We thank the Reviewer for their positive feedback and valuable inputs on the manuscript. Below, we provide a detailed response to each of the comments — reviewer comments are in black, and our responses are in blue.

Summary:

The manuscript makes use of the subset of models from CMIP6 that provided high temporal resolution SST / chlorophyll output to investigate the sub-seasonal dynamics of simulated phytoplankton. The model output is processed first using a decomposition methodology that breaks down the variability into different temporal modes. Subsequently, the output is further processed using spatial decomposition to identify whether the simulated variability has the correct horizontal length scales (e.g. to distinguish where models exhibit seemingly comparable variability to observations but on much coarser spatial resolution). The manuscript finds that none of the models realistically represents the seasonal and sub-seasonal patterns of variability observed (unlike the situation with SST). However, the spatial decomposition teases out patterns between models that allows them to be separated into three groups with better or worse representation of real-world variability. The authors note that one group is hampered by its spatial resolution, while the other two exhibit excessive sub-seasonal variability apparently from tightly-coupled predator-prey cycles. The manuscript concludes by noting the value of high temporal resolution output for identifying unrealistic model behaviour, and with a call for modelling groups to endeavour to provide this in CMIP7.

Review:

Overall, I found the manuscript interesting and quite convincing about the realism or otherwise of current generation CMIP models. I have no major comments on the content of the manuscript, but have made a small number of minor suggestions about improvements. I recommend accepting the manuscript following these minor corrections.

We thank the reviewer for their encouraging assessment of our manuscript.

Comments:

One overall comment I have is around the quality of the figures. There are some unhelpful choices here to my mind and I detail these below. However, I would accept that this is largely an aesthetic decision, and would not insist on my suggested changes being implemented. Another general comment I'd make is that it would be good to try to put the models examined into some sort of context within the wider CMIP6 ensemble – I've suggested an idea in what I say about Figure 2a, but there may be a more obvious or better solution.

Ln. 46: Inconsistent ordering of references; they're neither in chronological nor alphabetical order (I prefer the former).

The references are now arranged correctly – 'This is particularly critical for phytoplankton as it is characterized by large natural variability at diverse timescales, which often masks the long-term trends (Henson et al., 2010, 2016; Doney et al., 2014; Keerthi et al., 2022)'.

Ln. 64: Amend to "… produced by *a subset of* ESMs …".

The sentence is corrected – 'Capitalizing on high frequency global measurements of satellite ocean color SChl, we evaluated the performance of historical simulations produced by a subset of ESMs participating in the Coupled Model Intercomparison Project Phase 6 (CMIP6) to simulate global surface ocean phytoplankton dynamics across diverse temporal scales (sub-seasonal, seasonal, and multi-annual), with a specific focus on high frequency sub-seasonal variability'.

Table 1: The IPSL and CNRM models are lumped together (presumably because of a common ocean), but do they share a common atmosphere or atmospheric resolution?

The IPSL and CNRM models are grouped together because they share the same ocean model. However, they utilize different atmospheric models: **ARPEGE - Climat** for the CNRM models and **LMDZ** for the IPSL models. They are also based on different land surface models as well as different sea-ice components. All models are grouped based on their shared physical and biogeochemical (BGC) ocean components.

Table 1: The MPI rows have a missing border between HAMOCC6 and 150 km cells.

This has been corrected.

CMIP6 Simulations	**Physical Ocean Model**	**Ocean BGC Model**	**Horizontal resolution (Physical & BGC Model)**	**Model Simulations**	**References**
IPSL-CM6A-LR	NEMO-OPA	PISCES	100 km	Historical	Boucher et al. 2018, 2021; Séférian, 2018
IPSL-CM6A-LR-INCA					
CNRM-ESM2-1					
CESM2	POP2	MARBL	100 km	Historical	Danabasoglu, 2019a, b, c;
CESM2-FV2					
CESM2-WACCM-FV2					
MPI-ESM1.2-HAM	MPIOM	HAMOCC6	150 km	Historical	Neubauer et al., 2019 ; Wieners et al., 2019a, b, c; Jungclaus et al., 2019a, b; Schupfner et al., 2019
MPI-ESM1.2-LR				Historical, SSP5-8.5, piControl	
MPI-ESM1.2-HR			40 km		
NorESM2-LM	MICOM	HAMOCC	100 km	Historical, SSP5-8.5 piControl	Seland et al., 2019a, b, c ; Bentsen et al., 2019a, b
NorESM2-MM				Historical, SSP5-8.5	

Table 1. The CMIP6 Earth system models used in this study, their physical and biogeochemical ocean components, nominal horizontal ocean resolution and the simulations assessed.

