Replies to reviewer 1

General comments:

The addressed question of this paper is very interesting and the applied statistical package to detect shifts in phenology seems to be adequate. However, I am not convinced that the strategy of using an area-based average gives meaningful results. As mentioned by the authors, the North Atlantic shows a very heterogenous regional pattern between 30°-60°N. Longhurst classified at least 4-5 distinct oceanic biogeographical provinces. One might expect each biogeographic province to exhibit different temporal behaviour; e.g. Figure 5 clearly shows this for NorESM (a dipole structure for 2070-1850 and only an 1 day change of the mean peak NPP day over the entire domain, as stated in line 219). From my perspective, the analysis of peak NPP day averaged over the entire domain is therefore of little informational value.

I recommend to repeat the analysis with the kernel based model for smaller domains (Longhurst provinces or regions aligned with common characteristics of the ESMs?). Then it might also be possible to identify local drivers of the change of peak NPP day. I assume that the intention of the investigation on the "Day of MLD <= 40m" was to identify one of these potential drivers. This MLD analysis came as quite a surprise and its findings should be mentioned in the abstract. Again, I'm not convinced that the approach of using an average value of MLD day for the entire domain is reasonable. The relatively high cross correlation values with negative lag (in Fig. 9) might be related to this averaging.

However, I appreciate this data analysis and therefore recommend publication of this manuscript with major revisions. Please see also my specific comments.

Authors: We thank reviewer 1 for insightful comments and ideas that have greatly improved the manuscript.

We have separated the region into Longhurst provinces which better captures the spatial heterogeneity.

Specific comments:

L75: Please motivate the analysis of the change in MLD day already in the introduction.
Authors: We have added a motivation for the MLD analysis in the introduction as suggested (lines 85-89).

L96: There is a difference between “primary production” and “net primary production”. The latter is the daily growth of phytoplankton minus respiratory demand. Please use net primary production throughout the manuscript.

Authors: This has been corrected.

L101: Skyalls et al (2019) analysis did only cover a north-south transect east of 20°W from two cruises. Thus, the good agreement with observations is only shown for a very small part of the domain. Please add this information here.

Authors: This information has been added on line 126.

L119: Please add that phytoplankton growth is also a function of light and temperature in iHAMOCC.

Authors: This information has been added on line 144.

L 124: Please replace the subtitle “Observations”. CAFE is a model based on satellite data, which strictly speaking are also only results of an algorithm (i.e. a model) and not observations. Same for the subtitle 3.1

Authors: The titles have been changed.

L163 : typo, capitalize Gulf

Authors: This has been corrected

L171: typo, delete “the” in “For the the latter…”

Authors: The text has been removed.

L195: Statements about the multidecadal variability of an 18-year time series should be avoided. They are meaningless.

Authors: We have removed such statements.

L214ff: It is difficult to reconcile the results of Fig. 4, which shows a shift towards an earlier peak NPP day in NorESM at the end of the simulation and Fig. 5, which presents an extended area with a later peak NPP day in the Gulf stream region. In L219, you give a shift of only 1 day for NorESM between 2070 and 1850. From my point of view, the weakness of area-based averaging is clearly evident here. Could you also please provide the significance of the results of Fig. 5? Is ±10 days significant, especially for EC-Earth? It might be useful to also show the peak NPP day for the CAFE data set as a longitude-latitude plot. (see also comment L259)

Authors: We have separated the region into Longhurst provinces which better captures the spatial heterogeneity. Indeed the response is quite different between different provinces. The mean change between the last 30 yrs of SSP5-8.5 and the first 30 yrs of historical is summarized in Table. 2.

In the new version we added the day of peak NPP per region in Fig 3. and in Tab. 1. Given the large amount of figures, especially panels, with the new regions we thought

it was better not to add a whole new figure with full lat-long coverage on top of all the other new figures.

Regarding the significance we did add two new panels to Fig. 5 which shows the change normalized by the yearly standard deviation from the PI-control. It is readily evident that these results would be very unlikely to occur by chance.

L252: Please find different symbols for “l1” and “l2” for a better readability (e.g. capitalize L1 ?)

Authors: This has been changed as suggested*.*

L255: Could you please give an explanation why L1 and L2 do not identify the 2070 change point in NorESM? Would the results be more consistent if you decrease the penalty for L1 and L2 ?

Authors: The kernel based model does not provide us with information on the nature of the change that occurs in connection with a change point. L1 and L2 give us change points related to a change in median and mean respectively. The change picked up by the kernel based model may therefore be related to a higher order change (e.g. skewness or kurtosis) in the probability distribution.

