In this work, the authors use daily integrated net primary production data from two CMIP6 models and compare them against an observational dataset derived from satellite. They use change points analysis to highlight changes in the mean NPP and MLD time series. The model and observational data are compared but the comparisons are often subjective, leaving it up to the reader to decide by eye whether the models successfully reproduce the observed behaviour.

However, the authors do not describe the limitations of their work, leaving several key questions unasked or unanswered. In addition, the absence of a discussions section, means that results are often described but not fully interpreted or put into a wider context. This means that the overall conclusions

On the other hand, the language in this paper is excellent. It is very clearly written and there are very few spelling and typographical or grammar errors – at least that I have spotted. This is a very interesting topic and there is a lot that could be done with this high temporal resolution model data. With a bit more effort, this could be a very exciting and impactful paper and I'd encourage the authors to keep going!

Authors: We thank the reviewer for their thorough and extensive review. The excellent suggestions will help us deepen the analysis and improve the results. Our responses in bold italics below.

Major points:
Here are some questions that may encourage the authors to explore new ground. I don't expect all of these questions to be answered, but they may open up some new areas of interest.

Earth System Models s tend to have relatively simple representations of the marine Ecosystem, relative to dedicated ocean models. PISCES has two phytoplankton functional types and NorESM2 has only one. More complicated BGC models exist, such as BFM, Planktom, ERSEM and others, who have four or more phytoplankton classes. Furthermore, it is widely known that blooms are formed from cascades of multiples species. See for instance Kleparski 2021, which uses 24 species assemblages in CPR data to investigate the spatial distribution of north atlantic blooms in observations. How does the low PFT number of these models impact their ability to model blooms?

Kléparski, L, Beaugrand, G, Edwards, M. Plankton biogeography in the North Atlantic Ocean and its adjacent seas: Species assemblages and environmental signatures. Ecol Evol. 2021; 11: 5135– 5149. https://doi.org/10.1002/ece3.7406

Authors: Earth system models by necessity include simple representations of the biogeochemical system. All models will include fewer PFTs than reality and the PFTs that are included are aggregates of many taxa. So how does the number of PFTs impact the vertically integrated NPP? It is possible that more PFTs would have an effect on the NPP and on the seasonal timing. However, more PFTs also introduce challenges as more variables means more variables to tune and validate. We will add some discussion about this issue in the revised version

Similarly, marine BGC models are often run as standard-alone Ocean only runs. What do we gain by including the rest of the earth system when investigating phenology? Is there some feedback between the ocean and the atmosphere that improves bloom timing modelling?

Authors: The physics of the coupled atmosphere ocean system differs from that of the ocean forced at the surface. In the former, the atmosphere and the ocean continuously affect each other and the temperature, currents, stratification are therefore different from the uncoupled system. In particular the air-sea exchanges of momentum, heat, and freshwater are profoundly different in coupled and uncoupled models, affecting especially SST, SSS and mixed layer depths.

Furthermore, it has been demonstrated that interactive coupling significantly influences the variability of thermal variables like heat exchange, temperature etc. (e.g. Bhatt et al. 1998, Barsugli and Battisti, 1998). The coupling effect is less constrained for precipitation (salinity does not impact local precipitation while precipitation impacts directly on salinity).

These differences affect the biogeochemistry and the seasonal pattern of primary production is therefore not expected to be equal in the coupled vs the uncoupled case. As the coupled runs in this case are concentration driven, atmospheric pCO_2 is however not different from the uncoupled version.

The main technical criticism is: why only focus on these two models and this one scenario when all of CMIP6 is available? This work either needs a convincing argument as to why it uses this two model limit, or it needs to include other models from CMIP6. Similarly for the SSP5-8.5 scenario. A quick check of ESGF shows that while there is no daily primary production data, there are 380 datasets for daily Chlorophyll (chlos) from 12 different models with picontrol, historical, or Tier 1 ScenarioMIPdata. A tool like ESMValTool could help accumulate and prepare all the data, even if you don't actually use ESMValTool for the analysis. This would allow a comparison of the phenology (for chlorophyll) between several future scenarios and models.