Table 1: The MPI and NorESM2 rows mention piControl simulations, but I don't believe that these are mentioned elsewhere in the manuscript.

PiControl simulations are shown in Supplementary Figure 2 (Figure S2) and are described in line 485.

Ln. 134: A period of 8 months is mentioned here for the so-called “multi-annual component”. Why 8 months and not 12 months? I’m sure I’m not understanding something.

The multiannual component is defined as low-frequency variability characterized by timescales of approximately one year or longer. We did not impose a strict cut off at 12 months, thus this component encompasses variability with periodicities extending beyond 8 months. However, when analysing time series at specific locations, we observed that the signal within the 8–12-month range was relatively weak, suggesting that the dominant contributions to the multiannual variability arise from longer timescales. This flexible approach ensures a more inclusive representation of low-frequency variability without being constrained by rigid temporal boundaries. For more information on the temporal decomposition we applied here, please refer to Keerthi et al., 2020.

Keerthi, M. G., Levy, M., Aumont, O., Lengaigne, M. & Antoine, D.: Contrasted contribution of intraseasonal timescales to surface chlorophyll variations in a bloom and an oligotrophic regime. Journal of Geophysical Research: Oceans, 125(5), e2019JC015701, 2020.

Ln. 153: I’m not a fan of “Results and Discussions” sections, and would prefer the authors to properly separate results from discussion to improve the manuscript’s clarity. However, it can be difficult to separate them at this stage, so ignore this suggestion if it isn’t obvious to address.

In the initial stages of the manuscript, we attempted to present the results and discussion in separate sections. However, this approach led to some repetition of content. Consequently, we decided to combine them into a single "Results and Discussion" section. At this stage, separating them would be overly burdensome. Thank you for your understanding.

Figure 1: Conventionally, darker colours are used to indicate lower values while brighter colours are used to indicate higher values. The choice here is confusingly the reverse.

We have updated the figure, using a color scheme where darker tones represent lower values and brighter tones indicate higher values. The revised figure can replace Figure 1 in the manuscript.

