Review to

Phenological shifts in the North Atlantic net primary production detected in the 21st century. Results from two Earth system models.

Jenny Hieronymus et al.

General comments:

The authors did a great job in revising the analysis. By dividing the North Atlantic into Longhurst provinces, the results are much more consistent and meaningful. Interestingly, the most of the provinces show an increase of NPP until the end of the century except two regions in both models (disregarding the small changes in SARC and NADR in NorESM2-LM). This result was masked in the previous analysis where a declining trend in NPP was postulated for the entire domain over SSP8.5. This finding has to be included in the Summary and Conclusion section which still mentions an overall NPP decrease (L408). In addition, I do feel that many of the final findings are related to results from the EC-Earth, e.g. "phenological shifts occurring in the early 21st century " is not true for NorESM (in 6 out of 8 regions the changepoint is after 2048). Please critically review the entire manuscript to see if the final statements apply to both ESMs.

Authors: We have reviewed the manuscript and made sure that final statements apply to both ESMs.

A general remark on the quality of the figures:

- the increment of contour lines should be specified for the subplots in the caption; e.g. in Fig 3 each of the SON panels has different increment
- contour lines in Fig.8 are horrible – delete or omit the entire Fig. (see specific comments)
- I plea for a,b,c notation in the figures for more readability

  We tested making the figures with that notation during our revisions, but found that it became less obvious what the different panels contained and also that the figure caption became much less readable. We therefore think this is the least worst option for our set of panels.

Authors: We have specified the contour increment. We have removed the contour lines in Fig 8.

In general, I recommend the publication of the manuscript after my specific comments have been addressed.

Authors: We thank the reviewer for the thorough review that has greatly improved the manuscript.

Specific comments:

L25: Please correct: Net Primary Production (NPP) is the rate of photosynthetic carbon fixation minus cellular respiration

Authors: This has been corrected.

L82-83: “We divide the region into biogeochemical provinces (Longhurst et al., 1995) in order to see how localities with similar biogeochemical functioning differ across the region.” This sentence is confusing. What do you mean by “localities”? Do your provinces really have a similar biogeochemical functioning? Delete “Furthermore”.

Authors: This has been changed.

L85-88: Please motivate here the purpose of MLD analysis and reorder the sentences - first: change point analysis for MLD as for peak NPP; second: all about cross-correlation and what we learn from it.

Authors: We have rewritten as suggested.

L94: typo: in section 2.4 is the change point analysis L96: “maximum” instead of “max”

Authors: This has been corrected

L97: “found in your data” ESM data or CAFE or all data sets?
Authors: We were referring to the ESM data, but the statement is true also for the CAFE data.

L117: replace “external concentration in nutrients” by “nutrient concentrations of the ambient water”

Authors: This has been corrected.

L118: Please give the same information for both BGC modules. i.e. delete :“PISCES is suited for a wide range of spatial and temporal scales, including quasi-steady state simulations on the global scale.”
And add for iHAMOCC, that iHAMOCC also simulates the carbon system, as well as dissolved and particulate organic matter”
Authors: This has been corrected.

L119-120: “Net primary production is the growth of phytoplankton thus the term excludes mortality, excretion and grazing.” Why is this mentioned here? By definition, NPP excludes mortality, excretion and grazing. Don’t mix it up with NCP = net community production. Delete sentence?

Authors: The line has been removed.

L163: Rephrase your sentence to e.g. : “The seasonality of NPP depends, among other things, on local physical conditions of the ocean” ?

Authors: This has been changed.

L168: Longhurst defined the static boundaries – “made” is a strange word?

Authors: This has been changed.

L171: You never use “coastal, westerlies and polar” – delete; The North Atlantic domain is divided in the provinces shown in Fig. 1.

Authors: This has been changed.

L176 delete: The west wind regions;

Authors: We have deleted this.

L209ff: I recommend to show and discuss only MAM and JJA and omit SON. It shows a more or less a uniform pattern for the entire domain and complicates the data processing due the lack of data in CAFE in winter. SON gives no additional information. In addition, please find a better color scale. It is surprising, that your scale ends at 1000 but Fig.3 shows numbers higher than 1200. Please correct.

Authors: We have removed the SON panel and changed the colorscale.

L226: Instead of using daily ESM data, use a 8-day running mean for the comparison to 8-day mean data from CAFE. Results in Fig.3 are difficult to compare. Please reorder the seasonal cycles by region instead of data sets: e.g. BPLR+ARCT for CAFE and both models, and so on. Adjust axes to maximum values. Make sure that all lines have the same starting point if you mask the ESM data with available CAFE data.

Authors: We have done this (Fig. 3).

L264: make sure, that you don't use the word "region" for both, the entire North Atlantic and the provinces; use e.g. the words "domain" and "provinces" throughout the manuscript.

Authors: This has been done.

L265ff: could you improve the readability by shorten the name of the 3 periods: e.g. 1865s = 1850-1879, 2000s= 1985-2014, 2085s= 2070-2999? Then you can omit to write "period" or "early/late period"

Authors: We have changed the names of the periods as suggested.

Fig 5: Please use a standard statistical test (e.g. student's t-test) to determine the significance. With the given information, it is difficult/impossible to interpret the results. Please show results of EC-Earth on the left side as usual.

Authors: We have performed a Kolmogorov-Smirnov test and marked the significance to the 95th percentile in the bottom panels in Fig. 5. However, we do think that the pattern of change divided by the standard deviation of the piControl adds to the analysis as we obtain a measure of how far we are from the natural variability of the piControl.

