We thank the reviewers for their careful reading of the paper and their positive comments. We welcome the reviewer's suggestions and have considered each of these in the revised version of the manuscript. We have enumerated each comment/suggestion followed by our reply, and provide a corrected version of the manuscript with changed/added text in red and removed text struck out. The line numbers referred to in our reply correspond to the marked version of the paper.

Reply to Reviewer #1:

1. The ensemble generation method is not clear. How are the "separate" assimilation runs (ch.3, p.4, l.25ff) kept separated over time, when they are nudged to the same reanalysis fields?

   The second paragraph of section 3 "Forcing, initialization and ensemble generation" was rewritten to better explain the ensemble generation method and the spread in the assimilation runs. P5L23-26 of this section also addresses this question.

2. In my opinion, technical phrases could be used in an even more consistent manner, also in context what is used within the prediction community. I would like the authors to consider using only one phrase throughout the manuscript, including figure captions. For the initialized experiemnts, these phrases have been used: "historical decadal forecasts" - "hindcasts" - "retrospective forecasts" - "forecasts".
   For the uninitialized experiments, these phrases have been used: "historical" - "uninitialized" - "simulations". In particular the use of "simulations" for the unintialized experiments seems to be suboptimal.
   In context with the potentially predictable component, these phrases have been used: "noise-to-predictable variance ratio" - "noise-to-signal variance ratio"

   Following the reviewer's suggestion, we have made changes throughout the text to use the terms "hindcasts" when referring to retrospective forecasts initialized from observation-based climate states and "uninitialized simulations" for simulations that are not initialized from observation-based states. In a few cases the terms "historical" or "retrospective forecasts" are used to emphasize the historical period (e.g., P1L4, P1L7, P2L23, P3L17, P17L33). These are made clear from the context.

3. On the use of "predictability" or "prediction skill". From my perspective it is important to thoroughly keep apart "actual predictability" of the real world ((un)initialized experiment vs. observational product) and "potential predictablity" of the model world ((un)initialized experiment vs. own assimilation). For physical quantities, the authors

assess actual predictability, for primary productivity they assess potential predictability. I would like to ask the authors to state the "potential" when discussing primary productivity. However, I wished the authors could include maps of actual predictability for primary productivity (1997-present) as well.

We agree with the reviewer that this needs clarification. We have modified the text to clearly state that the correlation skill for primary productivity is computed relative to the assimilation runs, and therefore it provides a "potential" for actual skill (e.g., P3L15, P16L23, P19L12, caption to Fig. 19). We avoid however calling such a skill as "potential skill", since it is fundamentally different from our definition of "potential correlation skill" (Eq. 7), which is the correlation of the ensemble mean hindcats and the hindcasts ensemble members. "Potential predictability" is defined here within the "perfect model" framework (P7L12-14), and the "potentially predictable variance fraction" (Eq. 4) is defined as a measure of "potential" skill.

Regarding the inclusion of actual correlation skill maps for primary productivity, note that we have available observation-based data spanning the years 2000 to 2016 (for land) and 1997 to 2014 (for ocean). These sample sizes are suboptimal for a robust assessment of actual skill of decadal predictions. Moreover, the uncertainty associated to the land GPP datasets available make the assessment difficult (e.g., see Fig. 21b of the paper). Therefore, we prefer not to include such results in the paper. We nevertheless have computed the anomaly correlation coefficient for the data available and show it in Figs. 1-3 at the end of this document. The results correspond to linearly detrended data. The figures provide some evidence for local actual skill of primary productivity in CanESM5 decadal hindcats. The uncertainty in the GPP land products is evident for example over the Amazon, where the hindcast correlates with GOSIF in some grid cells but has very poor skill relative to MODIS in most of the region, and in eastern China where correlations based on the two products are generally opposite in sign.

4. Several times, the authors state that CanESM5 could have "interactive" carbon or a "carbon cycle", but for the experiments presented here, the interactivity is not used (ch.2, p.4, l.18). I am okay with having the possible "interactivity" mentioned in the beginning, but I would like the authors to thoroughly check that the actual non-interactivity is properly referred to whenever the results are discussed.