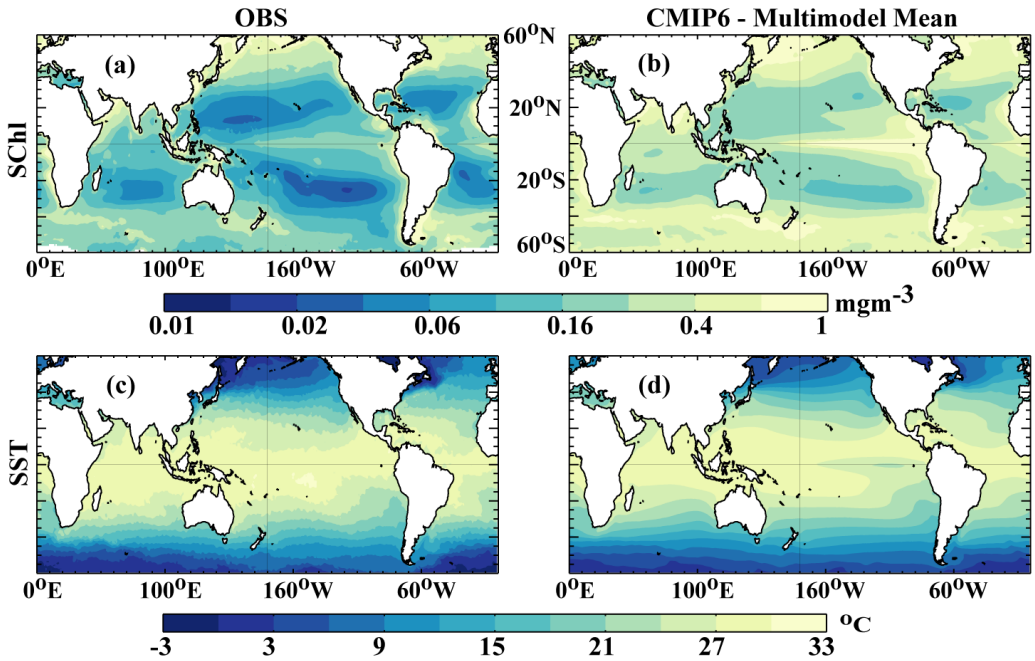

Figure 1: Mean state evaluation: Annual mean SChl (a) Observed (ESA OC-CCI product) and (b) CMIP6 multi-model mean for the years 1998-2014 and domain 60°N-60°S. (c & d) Similarly for SST.

Figure 2a: The models are distributed into two clear groups but the manuscript doesn't reflect on this. Is there any straightforward distinction to be drawn between them? For instance, what would the mean fields of the two groups look like? Would there be any clear distinguishing patterns.

MPI models exhibit significant overestimation in both mean field and variance across all temporal scales. A comparative figure showing the mean field of the MPI models against other models is attached. This figure shows that spatial patterns are similar but MPI models consistently overestimate the magnitude across the global ocean.

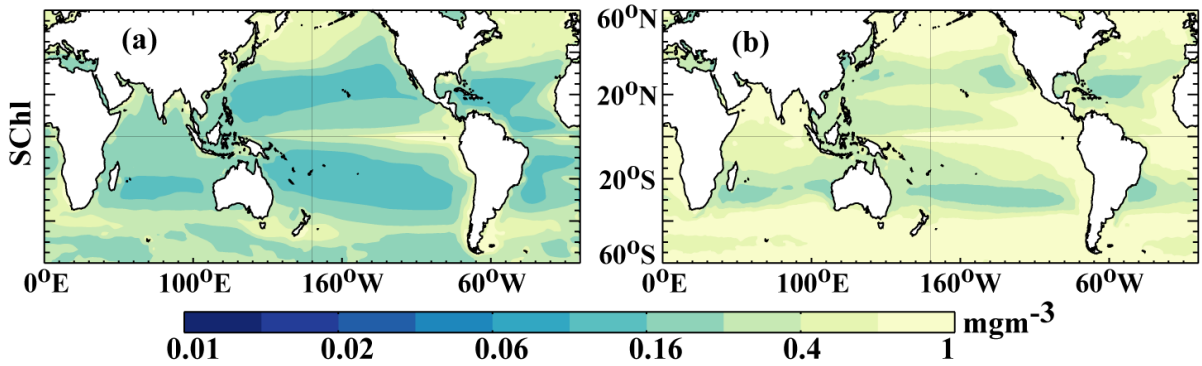

Figure 2: CMIP6 mean state evaluation: a) Ensemble mean SChl from the IPSL, CNRM, CESM, and NorESM models analyzed in this study; b) Ensemble mean SChl from the MPI models analyzed in this study.

Figure 2a: Since the models examined fall into only 4 "families", and given that they all perform fairly badly here, I wonder if it might be worthwhile somehow contextualising their performance against the wider CMIP6 ensemble? Possibly by adding other models that are

outside of the analysis here? Either in this figure, or in a supplementary version of this figure. Even without those models being analysed in detail as here, it would provide context for the representativeness of the models used here.

In this manuscript, we analyzed the SChl temporal variability simulated by the models, with a particular focus on the subseasonal timescale, which is often overlooked. So we included only models that provide SChl data at daily temporal resolution, necessary to properly evaluate subseasonal signals.

A more detailed analysis of the mean chlorophyll surface distribution predicted by a larger set of CMIP6 models is provided in Séférian et al. (2020). Figure 2 of this study displays the model-data deviations for this larger set of the same period (1998-2014) as the one used in our study.