L277 " size of NPP" – delete "size of"

Authors: This has been done.

L284: you don't average over different provinces, rephrase.

Authors: We have rephrased.

L285: Fig. 6 shows …. together with the largest (…. sentence incomplete

Authors: This has been corrected.

L304: The posed question was reasonable for the previous analysis, but I cannot see the benefit when using Longhurst provinces. Isolines in Fig. 8 should be removed, if not the whole figure is omitted or transferred to the supplement. In the supplement you could also add the discussion on the difference between the PELT method and Fig.8 and why one has blanks and the other not.

Authors: We have removed the isolines in the figure. We do, however, think that the figure adds to the analysis. Mostly because there can be considerable variations also within the Longhurst provinces and also because it gives an indication of the suitability of these regions for the ESM data. An interesting avenue of future research would be to let an objective algorithm find provinces and then compare those to the Longhurst ones.

L316: province averaged instead of area-averaged? Or just write: "between the time series in Fig 6 and 7" because it is clear how they were archived.
Authors: "area-averaged" has been removed.

L320: Typo? NADW is not defined
Authors: This has been changed to NASW.

L323: "Looking at Fig. 8 ….. you mean Fig 9 ?
Authors: We mean Fig 7. This has been corrected.

L329: you never show the "size of peak NPP". delete "size of peak" or explain what you mean
Authors: We have revised this statement.

L351: use "finding" instead of "observation"
Authors: This has been changed.

L352: replace "then that…" with "when the warming is the strongest in the SSP5-8.5"
Authors: This has been changed.

L392: replace "realistic physics" with "consistent physics"
Authors: This has been changed.

L400-401: rephrase: you don't use Longhurst provinces to look at spatial averages, but to account for the different areal conditions
Authors: This has been rephrased.

L408: As already mentioned above, the NPP increases for many provinces. Revise!

Authors: This has been revised.

Phenological shifts in the North Atlantic net primary production detected in the 21st century. Results from two Earth system models.

Jenny Hieronymus, Magnus Hieronymus, Matthias Gröger, Jörg Schwinger, Raffaele Bernadello, Etienne Tourigny, Valentina Sicardi, Itzel Ruvalcaba Baroni, and Klaus Wyser

Submitted to Biogeoscience, bg-2023-54

Review by Dr. Lee de Mora, Plymouth Marine Laboratory, Plymouth, Devon.

Summary

In this work, the daily Net Primary production of two Earth System Models in the Northern Atlantic are described, compared against satellite data and analysed using change point and cross corelation analysis for several regions of the North Atlantic. The timing of the peak of the bloom will shift earlier in the year in the Northern parts of the North Atlantic. The models disagree for the Southern North Atlantic, but it is less of a shift than in the northern regions. The change point analysis highlights that several regions are likely to have pass the change already and that nearly all regions will cross the change point in the 21st century. However, it's not clear how significant the scale of the change point will be.

The text is well written, the underlying science is well introduced in a clear way, the results are presented and described accurately, I did not spot any spelling mistakes and the grammar is almost always fine. At times, the style is a little colloquial and would be improved if certain parts were written in a more formal style. There's a few run-on sentences which need to be pruned. There were many formatting issues, described below, and there are likely many more that I missed.

The figures are generally clear, but I suggest a few improvements below which I think would benefit the paper as a whole.

I found that there were a few discussion points that were hinted at in the abstract and introductions, but never made it to the final draft. I have a few questions below, a few suggestions and a few possible additions.

In my opinion, the biggest weakness of the paper as it currently stands is that the key results are not sufficiently well articulated. It's crucial that the revision of this paper focuses on its unique results and explains why they are important as clearly as possible.

I would also recommend a careful and thorough readthrough (including the references section) before re-submission. Sometimes it's better to share this task with a more distant co-author, as the lead author is often too close to the work to spot these issues.

As the list of changes below has become rather long, I recommend major corrections before reconsideration. However, I don't want to come across as being harsh. This is a good, well written paper, with good scientific content, it fits within the remit of the journal, and most of these changes should be resolvable with minimal effort.

Authors: We thank Dr. Lee De Mora for the ambitious, in depth review that has led to a greatly improved manuscript.

Specific Comments

I expect a more forceful and direct tone in the title, abstract and the conclusions. At the moment, the abstract focuses on the methods, but it should effectively read: "This is our main result. This is why it

is important." Then once you've said that, only then you can describe the methods and models that were used to found out.

Similarly, the title could be a more direct and effective. Something like: "Net Primary Production Annual Maxima in the North Atlantic projected to shift in the 21st century" or something like that. (As an aside, I don't think you need to mention ESMs in the title – it's obvious that models were used if you're making projections of the future!)

Authors: The Abstract and Conclusions has been rewritten as suggested. The title has been changed to that suggested by the reviewer.

I'm not convinced that either model captures the observations over the historical period. The model-data comparisons in figure 1 and table 1 are subjective. I'd like to see a robust statistical comparison of the model and data over the historical period. A graphical version like a pair of Taylor or Target diagrams would be a solid improvement. Alternatively but probably less effective, you could add pattern statistics (bias, deviation, and correlation) to Table 1. Please bear in mind that you have made a new and unique piece of work, and people in the future will be glad to have a robust statistical benchmark to cite that they can compare their model quality against.

