Response to editorial comment

Dear Dr Di Capua,

Your paper has now been reviewed, and you will see that Reviewer 2 still has important concerns about the statistical analysis. Since this was the third round of review, I have reviewed the paper myself.

Based on my understanding of your methodology, I share the reviewer's concern that the comparison of the causal effect strength between ERA5 and SEAS5 (section 3.3) is not fully apples-to-apples; the selection procedure of the causal effect coefficients introduces a sampling bias, such that the ERA5 coefficients are expected to be higher than average. It is still possible that ERA5 does indeed exhibit stronger causal links than SEAS5, but a more careful analysis is needed to demonstrate that.

Reviewer 2 suggested two possible ways around the issue, which appear sensible to me. Please implement one (or both) in your revisions. Please also consider their additional comments, as well as my further minor comments below:

– Similar to the reviewer, I was unsure about the added value of Experiment B, discussed in section 3.4. Please make sure to more clearly motivate the usefulness of this approach, or alternatively include only Experiment A as suggested by the reviewer.

– I was unclear about the interpretation of Eq. 1. On the RHS, the superscript $k \neq l$ confused me – why is it necessary, and why on the RHS only? Is the l even needed – can't it just be k equals 1 or 2? Please add some clarification to the text.

Given that the paper has been reviewed three times, I will make a final decision about publication based on the revised paper and the reviewer's evaluation. Therefore, please make sure to address the reviewer's and my comments as carefully as possible.

Kind regards,

Paulo Ceppi

Dear Editor,

We sincerely thank you for taking the time to personally review our manuscript. We have revised the manuscript considering all editorial and reviewer #2 comments and we are confident that we could address all remaining concerns in a satisfactory way.

We now present the sensitivity analysis proposed by the reviewer and show that the results do not depend on the p-values chosen to calculate the beta values in SEAS5. Moreover, we have addressed all minor comments, further strengthening the clarity of the main message. A point-by-point response to the editorial comments can be found at the end of this document, while our point-by-point response to reviewer #2 comments is presented separately.

Finally, we would like to highlight that in our paper for the first time we apply causal maps based on the PCMCI algorithm to assess the skill of a general circulation model in forecasts mode to reproduce hemispheric tropical – extratropical teleconnection during boreal summer. This has implications for both the understanding of tropical – extratropical interactions, model development and weather patterns forecastability. Our analysis primarily shows that the model can capture the correct sign and direction of the links. This is a positive and encouraging finding and represents the main results of this work. Analyzing the difference in the magnitude of the links adds further information on where the model may have weaknesses in reproducing tropical – extratropical links, thus pointing to where potential teleconnections may be missing or too weak in the model.

We look forward to your decision,

Kind regards,

G. Di Capua

Potsdam, 24/04/2023

On behalf of all co-authors

Point-by-point response to editorial comments

(Our response in italic)

Based on my understanding of your methodology, I share the reviewer's concern that the comparison of the causal effect strength between ERA5 and SEAS5 (section 3.3) is not fully apples-to-apples; the selection procedure of the causal effect coefficients introduces a sampling bias, such that the ERA5 coefficients are expected to be higher than average. It is still possible that ERA5 does indeed exhibit stronger causal links than SEAS5, but a more careful analysis is needed to demonstrate that.

Reviewer 2 suggested two possible ways around the issue, which appear sensible to me. Please implement one (or both) in your revisions.

Please see our response to the main comment in the point-by-point response to reviewer #2.

Please also consider their additional comments, as well as my further minor comments below:

– Similar to the reviewer, I was unsure about the added value of Experiment B, discussed in section 3.4. Please make sure to more clearly motivate the usefulness of this approach, or alternatively include only Experiment A as suggested by the reviewer.

Please see our response to points 6 in the point-by-point response to reviewer #2.

– I was unclear about the interpretation of Eq. 1. On the RHS, the superscript k ≠ l confused me – why is it necessary, and why on the RHS only? Is the l even needed – can't it just be k equals 1 or 2? Please add some clarification to the text.

We thank the Editor for bringing our attention to this point. We have corrected the mistake, as it is not k≠l but i≠l. We further explain why it need to be so in lines 236-237 "Note that when conditioning on $MCA_{i\neq l}^{k}$, i needs to be different from l since when testing the influence of A on C(lat, lon) we want to remover the influence of B, thus if $MCA_{i}^{k} = MCA_{trop.OLR}^{1}$ then $MCA_{l}^{k} = MCA_{midlat.Z200}^{1}$.".

Response to Reviewer #2

We acknowledge the ample suggestions provided by reviewer #2 to strengthen the main findings of our manuscript and make it more convincing and useful to a broad scientific audience. In the additional round of revision, we have again attempted to convincingly address all suggestions made, and are confident that we have succeeded to provide as strong evidence for our previous claims as possible within the limits of a purely empirical study making exclusively use of the SEAS5 seasonal forecast ensembles.
Regarding the main issue highlighted by the reviewer, we now show that the reported underestimation effect of the beta values in the analyzed causal links indeed does not depend on the chosen p-value threshold. We have performed the corresponding sensitivity test as suggested by the reviewer, and our results demonstrate that the p-value does not play a role (see our answer to point the main comment). We therefore hope that the reviewer will consider our additional analyses and response as satisfactory, and thank the reviewer again for the time dedicated to our work.
A point-by-point response can be found below (in italic).

The authors have addressed some of the methodological issues that created confusion and potential bias in the results previously. Their results now seem to focus not on the relative magnitudes of the causal links they study in the reanalysis and seasonal forecasts but on what fraction of grid points have too weak links in the forecasts, which is simpler.
I'm still concerned that the methods have a bias that is unquantified and where it is difficult to understand how large the effect on the results is. This work is not publishable in my view without showing that the bias is not large enough to seriously affect the results. I have again focused on the method and main results and have not had time to read the rest of the manuscript in detail.
Main comments
The main result regarding comparing the reanalysis and forecasts now looks to be in sec. 3.3 and figs. 5,6, purporting to show that causal links in ERA5 are consistently towards the high end of those in the forecast ensemble. This works by selecting independent variables for performing the regressions based on those where the causal link magnitudes in ERA5 pass a statistical significance threshold. The same independent variables are used for ERA5 and the forecasts, alleviating the problem I mentioned last time that it is unclear how to compare regression coefficients when the independent variables are different. However, the method still seems like it will cause a selection bias where causal link coefficients in ERA5 that are large by chance will be selected more than those that are small by chance. In the forecasts, for the given independent variables, the random effects would not be biased high. So this effect will contribute to ERA5 having stronger causal links than the forecasts in this analysis.
Thoughts on ways to address this:
• One way would be, as I suggested last time, to look at the results of an equivalent analysis in figs.5,6 using an individual member of SEAS5 in place of ERA5 (defining the causal links to quantify based on that member) and verifying that the diagnosed causal links are not far from the 50th percentile of the rest of SEAS5.
• Another way (possibly better in that it also shows some sensitivity analysis that it would be a good idea to do) is to show how figs.5,6 appear if different p-value thresholds are used for selecting the causal links to be evaluated – including the case of using no threshold and examining all links, when there should be no bias from selection effects.

We thank the anonymous reviewer for their further suggestion.
We have adopted the second of the two suggested options since it was indicated as the most valuable by the reviewer.

We have run Experiment A for each p-value in the range P_range= {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0} for both MCA 1 and 2 and we have produced Figs. 5 and 6 for all the results. We have done so for both SEAS5 forecast data initialized on the 1st of May and on the 1st of March. The number of figures resulting

from this sensitivity exercise is 80, thus we have produced a figure which summarizes the results for SEAS5 initialized on the 1st of May. We also provide a selection of figures for p-values = {0.1, 0.5 and 1.0} in Appendix A at the end of this document, Figs RR3.A1-A8.

