Review 2 of Elsworth et al. (2022) "Anthropogenic climate change drives non-stationary phytoplankton variance", submitted to Biogeosciences.

The manuscript "Anthropogenic climate change drives non-stationary phytoplankton variance", summarizes projected changes in global and regional phytoplankton variability using the NCAR CESM1 Large Ensemble under a high emissions scenario. The authors explore the key drivers of declining phytoplankton variability, highlighting the importance of top-down, zooplankton grazing in potentially driving future phytoplankton response.

Generally, the article concisely represents its findings but there are several points of clarification I would recommend. In particular, the use of specific statistical terminology could be more accurate. Multiple times throughout the text, the term "variance" is used when, I think, "variability" is intended. In many cases this "variability" is being assessed via the standard deviation of the large ensemble members which is similar to the variance by not the same. Additionally, I am not proficient in MLR, but the comments made in the prior review are troubling especially considering the results are key to the paper's conclusions regarding top-down controls but these results seem underrepresented in the primary manuscript text. I've included several additional minor comments and suggestions below pertaining to clarity and organization.

We have clarified the use of "variance" and "standard deviation" in the text.

We have addressed Reviewer 1's comments regarding the MLR approach in our response above.

Specific Comments and Suggestions:

Lines 49-52: *Clarifying how variance in phytoplankton biomass may be changing over long time scales with climate change is important for fisheries management, especially at regional scales. Near- term predictions of phytoplankton biomass may also benefit from knowledge of the projected magnitude of internal variability, as the chaotic nature of internal variability hampers near-term predictions (Meehl et al., 2009, 2014).*

I think it's worth noting that the internal variability quantified using a large ensemble is Internal variability specific to the model and indicative of our uncertainty that results from its simplified representations of the real world processes and numerics. It doesn't necessarily have any bearing on real world manifestations of variability. Its primary utility to management and fisheries is in guiding our level of confidence in disentangling model signals from the noise.

This is an excellent point. We have included additional text to clarify this point.

Line 53: "… However, modeled internal variability may differ from that observed in the real world."

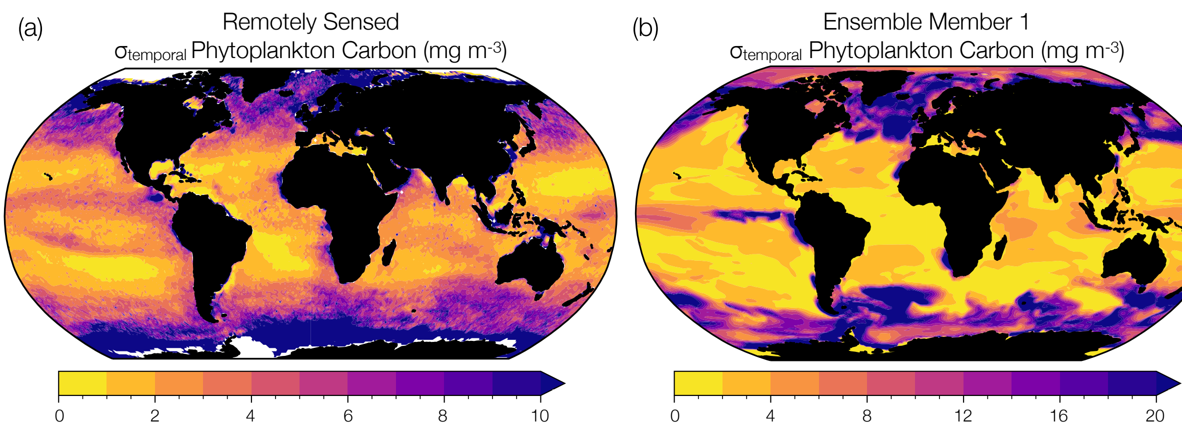
Lines 103-104: *Six CESM1-LE members had corrupted ocean biogeochemistry*

I'm curious, what does "corrupted ocean biochemistry" mean? it might help to explain what makes an ensemble member usable versus not.

The ocean biogeochemistry output fields of ensemble members 3 though 8 were corrupted during the saving process. Therefore, no information on biogeochemical variables is available for these ensemble members. However, the corruption of these ensembles affected only the biogeochemical output and, thus, other Earth system variables is preserved. Details are referenced here.

Figure 1. Add units: standard deviation should have the same units as the variable being assessed (i.e., phytoplankton carbon) but none appear in figure 1.

Thank you for bringing this to our attention. We have revised Figure 1 to include the units of mg m^{-3} and have modified the text of the figure caption.