Authors: We went through this also a bit in the first review, but we think it needs to be iterated why we don't believe this idea makes sense. Firstly, we are not comparing two models to a set of observations. We are comparing two earth system models to a different kind of model that utilizes satellite observations to infer phytoplankton biomass. These are not observations of phytoplankton biomass. This means of course that the CAFE model may have sizable errors, just as the ESMs might have. Moreover, while the CAFE model by construction models our recent history, the two earth system models do not. These are free running coupled models. The thing they have in common is that they are driven by similar greenhouse gas and aerosol forcing as our history. The climate variability is, however, not in sync. All relevant climate indices like NAO, AMO or AMOC will be out of phase in these three different models. Therefore a conventional Taylor diagram looking at temporal correlations would be meaningless. An early bloom in one model should not imply an early bloom in another during the same year, and so on. The only similarities we can hope for are climatological in nature. Long averages should hopefully be similar, but it is hard to say given the shortness of the CAFE model's time series, if that is indeed the case with our different models. One could in principle do a Taylor diagram with spatial correlations, but given that the temporal unsynced variability will also affect the spatial correlations, such a plot would be nearly impossible to interpret. Two twenty year periods could have rather different spatial patterns owing to different phases in the AMO for example. In conclusion, our unwillingness to put more advanced statistical methods into this comparison is because it would be a lot like looking at correlations between random numbers.

It would be valuable to include the analysis of the mean of the whole North Atlantic region in some of these figures (ie Figure 3, Tables 1, 2 and 3 as well?) I understand the value of the individual regions, but a clear result for the whole region would be a good headline result.

Authors: We have added values for the whole region in Tables 1 and 2.

I would be interested in seeing how different the phenologies of the various regions are on either side of the change point. Basically, a version of figure 3 which compares the climatological mean of ten (or some useful number of) years each side of the change point. This would be a clear and effective way of showing that there is indeed a real change between those two periods.

Authors:

We disagree with this approach. The change points found are real, in the only sense that change points can be real, that is, they are found by some algorithm to satisfy some objective, albeit arbitrary, condition imposed by us to define change points. One might think change points that are change

points according to multiple such definitions to be in some sense more real than others, which is one of the reasons for including the L1 and L2 method. The question you pose has more to do with whether they can be identified rather subjectively by a human eye, than whether they are real. We think such a figure could be useful only in very clear cut cases to illustrate some change that has indeed occurred, but for that we think Fig. 5 is a better illustration.

There is no daily net primary production data from CMIP6 on ESGF, but there is quite a lot of surface chlorophyll (chlos) and surface phytoplankton carbon (phycos). It would be interesting to place these two models against the rest of CMIP6 in the context of the phytoplankton carbon or chlorophyll. Are they typical or are they outliers? This would be a lot of additional work, so I leave it up to the authors to decide whether they can perform the additional analysis. If not, then maybe add it as a suggested extension in the discussion.

Authors: An article comparing these metrics across CMIP6 models would certainly be useful, but it is clearly material enough for an article in its own right. We don't think such an analysis really belongs here. We have added a line on the necessity of more models and scenarios in the conclusions. (Lines 470-473)

As I mentioned, several modelling centres have contributed daily surface chlorophyll and phytoplankton carbon to CMIP6, but no one has contributed daily NPP. Do you not want to make the case to include daily NPP (intpp) as a standard variable in future CMIP experiments? What do we gain from including NPP that we don't get from chlorophyll and carbon?
Authors: We have added a line on this in the Conclusions. (Lines 466-469)

At the end of the discussion section, I found myself asking several follow up questions like these. I have listed these below with the label "L394".

Typesetting & style comments

There is a tendency for sentences to be too long and complex, which makes them harder to read and parse - particularly in the abstract. While they may be accurate, they take more effort to understand. For this reason, I personally have a strong preference for simpler shorter sentences. I have pointed out a couple in the abstract and made some suggestions on ways to shorten and split them.
However, I'll leave it up to you from that point.
Authors:We have shortened long sentences.

Please try to be consistent with hyphenation and capitalisation. Change-point, time-series, cross-correlation, North-west, PI-control are all written in several different ways throughout the text.
Authors: This has been corrected.

There are several places where the superscript is lost for both the degree symbol, units and centuries (especially in the title!). Please be more careful with subscript and superscripts.

Authors: This has been corrected.

There are a few places where the text is stretched: L157, L543, L691.

Authors: This has been addressed.

For references in the body of the text, there are a few places where the name in the text either doesn't match the reference, or a reference does not exist. Similarly, there is some variability in typography of the "et al.", sometimes the period is missing (L336) and sometimes there's no space before the year (L140).

Authors: We have corrected this.

In the reference section, there are a lot of inconsistencies:

- Several references with strange characters that need to be corrected. Ie L471, 474.
- Some DOI's are links in blue and some are not, ieL 462 vs L464.
- Most references have the author initials, but L475 includes author's first names.
- Some are missing DOI's ie L480.
- Some place the year at the start (L461) and some at the end (L466).
- Some times the journal name is in italics (L468) and sometimes it is not (L464).
- Some references include et al. (L494 & L574) but most do not. Some instead use "…" (L501, L556, L562 & L594)

To me, this suggests that the author is manually writing the bibliography – which just makes your life more difficult! If you are, I strongly recommend instead using some kind of reference management tool to keep track of the bibliography and ensure that references are done properly and consistently. (I use bibtex for latex).

Authors: We have revised the references.

Without being an expert on the North Atlantic, I find the 8 different region names to be confusing. I'm constantly having to refer to figure 1 to check what region is being discussed. It may be clearer to write where things are happening more descriptively than just relying on the four letter acronyms. For instance, in line 253: "The largest standard deviation is found in NASE and the lowest in NWCS" could be clearer as "The largest standard deviation is found in the southeast of our domain in NASE and the lowest is in the Northwest Atlantic Shelf (NWCS)." This is closer to how you have described figure 5 in lines 264-274, which is much clearer.