Figure RR3.1 shows the summary figure for the ten histograms obtained with the selection of p-values in P_range for each of the original histograms shown in Fig. 5b,d,f,h and 6b,d,f,h in the main manuscript. Please see the figure caption for a detailed description of each panel. It can be seen that changing the p-value does not change considerably (i.e. by at most a few percent) the percentage of grid points for which beta_ERA is found in each quantile category (Fig. RR3.1). However, this figure is obtained considering only the causal patterns shown in Figs. 3, 5 and 6, which present robust causal teleconnections also shown in using ERA-Interim data in this publication: Di Capua et al. (2020)
https://wcd.copernicus.org/articles/1/519/2020/
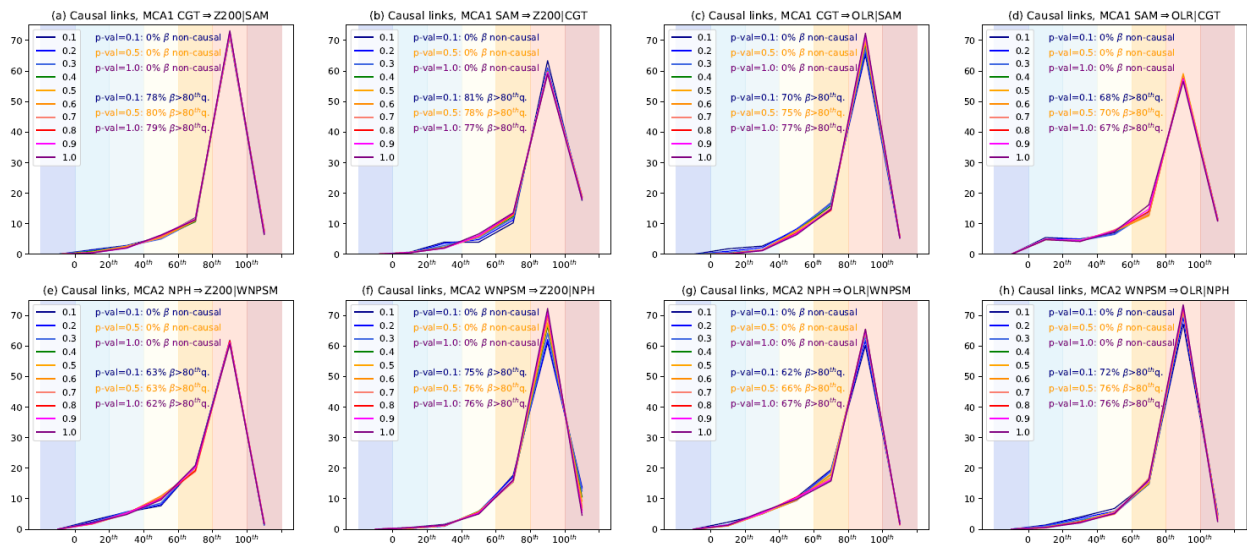
Figure RR3.1. Experiment A, sensitivity test. Panel (a): Histogram showing the percentage of grid points for which β_{ERA5} falls in a certain quantile range as obtained for β_{SEAS5} coefficients for link CGT→Z200|SAM. Panel (b): same as for Panel (a) but for link SAM→Z200|CGT. Panel (c): same as for Panel (a) but for link CGT→OLR|SAM. Panel (d): same as for Panel (a) but for link SAM→OLR|CGT. Panel (e): same as for Panel (a) but for link NPH→Z200|WNPSM. Panel (f): same as for Panel (a) but for link WNPSM→Z200|NPH. Panel (g): same as for Panel (a) but for link NPH→OLR|WNPSM. Panel (h): same as for Panel (a) but for link WNPSM→OLR|NPH. This figure is obtained taking into account only the grid points that show significant causal links in Fig. 4 in the main text. In each panel, the percentage of non-causal grid points analyzed and the percentage of grid points for which the beta values exceeds the 89th percentile is highlighted for p-values in the range {0.1, 0.5, 1.0}.

For the sake of completeness we also demonstrate that even using non-causal beta values, i.e. also those that do not represent causal links and are therefore not significant in Figs. 3,5 and 6, the underestimation by SEAS5, though less strong, still dominates all causal maps for all chosen p-values (Fig. RR3.2). We would like to highlight that using a p-value =1.0 (i.e. no threshold), up to 80% of the obtained beta values no not represent causal links. Nevertheless, a clear peak between the 80th and 100th percentiles is still clearly visible also for very high p-values and the number of beta values that fall above the 80th percentile (which should be 20% in case an underestimation effect of the beta_ERA5 is not present) is still found in between 24% and 42% even in the most extreme case (p-value=1.0). Thus, we can confidently show that the underestimation of beta_ERA by SEAS5 does not depend qualitatively on the chosen p-value.
We have reported this explanation and figures RR3.1-2 in the Supplementary material as Text T1 and Figs. S9 and S10. We refer to those in the main text in lines 425-426 *"Finally, we check that the underestimation effect does not depend on the chosen p-value threshold (see Supplementary Material, Text T1 and Figures S9 and S10)."*

If possible on the WCD system, we would be willing to upload a zip folder containing all additional figures from the described sensitivity experiment.
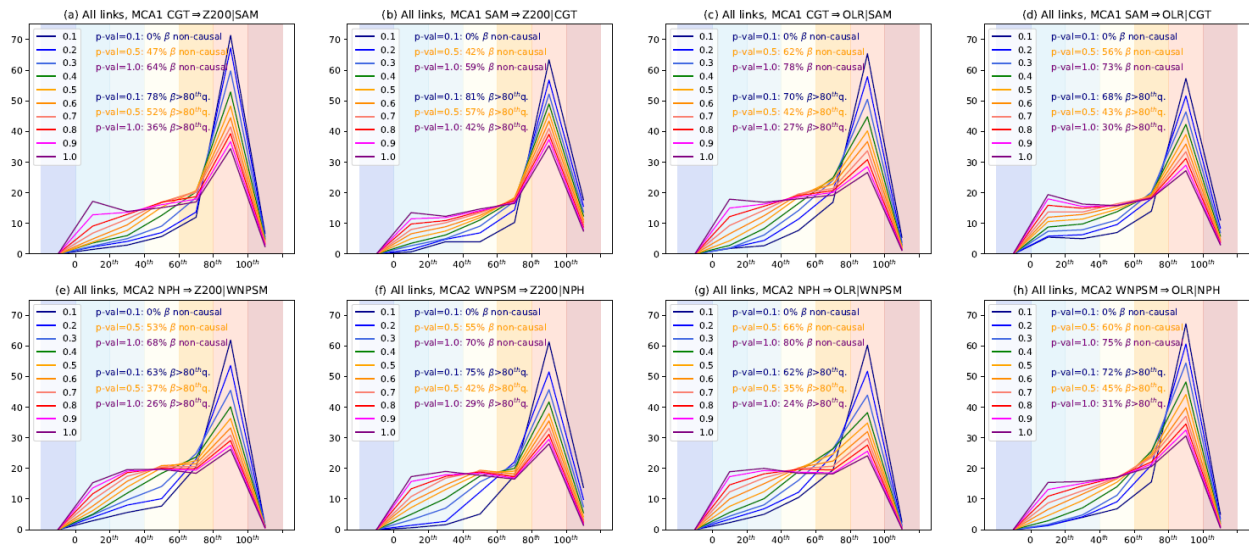
Figure RR3.2. Experiment A, sensitivity test. Same as for Fig. RR3.1 but obtained taking into account the beta values of all grid point, thus also those not showing any causal link.

Other comments
1. L48-9 The results haven't been shown in a way that allows fair quantitative comparison of the beta coefficients in the two models.

This point is addressed as part of the main comment above. We are confident that our two subsampling experiments along with the additional sensitivity analysis can be considered a fair comparison to the maximal possible degree achievable from the exploitation of the existing SEAS5 dataset.

2. L250-2 It's confusing to have multiple different methods used in different parts of the paper. I would pick one to focus on, and only use others if necessary to make a particular point, which should be made clear.

We follow the suggestion of the reviewer and implement only one significance method throughout the entire manuscript (as described in lines 245-248). Affected by this change are the former Figs. 3-4 in the main text and S1, S2 and S4 in the Supplementary Material. The updated figures can be found in the revised version of the manuscript as Figs. 3-4 and Figs. S1-S2 and S5 in the Supplementary Material.

3. L327 It's good to see the MCA-1 results looking consistent. I think the same should be shown for MCA-2.

The same figure as for Fig. S3 but for MCA2 is now available in the revised version of the Supplementary Material as Fig. S4.

4. L371-4 I'd delete this part and just say the strengths of the links can't be compared when using different-length datasets given the use of a statistical significance threshold. Else it's confusing. I think the point of this section is to say SEAS5 produces causal link coefficients with a similar spatial structure? If so, this could do with being made clearer.

We have removed the former lines 371-374 *as suggested by the reviewer. We further highlight that the spatial structure and sign of the links are the main points of this Section and first key result of the paper in the revised version of the manuscript in lines* 366-375 *"In general, the sign and the geographical position of the causal links detected in SEAS5 are consistent with those found in ERA5, meaning that the effect of each MCA mode on the analyzed fields is consistent in sign and spatial location between the two datasets. For example, the link $SAM_{\tau=-1} \rightarrow Z200_{\tau=0} \mid CGT_{\tau=-1}$ has a positive causal link on the Sahel region both in ERA5 and SEAS5 (Fig. 3e and 4e). Thus, the first key result obtained in this Section is that the main tropical – extratropical intraseasonal causal relationships in boreal summer in the Northern Hemisphere are at least qualitatively well represented in the SEAS5 system. These causal maps also show that the two-way causal pathway between tropical convective activity and extratropical circulation is captured by the seasonal forecasts. Thus, on the one hand we gain confidence in the interpretation of the earlier ERA-Interim and ERA-S/L causal map analysis, which is reproduced by SEAS5, and on the other hand we show that, to a first approximation, seasonal forecasts can reproduce such causal links."*

5. L411-3 "As for MCA1…" – I don't see these points made for MCA1 before. It also doesn't seem clear to me that there is a big difference between coefficients for the tropics and mid-latitudes in figs.5-6.