L259: The finding of the kernel based model for EC-Earth is consistent with the findings of Fig. 5 (12 days change in peak NPP day compared to 11). However, the result for NorESM is quite different (10 days instead of 1). In addition, Fig.6 seems to give a much larger change for the end of the simulation than 1 or 10 days. Could you please elaborate this issue for NorESM? Furthermore, why didn’t you already show the period 2010-2029 in Fig. 5? It seems to contain a strong change pattern. Why do you use 30-year averages for Fig. 5 and here you use a 50-year average?

Authors: This problem has been corrected. The results are now also summarized in Table. 2.

L273: Please give this motivation to look at the MLD already in the introduction
Authors: We have added a motivation for the MLD analysis to the introduction (lines 85-89)

I have to admit that I’m not familiar with the used kernel based model. However, results shown for day of peak NPP in Fig.6 seem plausible: the decreasing penalty increases the numbers of change points that are found, but the change point with highest penalty is identical with one of the two changes points (there is an overlap of the pink line with one red lines in both models). I became confused when I looked at the results of “Day of MLD <= 40 m”. There is no alignment between pink/red lines. Why does the kernel based model miss the largest change point with a decreasing penalty?

Authors:

The largest changepoint using one set of constraints is not necessarily also the largest when another set of constraints are used. This has to do with the fact that changepoints, although defined as local properties, are dependent on an optimal segmentation, a global property. The problem of finding the changepoints is to find the optimal segmentation of a signal given some criterion. A useful criterion codes for example for, the type of change that is looked for and the number of changes. One may for example wish to find the best segmentation into two segments in terms of distance from the mean of some signal. One could then calculate the sum of the deviation from the mean for all possible segmentations by simply looping through the time series, and the smallest one would be the change point. However, if one instead had three changepoints, the set of possible segmentations is different,the optimal segmentation is by necessity also different (as you have more segments) and the largest change point could be, but is not necessarily, different.

L292: Could you explain why both models have relatively high cross correlations for negative lag (-3 to -1) years?

Authors: Our interpretation is that the fact that there is strong correlation on many both positive and negative lags is owing to the changes seen in the two variables having the same drivers (i.e. global warming). Both mixed layer depth and NNP are directly affected by warming, so there is not just a simple NNP=f(MLD) but also at least one lurking variable in temperature. There are, of course, even more variables at play, but the reverse coupling that could exist i.e. NNP affecting light penetration and thus MLD, is not used in the models. We have added some discussion on this on lines 372-375.

L313: Here just to summarize that the change in peak NPP day is 1 day for NorESM is not convincing. For a different time period (1960-2010) versus (2010-2060) it was 10. It would be helpful to provide a little more discussion.

Authors: This problem has been corrected (review comment relating to L259).

L316 Please share your view on the results of the cross correlation analysis. Does it support the theory of Behrenfeld and Boss?

Authors: The cross correlation between the day of peak NPP and the first day of MLD above 40 m indicates a high correlation between the two on an annual basis for most provinces. The new results do however also show provinces without notable correlation and even anti-correlation. Even though the correlation of course is not an indication of causality the results are in line with the Disturbance Recovery Hypothesis (Behrenfeld, 2010; Behrenfeld and Boss, 2014) as well as The Critical Depth Hypothesis of when a bloom has the potential to start (see review by Behrenfeld and Boss, 2017).

We have added discussion around the correlation and anticorrelation between day of peak NPP and day of MLD shallower than 40 m on lines 360-377.

Figure captions:

Fig. 1 Give period for the data averaging; is it 2003-2021 as in Fig 2 or 2002-2021 as stated in L135? Did you also masked the model data as in Fig 2? Please indicate in the text if you always use masked ESM data for comparison with CAFE. How large is the difference between masked and unmasked ESM data ?

Authors: The masking is now applied on both Fig. 2 and Fig 3. The maximum latitude of extant CAFE data is displayed in Fig. S1. The masking affects Nov, Dec, Jan and Feb values.

Fig. 3 The overlapping colour coding is difficult. Please find a better solution (e.g. instead of full time series show only mean/min/max for each ESM model.)

Authors: We have removed this figure and added instead a figure (Fig. 4) showing annual mean NPP for each Longhurst province.

Fig. 5 Please add significance level

Authors: We have added the 2070-2099 anomaly divided by the standard deviation over the PI-control run in the lower panels of Fig. 5.

Fig. 6 I assume that the centre of the symbols (circle and triangle) corresponds to the year of the change point. Please add this information. Please use the same names for l1 and l2 through the manuscript and in the figures. E.g. L1 and L2, not "model l1" or "model=l1"

Authors: This has been changed as suggested.

Fig. 7 I assume you used the highest penalty level (corresponding to the pink line). Please note. Please also change order of figures – usually EC-Earth is the top figure.

Authors: We have changed the search method to optimal detection where we predefined the amount of changepoints. The corresponding results using the Pelt search method and a penalty tuned to pick up only one change point is presented in the Supplementary material. The order of the figures has been changed.