   We have modified or removed the statements alluding to an "interactive carbon cycle", except for the introduction (P1L19) where we specify that CanESM5 "has the capability" to incorporate an interactive carbon cycle. We emphasize that land and ocean CO_2 do not feed back on the simulated

physical climate (P4L18-20), and so carbon cycle variables are purely diagnostic. We avoid using terminology such as "prediction of the carbon cycle" and favoured expressions such as "prediction of carbon cycle variables".

5. Overall, there is a rather high content of abbreviations in the text. In particular, mathematical symbols are in parts heavily used, e.g. r_{XY}, q_{e_i}. This sometimes renders the text more difficult to read, especially when in-text equations are used. Nevertheless, the text remains understandable, but perhaps the authors could check if some of the in-text equations could be made obsolete.

   We appreciate that reading mathematical symbols and equations could be in some cases demanding. Following the reviewer's suggestion, we have kept the mathematical symbols that we believe are strictly necessary for clarity and accuracy, and removed the various in-text equations just below Eq. (2) and in the second paragraph of section 8 of the original paper. We also agree that most equations in section 4 and appendix A can be found in equivalent forms in previous publications. These publications are cited accordingly. We include those equations in the methods and appendix sections for completeness, so as to avoid the reader having to look for the relevant information elsewhere.

Specific comments: Together with the general comments, a bunch of specific comments, which the authors may or may not consider, can be found in-text in the uploaded pdf.

The reviewer's specific comments were very helpful. We made several changes in the text following the suggestions. The following are answers to the questions not discussed previously:

a) decadal = 2-10 years?

Right. As mentioned in P8L18-19, we exclude Year 1 to assess forecast ranges beyond seasonal lead times.

b) "several years", some systems only integrate 5 years, some also 20.

Changed

c) Why mentioning bias correction for forecasts here in the initialization paragraphs?

Deleted

d) "... mean square skill score" equation number ?

Done

e) What about using "PPVF" over "ppvf" as an abbreviation, it would greatly enhance the readability in the skill chapters.

Italics are now used to highlight this abbreviation.

f) This seems to be the “ratio of predictable components” of Eade et al. (2014). Please mention this ”name” in the text.

Done

g) In Fig.5d-f, what does “contribution” mean in comparison to Fig.5g-i?

This has been clarified in P10L10-12. Emphasis is made on the relationship between r_{XY_i} shown in Fig. 5g-i and $r_i = r_{XY_i}\sigma_{Y_i}/\sigma_Y$ shown in Fig. 5d-f.

h) “These variations and unrealistic trends are imprinted on CanESM5 assimilation runs as they are nudged toward ORAS5 temperature and salinity fields to initialize the hindcasts (section 3).” How does this influence the simulated AMOC in CanESM5 hindcasts? What is the authors stance on the importance of AMOC for inter-annual predictions in the North Atlantic? If the AMOC is wrongly initialized in CanESM5, how important would that be for North Atlantic predictions?

A complete answer would require a study of the representation of AMOC in CanESM5, which is out of the scope of the present paper. The importance of AMOC on decadal and possibly interannual predictions in the North Atlantic and elsewhere has been discussed previously e.g., Zhang et al. (2019) and the references therein. On seasonal time scales, Tietsche et al (2020) provide evidence that AMOC initialization can contribute to forecast skill. To briefly address the effect that wrongly initialized western subpolar North Atlantic (WSPNA) temperature and salinity fields might have on the representation of AMOC in CanESM5 hindcasts, we show in Fig. 4 of this reply the maximum of annual mean AMOC streamfunction at 26.5°N for CanESM5 assimilation runs and decadal hindcasts. As a reference, we also include the uninitialized simulations and observation-based values from the RAPID dataset (Moat et al, 2020). AMOC at this latitude is strongly influenced by conditions further north in the Atlantic. The steep decrease of AMOC in the assimilation runs and hindcasts during the late 90s and the resulting bias afterwards (Fig. 4) suggests a link with ORAS5 anomalous water mass and heat transport before the 2000s (P10L32-33). This behaviour is also consistent with the winter SST cooling of Year 2 hindcasts in the WSPNA region during the same period (Fig. 7a,e).