We have corrected this sentence in the revised manuscript, the new statement can be found in lines 415-418*: "Both for MCA1 and MCA2, the underestimation of β_{SEAS5} values is more pronounced in tropical regions, where values exceed the 100th percentile more often than in the mid-latitudes, where β_{ERA5} values are commonly in between the 60th and the 100th percentile of their β_{SEAS5} counterparts but do not exceed the latter.".*

6. Sec.3.4 I don't understand the motivation for using a different method of estimating the causal coefficients in this part – why not just use the experiment A samples? I can see it might be interesting to compare results when using the experiment B method, but I'd suggest computing the results for both experiments in this part and then this allows the comparison (perhaps with results for the second method as supplementary info). Currently it's hard to tell what effect changing the method has made and therefore how to consider the results from each experiment.

Experiment B represents a crucial step of this analysis as we want to assess whether the direction sign and spatial patterns of the causal links shown in ERA5 are detected in SEAS5 as well. We clarify why we use these two experiments in lines 261-272*. To further clarify this point we have added lines* 265-268 *"However, we cannot a priori assume that all the causal links detected in ERA5 will be reproduced by the SEAS5 forecast with the same sign, direction and strength (which would be the equivalent to assume that the model does a perfect job in reproducing all observed teleconnections, while instead well-known biases between the model and the reanalysis products are observed; e.g. Johnson et al., 2019).".*
Following the reviewer's suggestion, we now provide Figs. 7 and 8 also for Experiment A. These new figures are found in the revised version of the Supplementary Material as Figs. S11 and S12. They are described in the revised version of the main text in lines 481-490*: "*

It should be noted again that the β values obtained in Experiment B do not refer to the same set of causal links as shown in Figs. 3, 5 and 6. Thus, we provide Figs. 7 and 8 also using the causal maps obtained in Experiment A (Figs. S11 and S14 in the Supplementary Material). In general, β values obtained from Experiment A for the analysed regions show a good agreement for MCA mode 1, where like in Fig. 7, the link $SAM \rightarrow Sahel\ Z200 \mid CGT$ shows the strongest bias with β_{ERA5} falling outside the β_{SEAS5} distribution (above the 100th percentile), while β values for SE-Asia Z200, India OLR and Mediterranean OLR fall below the 90th percentile (Fig. S11). For MCA mode 2, all β_{ERA5} fall between the 90th and the 100th percentiles. Thus, the underestimation effect, which in Experiment B is limited to the NPH $\rightarrow$ NW-US Z200 $\mid$ WNPSM and the WNPSM $\rightarrow$ M.Cont. OLR $\mid$ NPH causal maps (Figs. 8d,f), affects also causal links WNPSM $\rightarrow$ Japan

Z200 | NPH and NPH → CE-EU OLR | WNPSM in Experiment A (Figs. S14 in the Supplementary Material). These results further support that, while the spatial pattern and sign of the causal links is fairly well reproduced both in SEAS5, the underestimation of the strength of the β values is found in both Experiments A and B.

*   "*

*   For completeness, in Appendix A we also show a sample of Fig. 7 and 8 obtained with the sensitivity test described in our response to the main comment in this document for p-values in the range {0.1, 0.5, 1.0} (see figures RR3.A9-A12 in Appendix A at the end of this document. From these figures we would like to highlight that not only all results are consistent with Figs. 7,8, S11 and S12, but that in general increasing the p-values also increases the bias between beta_ERA and the beta_SEAS5 distribution.*

7. It should also be made clear here that in each regression the independent variables will often differ between the ERA and SEAS5 analyses, which will generally affect their meaning.

We clarify this point in lines ==481-482== *(see point 6 in this document).*

8. L434 "We identify these regions based on…(ii) the misrepresentation of the strength of the β values in Figs. 5 and 6" – for 4/8 of the chosen regions, the ERA coefficient looks very close to the centre of the distributions of SEAS5 coefficients, so there doesn't appear to be particular misrepresentation in those.

We have clarified this point in lines ==441-442==: *"We identify these regions based on (i) the prominence of the signal in Figs. 3 and 4 and/or (ii) the misrepresentation of the strength of the β values in Figs. 5 and 6."*

9. The position of the ERA coefficient relative to the SEAS5 pdfs in figs. 7,8 also seem to have changed quite substantially since the last submission (e.g. SAM -> OLR in India, fig.7e,f) and it's not clear to me why.

The shape of the probability density functions (PDFs) shown in Figs 7-8 in the second version of the revised manuscript is exactly the same as that for the PDFs shown in the first version of the revised manuscript. The shift along the x-axis is caused by the fact that we no longer show the PDF of the standardized beta values, but we now show the absolute value of the beta coefficients. This was done following the suggestion of the reviewer (as explained in point 6 of the previous point-by-point response). Using standardized beta values centered around zeros implies by construction that 99% of the values will be limited between +/-3 s.d. and all PDFs will be centered around the same position on the axis. On the contrary, using the absolute value of the beta coefficients, which are found between 0 and ~0.5, and whose mean values depend on the analyzed region, produced PDFs that are not centered around zero anymore, more around the same position on the axis. As a consequence the PDFs get compressed, however, their shapes do not change. To better appreciate these similarities, we provide the PDFs from the second (equal to the current) and first revision side by side for both MCA modes in Figs. RR3.3. and RR3.4.
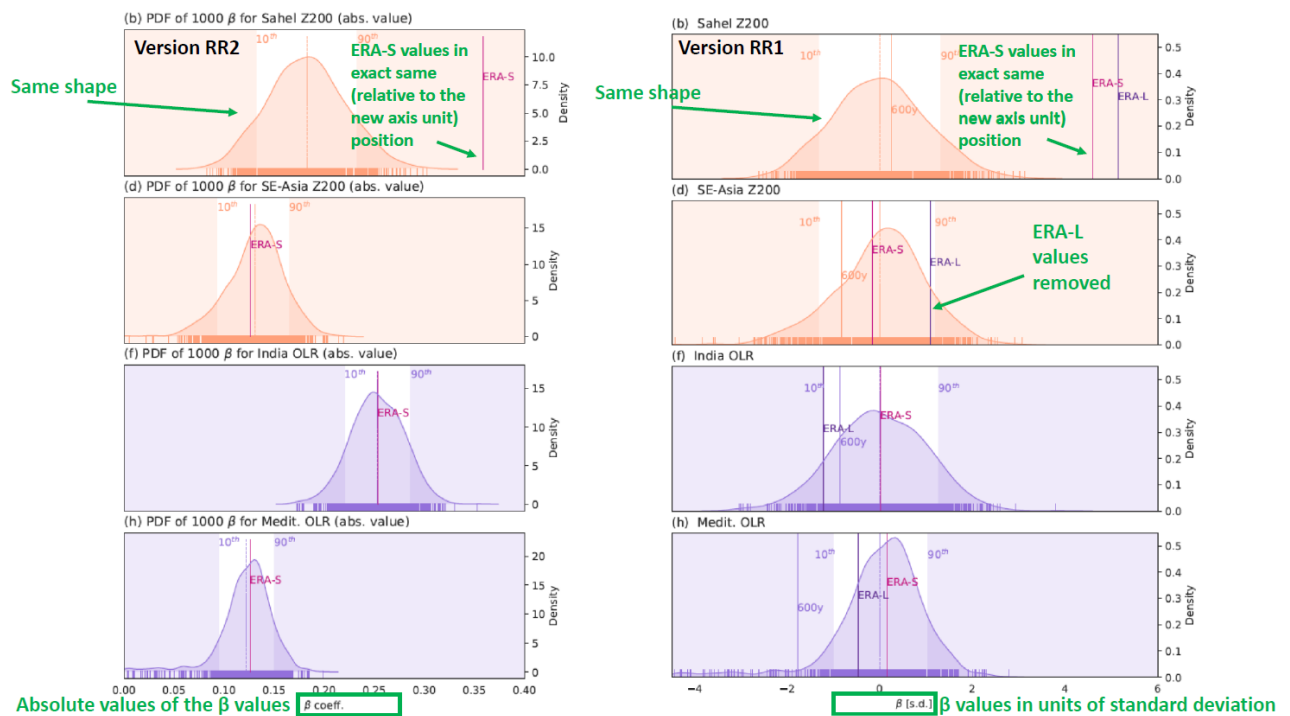
Fig. RR3.3 Left column: PDFs as shown in Fig. 7 in the latest and current round of revision. Right column: PDFs as shown in Fig. 7 in the first round of revision.
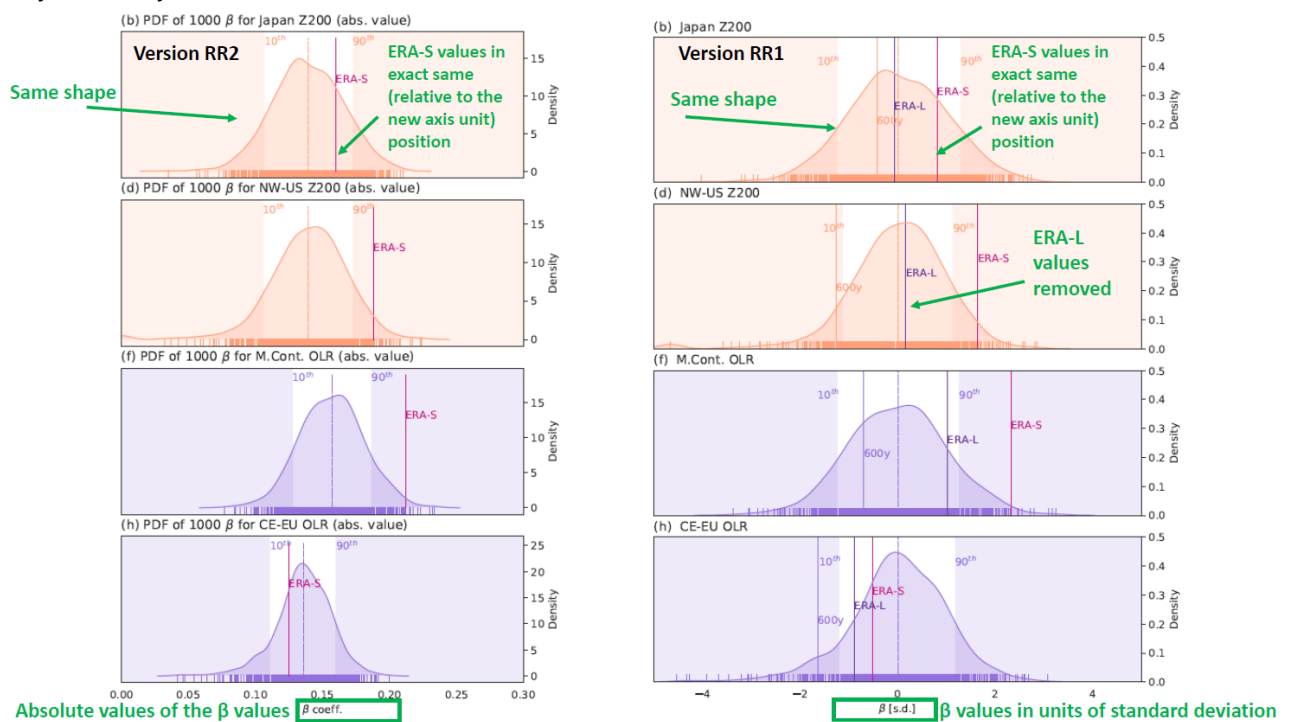
Fig. RR3.4. Left column: PDFs as shown in Fig. 8 in the latest and current round of revision. Right column: PDFs as shown in Fig. 8 in the first round of revision.

