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 will greatly improve the manuscript.

The region is indeed heterogeneous and the change point analysis was therefore made for every grid point to ensure a robust pattern. However, we agree with the reviewer and will divide the region into Longhurst subregions.

Specific comments:

L75: Please motivate the analysis of the change in MLD day already in the introduction.

Authors: We will motivate the MLD analysis in the introduction as suggested.

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 will be corrected. The confusion arises from the fact that these models do not have that distinction (no explicit phytoplankton respiration).

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 will be added.

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

Authors: This information will be added.

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: Agreed, the titles will be changed.

L163 : typo, capitalize Gulf

Authors: This will be corrected

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

Authors: This will be corrected.

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

Authors: We will remove 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 will add a plot showing the day of peak NPP in a longitude-latitude plot as suggested. We will also calculate the yearly standard deviation of the peak NPP day from the PI-control, so that these changes can be compared to the range of the natural variability. Indeed there is a strong spatial heterogeneity in the response and we hope the analysis broken down into Longhurst provinces will reveal interesting regional features. Thanks for the suggestion.

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

Authors: This will be 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: Thanks for noticing! In part this is likely due to different periods being used. However, it is likely also the result of us applying unfortunate averaging. In the case of Fig. 6 and the 50yr means, we first take the spatial mean of NPP and get a time series of daily NPP. After that we take the day of the year of max NPP.

In the case with the 30yr means, we have first calculated the temporal mean of daily NPP over a 30 year period in every grid point so that in each grid point we have 365 NPP values (Jan 1 is the temporal mean Jan 1 over 30 years and so on). After that we calculate the day of peak NPP in each grid point and obtain Fig. 5. Only after that we do the area averaging to get the 30yr mean values you are referring to. Thus Fig. 6 shows the spatial mean of the day of max NPP, while Fig 5. shows the day of max NPP of the mean seasonal cycle of NPP over the 30 yr period.

Of course, the order of operations matter when the max function is used. Since our focus is on the day of max NPP, we favor having only the mean of the day of max NNP in the revised manuscript. However, the difference between the two models in this respect is interesting and the implications of having the differences we see in NORESM might be worth investigating. One possible interpretation is that the larger signal toward the end of the simulation in NORESM in Fig. 5 than in Fig. 6 could be due to large interannual variability in NNP during those late years.

We will also include a running mean in Fig. 6 so that differences in the mean over different periods can more easily be distinguished.

L273: Please give this motivation to look at the MLD already in the introduction

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: We believe this is a sort of nonlinearity that materializes. Imagine that we have a dataset that changes in mean in two equal steps at times t1 and t2 from M1 to M2, so that each change is (M2-M1)/2. Having two change points you expect to find t1 and t2, but having only one, what is the expectation (t2-t1)/2 perhaps? However, admittingly, the routine we used for change point evaluation is not easily adaptable to make quantifications of the sort that could verify or refute this hypothesis, so the explanation is clearly hand-wavy.

As an antidote we will in the revised manuscript also try an alternative search method where the number of change points is a user input and see if the phenomenon persists.

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.

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 was discussed above (review comment relating to L259) and will be addressed in the revised version.

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. 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). Some more discussion around this will be added to the manuscript.

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 only applied in Fig 2. However, as the November values are in fact influenced by low winter light we will apply the mask in the SON maps in Fig. 1 as well. Furthermore, we will add curves showing the unmasked data to Fig. 2 or the corresponding figure for the different subregions that will be implemented in the revised version.

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 will instead use min and max levels for each ESM.

Fig. 5 Please add significance level

Authors: We plan to calculate the standard deviation over the PI-control run which will give us a good indication of the natural variability.

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 I1 and I2 through the manuscript and in the figures. E.g. L1 and L2, not “model I1” or “model=I1”

Authors: This will be 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: The penalty in every grid point was tuned to, if possible, pick up only one change point. The level of the penalty is dependent on the time-series in every grid point. The order of the figures will be changed.

Fig. 8 The figure caption is incomplete ; add …in the median (model I1). Or just write : same as Figure 6, but for first day of ….

Authors: The caption will be 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