Séférian et al.: Tracking improvement in simulated marine biogeochemistry between CMIP5 and CMIP6, Current Climate Change Reports, doi:10.1007/s40641-020-00160-0, 2020.

Figure 3: This chart makes a sensible comparison between the variability modes of the obs and models. However, I wonder if there's a way to put the information it presents onto a single axis where the models and observations can be seen together. For instance, "total" variability on the x-axis, and the fraction that's sub-seasonal on the y-axis? You may have tried something like this already.

Thank you for your suggestion. We did explore several alternative visualization methods. However, after multiple iterations, we found that the current figure is the most effective way to clearly and accurately convey the variability modes.

Figure 4: Add in the caption which models, and why, are missing here. Presumably data availability?

The caption has been modified to – '**Figure 4: SST variability across timescales.** This figure is similar to Figure 3 but focuses on SST. Note that NorESM2-LM and NorESM2-MM are excluded, as daily SST data for these models is not available on the CMIP6 data portal.'

Figures 4, 5: A bit more consistency in style would be good for these bar chart figures. Figure 3 seems to make use of the space best, with Figures 4 and, especially, 5 using it less well (i.e. the bars are thinner).

Figures 3 and 4 are of the same size and have the same bar widths. The size discrepancy occurred when pasting the TIFF files into the word document. In Figure 5, we attempted to include both SChl and SST, as the information provided by these figures is less detailed than that in Figures 3 and 4.

Figure 7: I think this could be a much better figure if pie charts weren't used. Each model (and possibly model region) could be given a simple x-y subplot in which the x-axis is period and the y-axis is geographical area or frequency. Each subplot could then also contain the same information for the observational data. This would add information currently hidden by the limited number of periods selected for the pie charts, and would make it easier to compare with the observational data. At present the reader has the unenviable task of squinting to try to work

out how similar / different one pie chart is from another. Line plots would – I suggest – be much better here.

Thank you for this suggestion. We agree with the reviewer that the pie charts can be difficult to compare. As a result, we have changed it into a bar diagram for, we hope, better clarity and comparison.