10. L618-9 How are the results relevant for assessing meaningfulness of the patterns coming from PCMCI? The results just show commonalities between ERA and for SEAS5 - but this would be true for any analysis if the simulations are decent, regardless of whether the results are "meaningful".

The term meaningful here refers to the ability of PCMCI to identify teleconnection patterns both in ERA5 and SEAS5, as explained in lines 589-593 in the discussion section: "Despite consistent underestimation of causal link strength in certain regions (Figs. 5 and 6), these results imply the ability of the SEAS5 forecast

system to reproduce the sign and the spatial distribution of the observed causal patterns for boreal summer intraseasonal variability in the Northern Hemisphere (Figs. 4-8). Although this analysis does neither rely on nor imply a skilful forecast, the causal effect of tropical and mid-latitude patterns on circulation and convection in the Northern Hemisphere in SEAS5 seasonal forecasts is qualitatively well comparable with that shown in ERA5 reanalyses.

11. L618 As before, I don't see clearly larger biases in the tropics.

We have modified this sentence which now reads "… this negative bias is actually contained in the spread of the SEAS5 seasonal forecasts, and that the bias is stronger in tropical regions" (lines 640-641). Also see our answer in point 5 in this document.

12. L618-20 "our confidence in…the ability of the SEAS5 forecasting system to correctly represent those causal links is increased" – but isn't the main claim that SEAS5 underestimates the strength of the links?

*The main point of this manuscript is dual, as expressed in the Discussion section in lines 540-563. From these two paragraphs, we would like to highlight that the "correct representation" of causal links in SEAS5 when compared to ERA5 refers to their **qualitative** characteristics, e.g. the sign and direction of causality, as expressed in lines 543-546: "In general, causal maps obtained with SEAS5 correctly reproduce the sign and the spatial patterns of ERA5 causal maps, though with weaker magnitudes (Fig. 4). Thus, spatial patterns shown in SEAS5 seasonal forecast causal maps are validated by those extracted from ERA5: since the SEAS5 forecasting system can qualitatively reproduce the patterns seen in ERA5 reanalysis, we gain confidence that observed causal maps represent actual physical mechanisms". Only afterwards we analyze the differences in the strength of the beta values, and we highlight that those are generally weaker in SEAS5 when compared to ERA5 both when imposing the same causal links detected in ERA5 and when letting the PCMCI algorithm free to detect causal links in SEAS5 without any a priori assumption. From the second paragraph, we would like to highlight the following passages: "We have imposed the same set of causal links as observed in ERA-S (Fig. 3) and calculated the causal effect in the 1000 subsampled SEAS5 data (Experiment A), showing that in general ~70% of the grid points show a β_{ERA5} value above the 80th percentile of the β_{SEAS5} distribution." (lines 552-554); "Then, we ran the 1000 subsampling experiment a second time but leaving it to the PCMCI algorithm to identify the causal links characteristic of SEAS5 without further constraint (Experiment B) and identified eight key regions for which we compared the observed ERA5 causal link strength with the range of SEAS5 values obtained from the subsampling ensemble (Figs. 7 and 8)."(lines 555-557-) and "Thus, SEAS5 has difficulty generating high values of the teleconnection strength especially over North Africa, North America and the Maritime Continent (when the ERA5 reference values exceed the SEAS5 90th percentile). In the other analysed regions, we have shown that for a correct estimation of the strength of the causal links, using time series of the same length is crucial to avoid underestimation effects due to the length of the time series." (lines 560-563).*

13. Fig.7 caption needs to say the values shown in the distribution are absolute values. The fig.8 caption can refer back to this one without repeating lots of the same information.

We now specify in the caption of Fig. 7 that the values in the PDF are absolute beta vales. We have reduced the length of the caption for Fig. 8 as suggested.