References:

Zhang et al. (2019), A review of the role of the Atlantic Meridional Overturning Circulation in Atlantic Multidecadal Variability and associated cli-

mate impacts. *Reviews of Geophysics*, 57, 316–375, https://doi.org/10.1029/2019RG000644

Tietsche et al., (2020), The importance of North Atlantic Ocean transports for seasonal forecasts *Climate Dynamics* 55:1995–2011, https://doi.org/10.1007/s00382-020-05364-6

Moat B.I. et al, (2020). Atlantic meridional overturning circulation observed by the RAPID-MOCHA-WBTS (RAPID-Meridional Overturning Circulation and Heatflux Array-Western Boundary Time Series) array at 26N from 2004 to 2018 (v2018.2). British Oceanographic Data Centre, National Oceanography Centre, NERC, UK. doi:10/d3z4**.**

i) "The poor skill in WSPNA and Labrador Sea can potentially impact predictions of surface climate". Is this addressed in the discussion? How would you solve this negative dependence on ORAS5?

This is not addressed in the discussion as it would require further analyses on the impact of WSPNA and Labrador Sea on surface climate in CanESM5, which is out of the scope of this paper. Data post-processing provides a means to mitigate the impact of biases on forecast skill due to erroneous initial states.

j) "The noise-to-signal variance ratio of forecast and simulation ensembles..." in Fig.15 this is called "noise-to-predictable variance ratio"

Changed accordingly

k) Figure 5. At first glance, it is hard to understand the difference between d-f and g-i.

This has been clarified in P10L10-12 where Figs. 5d-i are introduced.

l) "Cross-hatched regions indicate values significantly different from zero at the 90% confidence level." They seem to be identical in d/g, e/h, f/i. Is this expected?

This is consequence of the relationship between $r_i = r_{XY_i}\sigma_{Y_i}/\sigma_Y$ (Fig. 5d,e,f) and the correlation r_{XY_i} (Fig. 5g,h,i). The sign of r_i is determined by r_{XY_i}.

Reply to Reviewer #2:

1. Abstract L4: sounds like the hindcasts were started in 1961 and then run continuously until present.

   Corrected

2. P2 L31: "at the end" maybe be more specific?

   Corrected

3. P5 L5: “augmented” replace with more specific phrase, please.

   “augmented with” was changed to “merged with weekly”

4. P5 L16-18: I don’t understand what initialization through response means. Are the carbon cycle components running during the assimilation phase and the initial state for each hindcast corresponds to their states at that point in the assimilation? Please clarify in the manuscript.

   This has been clarified in the second to last paragraph of section 3 “Forcing, initialization and ensemble generation”.

5. P6 L9: Suddenly a subscript ”e” appears in some equations, what is that? Not defining this makes it hard to follow the rest of the derivations.

   The text below equation (2) was modified to introduce the subscript e.

6. P8 L9: “secular” is a strange timescale (is it an astronomy term?), do you mean centennial?

   The expression “secular” refers to the long-term non-periodic variation of time series. To avoid confusion, we changed “secular” to “centennial or longer”.

7. P9 L5: Why are you excluding the Arctic? The skill seems to come from initialization, but you’d think that it doesn’t vary much under the ice, so that’s a contradiction.

   The reviewer is correct that the potential predictability variance fraction in the Arctic mainly results from initialization, as seen in Fig. 3d-f. We have changed the text to reflect this. As in sectors of the Southern Ocean and in the WSPNA and Labrador Sea regions, the uninitialized simulations display positive trends in the Arctic (Fig. 4c), whereas ORAS5 does not (Fig. 4b). These negative trends in ORAS5 are imprinted in the hindcasts (Fig. 4d) during the initialization process, which affect the predictable signal. As suggested by the reviewer, there is not much internal variability in the Arctic, therefore a large portion of the total variance result from the predictable variance derived from these negative trends. As a consequence, the potentially predictable variance fraction in the Arctic is high (Fig. 3a-c), and is largely attributed to initialization (Fig. 3d-f).

8. P9 L7: What is the relevance of mentioning the strong linear trends? What is the impact of them?

   This has been clarified in P10L1-3. The negative linear trends, which are strong in sectors of the Southern Ocean, the WSPNA and the Labrador Sea,

result from the ocean reanalysis product used to initialize the forecasts. As explained in item 7 above, these trends drive the predictable variance attributed to initialization.