Authors: Daily surface chlorophyll (chlos) has the advantage that it is easily validated against satellite data and it is a common output from ESMs. However, NPP and surface chlorophyll are not connected in a simple manner. NPP is the rate of photosynthetic carbon fixation and it is integrated over the water column. It is thus a measure of the total water column production. This means that the primary production maxima can occur deeper down in the water column but still be picked up by the NPP variable (Richardson and Bendtsen, 2019). Furthermore, chlorophyll is not directly related to biomass or primary production and in PISCES chlorophyll in phytoplankton is modeled as a separate variable in accordance with Geider et al (1997). In that sense, primary production is more straightforward and more easily relatable to carbon dioxide fixation.

Moreover, the two runs used in the manuscript, in particular their output, were purposely made as part of the COMFORT project to investigate abrupt changes in ocean biogeochemistry. To this end we save not only high temporal resolution NPP, but also MLD, SST and some other biogeochemical variables. An important part of the work presented was to investigate whether NPP changes could be connected to changes in other variables. This would not be possible with data from the CMIP archive, as most of those variables will not be saved with high temporal resolution.

In the revised manuscript, we will include a motivation for the use of vertically integrated NPP.

You've located several change points for bloom timing and MLD. What are the differences (if any) between the period before this point and after this point? For instance, you could revisit fig 1, 2 or 4 comparing the "pre- change" to the post-change for these models.

Authors: The largest change point in the time series marks the point in time of the largest change of the probability distribution. This could be a change in mean or median but could also be higher order changes like skewness or kurtosis. Since the kernel based model does not generate information on the type of change, the analysis was complemented with the L1 and L2 models that give change points in mean and median respectively. The alignment or non alignment of these different methods therefore give an indication of the type of change. From a more biogeochemical perspective it seems clear that the early change point is a change toward earlier peak NPP. We will add more discussion about this in the revised manuscript.

On L41, you mention that NPP is affected by precipitation, wind patterns, temperature and light, but do not investigate any of these fields. You only investigate MLD as a proxy for reduced nutrient availability. I'm not fully convinced that the argument has been made that MLD shallowing is the cause of the bloom in these models. Daily data for several of these variables should be available in CMIP6 as well.

Authors: We have looked at daily SST and MLD both absolute values and phenological indicators and their relation to NPP. No clear cut correlations were found between SST and NPP or their phenology. Note also that MLD is directly dependent upon all the mentioned variables, so these different drivers should not be considered independent. We will add a deeper discussion about how different drivers affect NPP in the revised version.

The paper is also missing a discussions section. This could include things like:

- How well this work fits into the current state of research for this field.
- What mechanisms makes one model better than the others? Is it purely their physical behaviour or is there something about the parameterisation of the marine BGC model?
- Can you estimate how many years of data would be needed to detect these change points in CAFE?

- Could you use longer data sets (ie Continuous Plankton Recorder data) to detect change points in these regions?
- A limitations of this method section

A more thorough investigation of phenology may also be interesting, not just the bloom initiation or maxima, but the intensity and duration of blooms may also be changing in the future. See for instance:
https://www.sciencedirect.com/science/article/abs/pii/S1470160X11002160

Authors: The discussion section is currently put together with the results section in Section 3, Results and discussion. We will add a separate Discussions section in which we will add more discussion as indicated in the following responses.

In addition, other sources of observational data are available, and would strengthen the arguments around the presence of change points:

Floats: https://www.frontiersin.org/articles/10.3389/fmars.2020.00139/full

CPR (in North Sea): https://aslopubs.onlinelibrary.wiley.com/doi/10.1002/lno.11351

Bermuda time series: https://bats.bios.edu/

Authors: These are great references that we will add to the discussion. However, comparing vertically integrated NPP to observations is difficult. Most sources use phytoplankton biomass or chlorophyll as in the suggested references. This is also the reason why we choose to use the satellite based CAFE model for the comparison.