APPENDIX A
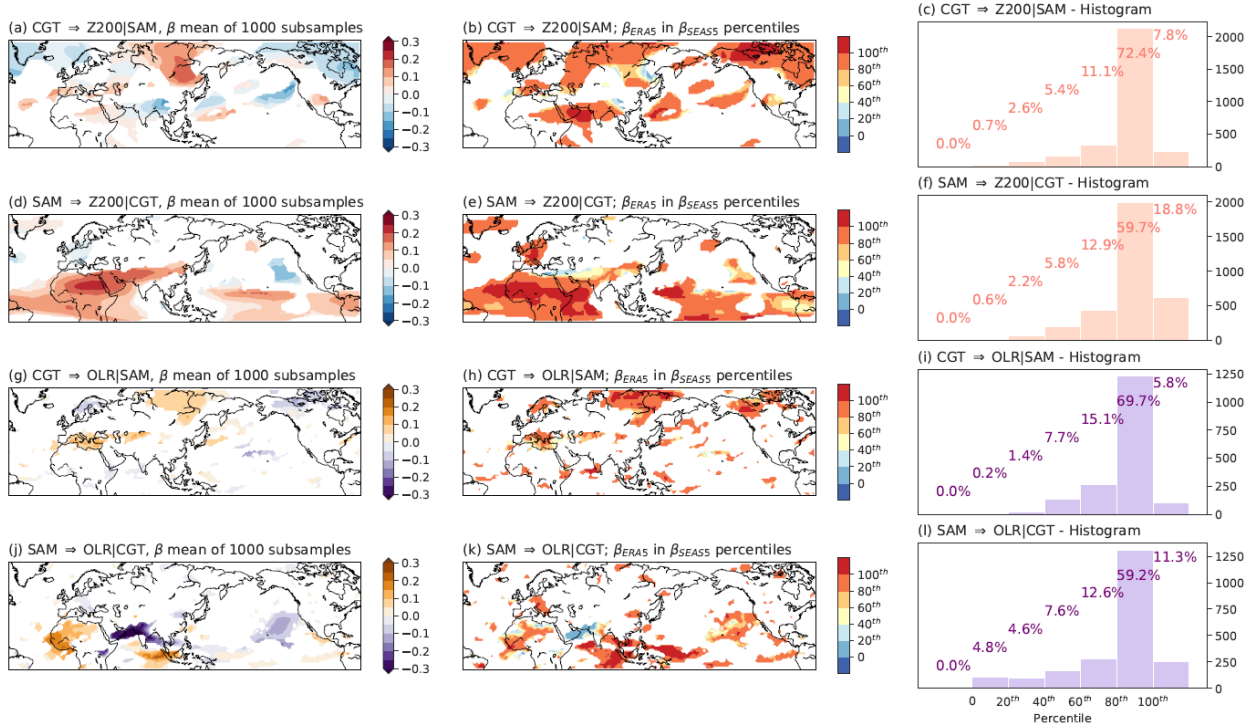
Figure RR3.A1. Same as for Fig. 5 in the main text but for a p-value = 0.5 and showing only grid points where causal links are present.
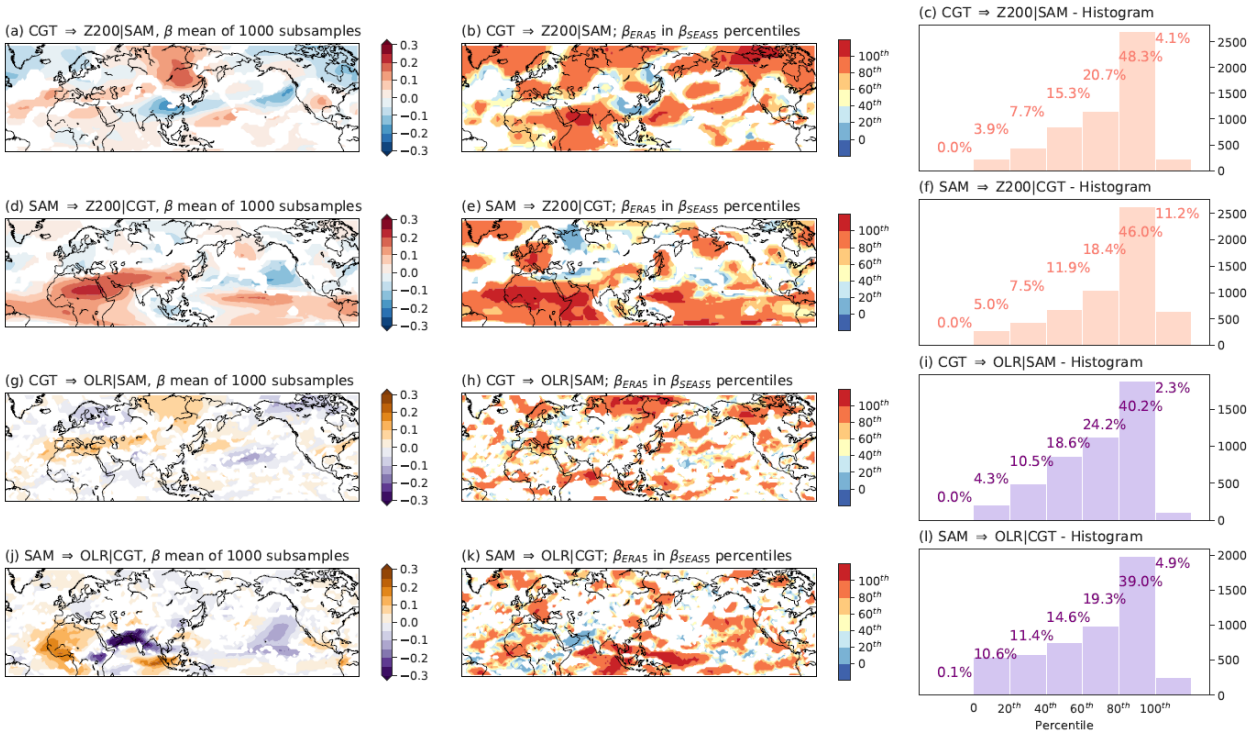
Figure RR3.A2. Same as for Fig. 5 in the main text but for a p-value = 0.5 and showing all grid points (also those where no causal links are present).
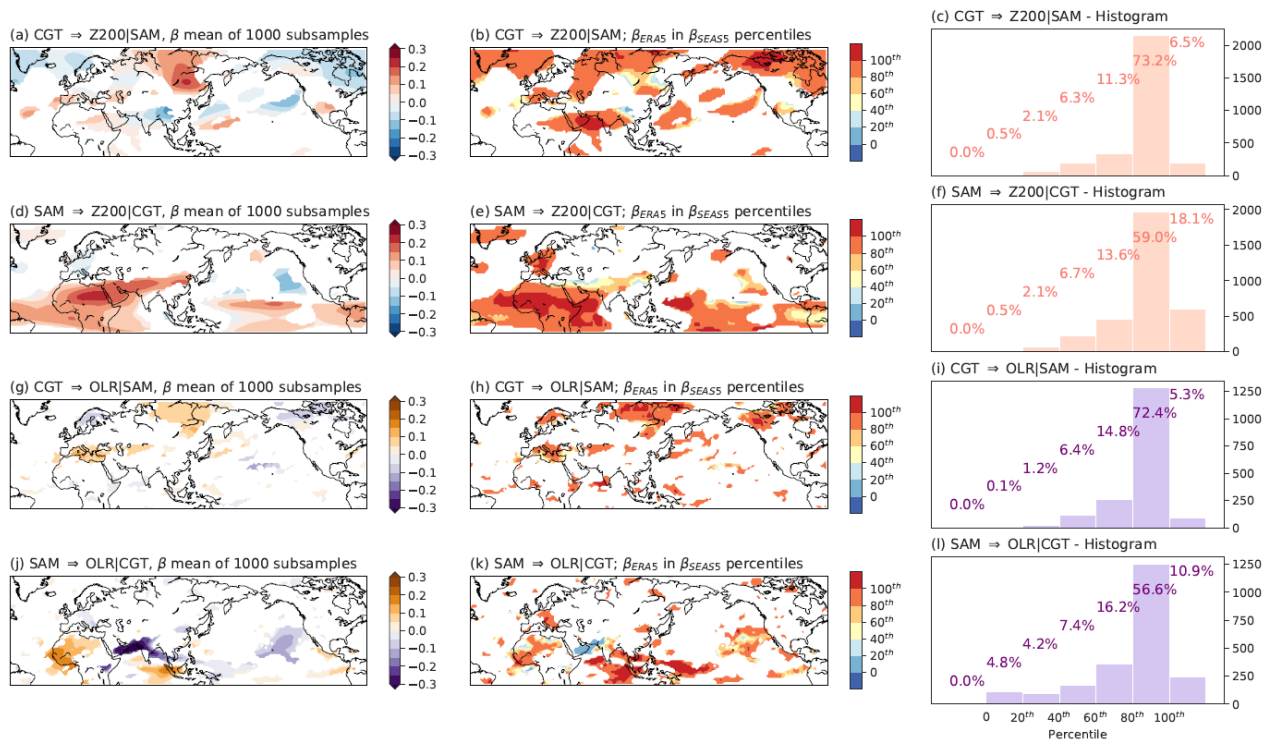
Figure RR3.A3. Same as for Fig. 5 in the main text but for a p-value = 1.0 and showing only grid points where causal links are present.