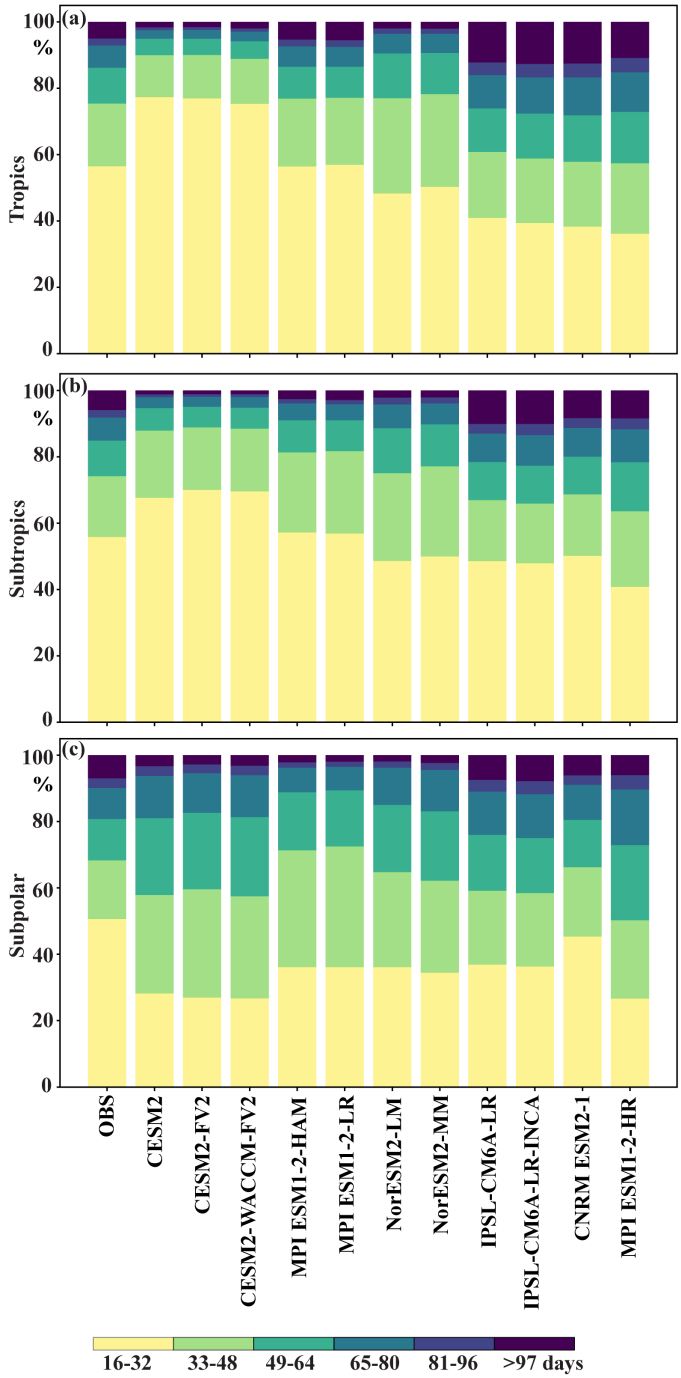

Figure 7: Sub-seasonal SChl variability across temporal subperiods: Bar plot showing the relative contribution of each temporal period to the total SChl sub-seasonal variance in the observations and different CMIP6 historical simulations.

This bar diagram can replace Figure 7 in the manuscript.

Figure 8: This is a horrible colour map. Not only is it a single colour, but the different shades of that colour are very difficult to discern, with an emphasis on darker shades that make any distinctions in the plots fairly invisible. Why not use one of the colour maps used elsewhere to make discerning the distinctions easier?

We utilized the same color scale as in Keerthi et al. (2022) to facilitate direct comparison between the results of this study and those presented by Keerthi et al. (2022).

Keerthi MG, Prend CJ, Aumont O, Levy M.: Annual variations in phytoplankton biomass driven by small-scale physical processes. Nature Geoscience :1–14, 2022.

Ln. 502: The structure of biogeochemical models is alluded to here but no evidence is presented. Perhaps illustrating with time-series plots of representative differences between models might help clarify this here. Or even examine the low frequency output of the models involved to determine if they differ in their phytoplankton-zooplankton relationships. However, this is only a suggestion as it might be sending you on a wild goose chase.

Thank you for your suggestion. We agree that illustrating the differences between models using time-series plots or examining the low-frequency output to assess differences in phytoplankton-zooplankton relationships could provide valuable insights. However, this is a challenging aspect to address, and adding this analysis would significantly extend the manuscript. We refer to Rohr et al. (2023), which discusses the largest source of inter-model uncertainty in marine biogeochemical models, specifically regarding phytoplankton-specific loss rates to zooplankton grazing. Rohr et al. (2023) found that this uncertainty is more than three times larger than that of net primary production and is driven by large differences in prescribed zooplankton grazing dynamics. Given these findings, further exploration of phytoplankton-zooplankton interactions across the models may indeed provide a deeper understanding, but we feel this would require substantial work beyond the scope of the current manuscript.

A likely better way to study the impacts of coupling assumptions between phytoplankton and zooplankton would be to use a single modeling framework to explore the major differences highlighted by Rohr et al. (2023). Otherwise, the many differences in the representation of marine biogeochemistry and other components of Earth System Models, would almost certainly prevent attribution.

Ln. 515: The authors advocate for CMIP modelling groups to submit daily outputs of biogeochemistry variables but don't mention which ones specifically. Obviously chlorophyll but, per the preceding point, would they advocate for others like surface zooplankton too? This is a good opportunity to advocate for them.

According to Rohr et al. (2023), there are significant differences in the prescribed zooplankton grazing dynamics among CMIP6 simulations, which leads to considerable variations at higher frequency timescales. In particular, predatory-prey oscillations are suspected here but proved to be extremely difficult to evidence without corresponding zooplankton information. Having this information, i.e. zooplankton concentrations and grazing rates at daily resolution would be very useful in this regard.