Note also that the presence of change points in the models is independent of the models ability to reproduce observations. What comparisons to observations can give us is some grasp of the model's skill. Although, even that is hard given that we compare historical simulations to our recorded history (i.e. one realization of all possible histories) and also given that the observations used are, in fact, also model dependent. However, the silver lining is that understanding and describing the behaviour of ESMs is interesting and important even in itself given their extensive use in science and their impact on policies. We will add some of these points to the introduction.

Specific comments:

Title: To get more impact, state your main result in the title. Peak Net Primary Production will shift earlier in the year over the 21 st century. (or something like that).

Authors: The title will be changed so that our main result is indicated there.

Abstract:

L11: Similarly to the Title – if you open the abstract with the main result, it can be more eye catching. Try not to hide your exciting result at the bottom of the abstract!

Authors: We will change this.

L16: “The low spatial resolution of the earth system models can explain much of such difference”: Can it? How so? I’m not convinced that this argument has been made.

Authors: No, you're right. This has not been properly demonstrated. We will change this statement.

L20: Using SSP585 as a forecast needs to be treated carefully, as this is not a “realistic” scenario. SSP585 is the scenario with enhanced fossil fuel usage – meaning that the rate of fossil fuel emissions accelerates beyond “business as usual” - something we have fortunately not seen in the previous 8 years since the end of the CMIP6 historical period.

Authors: It is true that SSP5-8.5, in that sense, is not the most plausible scenario. However, this also gives us kind of an upper end on the size of change. Furthermore, one of the largest change points in the timeseries occurs in both models at the end of the historical simulation.

Introduction:
L27: See Le Quéré et al for more up to date reference on the carbon budget.
Le Quéré, C., et al : Global Carbon Budget 2018, Earth Syst. Sci. Data, 10, 2141–2194, https://doi.org/10.5194/essd-10-2141-2018, 2018.

Authors: This will be updated.

L35 and elsewhere: CO2: 2 should be subscript

Authors: This will be corrected.

L41: Can you add some references for this? How is NPP impacted by precipitation in the North Atlantic?

Authors: Increased precipitation and changed wind patterns affect the water column stratification. In this way, NPP is indirectly affected by these variables. Furthermore, precipitation can have a fertilizing effect by transporting nutrients to the surface waters (Myriokefalitakis et al., 2020). In particular, MLD is affected by precipitation. Freshwater added to the top of the ocean increases vertical stability through its effect on density. Fresher water needs to be cooled more to get the same density as saltier waters.

L55: “Depending on the onset…” This sentence needs to be more explicit. Ie, If thermal stratification occurs, then spring bloom may start earlier…” or similar.

Authors: We will change the sentence.

L57: One alternative theory about the causes of bloom timing changes is the switch from positive net heat flux to negative net heat flux:
Smyth TJ, Allen I, Atkinson A, Bruun JT, Harmer RA, Pingree RD, et al. (2014) Ocean Net Heat Flux Influences Seasonal to Interannual Patterns of Plankton Abundance. PLoS ONE 9(6):e98709. https://doi.org/10.1371/journal.pone.009870

Authors: We will add this reference

L72: What makes it unique? Why only these two models? Why not a full CMIP6 ensemble? What do you gain by using the years 1750-1850?

Authors: The Pi-Control is an important addition that underlines the uniqueness of the largest change-points. Furthermore, the benefit of using long time series is a more robust estimation in the tails of the probability distribution. Consequently, rare events such as extreme conditions are better reflected in the data set. The data used in this work was produced in the H2020 project COMFORT and the two models are the only ones that have saved daily vertically integrated primary production for the full 1750-2100 period.

Methods:
L76: Just to confirm, are you are using the CMIP6 dataset, intpp, from ESGF? When you calculate the mean, are you taking the cell area weighted mean?

Authors:
The EC-Earth-CC and NorESM2-LM historical and SSP5-8.5 used in this study were performed according to CMIP6 protocols (with additional daily output activated) but have not been published on ESGF. A minor update was made to the terrestrial vegetation model in EC-Earth-CC after the PiC was performed but is not something that is expected to affect the results.
The mean is the cell area weighted mean.

L87: Nemo should be NEMO
Authors: This will be corrected.

L89: pCO2: 2 should be subscript
Authors: This will be corrected.