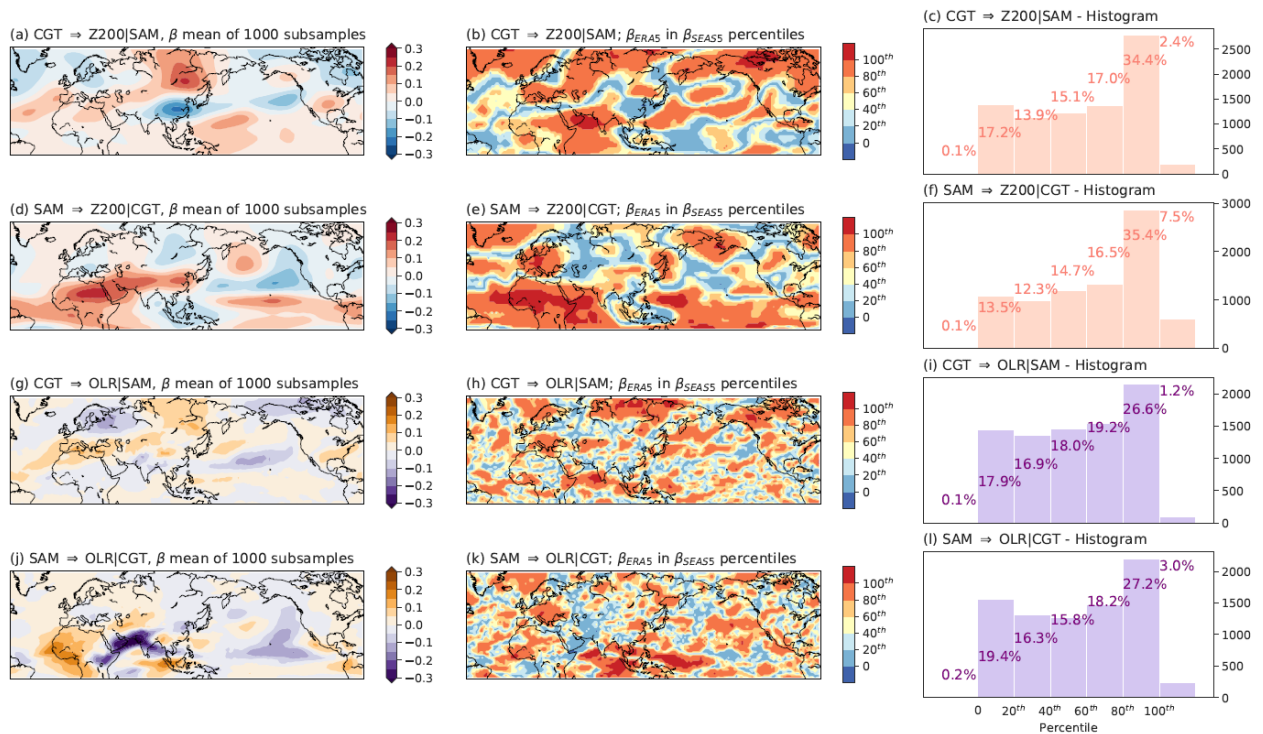
Figure RR3.A4. Same as for Fig. 5 in the main text but for a p-value = 1.0 and showing all grid points (also those where no causal links are present).
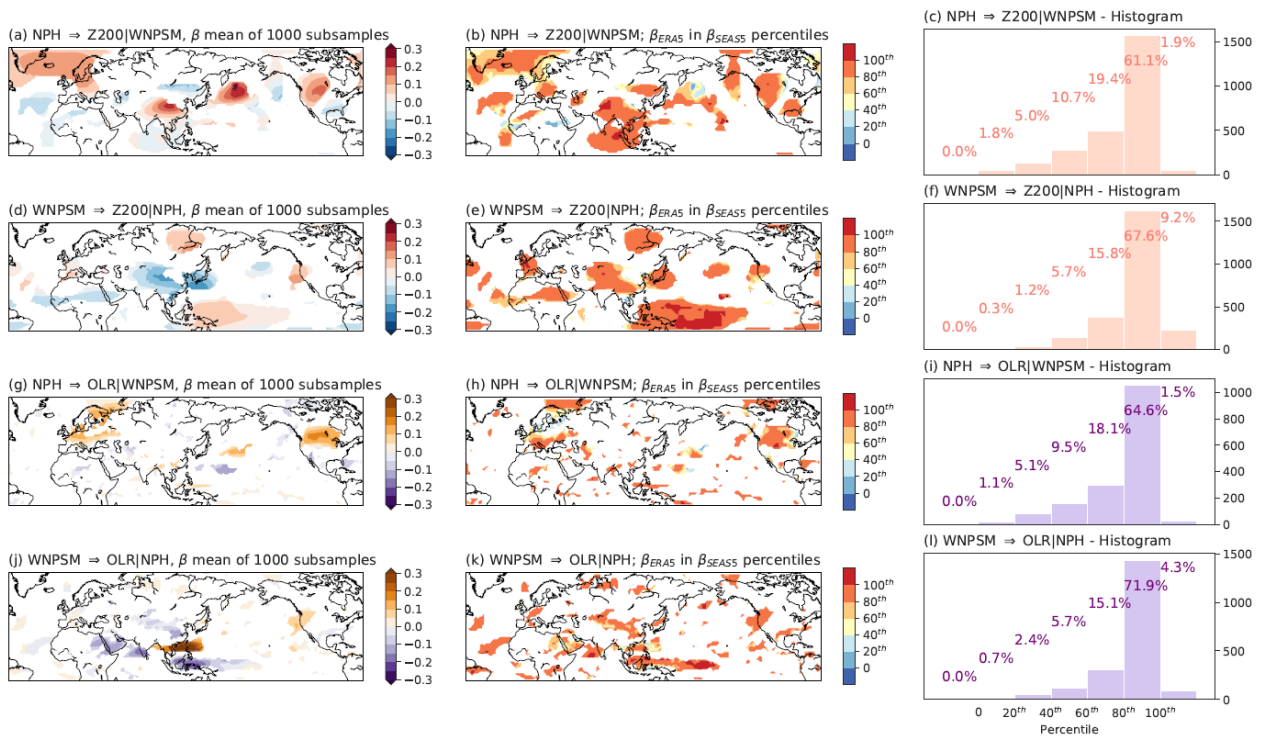
Figure RR3.A5. Same as for Fig. 6 in the main text but for a p-value = 0.5 and showing only grid points where causal links are present.
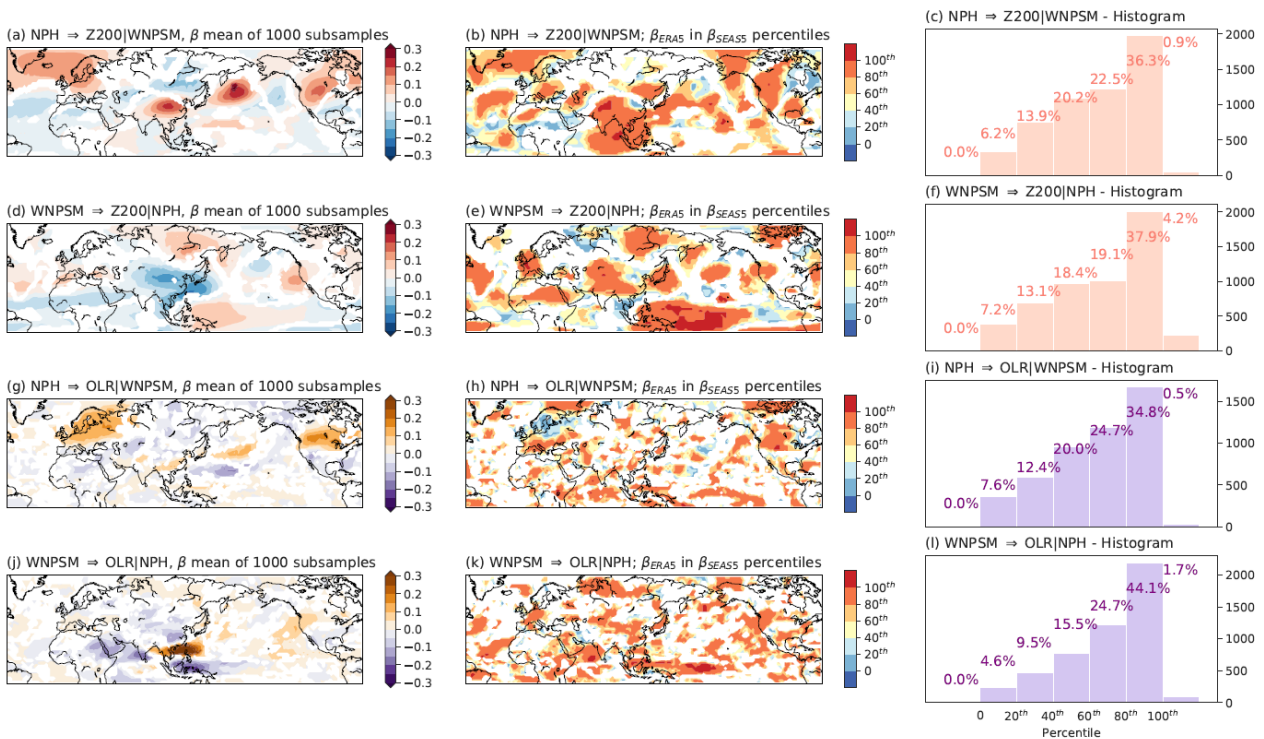
Figure RR3.A6. Same as for Fig. 6 in the main text but for a p-value = 0.5 and showing all grid points (also those where no causal links are present).