9. P9 L33: How can the errors be fully attributed to initialization and then in the same sentence also attributed to the response to external forcing?

   This is because initialization can affect the model response to external forcing. Note that the response to external forcing in the hindcasts is generally different to that in the uninitialized simulations (Eq. 1). We now emphasize this in P10L30-32. For the WSPNA and Labrador Sea regions, the erroneous trends in SST hindcasts are imprinted by the ocean reanalyses used for initialization. We show that these errors are "fully" attributed to initialization because the correlation skill satisfies $r_{XY} = r_i$ (see Fig. 5a-f). This is consequence of a mismatch in the forced responses of hindcasts and uninitialized simulations, since we found that $r_{YU} < 0$ which implies that $r_u = 0$ and so $r_{XY} = r_i$ (see Eqs. A16 and A17). In addition to the analysis provided in the text, we cite Sospedra-Alfonso and Boer (GRL 47, 2020) where this is discussed in detail.

10. P10 L35: "can potentially" - I feel you should have a little more certainty than this about the impacts of the Atlantic SST problem. Perhaps if sections 6 & 7 were swapped, you could discuss here what has been shown for skill over land.

    Although many studies such as those cited in section 6 have established a relationship between the North Atlantic SST and surface climate elsewhere, we prefer to be cautious as we have not fully established these relationships for CanESM5. We do expect this to be the case for CanESM5 and have therefore changed "can potentially" to "are likely to". Also, we prefer to keep the order of sections 6 and 7, as we would like to assess decadal prediction skill in the ocean first (section 6), which is expected to contribute to the decadal prediction skill on land (section 7).

11. P11 L32: Perhaps it is worth having a note here that there are several papers link the Sahel precip to the AMV and that this is discussed at the end of Section 8 in the paper.

    Following the reviewer's suggestion, we have made a reference in the text to the discussion and bibliography cited at the end section 8

12. P12 L13: I think the description of the volcanic experiments needs to have a few more details and be placed in the methodology section of the paper.

    A brief description of the volcanic experiments has been included in the last paragraph of section 3 "Forcing, initialization and ensemble generation".

13. P12 L16: “volcanic forcing seems to be” I think the evidence presented is only strong enough to say “volcanic forcing could be”. Initialisation seems to be quite important too!

    We agree with the reviewer and have changed the text to reflect the evidence presented here for the contribution of initialization and possibly volcanic forcing to precipitation skill.

14. P15 L14: “time pan”

    Corrected

15. P17 L11: “Strong warming” Where does the climate sensitivity of CanESM5 lie compared to CMIP5/6 estimates of the real world probable range?

    Meehl et al. (Sci. Adv. 2020) report that Earth system models participating in CMIP6 (CMIP5) have equilibrium climate sensitivity (ECS) ranging from 1.8 K to 5.6 K (2.1 K to 4.7 K), and transient climate response (TCR) ranging from 1.3 K to 3.0 K (1.1 K to 2.5 K). CanESM5 is towards the highest end with ECS = 5.6 K and TCR = 2.7 K. Historical warming trends in CanESM5 are also strong, as seen in Figs. 4c and 10c of the paper, and as reported in Figs. 25a and 26 by Swart et al (Geosci. Model Dev., 12, 4823–4873, 2019). For context, this information has been included in the first paragraph of section 7 “Predictability and skill of surface climate on land”.

16. Figure 1: This color scale looks problematic for color blind people (https://www.color-blindness.com/coblis-color-blindness-simulator/). Additionally, if the color scale for Fig 2 was used here, it would be easy to flick between the figures to see how much of the potential predictability has been realised (this applies to the other Figures showing the same thing for other variables).

    We followed the reviewer’s suggestion and changed Figs. 1, 3, 4, 9, 10, 12, 14 and 15. Note that the results are the same and only the appearance of the figures have changed.

17. Figure 10: (b) Year 1 and (d) Year 2 look very similar. Year 1 is not shown in Fig 9, so perhaps use (b) Year 2 and (d) Years 2-5? This would help with backing up the conclusions in the text too by making the trends later in the forecast clearer.

    We changed Fig. 10 following the reviewer’s suggestion.