Authors: We have expanded the text in accordance with the reviewers suggestion.

In the figures, I think there's scope for some additional consistency which would make it easier to interpret. For instance, you can use the same regional colours from figure 1 again in figure 3. Instead of the blue/orange colour scheme in figures 4, 6 and 7, you could use the regional colours again, but have a lighter one for EC-Earth and a darker one for NorESM2 (for instance). Alternatively, you could keep the same two line colours, but change the regional labels to match figure 1. Similarly, I see no reason for figure 9 to be different to figures 6 and 7. Finally, and this is purely subjective, but I'm not crazy about the blue and orange colours – I think a more aesthetic colour pairing could be found (https://colorbrewer2.org/ is a good resource for something like this).

Authors: We have labelled the different plots with province names in the same color as in Fig. 1. We do think the blue/orange/green works though and will keep them. Different shadings would not work as we use shading to discern between one and two change-points in Figs. 6 and 7.

As a added suggestion, daily data looks great in video format – much better than monthly and annual data. Have you considered including a supplementary video of the daily climatological mean for the two models and the observations? This would be a great resource to show when presenting this work in person. Alternatively, it could contribute to a great video abstract.

Authors: Thanks for the suggestion. The timeframe of this review does not allow us to explore this option currently, but we will keep it in mind for future presentations of this work

Specific Points

Abstract:

L14: This is a long sentence which could be clearer: “The majority of the region displays the largest change point in the day of peak NPP occurring after the year 2000 indicating a shift towards earlier peak NPP with the most change occurring in the northern parts of the domain.”

Suggested change: “Most of the region has the largest change point in the day of peak NPP after the year 2000. This indicates a shift towards earlier peak NPP and the most change occurs in the northern parts of the domain.”

Authors: The sentence has been changed as suggested.

L18: long sentence: “Furthermore, the occurrence of the first day with MLD shallower than 40 m shows positive correlation with the occurrence of the day of peak NPP for most of the domain and, similar to the day of peak NPP, displays the largest changepoints occurring around or after the year 2000.”

Suggested change: “Furthermore, the occurrences of the first day with a MLD shallower than 40 m and the day of peak NPP are positively correlated over most of the domain. As was the case for the day of peak NPP, the largest changepoints occur around or after the year 2000.”

Authors: The sentence has been changed as suggested.

L20 and elsewhere: Is it **changepoint** or **change point** or **change-point**? Beaulieu (2012) uses change-point so maybe that is the best option?

Authors: It is now “change-point” throughout.

Introduction:

L26: Turnover time is not defined and never referenced again. Why is it important?

Authors: Turnover time has been removed from the text.

L27: “Almost equals”? What is the land NPP value?

Authors: There is a lot of uncertainty in these estimations. Field et al. (1998) calculates marine NPP to 48.5 and the terrestrial NPP to 56.4 Pg yr^{-1}. We have added this reference and changed to “similar in size”.

L27 – and elsewhere: I prefer Pg instead of Gt. Ton is not an SI unit and there are many definitions of tons/tonnes/imperial tons and it can be confusing for international readers. Also, /yr should be yr^{-1}

Authors: This has been changed.

L28: “constitutes” isn’t the right word. NPP is the act of fixing the carbon, while the phytoplankton themselves are the basis of the food chain.

Authors: The sentence has been changed.

L30-31: Add a reference for this.

Authors: Reference has been added.

L35: north -> North

Authors: This has been corrected.

L37: remove “here, ”, but also consider simplifying this sentence.

Authors: The sentence has been simplified.

L53: “The seasonal cycle of phytoplankton blooms has been explained with various theories” -> “Several mechanisms have been hypothesized to explain the seasonal cycle of phytoplankton blooms.” (Suggestion)

Authors: The sentence has been changed as suggested.

L53-59. There’s a bit of a confusion here about theories vs hypotheses. A theory is specifically a widely accepted and tested hypothesis (like gravity, evolution or similar). So by definition, these competing explanations can’t all be theories! Also, I don’t think that the critical depth hypothesis can be both a hypothesis and a theory. I recommend rephasing this paragraph so that these are called hypotheses, explanations or mechanisms, instead of theories.

Authors: We have changed to “hypothesis”.

L62: Please be more explicit with your definition of phenology. It’s the core of the paper and its in the title. It deserves a full definition.

Authors: We have added this to the introduction (Lines:64-67)

L76: “ a maximum temporal resolution of not more than 20 days is required.” -> “a temporal resolution of 20 days or less is required.”

Authors: This has been changed as suggested.

L79 “In this paper…”: long sentence

Authors: We have changed the sentence.

L88: “highlights at which leads and lags” is there a missing word here? Consider simplifying this sentence

Authors: A misplaced comma resulted in a strange sentence. We have also divided the sentence in two.

Methods

L91: Is the NPP integrated to the sea floor or to some other depth?

Authors: In EC-Earth3-CC NPP is integrated to the sea floor while in NorESM2-LM, NPP is integrated over the top 100 meters. We have added this information on lines 101-102.

L92: “100 years pi-control”- > “100 years of Pre-industrial Control (piControl)” Note that there several spellings of PI-control here, you should choose one.

Authors: This has been changed as suggested and to “piControl” throughout.

L92: “Kriegler et al., 2017” doesn’t exist in the references.

Authors: The reference has been added.