Fig. 8 The figure caption is incomplete ; add …in the median (model l1). Or just write : same as Figure 6, but for first day of ….

Authors: The caption has been expanded as suggested.

References:

Behrenfeld, M. J. (2010). Abandoning Sverdrup ' s Critical Depth Hypothesis on phytoplankton blooms. Ecology, 91(4), 977–989.

Behrenfeld, M. J., Boss, E. S. (2014). Resurrecting the ecological underpinnings of ocean plankton blooms. Ann Rev Mar Sci.;6:167-94. doi: 10.1146/annurev-marine-052913-021325. Epub 2013 Sep 25. PMID: 24079309.

Behrenfeld, M. J., & Boss, E. S. (2018). Student's tutorial on bloom hypotheses in the context of phytoplankton annual cycles. Global Change Biology, 24(1), 55–77. https://doi.org/10.1111/gcb.13858

Replies to reviewer 2

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 that has helped us to greatly improve the manuscript.

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: The community structure of the biogeochemical model has been shown to affect the NPP with the more complex models generating a more non-linear response to climate change (Fu et al., 2016). We have added some discussion around this on lines: 365-370.

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.

We have added some discussion around this on lines:372-379

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:
We have added some discussion on this on lines:338-348.

Although daily surface chlorophyll (chlos) has the advantage that it is easily validated against satellite data and it is a common output from ESMs, it is not in a simple manner related to vertically integrated NPP. The vertical structure of chlorophyll differs between regions and peak NPP may occur below the surface layer (Sathyendranath et al., 1995, Richardson and Bendtsen, 2019). Therefore, peak

surface chlorophyll does not necessarily occur at the same time as peak vertically integrated chlorophyll or the NPP.

Furthermore, the model runs presented in this work were made within the H2020 project COMFORT and daily output of NPP as well MLD, SST and some biogeochemical variables are available. It is therefore possible to compare the model output of NPP to these variables which would not be possible using the CMIP archive as most of these variables are not saved at daily resolution.

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 kernel method for change points used here is not simply picking up changes in a given moment of a distribution, like mean, variance or skewness. It can detect any type of change. This has obvious benefits in that the palette of changes that can be detected is much larger. However, it also makes interpretation more difficult. This ambiguity is the reasoning behind using the L1 and L2 models that focus on specific moments. Most often, the L1 and L2 models find the same change points as the kernel model, suggesting that the changes found are common to many statistical moments (line 356).

However, the information that we are most interested in is really more tied to physical and biogeochemical reasoning than to statistical moments. From this perspective, it is instructive to use a change point analysis mostly as a detection algorithm, but to leave the interpretation of the nature of the change to physical and biogeochemical reasoning. This is the essence of why we combine e.g. MLD analysis with the phenology, it also means that only some change points will be meaningful to us in this sense. Of course, also in the statistical sense the meaningfulness of a change point is down to arbitrary measures, like the size of a chosen penalty or one's favorite level of confidence.

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 have added discussion about how different drivers affect NPP in the revised version (From line 362).

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 was put together with the results section in Section 3, Results and discussion. In the new manuscript we separated them and added more discussion.

About the CAFE data and the length needed to detect these change points, it is important to realize that all change points that are down to natural variability will be different in CAFE and the two ESMs. So we should not expect, generally speaking, to find change points from an ESM in CAFE unless we can tie them to changes in a common forcing like greenhouse gases or aerosols. This is true regardless of the length of the CAFE time series.

However, it is also true that generally speaking that if a single time series, T, of length n has only two change points and both are in the latter half of the series (i.e. in T(n/2:n)), then the time series T(n/2:n) still could have two different change points. That is, running T(n/2:n) through the same change point algorithm as T does not necessarily give the same change points, because the optimal segmentation is a global property. In other words, the optimal segmentation of T and T(n/2:n) is not necessarily the same. Therefore, even for perfect replica time series, the time series has to have the same length to ensure that they have the same change points.

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. 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 in terms of having a reasonable climatology. 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 a model. However, the silver lining is that understanding and describing the behavior of ESMs is interesting and important even in itself given their extensive use in science and their impact on policies. We have added a comment on this (line 215).

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 is changed so that our main result is highlighted: Phenological shifts in the North Atlantic net primary production detected in the 21st century. Results from two Earth system models.

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 have rewritten the abstract in order to showcase the main result higher up.

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: We have removed 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 estimate on the size of change. Furthermore, many of the largest change points in the timeseries occur in both models at the end of the historical simulation (Line 351).

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 reference does not give explicit numbers for NPP.

L35 and elsewhere: CO2: 2 should be subscript

Authors: This has been 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.

We have added some references on line:43.

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: The text has been changed starting from line: 54.