"Figure 1: Comparison between observed and modeled phytoplankton biomass interannual variability. (a) Temporal standard deviation in annual mean phytoplankton carbon concentration (mg m^{-3}) reconstructed from remotely sensed chlorophyll concentrations, backscattering coefficients, and phytoplankton absorption (1998 to 2019) (Bellacicco et al., 2020). (b) Temporal standard deviation in annual mean phytoplankton carbon concentration (mg m^{-3}) simulated by ensemble member 1 of the CESM1-LE over the same observational period (1998 to 2019). Note the different magnitudes on the colorbars."

Lines 121-122: *Internal variability at each location (x,y) is approximated as the standard deviation across ensemble members (EMs) at a given time (t)*

The method described here indicates that the standard deviation is being used to quantify variability. However, throughout the paper, the authors reference the "variance" when I think they mean "variability". This is problematic because "variance" and "standard deviation", while related, are two different values and the way they are interchanged throughout the text is confusing. Please check all instances of "variance" in the paper for intended meaning and replace with "variability" where appropriate. I suggest including a description of the "coefficient of variance" method here, too.

We agree. We have included a description of the coefficient of variance in the methods section.

Line 130: “The coefficient of variance (CoV) is calculated as the standard deviation across the ensemble members divided by the ensemble mean,

$$\mathrm{CoV}\,(x, y, t) = \frac{\sigma(\mathrm{EM}\,(x, y, t))”}{\mathrm{LE}}$$

Lines 142-143: *However, while the model ensemble captures regional patterns of observed variability, the CESM1-LE overestimates the magnitude of observed interannual variability.*

I may be mistaken but it seems this was determined using only a single ensemble member - is it appropriate for conclusions to be drawn for the full ensemble when only considering one ensemble member?

We tested the temporal standard deviation for all ensemble members and report the (small) difference for the reviewer. The figure below shows the temporal standard deviation of 34 ensemble members of the CESM1-LE across the observational period (1998 to 2019). The difference in temporal standard deviation between ensemble members is small over this period.

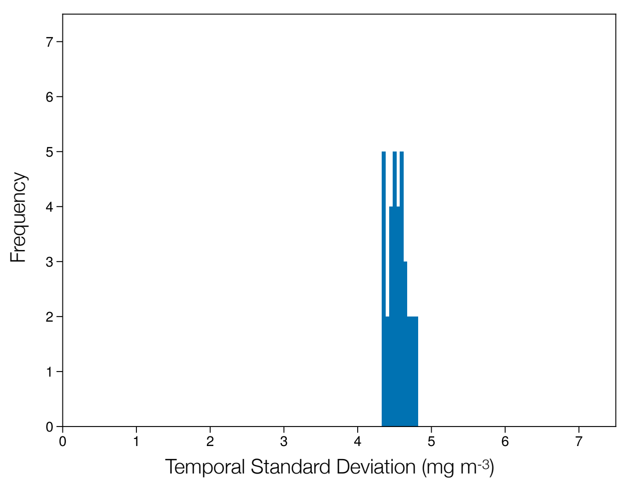


Figure: Histogram of global average temporal standard deviation (mg m^{-3}) for each of the 34 ensemble members of the CESM1-LE over the observational window (1998 to 2019).

Lines 147: *A synthetic ensemble is a novel technique*

I don’t think this technique can be called “novel” if it appears in two prior references

This is a valid point. We have removed the word “novel” from the text on Line 147 and Line 590.

Lines 149-151: *Compared to the internal variability over the observational period (2002 to 2020) (purple circle, (Figure S2), the model ensemble slightly overestimates the magnitude of internal variability in chlorophyll observed in the real world.*

This seems like a result/ should appear in the result section. Also, it makes an assessment of the ensemble as a whole, but isn't it still based on the results from the single ensemble member? If not, this was a point of confusion for me, and I suggest clarifying.

This is our second model validation exercise, and thus we opted to keep it in the methods section. We note for the reviewer that this is an assessment of internal variability (ensemble spread), as compared to a synthetic ensemble generated from observations. In response to this comment, we added a paragraph describing the interpretation of the results from the interannual and internal variance validation exercises to the methods section.

Line 163: "Some regions of the global ocean display a substantial mismatch in temporal variability between the model and that estimated from observations (Figure 1, Table S1). While the differences can be quite large in some regions, we note that this is an evaluation of temporal variability (rather than internal variability, the focus of this study), and that estimates from the satellite product derive from a collection of data products which may also display biases (Table S1)."