L92: Can you justify why only one scenario and why you chose SSP5-8.5?

Authors: The reason is mainly that this gives us a sort of upper boundary on potential shifts. It is also interesting to note that most of the change points, especially in EC-Earth3-CC are located before SSP5-8.5 warming deviates too much from lower SSPs, as stated on lines 382-385. The reason for using only one is mainly that these kinds of runs, where daily data is saved, require a lot of resources and SSP5-8.5 was therefore the only scenario run within the H2020 project COMFORT.

L94: Please check the submission guidelines for referencing sections. This should be: “The models are described in Sect. 2.1. Section 2.2 describes the observational data set and Sect. 2.3…”
https://www.biogeosciences.net/submission.html
Authors: This has been changed.

L96: “which is calculated as a simple max of NPP” -> “which is calculated as the annual maximum of the daily means NPP, in units of mgC m^{-2} d^{-1}.”

Authors: We have changed the sentence although since the focus is on the “day of peak NPP” we did not see the reason to include the NPP units.

L96: Do you calculate the regional mean first and then the annual maximum? Or do you calculate first the spatial distribution of the annual maximum and then the regional means?

Authors: First we calculate the day of max NPP in every grid point and then we average over the area of each province.

L100: comma after EC-Earth3-CC and NorESM2-LM.

Authors: Commas have been added.

L100 and L104: unit superscripts: cm2 should have a superscript cm^2, s^{-1} and kg m^{-3}.

Authors: This has been corrected.

L104: I don’t think that these two methods are compatible with each other. While I note that you never compare the two MLD datasets, do you expect the differences here to impact your conclusions? If not, please explain why.

Authors: Both methods are classical ways of finding the mixed layer depth. The turbulent mixing coefficient in NEMO is typically orders of magnitude larger in the surface mixed layer than in the stratified layers below. Both methods thus essentially find the depth where stratification starts. In a perfect world we would have used exactly the same criterion, but mixed layer depths are computed online at each timestep. An offline diagnostic using the same criterion, but not done at full temporal resolution would surely be worse. The two methods are thus different proxies for the same thing. Note that apart from differences in these criterion there are plenty of other differences that can affect the mixed layer depth, like for example the vertical resolution.

If you change the MLD criteria a little the MLDs obviously also change a little. However, the connection between MLD and other variables are robust to such changes. Note, for example, that we have tried many different thresholds for the MLD with similar results. Thus, there is no reason to expect neither the two methods to be none-compatible nor to expect the use of different criteria to affect the conclusions.

L110: This reference is authored by “Gurvan Madec and the NEMO team”, so perhaps should be an et al, or the NEMO collaboration or similar.

Authors: We added “et al.”

L114: please use colons: “PISCES is a mixed Monod-quota model simulating two different phytoplankton functional types: diatoms and nanophytoplankton, two size classes of zoo- plankton: micro and meso, and the nutrients: nitrate, ammonium, phosphate, iron and silicate.”

Authors: Colons have been added.

L116: Add a reference to Redfield.

Authors: Reference has been added.

L124: remove comma after EC-Earth3.

Authors: This has been corrected.

L125: North-west -> Northwest.

Authors: This has been corrected.
L124 and elsewhere: “ocean only” -> “ocean-only”
Authors: This has been corrected.

L124: “Skyllas et al. (2019) validated EC-Earth3, in an offline ocean only NEMO-PISCES version, for a north-south (29-63oN) transect in the North-west Atlantic using cruise data of temperature, salinity and nutrients and chlorophyll-a and found a good agreement with observations.”

Suggest re-writing this as: “Skyllas et al. (2019) showed a good agreement between EC-Earth3 and temperature, salinity and nutrients and chlorophyll-a observations in an offline ocean-only version of NEMO-PISCES, for a north-south (29-63°N) transect in the Northwest Atlantic.”

Authors: The sentence has been changed as suggested.

L138: 2o -> 2°

Authors: This has been corrected.

L140: Assman et al.(2010). This should be Assmann, and add a space after et al.

Authors: This has been corrected.

L141: replace “with one phytoplankton and one zooplankton compartment and” with “with one phytoplankton functional type, one zooplankton functional type and”

Authors: This has been changed as suggested.

L142: Is this nitrogen and phosphorus or nitrate and phosphate?

Authors: It should be nitrate. This has been corrected.

L158 and L168: move the weblink to a reference.

Authors: We have moved the weblink to Data availability.

L159 please define MODIS.

Authors: This has been done.

L166: “Division of the global ocean into biogeochemical provinces has been done in a number of references” –> “The division of the global ocean into biogeochemical provinces has been attempted several times” (you can probably find more recent attempts as well.)

Authors: The sentence has been changed as suggested.

L176: Replace “and” after (NADR) with a comma.

Authors: This has been done.

L184: C-> Carbon
Authors: This has been done.

L188: “Generally speaking” is colloquial.

Authors: This has been changed to; In general,

L191: “we directly pick the number of change points to find” -> “we directly pick the desired number of change points”

Authors: This has been changed as suggested.

L191: Remove “in fact,”… This should all be facts, lol.

Authors: This has been done.

L193-194: extra space between “to instead” and “that is”? L200:
remove “in the following”

Authors: This has been done.
L201: “all sorts” is colloquial.

Authors: This has been change to “all types”

L206: ruptures was previously capitalised: Ruptures.

Authors: This has been changed.

L207: You haven’t described the cross-correlation method yet.
Authors: We have used matlabs crosscorr routine. We have added a reference to this. As the cross correlation is a standard statistical method we see no reason to provide a deeper description here.