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 have added this reference (Line 60)

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 a very important addition that gives us a good gauge of the range of the natural variability, something that the forced changes can be compared against, see Fig. 5 bottom panel (new addition). 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 has been corrected.

L89: pCO2: 2 should be subscript
Authors: This has been 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 have changed to "net primary production".

134: Modis should be MODIS
Authors: This has been 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: We have expanded this section.

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. We clarified that it is total water column NPP on line 153.

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 relevant questions. In essence, what we have here is not a well constrained model**

data fit problem. Any method of comparison is only as good as the data admits, and this data admits very little in terms of direct comparisons. Put another way; say we calculate a spatial correlation between the CAFE data and our two models and find that r(EC-Earth,CAFE)<r(NORESM,CAFE), or that a trend in some area in EC-Earth is more similar to that in CAFE than to that in NORESM, what, of value, would that tell us? The answer unfortunately is nothing of great value, as all such metrics would be strongly influenced by natural variability during the short CAFE period. This type of variability is not in temporal sync in these different models. A comment on this has been added on line 213-215.

L176 – If the mask is important, you may need to indicate it in this figure (or elsewhere). ***Authors: We have added the seasonal cycle of the maximum latitude of the Cafe data that has been used to mask the model data to the supplementary material (Fig. S1).***

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

Authors: We have removed this statement. 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 these data are drawn from different distributions. Moreover, given that they are sampling different histories it is not known what a good fit would be.

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 have removed this statement in the new version that includes separate biogeochemical provinces.

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. See comment on lines 213-215.

Fig2. The coloured lined are the multi-year mean, but what does the lightly shaded area represent?
Authors: (now Fig. 3) The shaded area has been removed as the figure that now includes eight different biogeochemical provinces would become too messy.

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

L195: CAFE (uppercase)
Authors: This has been 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: (Now Fig. 4) We have removed the shading as the figure that now contains eight different biogeochemical provinces would get too messy. The lines show the annual means from the different models.

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: We have removed this figure.

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.

We have added some text on this on line 96.

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 have changed the labels. We have added the mean day of peak NPP for the different provinces and for both CAFE and the ESMs to Table 1.

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 have redone the analysis for Longhurst provinces.

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-Earth3-CC is not the same as 2010 in NorESM2-LM 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. Moreover, we note now with the regional analysis in place that a 2010 change point is not a common feature between models and regions.

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: We have removed those.

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 much shorter and the change points depend on the global segmentation the results would not be comparable.

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. Moreover, the bottom panel in Fig 5 now shows the change normalized by the yearly standard deviation from the PI-control experiments. From this it is very clear that these changes did not occur by chance.

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: We have edited the figure. The color map is now exchanged for a diverging centering around the year 2000. We find that this gives a clearer view of the conclusion that the largest changepoint for most of the grid points occurs after the year 2000. We changed the land color. Furthermore, we changed the search method so that there are no nans in the data. A corresponding map containing the old search method is, however, included in the supplementary material (Fig. S6.). In this figure, previously white nans have been colored blue.

Fig 7 caption: rephrase "White spaces are areas where a single change point could not be found."
Authors: We have replaced this figure with a corresponding one using a different search method. The white spaces are not present in the new figure. A figure using Pelt search method has been added to the supplementary material (Fig. S6).

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 have added a discussion around the connection to other variables on L339. We have also added the Smyth et al reference (line 60). However, as we did not have daily output of net heat flux, this was not possible for us to examine.

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 have changed 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 definitive 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. Several other numbers have been tested with similar results (Supplementary material Figs. S3-S5).

Fig 8: This figure could be merged with figure 6, and it would drive the MLD discussion earlier in the results section.
Authors: We found it difficult to merge the figures as they both now contain data from all eight biogeochemical provinces. We have, however, switched the figures so that the MLD figure (now Fig. 7) comes right after the day of peak NPP (Fig. 6).

Conclusions:

L300: Unresolved Eddies? How do you know this? I’m not convinced that you’ve demonstrated this conclusion.
The line is no longer 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 has been changed. We have also included a table (Tab 2.) that summarizes the change in the different provinces.

L314: 31/33 should be “31 and 33”
Authors: This has been changed.

L320: First mention of fish! Maybe put something in the introduction or the discussion sections.
Authors: Fish is also mentioned in the introduction on lines and 63.

L322: First mention of ecosystem structure! Maybe put something in the introduction or the

discussion sections.
Authors: The effect of model community structure on the NPP is mentioned from line 380.

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

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

Authors: We have rewritten.

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.

Fu, W., Randerson, J. T., & Keith Moore, J. (2016). Climate change impacts on net primary production (NPP) and export production (EP) regulated by increasing stratification and phytoplankton community structure in the CMIP5 models. Biogeosciences, 13(18), 5151–5170. https://doi.org/10.5194/bg-13-5151-2016

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