Line 176: "Taken together, our model validation exercises demonstrate that the model tends to overestimate both the temporal (interannual) variability and the internal variability in phytoplankton as compared to satellite observations on both global and regional scales. Thus, we must interpret our findings with this caveat in mind. The change in variance that we model is likely an upper-end estimate."

Lines 153-154: *Annually averaged, global mean, upper-ocean (top 150m) integrated phytoplankton biomass across the model ensemble decreases from 76.1 mmol C m-2 to 66.2 mmol C m-2.*

It's not clear what timeframes these values represent. Is it 2006 vs. 2100? If so, it seems that such a narrow, 1-year window would risk aliasing higher frequency variability and potentially under- or overestimate the change in mean state. This is somewhat compensated for by the size of the ensemble but differs from the 10-year averaging described later in Line 223

This is an excellent point. The decline in phytoplankton biomass is calculated as the difference between the average of the first (1920 to 1930) and last (2090 to 2100) decades across the historical and the RCP8.5 forcing scenario. We have clarified the time windows used in this calculation in the manuscript text.

Line 181: "The change in the mean is calculated as the difference between the first (1920 to 1930) and last (2090 to 2100) decades across the historical and the RCP8.5 forcing scenario."

Lines 177-178: *we calculated the coefficient of variance as the standard deviation across the ensemble members for a given year (ensemble spread) divided by the ensemble mean.*

I suggest including this description in the methods section rather than the results.

Thank you for this suggestion. We have included a description of the coefficient of variance in the methods section.

Line 130: "The coefficient of variance (CoV) is calculated as the standard deviation across the ensemble members divided by the ensemble mean,

$$CoV(x, y, t) = \frac{\sigma(EM(x, y, t))}{LE}$$"

Lines 178-180: *Figure 2b illustrates the change in the coefficient of variance from the historical period through the RCP8.5 forcing scenario (1920 to 2100).*

The results seem to jump from Figure 2a, to Figure 3, then back to 2b which is a bit confusing.

This is an excellent point. We have modified the text to enhance the flow of the manuscript.

Line 180: *The coefficient of variance is relatively constant across the historical period (1920 to 2005), and then significantly declines by ~20% from 2006-2100.*

I'm not sure I agree with the assessment that the coefficient of variance is relatively constant across the historical period. 1920-1980 appears to have a positive trend with a range of about 6.1 to 7.3, which appears similar to the range of the time period covered by the dashed line in Figure 2b. I suggest testing the significance of the 1920-1980 trend. Also, could the drop in coefficient of variance instead be explained by temporal distance from the perturbation that differentiates the ensemble members? If the 34 ensemble members differ in initial air temperature conditions, would the spread perhaps be expected to decrease as the simulation integrates further away from that initial discrepancy (i.e., solutions start to converge)?

This is a good point. We have tested the significance of the 1920 to 1980 trend and find that it is not significant.

However, the decrease in the coefficient of variance over the course of the simulation is not due to an increase in the time since the ensemble members were perturbed. This has been demonstrated in the study "An Ensemble Covariance Framework for Quantifying Forced Climate Variability and Its Time of Emergence" published by Yettella et al., 2018. This is illustrated by different responses of ocean and land variance over the 21st century, with ocean variance declining and land variance increasing over time.

Lines 190-193: *From 2006 to 2100, the coefficient of variance decreases by 3.3 x 10-5 yr-1 in the CESM1-LE, 2.0x10-4 yr in the MPI-ESM-LR1, 5.2x10-5 yr-1 in the CanESM2, and 3.9 x10-4 yr-1 in the GFDL-ESM2M. These declines are statistically significant in all model ensembles with the exception of the MPI-ESM-LR1 (Figure S2).*

It's not clear how these values across models are calculated, whether the end points of the time series or a range of years - the latter would be more appropriate (as done in Line 223) to avoid higher frequency variability and thus under- or overestimating the nature of the change. I also suggest reporting the specific statistical testing methods in the text if stating that the changes are significant.

Thank you for mentioning this. Changes in the coefficient of variance are calculated using averages of the first (2006 to 2016) and last (2090 to 2100) decades of the RCP8.5 forcing scenario. We have clarified this in the caption of Figure S2 (now Figure S4) and in the manuscript text. Significance of the trends are determined by a t-test with a p-value less than 0.05. We have also included this in the caption of Figure S2 (now Figure S4).

Line 220: "… The change in the coefficient of variance is calculated using averages across the first (2006 to 2016) and last (2090 to 2100) decades of the RCP8.5 forcing scenario."

Figure S2 (now S4): "… Trend significance is determined by a *t*-test with a p-value less than 0.05."