ResultsAuthors: We have used matlabs crosscorr routine. We have added a reference to this. As the cross correlation is a standard statistical method we see no reason to provide a deeper description here.

L211: “for March” -> “for the March”

Authors: This has been changed.

L213: “the internal variability of the climate system as modelled by the two ESMs is not in sync with that in reality or with each other.” -> “the internal variabilities of the two ESMs climate systems are not synchronised with nature or each other.”

Authors: This has been changed.

L222: Is the yellow blob of overestimated NPP in the Spring EC-Earth3 related to the Döscher et al (2022) result that the model has too much active convection in the Labrador Sea?

Authors: It is possible, although speculative. Convection in the Labrador sea does affect the AMOC and at least to some degree the Gulf Stream. However, the Gulf Stream is mostly a wind driven affair, a consequence of Sverdrup balance and lateral (Munk) or bottom (Stommel) friction. However, the blob could also have other causes, like large riverine nutrient fluxes or large lateral tracer spreading in the area. With the experiments and CAFE data we have at our disposal there is no way to prove such a connection. In fact, it is not even a given that NPP is overestimated in EC-Earth in this area during MAM. Perhaps the other two models underestimate the NPP, or perhaps NPP has considerable multidecadal variability.

L240: “In general, EC-Earth3-CC is closer to the CAFE data in size but NorESM2-LM is closer in timing.” Can you back this up with some kind of objective statement?

Authors: We have removed this statement and added some text on lines:247-252.

L246: New paragraphs is missing a blank line. L247 and elsewhere: Please fix the units.
Authors: This has been done.

L257-258 and elsewhere: yr-> year. Please use “year” in prose and use “yr” when it is a unit.
Authors: This has been changed.
L265 & 309: Fig -> Fig. (Maybe elsewhere too)
Authors: This has been corrected.
L296: What is going on in EC-Earth3-CC ARTC in figure 7? There’s such a wide range of behaviour there, it’s hard to see why the charge point is placed where it is. Is there any way to quantify the quality of the fit – because this looks like a poor fit!
Authors: Not sure why you think so, but a common misconception about change points is that many think of them as purely local attributes, while in fact they signify, in some sense, optimal segmentations of a time series. That is, they depend on the global properties of the time series.

Here it looks to us like the time before and after the change point have very different means, which seems to be what the algorithm captures. Note also that it is not so much about a fit as it is about how much less optimal the second most optimal changepoint is compared to the most optimal and so on. In time series with a lot of variability and large jumps it is often harder to find these optimal segmentations by eye.

L315: I don't fully understand the value of the cross correlation analysis. Surely it's obvious that the current MLD is going to have the most impact on the current year's NPP? Maybe there's more to it here, but I think if you must include this analysis in your final paper, it needs to be explained with a bit more precision, details and specificity for those of us that don't use it on a regular basis.
Authors: The basic idea with the lagged correlations is to test the assumption that it is obvious that the timing of MLD shoaling is the most important. We find this to not always be the case. Had it been a simple story of the current MLD controlling day of peak NPP, one would have seen a big peak at lag zero and nothing else. We find instead strong correlations at a large range of lag and lead times. This we interpret as a lurking variable. That is, instead of simply having MLD controlling NPP, we have more hidden variables controlling both MLD and NPP. This hidden "variable" is climate change that for example through temperature, salinity and sea ice changes affect both MLD and NPP and gives rise to covariance over much longer than yearly periods.

L324: "lurking variable" could use a bit more explanation.
Authors: We have changed the sentence in accordance to the reply above on lines 343, 401-403.

Discussions

L329: "the size of peak NPP was well captured by the ESMs". I'm not sure that this case has been made. You could just as easily argue from Figure 1 and Table 1 that neither model captures the observations very well. I'd like to see a robust statistical comparison of the model and data over the historical period. A graphical version like a pair of Taylor or Target diagrams or add the bias, deviation, and correlation of the annual time series to Table 1. (I repeat this in the general comments above).
Authors: We have revised this statement on lines:348-352. For the issue of deeper statistical comparison with CAFE we refer to the answer to the reviewers general comment above.

L336: north -> North
Authors: This has been corrected.

L345: Several modelling centres have contributed daily surface chlorophyll to CMIP6, but no one has contributed daily NPP. Do you want to make the case to include daily NPP (intpp) as a standard variable in future CMIP experiments?
Authors: As answered above: We have added a line on this in the conclusions (line: 467)

L351: Remove "A noteworthy observation is that" L352 and elsewhere: 21st should be 21st
Authors: This has been done.

L355: IPCC 2022 is not an appropriate reference here. It's 3000+ pages long! At the very least, please cite an individual chapter. Also, this citation is for WG2, and the information you cite is more likely to be in WG1. You should probably instead cite O'Neill 2016 https://gmd.copernicus.org/articles/9/3461/2016/ and Riahi 2017 https://www.sciencedirect.com/science/article/pii/S0959378016300681.
Authors: We have referenced the above instead.

L372: "NPP and its timing is, of course, both in the models and in reality dependent on many other factors in addition to the MLD. Some examples are light availability, nutrient concentrations and temperature." -> "In both models and in nature, NPP and its timing is dependent on many other factors beyond the MLD, include light availability, nutrient concentrations and temperature."
Authors: This has been changed as suggested.

L374: remove "it is clear that"
Authors: This has been done.

L379 & L387: earth -> Earth
Authors: This has been done.