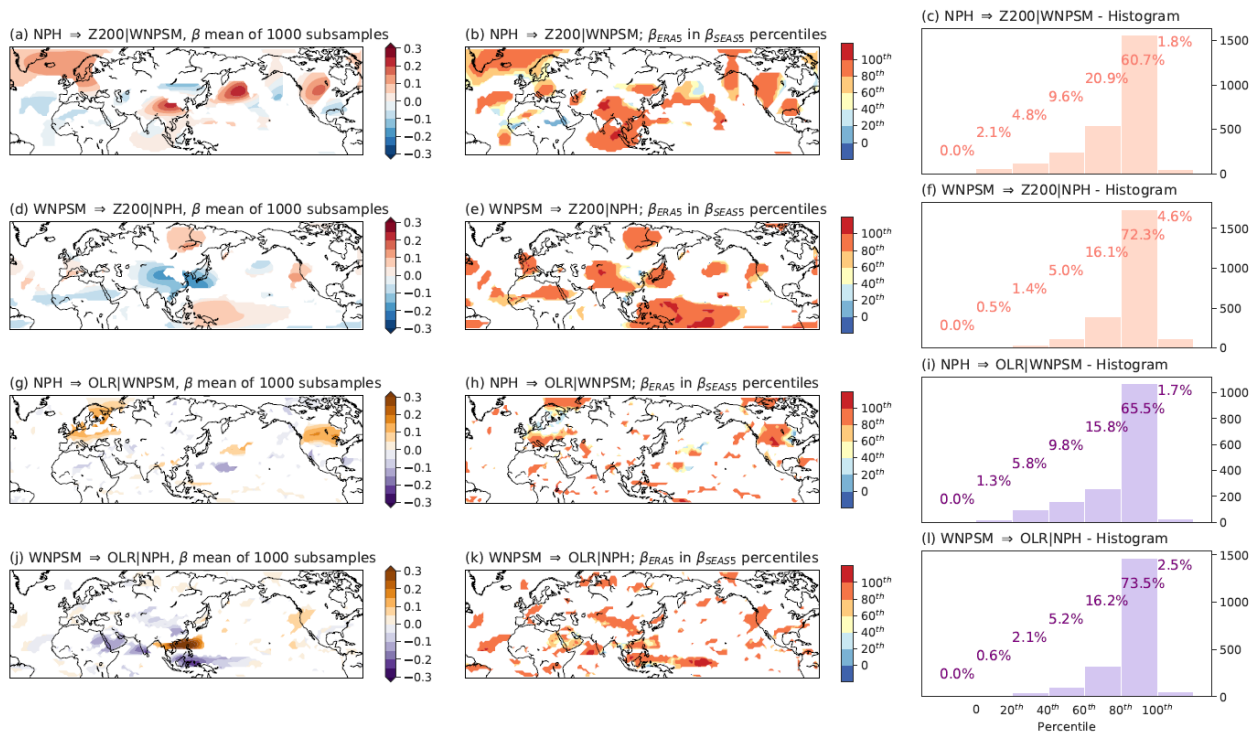
Figure RR3.A7. Same as for Fig. 6 in the main text but for a p-value = 1.0 and showing only grid points where causal links are present.
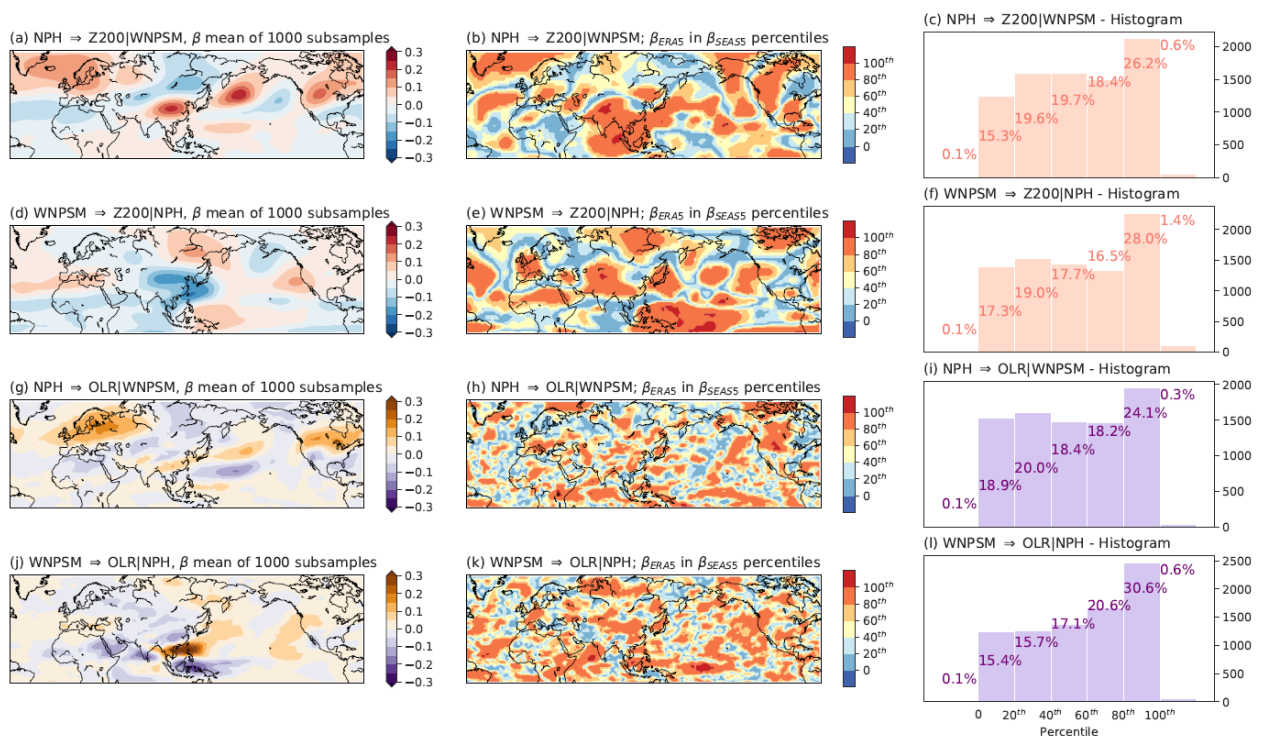
Figure RR3.A8. Same as for Fig. 6 in the main text but for a p-value = 1.0 and showing all grid points (also those where no causal links are present).
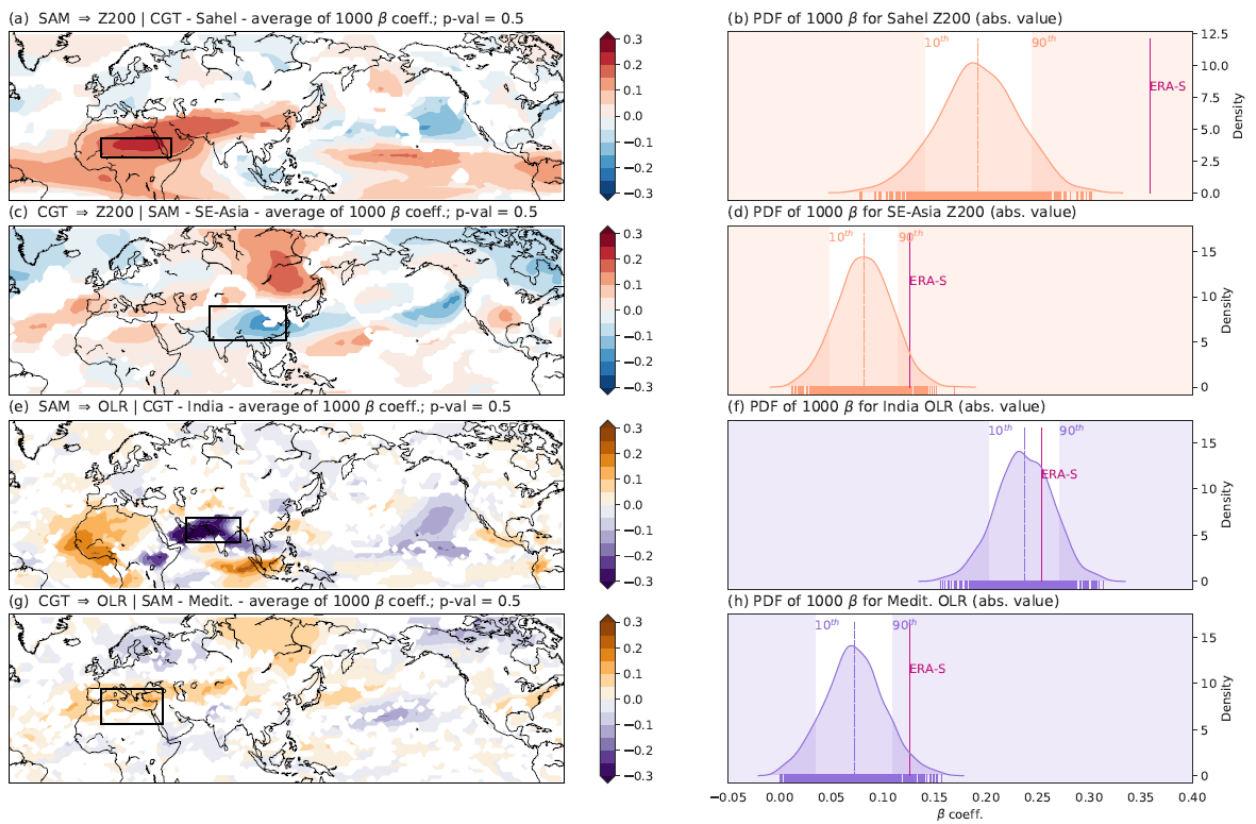
Figure RR3.A9. Same as for Fig. S11 but obtained with a p-value = 0.5 and showing all grid points (also those where no causal links are present).