18. Figure 14: Why is this years 2-4 and not years 2-5 like in other figures in the paper?

    We look at years 2-4 since the precipitation response to volcanic eruptions is expected to peak during this time window. This is noted in P13L17.

19. Figure 16: This is similar to Fig 15, would years 6-9 be more interesting?

We are not sure about this question. Figure 16 shows the dependence of correlation skill on ensemble size and Fig. 15 shows maps of noise-to-predictable variance ratio for hindcasts and uninitialized simulations. Perhaps the reviewer is referring to Figs. 16 and 17. If so, note that the impact of initialization is much reduced for Year 6-9 (Fig. 13 f), which would make the discussion less meaningful. As pointed out in P14L22-35, the impact of initialization for the Year 2 annual precipitation forecast in Central South West Asia is detected for $\gtrsim 15$ ensemble members (Fig. 16d), whereas $\gtrsim 35$ are required for Year 2-5 forecasts (Fig. 17d).

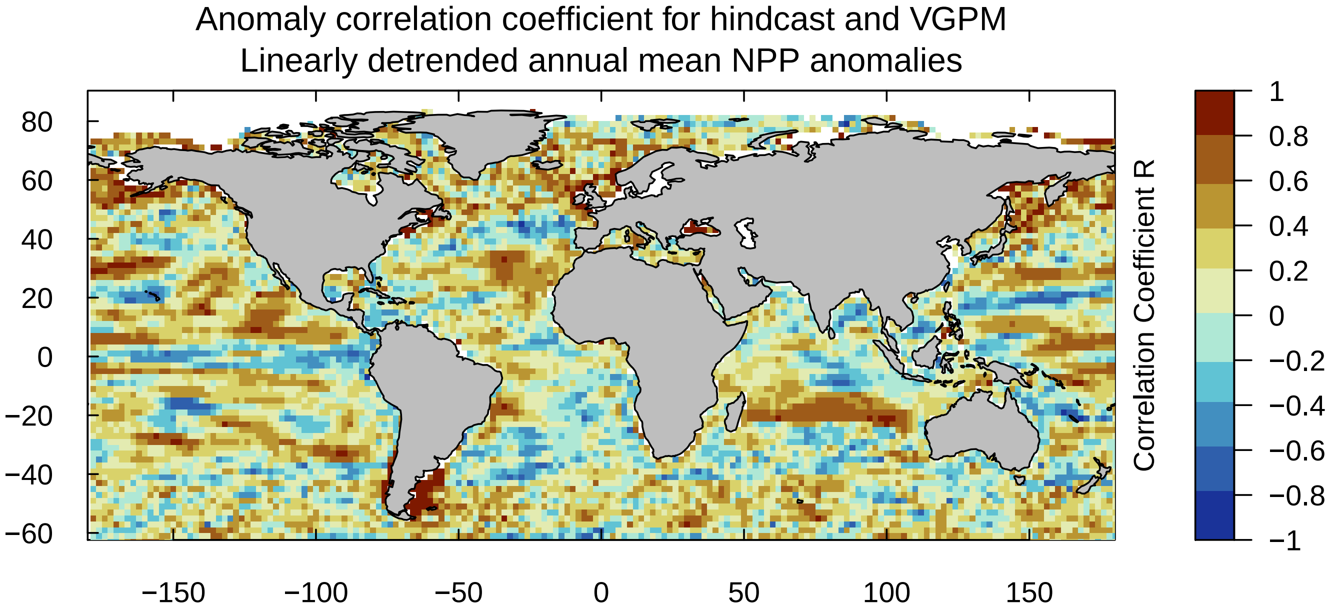
Anomaly correlation coefficient for hindcast and VGPM
Linearly detrended annual mean NPP anomalies
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150
Correlation Coefficient R
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 1: Anomaly correlation coefficient of annual mean Year 1 NPP hindcasts and VGPM (Table B2 of the paper) for 1997–2014. Anomalies are linearly detrended. This figure is linked to comment 3 of Reviewer # 1.