L96: Primary production is indeed growth of phytoplankton, but elsewhere you talk about net primary production. Net PP usually does include some loss terms – otherwise you mean Gross Primary Production (GPP).
Authors: We mean NPP, that is GPP minus respiration. The models only generate NPP as they do not explicitly model respiration. This will be clarified.

134: Modis should be MODIS

Authors: This will be corrected.

136: The Change point analysis method description section (sect. 2.3) does not sufficiently explain how the method works, and gives the impression that the authors have used the Ruptures packageas a “black box”. Can you give more detail on how this method works here (or perhaps in an appendix)? Same for L1 & L2 methods.
Authors: This method has been described in the cited references. However, we understand the criticism and will extend this section with a more thorough description.

Results:
L159 and figure 1: Is it sensible to compare depth-integrated Net Primary production to satellite (surface) Net Primary Production? Are these data comparable?
Authors: CAFE is a model that utilizes ocean color and other optical properties to estimate the vertically integrated NPP (Silsbe et al., 2016). These values are not just for the surface but for the entire water column in similarity with the ESM results. The vertically integrated properties is a strong argument for the method used in this manuscript.

L169: Some statistical tools would help give an objective estimation of model data fit, something like some pattern statistics or even a linear regression?

Authors: **Given the short time series and the fact that these series are from different histories, we would argue that applying more statistical methods would not answer the relevant questions. Any method of comparison is only as good as the data admits, and this data admits very little. Put another way; say we calculate a spatial correlation between the CAFE data and our two models and find r1(EC-Earth-CAFE)<r2(NORESM-CAFE), what, of value, would that tell us?**

L176 – If the mask is important, you may need to indicate it in this figure (or elsewhere).
Authors: We will add the unmasked data to the figure. So that the seasonal cycle for the entire region from the ESMs is visible.

L180: “Reasonable”. Once again, an objective measure of goodness of fit may be useful here.

Authors: We will reformulate the sentence. We would, however, argue that it is easy to tell by eye that these distributions are quite different. Of course we can calculate the p-value of a two-sample Kolmogorov-Smirnov test and include that, but really it is evident that the data are drawn from different distributions.

L186 – is a version of the model with higher resolution and better agreement with observations exists, then why not use data from that one?

Authors: The model results were more similar to what is seen in EC-Earth-CC but not necessarily better. EC-Earth-CC displays a too early peak NPP and a too strong

decline after this peak. Model development of NorESM targeted this behavior resulting in a better timing of the peak. We will rewrite this to make this clearer. Note also that daily data from the higher resolution model version was not available.

Fig1: The MAM bloom in EC-Earth-CC seems unrealistically high, but I'm not convinced that either model is able to reproduce the observed behaviour over this time period. A statistical comparison would allow you to state how well these models reproduce observed behaviour. We can't expect ESMs to reproduce specific observations perfectly, but broad scale decadal means should be feasible.

Authors: By construction, neither of these models can replicate observed behavior over specific time periods as the internal variability of the two ESMs is not in sync with the observed one. An ocean only model forced with reanalysed atmospheric forcing could conceivably replicate observed behavior over specific time periods, coupled climate models on the other hand can be in completely different phases on NAO, ENSO, AMO and so on owing to the chaotic nature of unforced climate variability.

Fig2. The coloured lined are the multi-year mean, but what does the lightly shaded area represent?
Authors: The shaded areas show +- 1 standard deviation over the period 2003-2021. This information will be added to the figure caption.

Section 3.2:
L193 – Try to avoid single sentence paragraphs like this.
Authors: This will be changed.

L195: CAFE (uppercase)
Authors: This will be corrected.