L379-385: Is there any indication of a difference between PISCES's phytoplankton functional types in terms of how climate change will impact their bloom onset phenology? Either here, in the monthly data, or in literature?
Authors: We have not checked this as we do not have any daily data of the two functional types.We do expect that different biogeochemical models as well as different functional types will respond differently to the same physical forcing due to the differences in their response to temperature, mixed layer depth etc. In a recent article Kléparski et al. (2023) investigated the climate change response of different functional types in six CMIP6 ESMs. Their main result is that the seasonal duration as well as the biomass of the diatom bloom will decline and the biomass of the slower sinking dinoflagellates will increase with possible impacts to carbon sequestration in the deep ocean. We have added some text to the discussion (lines:423-426)

L391-393: While I more or less agree with this sentiment, I think it's a bit of a reach. You may need more to back up this statement. I'm not convinced that understanding bloom phenology will be better served by ESMs with only one PFT than an ocean-only model with multiple PFTs. This could be my personal bias talking – but I think there's definitely scope for both approaches.
Authors: Nowhere in that section is it suggested that ESMs with one PFT are superior to ocean only models with multiple PFTs. Nor do we advise against using either more or less comprehensive models. Rather, we think that a considerable range of different models are useful to tackle questions of the nature addressed here. Complexity usually comes at the cost of interpretability and computational time; what's optimal depends on one's means as well as analytical strengths and presumably several other things. However, keeping the ocean model constant we think few would argue against our assertion that coupled effects could be important, and that coupled simulations are more physically realistic (boundary conditions are more inline with those of the real atmosphere-ocean system) than uncoupled ones.

Additional discussions to consider:

L394: The abstract concludes with "This highlights the need for long term monitoring campaigns in the North Atlantic.", but this is never discussed or mentioned again. Please add some discussion around this idea.
Authors: We have added this in the discussion (lines:387-388).

L394: I'd also like to see a discussion around the consequences of the changes these models have projected. As you mention in L35-L39, "the north Atlantic is a region of particular importance for carbon sequestration in the deep ocean." How will the changing phenology impact Carbon sequestration and deep mixing? Which regions are the most important and how will they change? Is the drop in NPP likely to affect higher trophic levels? How does this interact with biodiversity and marine policy?
Authors: As we state in the introduction, changed seasonality may impact entire ecosystems through trophic level decoupling. This could have an impact on carbon sequestration through a reduced strength of the biological pump. We have added some text on this on lines:436-440.

L394: Several of the figures are not explicitly mentioned in the discussion. Please add links to the relevant figures where you discuss them, and make sure that all figures (except maybe fig 1) are mentioned in the discussion.
Authors: All figures are now referenced in the discussion.

L394: What are the limitations of change point analysis? What does it mean when we pass a change point? How should this be interpreted? Is it like a regime change?
Authors: The main limitations or perhaps better pitfall, we would say is that it is easy to overinterpret statistical measures as if they also by necessity have a physical or biogeochemical meaning. This is not exclusive to changepoints. In fact, by far the debate has been most lively about statistical significance and its interpretation, which in some fields (fortunately seldom our

own) is very often wrongly interpreted. All the same mistakes done with statistical significance can also be done with change points. The interpretation we find most valuable is that they give in some, well defined albeit arbitrary, sense an optimal segmentation of one's data. Change points can mark regime shifts, although we think of regime shifts as a more physical or biogeochemical than a statistical trait. However, how a regime shift is defined is purely a lexical semantics question, of little concern for the current study.

L394: Figures 2 and 3 shows that neither model is amazing at representing the historical behaviour. How much can we trust the projections of models that fail to capture historical observations? (I realise that we have no other tools available – but allow me to play devils advocate here!)
Authors: Since the earth system models are not synchronized with real world internal climate variability, they can not be expected to reproduce the historical patterns correctly. Furthermore, the models are not eddy permitting which will generate discrepancies. The indications the models give about future behaviour are therefore taken to be indications based on future climate change effects.

L394: Can you discuss the long-term projection of NPP in figure 4? In most regions (except NASE), it looks like both models project a rise in NPP? Can you compare these two models against the CMIP6 mean, for instance from https://www.frontiersin.org/articles/10.3389/fclim.2021.738224/full, which shows a decline in North Atlantic NPP in both CMIP5 and CMIP6 multi-model means.
Authors: We have added the total change averaged over the entire domain between the final 30 yrs of SSP5-8.5 and the first 30 yrs of historical in Tab. 2. We have also added some discussion on lines:356-366.

L394: Similarly, considering that SSP5-8.5 is the most extreme climate change scenario (where future fossil fuel emission grows to 5x current values!), it barely impacts the NPP in figure 4. Is it possible that these two modelled ecosystems are particularly insensitive to climate change? Are they suitable for this type of analysis or are more flexible BGC models neccesairy for projecting the impact of
climate change on the marine ecosystem?
Authors: As noted further down. The ECS of our two models pretty much span the likely range (66% percent of the probability range) assessed by the IPCC for the ECS. So in that sense, at least, they are not clear outliers in a physical modelling sense. Moreover, the assessment that the extreme climate scenario barely impacts NPP is subjective. That is, there is no good answer to the question of what the expected range of NPP changes for a scenario like SSP5-8.5 should be. Thus, there is no real baseline to which our projected changes can be seen as large or small. The paper referenced above also shows that there are large variations in NPP in the CMIP6 models. In light of this, we don't see why a different, more flexible, model type should be called for. That is not to say that models do need to be improved in many different ways. They certainly do, but the need of a different model type does not seem to follow from the presented arguments.