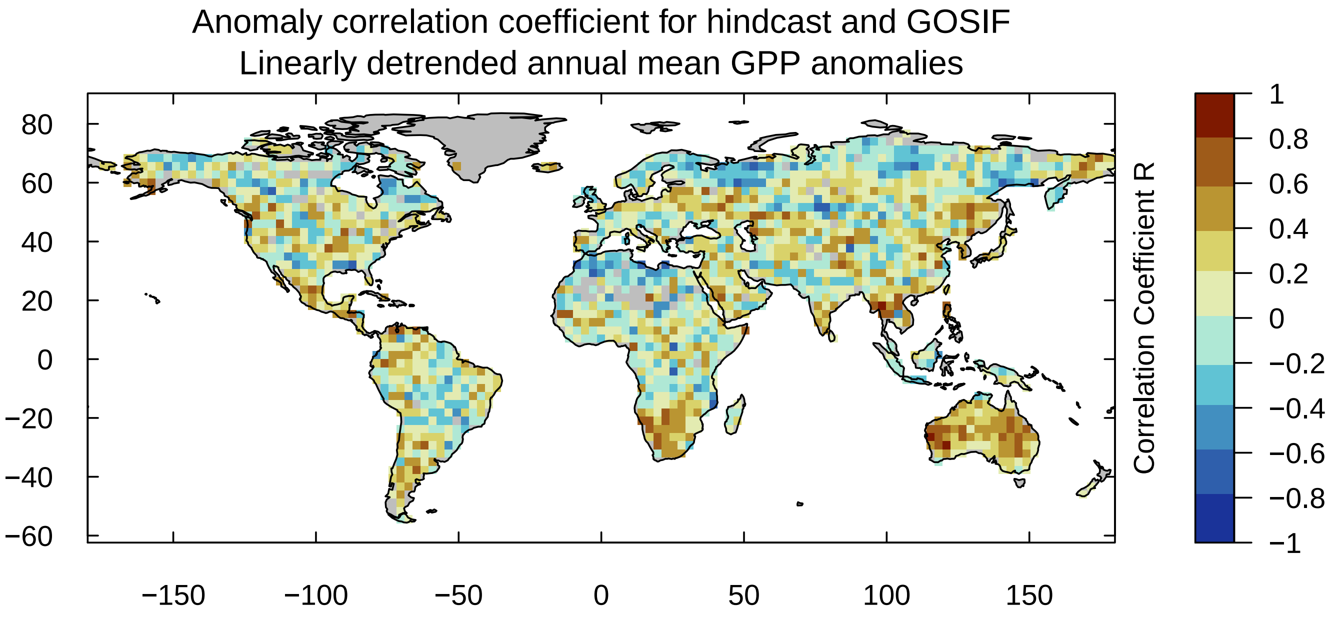
Anomaly correlation coefficient for hindcast and GOSIF
Linearly detrended annual mean GPP anomalies
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150
Correlation Coefficient R
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 2: Anomaly correlation coefficient of annual mean Year 1 GPP hindcasts and GOSIF (Table B2 of the paper) for 2000–2016. Anomalies are linearly detrended. This figure is linked to comment 3 of Reviewer # 1.