Fig 3: Are you actually showing 8 day means on a figure that spans over three centuries? It looks like the shaded areas are the 8 day mean's annual minimum and maximum. Naively, from figure 2, I would expect the minimum value of NorESM2-LM to be around 50 mgC/m2/day, but it appears to be lower than that? Similarly, the range in EC-Earth-CC has a minimum around 200 in figure 2, but it looks closer to 150 in figure 3. I'm not convince that showing the range is useful here, and it's not clear to be me what it represents. Perhaps it may be easier to show the 5-95 percentile ranges (once again – weighted by area) instead, to avoid erroneously high or low values?
Authors: The shaded area shows the full area weighted mean time series of daily and 8 daily data for the ESMs and CAFE respectively.Whereas the lines shows the yearly means from the different models (i.e. the lines are a time filtered version of the full data set). We will add figures for the different subregions in the revised manuscript as suggested by both reviewers. As for the idea about the percentile ranges. Note that these are, non stationary, time series from single models, not distributions of outcomes from a model intercomparison experiment. That said, we don't understand what distribution those percentiles would show?

Figure 4 hides a lot of the important information, only showing a little of the earlier years

underneath the later years. Perhaps you could instead show some decadal averages (or variouschange point regimes) as semitransparent bands?
Authors: Yes, it is true that a lot of information is hidden in this figure. We will make 30 yr mean seasonal cycles instead.

208: Is the peak NPP the best metric for this? In the past, I've seen bloom timing calculated using the maximum in the first derivative, ie,when phytoplankton is growing the fastest, or when the chlorophyll concentration rises above the long term median: See for instance: Philippart, C.J.M., van Iperen, J.M., Cadée, G.C. et al. Long-term Field Observations on Seasonality in Chlorophyll-a Concentrations in a Shallow Coastal Marine Ecosystem, the Wadden Sea. Estuaries and Coasts 33, 286–294 (2010). https://doi.org/10.1007/s12237-009-9236-y ,Marie-Fanny Racault, Corinne Le Quéré, Erik Buitenhuis, Shubha Sathyendranath, TrevorPlatt, Phytoplankton phenology in the global ocean, Ecological Indicators, Volume 14, Issue 1, 2012, Pages 152-163, ISSN 1470-160X, https://doi.org/10.1016/j.ecolind.2011.07.010.
Authors: We claim no optimality for this metric. There are different phenological indicators that could be used. The timing of the peak is a well defined property that has been used before (eg. Nissen and Vogt, 2021; Henson et al., 2013) while several different methodologies, giving different results, exist for the timing of the start (Thomalla et al., 2015). We have tried some other phenological indicators but found the day of peak NPP to be a more robust metric for this data set. Max is generally robust unless the distribution is bimodal, and two peaks have similar magnitude, which we did not see. First derivatives are typically spiky, concentration based cut-offs are a bit arbitrary, and different cut-offs would likely be needed for the two models.

Figure 5: The pane labelled 1850 is the mean of 1850-1879. Perhaps these labels should be 1850-1879 mean, 1970-1999 anomaly, and 2070-2099 anomaly. I also think that a discrete colour scale would be useful here. I'd also like to see the CAFÉ data peak NPP day.
Authors: We will change the labels and experiment with discrete color scales. We will also add CAFE day of peak NPP.

L241: I'm not convinced how representative the mean over the whole region is here, especially after seeing how heterogenous the phenology behaviour is in fig 5! Perhaps it would be better to define sub regions within the domain and see how they behave (ie Labrador Sea, Gulf Stream, Southern NA, Central NA, Northern NA... etc.)?
Authors: We agree and will define subregions as was also suggested by reviewer 1.

L254: The fact that several methods agree that 2010 is a change point for EC-Earth-CC should give you confidence that this is a real change. However, there isn't the same agreement in NorESM2-LM. How would you interpret this? Are the changes perhaps more real in EC-Earth-CC than in NorESM2-LM? Can this shift at 2010 be seen in any observations? How do the phytoplankton bloom and the physical drivers of the bloom differ in EC-Earth-CC before and after 2010? The second EC-Earth-CC change point is around the year 2090 – how many years after the change are required to register the change?