More importantly, the many questions of realism posed: similarities to observations, feasibility of SSP5-8.5 and differences between CMIP6 models, are in many ways good questions. However, they are not the main target questions of our research. The questions we pose about phenological changes under strong climate change in our models, we find to be interesting basic science regardless of the degree to which these model simulations will capture our future changes. That is, the value of our results is not so intimately tied to the likelihood of e.g. SSP5-8.5 coming to pass as one might think.

L394: I'd like to see some discussion about the suitability of the SSP5-8.5 scenario. It has extremely high fossil fuel emissions and subsequent warming, which is likely to move these change points earlier in the simulation than you would see in other scenarios.

Authors: We have added some text on this to the discussion (lines:384,392)

L394: NorESM2-LM has a particularly low (but feasible) Effective Climate Sensitivity or 2.54K

(https://gmd.copernicus.org/articles/13/6165/2020/), while EC-Earth3 has an ECS of 4.3K (https://gmd.copernicus.org/articles/13/3465/2020/). Does the difference in their sensitivity to carbon impact the overall conclusions? For instance, is it possible that the surface waters of EC-Earth3 will warm more than NorESM2-LM, and this may shift the locations of habitats in these models.
Authors: It is possible, but it is far from the only difference between these models. Regional climate, climate variability, Arctic amplification and differing phytoplankton growth model's are just a number of other things that might play a role. With the simulations we have at our disposal it is not possible to attribute differences specifically to ECS differences. Note also that EC-Earth rather than NorESM is the outlier here; the likely range given by AR6 is 2.5-4, with a best estimate 3 degrees warming for a doubling of CO2. Our two models span this likely range fairly closely, but NORESM is much closer to the central estimate than EC-Earth.

Conclusions:

L402: This is the first mention of the growing season. Please add a description of these around your figure 2 results in lines 220.
Authors: We have removed this statement here as it is not related to our main result. We did however add a line on this in the results (line:234).

L416-419: Please explain why this is important?
Authors: We have added this on lines:463-465.

Figure captions:

L655: Seasonal mean vertically integrated NPP-> “Vertically integrated seasonal mean NPP”
Authors: This has been changed as suggested.

L686: There are several issues with white space here. Including an extra space in (L1) and (L2), here and in the caption for figure 7.
Authors: This has been corrected.

L694: remove period from “.Figure 7”.
Authors: This has been corrected.

L698: No period at the end.
Authors: This has been corrected.

Figure 2:

- This colour bar should be made with a pointed end at the maximum value (use extend = ‘max’ in matplotlib), to indicate that the highest values shown are beyond the end of the scale.
- It’s a shame you don’t include DJF, as it looks like both models really struggle to capture the behaviour then in figure 3.

Authors: We have changed the colorbar so that it shows the full range. Unfortunately, data for DJF is not present in CAFE as the satellites do not see the study area in winter. We have also removed SON as was suggested by reviewer 1.

Figures 1, 3 and 9:

- It would be nice if the line colours matched the colours in the map (and the region labels in figure 4). See my comments earlier.

Authors: We have colored the province labels in accordance with Fig. 1. Fig. 3 has been separated into provinces instead as suggested by reviewer 1.

Figure 6 & 7:

- While I understand why all figures share the same y axis, perhaps this figure would be better served by each region showing it’s own bespoke range so that the change point is easier to see. If you

have to, you can move the region label above the axes.

- Please add ticks to the x axes of the top 6 panes.
- Move the legend outside the figure – ideally below the main figure.
- Can you also the solid and dashed lines to the legend.
- Replace the sidewards pointing triangle with a vertical pointing triangle.
- Is there any observational data for this that can be added? If you can't find anything, WOA has monthly MLD: https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18.pl?parameter=M

Authors: We don't believe observational estimates of MLD would add anything useful to the investigation made here. Similarities would at best be climatological, as the internal variability is not in phase between our history and that modeled by the ESMs. Moreover, the ability of our ESMs to model MLDs, although an interesting question in its own right, is rather far from the topic of this article.

In the past, I've seen change point analysis that included a trend line either side of the change point. Is there any reason why this was not included here?

Authors: We have given the subplots individual y-axes. We have added x-ticks, the legends have been moved outside the main figure, solid and dashed lines have been added to the legend and the triangles are now pointing upwards in accordance with the reviewers suggestions.

Regarding the trend line we find it much harder to argue for the plotting of a trend line on either side of the change point than against. Firstly, because the presence of a change point, an optimal segmentation of a time series, does not to our knowledge imply the presence of trends in the segments it separates. Secondly, because the kernel method identifies change points of many different kinds it would not be clear if one should look for a trend in mean, variance or something else.

Figure 8

- You don't need the thick black contours here, just the colour scale might be more readable. As it stands, the figure really emphasizes the regions where the first change point is before the year 2000.

Authors: We have removed the contours from Fig. 8.

Figure 9:

- Can you remake this plot with the same style as figure 6 and 7?
- Can you highlight the times when you're within the 95% confidence bands, perhaps by making the line or dots thicker, or making them thinner when you're outside the confidence bands?

Authors: We have added colored province labels to the plot and changed the colors. We have increased the thickness of the 95% confidence band.

Supplementary data

There's no readme to describe the contents of each file.
Authors:We have added Readmes

Supplementary Model

The directory structure is not straightforward to understand. The structure described in the Readme doesn't match up with the directories in the zip. Please make the leading directory names more explicit.
Authors: This is the officially released source code of NorESM2 that was used for CMIP6 simulations. In the README file there is a link to an extensive documentation web-page. We are sorry but it is not possible to change this release.