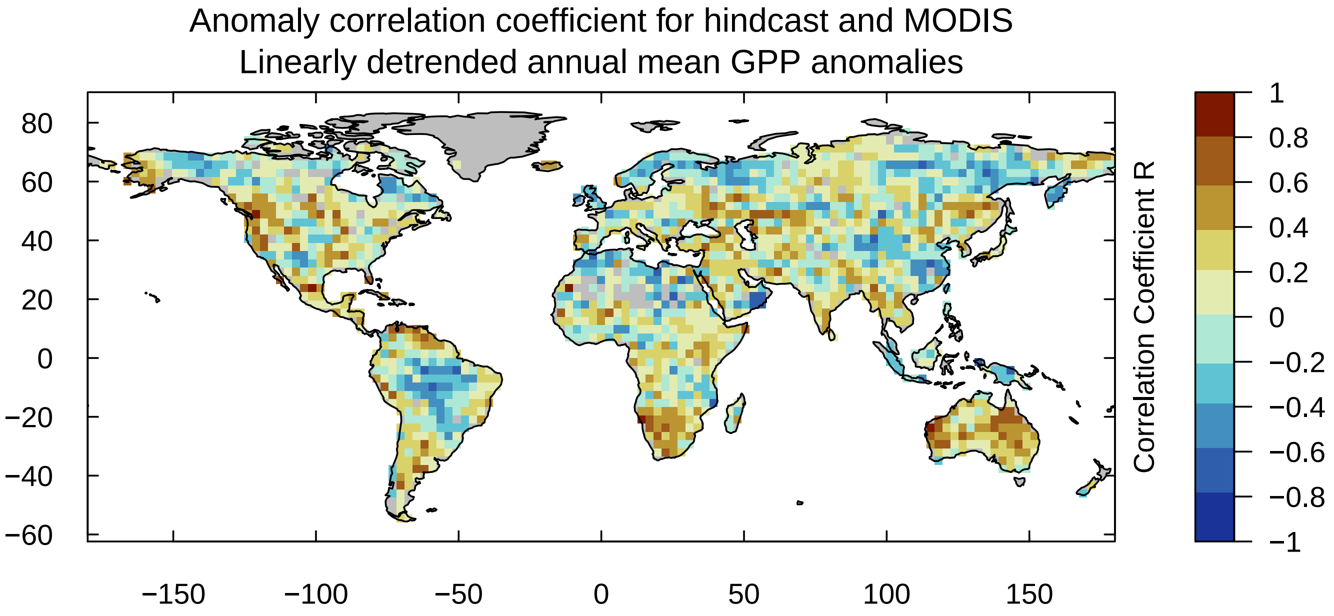
Anomaly correlation coefficient for hindcast and MODIS
Linearly detrended annual mean GPP anomalies
80
60
40
20
0
−20
−40
−60
−150
−100
−50
0
50
100
150
Correlation Coefficient R
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 3: Anomaly correlation coefficient of annual mean Year 1 GPP hindcasts and MODIS (Table B2 of the paper) for 2000–2016. Anomalies are linearly detrended. This figure is linked to comment 3 of Reviewer # 1.

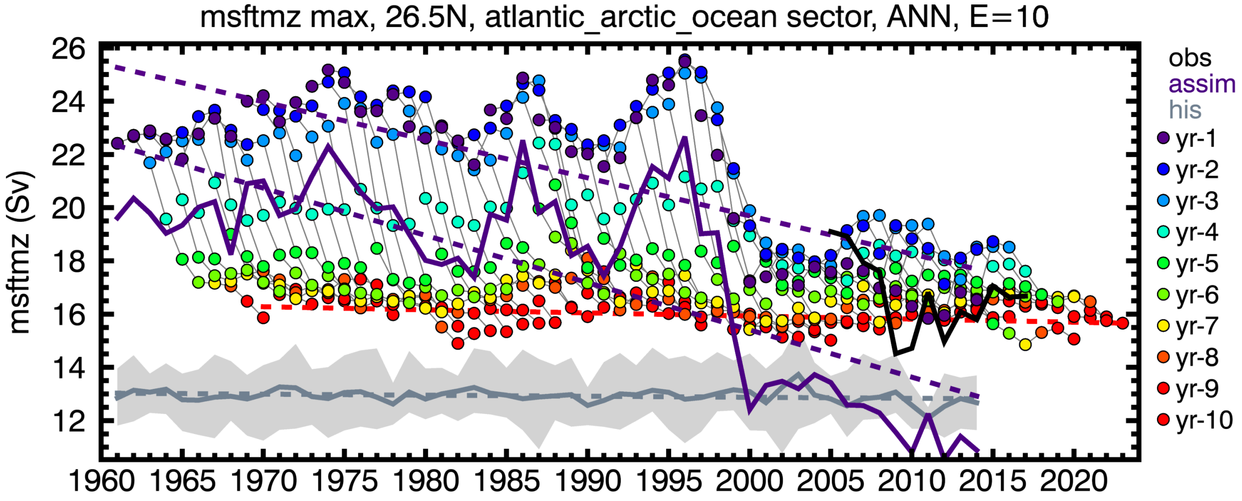

Figure 4: Maximum of annual AMOC streamfunction at 26.5°N from a 10-member ensemble mean CanESM5 (purple) assimilation runs, (colored dots) hindcasts and (gray) uninitialized simulations. Dashed lines represent linear trends. Black curve is from the RAPID observational dataset. Gray band represents the spread of the uninitialized ensemble. This figure is linked to comment (h) of Reviewer # 1.