Line 201: *We observe the largest magnitude decline in total phytoplankton carbon variance*

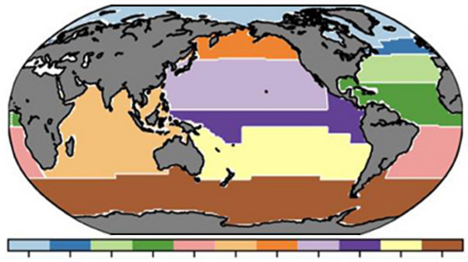
The table is reporting change in standard deviation, not variance. Standard deviation is expressed in the same units as the analyzed variable while variance is reported in the square of those units.

Thank you for clarifying. We have changed "variance" to "standard deviation" in the text to be more precise.

Line 229: "Global changes in total phytoplankton biomass standard deviation are a manifestation of changes in diatom and small phytoplankton variability (Table 1). We observe the largest magnitude decline in total phytoplankton carbon standard deviation in the subpolar Atlantic (ASP) region, where diatom standard deviation declines by ~10 mmol C m^{-2} and small phytoplankton standard deviation declines by ~2 mmol C m^{-2} (Table 1). The CESM1-LE simulates a moderate magnitude decline in total phytoplankton standard deviation in the subarctic Pacific (SAP) region, driven by a decrease in small phytoplankton standard deviation (~2 mmol C m^{-2}) with minor contributions from declines in diatom standard deviation (~1 mmol C m^{-2}) (Table 1). Moderate declines in standard deviation are also simulated in the Arctic (ARC), North Atlantic subtropical gyre (NAS), Southern Ocean (SOC), and Equatorial Pacific (EQP) regions, driven by declines in diatom carbon standard deviation in the SOC region and declines in small phytoplankton variance in the EQP region (Table 1)."

Figure 4: It's not clear what this figure adds to the discussion - it seems to be redundant with information in Figure 5. Perhaps if the outlines of the ecological regions were included?

This is an excellent suggestion. We have included a map of the aggregated biomes as cited in Tagliabue et al., 2021 in the Supplemental Information as Figure S2. We have also referenced this map in the text.



"Figure S2: The 11 ecological provinces defined in Tagliabue et al., (2021) and Vichi et al., (2011). Provinces were aggregated using multivariate statistical analysis of physical (i.e., salinity, temperature, mixed layer depth) and biological (i.e., chlorophyll concentration) ocean parameters to group ocean regions with similar physical and environmental conditions. Figure adapted from Tagliabue et al., (2021)."

Line 113: "We classified the marine environmental into 11 ecologically cohesive biomes as in Tagliabue et al., (2021) and Vichi et al., (2011) (Figure S2), which are a consolidation of the 38 ecological regions defined in Longhurst et al., (2007)."

Lines 219-221: *We quantified the relationship between phytoplankton carbon and the variables which contribute to changing phytoplankton biomass and its internal variability by performing a multiple linear regression (MLR) analysis. The MLR analysis was performed on linearly detrended annual anomalies using the ordinary least squares function of the Python package statsmodels.api*

This and the associated equations seem to belong in the methods section.

Thank you for this suggestion. We have moved this and the associated equations to the methods section.

Line 134: "… We quantified the relationship between phytoplankton carbon and the variables which contribute to changing phytoplankton biomass and its internal variability by performing a multiple linear regression (MLR) analysis. The MLR analysis was performed on linearly detrended annual anomalies using the ordinary least squares function of the Python package statsmodel.api."

Line 274: *…and important global biogeochemical regions…*

What is considered an important biogeochemical region? This seems somewhat vague - I suggest elaborating to be a bit more specific.

We appreciate this comment. An important biogeochemical region is an ocean region characterized by coherent physical and environmental conditions, which support unique marine ecosystems with an outsized role on ocean biogeochemical cycling. We have added a sentence in the text to clarify this point.

Line 123: "… Important biogeochemical regions are those characterized by coherent physical and environmental conditions, which support unique marine ecosystems with an outsized role on global ocean biogeochemistry."

Lines 278-280: *As such, the magnitude of changes in phytoplankton internal variance derived from the model ensemble should be interpreted as an overestimate when considering changes in phytoplankton internal variance driven by anthropogenic warming.*

Again, my impression was that this conclusion was derived from analyzing a single ensemble member which seems insufficient for assessing the entire ensemble.

This conclusion was derived from our second model validation exercise, where we compare the spread across all modeled ensemble members with that of a synthetic ensemble derived from satellite observations.