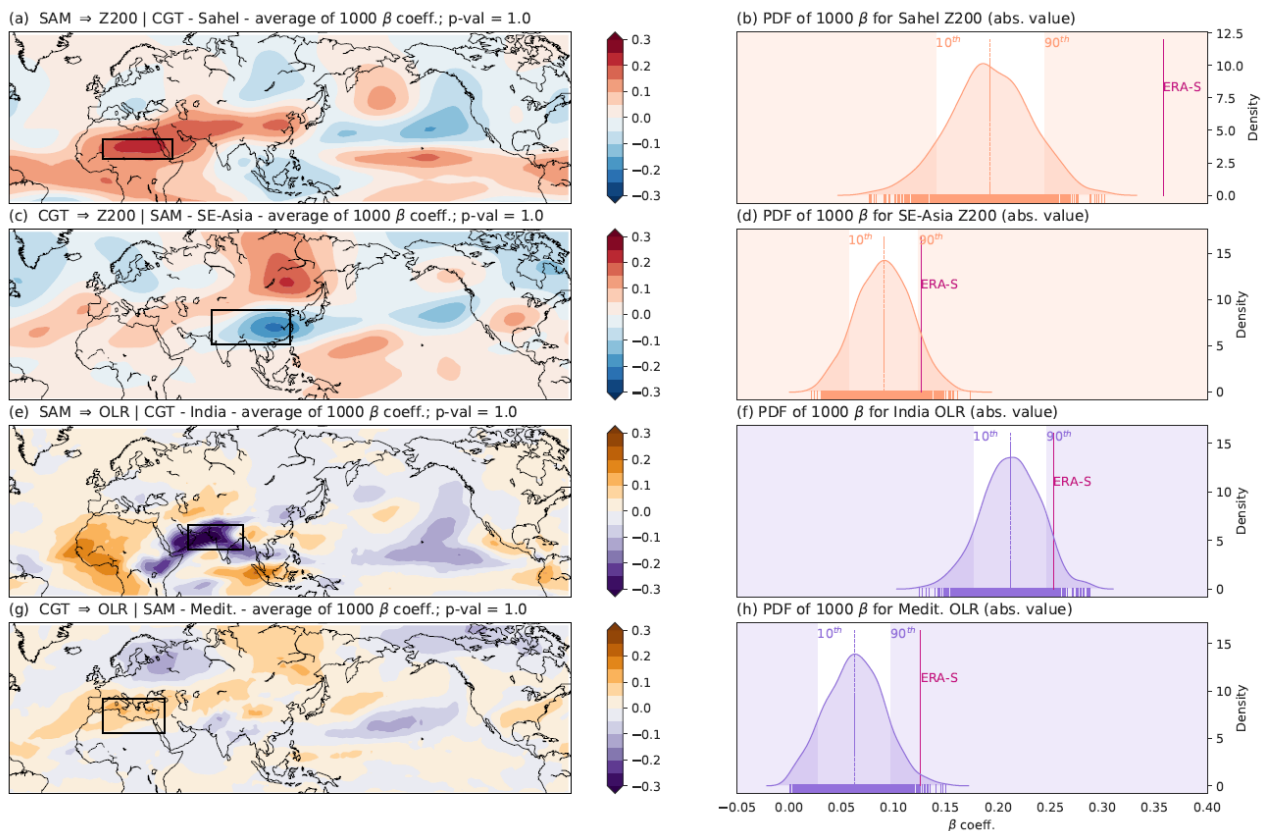
Figure RR3.A10. Same as for Fig. S12 but obtained with a p-value = 1.0 5 and showing all grid points (also those where no causal links are present).
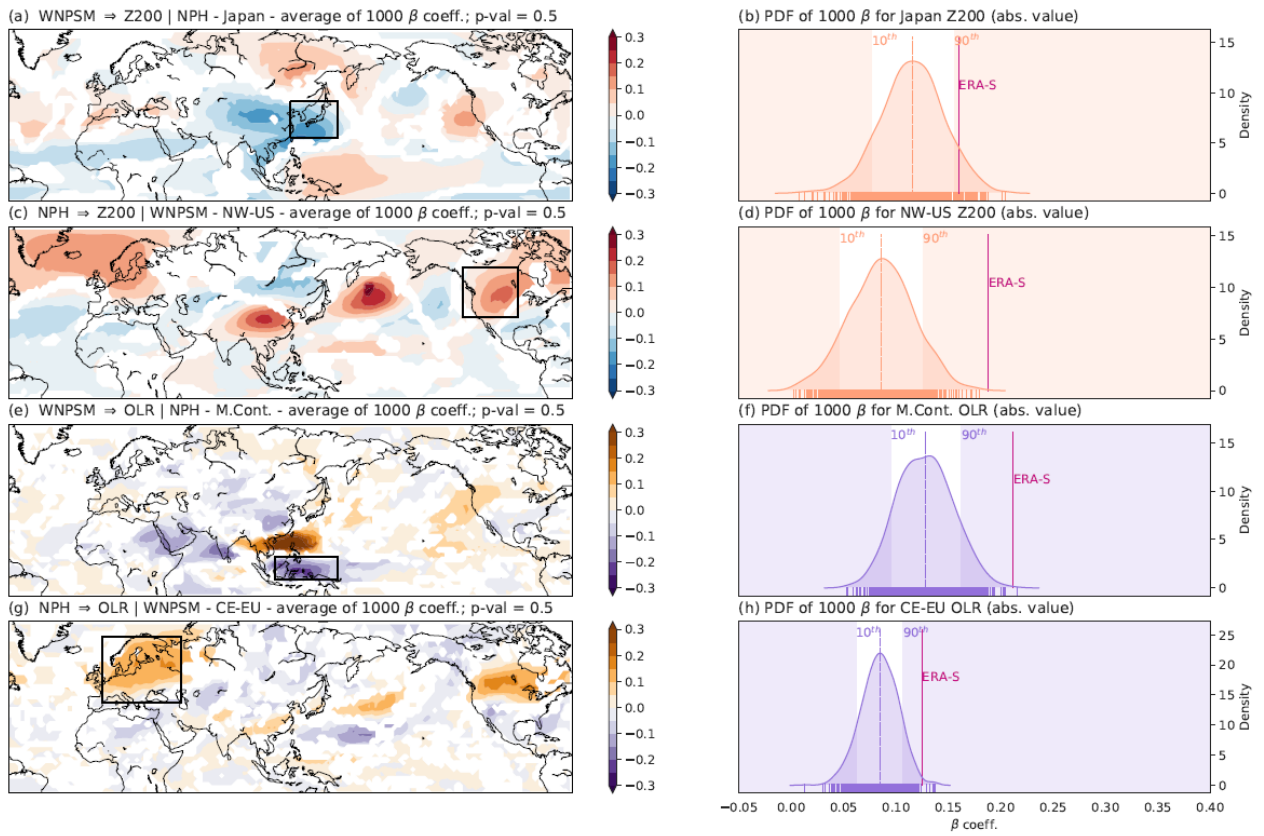
Figure RR3.A11. Same as for Fig. S12 but obtained with a p-value = 0.5 and showing all grid points (also those where no causal links are present).
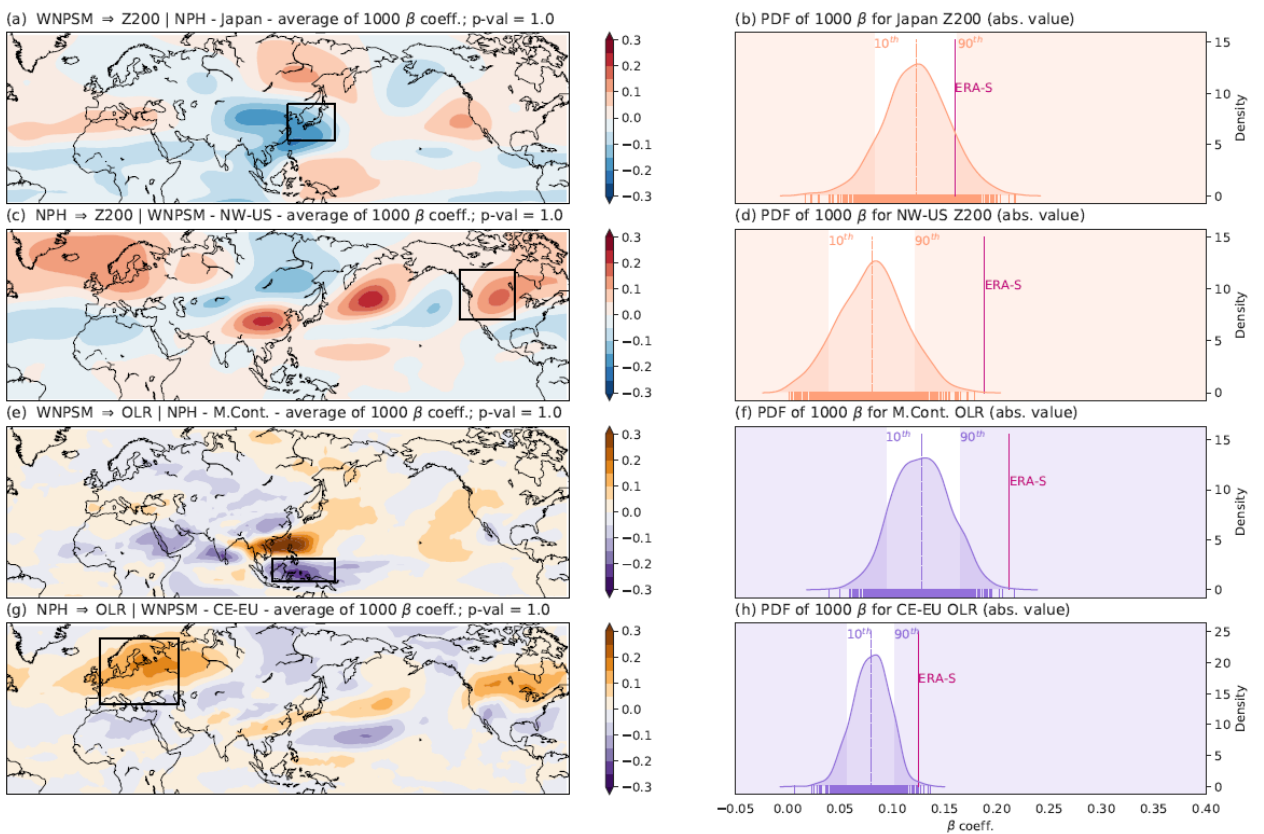
Figure RR3.A12. Same as for Fig. S12 but obtained with a p-value = 1.0 5 and showing all grid points (also those where no causal links are present).