Authors: The kernel based model finds all changes in the probability distribution. This means that the exact nature of the change is not clear. We therefore complemented the analysis with the L1 and L2 method that finds changes in median and mean respectively. That the change points are not at the same location just means that the changes in mean, median and higher order changes do not occur at the same point in time. We think it is quite important to not overinterpret change points. In the sense that the statistics of the time series changes in some appreciable manner, all change points are real. In the sense that there is necessary a co-occurring change in the physical drivers, perhaps no change point is real. The observational time series is not long enough for it to make sense to look for change points. Moreover, the year 2010 in EC-Earth is not the same as 2010 in NORESM or in reality. What they have in common is similar levels of greenhouse gasses and aerosols in the atmosphere. Climate variability is not in phase in the two (or any other CMIP6) models nor is either model in phase with reality. Therefore, it is unlikely that 2010 has that significance. There is one possibility, however, 2010, was the year Eyjafjallajökull erupted. Perhaps this could have an imprint both in the models and in reality.

The algorithm generally always picks up change points at the end points of the time series. So the constraints of the kernel based method together with the pelt search algorithm used does not impose a minimum distance from the boundary for where change points can be picked up. So in a mathematical sense, change points can be found anywhere. However, our own, more subjective, constraints lead us to disregard change points at the end points of the time series. However, we have not decided on a specific distance from the end-point where change points should be dismissed. The latest change points found by the algorithm (except the end point) at year 2090, we subjectively judge to be far enough away from the boundary. The subjectivity may seem unwanted. However, a degree of subjectivity is no doubt inevitable when trying to translate a purely statistical result into a result that can have a biogeochemical meaning.

Figure 6: What's the value of using 8 change points? I can't see them mentioned anywhere in the text, I recommend removing them from this figure if they aren't discussed.
Authors: The reason for the inclusion of the 8 change points was to illustrate that decreasing the penalty gives change points all over the time series. We will reevaluate this after having remade the figures for subregions instead as suggested by both reviewers.

Figure 6: What does it look like if you apply this same method to CAFE? There isn't as much data but there is still a couple decades. Is that enough to detect changes?
Authors: We can for sure get change points in the time series as we are able to obtain change points all over the 350 yr model time series for a low penalty. However, as the time series is very short it is difficult to say if the changes picked up are at all significant.

L271 and figure 7: How do you interpret change points in the pre-industrial period? If they can occur using this method, then it's hard to justify that later change points are linked to climate change without additional analysis.

Authors: It is generally speaking true that the change point analysis does not say anything about the cause of the change. Therefore, regardless of when a change point occurs you always need some additional arguments to tie it to e.g. climate change. Nevertheless, our main result is that the largest changes over this 350 yr timeseries occur in the late historical simulation or in the future scenario for the vast majority of this region as seen in Fig. 7. This is true in both models and their natural variability is not in sync, climate change is thus a very plausible candidate as the cause of these changes. We will add some more discussion around this in the revised manuscript.

Figure 7: I really like this figure, but I think it could be more effective. Perhaps you could focus in on the recent past and the future regions by limiting the colour scale to (ie) the years 2000-2100? Do we really expect change points to occur earlier than say 1950? If white means no change single point could be found, the land colour needs to be a different colour – light grey perhaps. (You also have some ocean points occurring over the land mask, so make sure you set land zorder to be higher than the pcolormesh here and elsewhere. Similarly, the contour lines are the same style, thickness and colour as the same surface, and this is confusing.
Authors: Change points are indeed occurring before 1950 in some grid points seen as blue colors in the figure. We thought this was interesting to show. We do however see the reviewers point that it is difficult to see variations in the green colors which constitutes the majority of the figure and we will change the time span of the colorbar. We will also change colors and line styles in accordance with the reviewers suggestions.

Fig 7 caption: rephrase “White spaces are areas where a single change point could not be found.”
Authors: We will rephrase.

L278: I find the Smythe theory about Net heat Flux (linked above) particularly compelling – even if it may not be applicable to the open ocean. Could heat (or temperature as there is a lot of CMIP6 SST data!) play a role in bloom timing?
Authors: We have compared to SST but did not find any convincing correlations. We will add a discussion around this.

L279: Please be cautious with using “more and less” to describe depths. As closer to the sea floor means larger values of depth, “40m or more” can mean: “deeper than 40m” or “shallower than 40m”!
Authors: We will change this.

