplankton biomass or Chl a, do you obtain similar results (i-e does the importance of phytoplankton matches that of zooplankton as a predictor)? If so, it would suggest that top-down control is unlikely.

As mentioned above, we withheld the zooplankton terms and recalculated the RMSE of the testing dataset when developing a BRT model in the Equatorial Pacific, where zooplankton plays a key role. When zooplankton grazing terms were withheld in this region, the RMSE increased by 7%, indicating that the predictive model performs slightly worse without zooplankton grazing included.

- How are you defining "grazing pressure"? Is it the total amount of grazed phytoplankton, or is it normalized by phytoplankton biomass? I believe the second option might be more suitable to account for zooplankton's top-down influence.

In this context, grazing pressure is the fraction of phytoplankton biomass grazed. We have defined grazing pressure in the discussion to clarify this point. We also point the reviewer to the methods and supplemental information where we have discussed the functional form of zooplankton grazing in the CESM1-LE.

L30: "Previous studies of phytoplankton change with climatic warming have demonstrated that grazing pressure, the fraction of phytoplankton biomass grazed, is a contributor to biomass decline in low to intermediate latitude regions across a suite of model simulations with different marine ecosystem models (Laufkötter et al., 2015)…"

- How does the performance of the ML model improve when you include zooplankton/grazing compared to an ML model with only bottom-up controls.

Thank you for this comment. To test the effect of the zooplankton grazing controls, we withheld all zooplankton terms (zooplankton carbon, diatom grazing, and small phytoplankton grazing) when performing the predictor importance analysis in the Equatorial Pacific (the only region with a strong zooplankton dependence). When zooplankton grazing terms were withheld in this region, the RMSE increased by 7%, indicating that the predictive model performs slightly worse without zooplankton grazing included.

- Does the BRT's performance show enhancement when you train the model regionally compared to using a global scale?

Thank you for this suggestion. We chose not to train the machine learning model on the global scale as regionally specific processes dominate in each ecosystem. However, our regional analyses allow us to identify predictive drivers in discrete regional ecosystems with cohesive ecological and biogeochemical dynamics.

- In a broader context, similar to Denvil-Sommer et al. 2023, you could experiment with different sets of predictors to observe how the model performs and gain insights into the most crucial drivers. Given the critical nature of predictor choice in ML, this could be particularly informative for testing the role of zooplankton.

As mentioned above, we withheld the zooplankton terms and recalculated the RMSE of the testing dataset when developing a BRT model in the Equatorial Pacific, where zooplankton plays a key role. When zooplankton grazing terms were withheld in this region, the RMSE increased by 7%, indicating that the predictive model performs slightly worse without zooplankton grazing included.