L287: You have tested for the first day below 40m but comparing figures 6 and 8 makes it look like this threshold generally occurs after the peak NPP in EC-Earth-CC. The MLD lines average around day140, while the peak NPP seems to be around 130-135. How can MLD shallowing occur after the bloom peak if it is the main cause of the bloom initialization?
Authors: In reality MLD shallowing in the model is gradual, but the choice of a 40 m threshold gives it a discrete date. In other words, we suggest that blooms may start when the MLD becomes sufficiently thin, not that 40 m is the magic number. The chosen number, 40 m, is just a proxy.

Fig 8: This figure could be merged with figure 6, and it would drive the MLD discussion earlier in the results section.
Authors: We will merge those. However, the figures will change as we will separate the analysis in different subregions as suggested by both reviewers.

Conclusions:
L300: Unresolved Eddies? How do you know this? I'm not convinced that you've demonstrated this conclusion.
Authors: It is true that this has not been properly demonstrated. Directly demonstrating this would amount to running also eddying versions of the models and compare the results, that we have not done. The sentence is intended to be more of a well founded hypothesis for future investigations. The idea is simply that the region is very much more dynamic in reality than in the models. The Gulf Stream meanders, rings are detached, etc. Such dynamics lead to spreading of water properties. In a long time-average like in fig.1 we therefore expect the CAFE data to have more spatially homogenized properties than the non-eddying models. We will rephrase and elaborate on this in the manuscript.

L313: "1/11 day/s in NorESM2-LM/EC-Earth3-CC" should be: "1 day in NorESM2-LM and 11 days in EC-Earth3-CC"
Authors: This will be changed.

L314: 31/33 should be "31 and 33"
Authors: This will be changed.

L320: First mention of fish! Maybe put something in the introduction or the discussion sections.
Authors: We will put something about this in the introduction and discussion.

L322: First mention of ecosystem structure! Maybe put something in the introduction or the discussion sections.
Authors: We will put something on this in the introduction and discussion.

Code Availabililty:
L331: What about the Ruptures python package?
Authors: A link to Ruptures will be included.

L340: This link did not work for me, nor could I find it using the zenodo search bar.

Authors: The link works but a space is missing after "Zenodo:" which makes it look like that word is supposed to be included in the link. We will change this.

References:

Bhatt, U. S., M. A. Alexander, D. S. Battisti, D. D. Houghton, and L. M. Keller (1998). Atmosphere–Ocean Interaction in the North Atlantic: Near-Surface Climate Variability.

J. Climate, 11, 1615–1632, https://doi.org/10.1175/1520-0442(1998)011<1615:AOIITN>2.0.CO;2

Barsugli, J. J., and D. S. Battisti (1998). The Basic Effects of Atmosphere–Ocean Thermal Coupling on Midlatitude Variability. J. Atmos. Sci., 55, 477–493, https://doi.org/10.1175/1520-0469(1998)055<0477:TBEOAO>2.0.CO;2.

Geider, R. J., MacIntyre, H. L., and Kana, T. M. (1997). A dynamic model of phytoplankton growth and acclimation: responses of the balanced growth and Chlorophyll a : carbon ratio to light, nutrient limitation and temperature, Mar. Ecol.-Prog. Ser., 148, 187–200.

Myriokefalitakis, S., Gröger, M., Hieronymus, J., and Döscher, R. (2020): An explicit estimate of the atmospheric nutrient impact on global oceanic productivity, Ocean Sci., https://doi.org/10.5194/os-16-1183-2020

Nissen, C. and Vogt, M. (2021). Factors controlling the competition between Phaeocystis and diatoms in the Southern Ocean and implications for carbon export fluxes, Biogeosciences, 18, 251–283, https://doi.org/10.5194/bg-18-251-2021.

Richardson K, Bendtsen J (2019). Vertical distribution of phytoplankton and primary production in relation to nutricline depth in the open ocean. Mar Ecol Prog Ser 620:33-46. https://doi.org/10.3354/meps12960

Thomalla et al. (2015). High-resolution view of the spring bloom initiation and net community production in the Subantarctic Southern Ocean using glider data, ICES Journal of Marine Science, Volume 72, Issue 6, Pages 1999–2020, https://doi.org/10.1093/icesjms/fsv105