Response to reviewer #2

We thank the anonymous reviewer #2 for the careful and insightful review comments. We have revised our manuscript taking into account all suggestions and we find that the manuscript has improved in its clarity and that the analysis itself is now more robust. The most critical comment is #1, where the anonymous reviewer suggests testing the subsampling using 24 years instead of 60 years as previously done. Addressing this comment shows that the strength of the β values in SEAS5 is less underestimated than what has been shown in the first version of the manuscript. Thus, considering causal maps obtained with time series of the same length does represent an important improvement in the manuscript, i.e., we can be more optimistic about the capability of SEAS5 in reproducing the strength of the observed causal links. Nevertheless, qualitatively the main message of the paper does not change, i.e. (a) SEAS5 can reproduce the sign and spatial patterns of the tropical – extratropical teleconnection in boreal summer although (b) in some regions, e.g. North Africa, the Maritime continent and North America, the strength of the β values is strongly underestimated. A detailed response to each comment is found below.

Most significant comments

1.

The most important part of the analysis looks to be the comparison of causal effect strengths in ERA5 and SEAS5 (starting around L346). It is concluded that the coefficients in SEAS5 are too weak. This is on the basis of comparing the magnitudes of coefficients found to be statistically significant in each dataset. However, the significance threshold is higher for smaller datasets (ERA5 in this case), so I think it would be expected a priori that coefficients in the large SEAS5 dataset would typically be found to be smaller even if SEAS5 were perfect. So the finding of smaller coefficients in SEAS5 does not clearly allow conclusions about biases to be drawn.
I think a simple way to address this would be to calculate differences between the best estimates of the coefficients in SEAS5 and ERA5 without masking the values found not to be statistically significant – then there wouldn't be a bias due to the different dataset sizes as far as I can see (this is presuming that there is no bias in the method of estimating the coefficients). Then grid points where the differences are statistically significant could be marked in the plots e.g. by stippling/hatching (or masking points based on statistical significance of the differences if preferred – but I think it's better for preserving information to show data even where it is not statistically significant – but that's something of a matter of taste). This statistical significance could be estimated in a similar way to the resampling method used later on in the paper.

We thank the anonymous reviewer for bringing up this point, which needs to be further clarified. The length of the time series affects the significance and, in some cases, also the strength of the links. Although it is not the significance itself that affects the strength of the causal links, the significance level determines which set of causal precursors are identified for a certain grid point. Not considering the significance level would, in practice, reduce the concept of causality to that of a multiple linear regression in which <u>all</u> potential precursors are used to provide a linear model of the response variable. Instead, PCMCI identifies first the causal precursors and then performs the multilinear regression with only the set of causal precursors. This process requires, by definition, the use of a significance threshold. To further clarify our point, we provide in this response the maps showing the multilinear regression assuming that the Z200 (or OLR) time series for each grid point depend on all potential precursors at lag -1, i.e., Z200 itself, CGT (or NPH) at lag -1 and SAM (or WNPSM) at lag -1. We provide the multilinear regression maps for both ERA5 and SEAS5 and show that using the multilinear regression instead of the causal tool results still shows a bias in the strength of the regression coefficients, with smaller values in SEAS5 (about half) compared to ERA5 (Figs. R2.1 and R2.2 in this response). After careful consideration, we conclude that it is indeed more relevant to the analysis that the time series in SEAS5 have the same length as those in ERA5 and we have adjusted the analysis accordingly (see the point below).

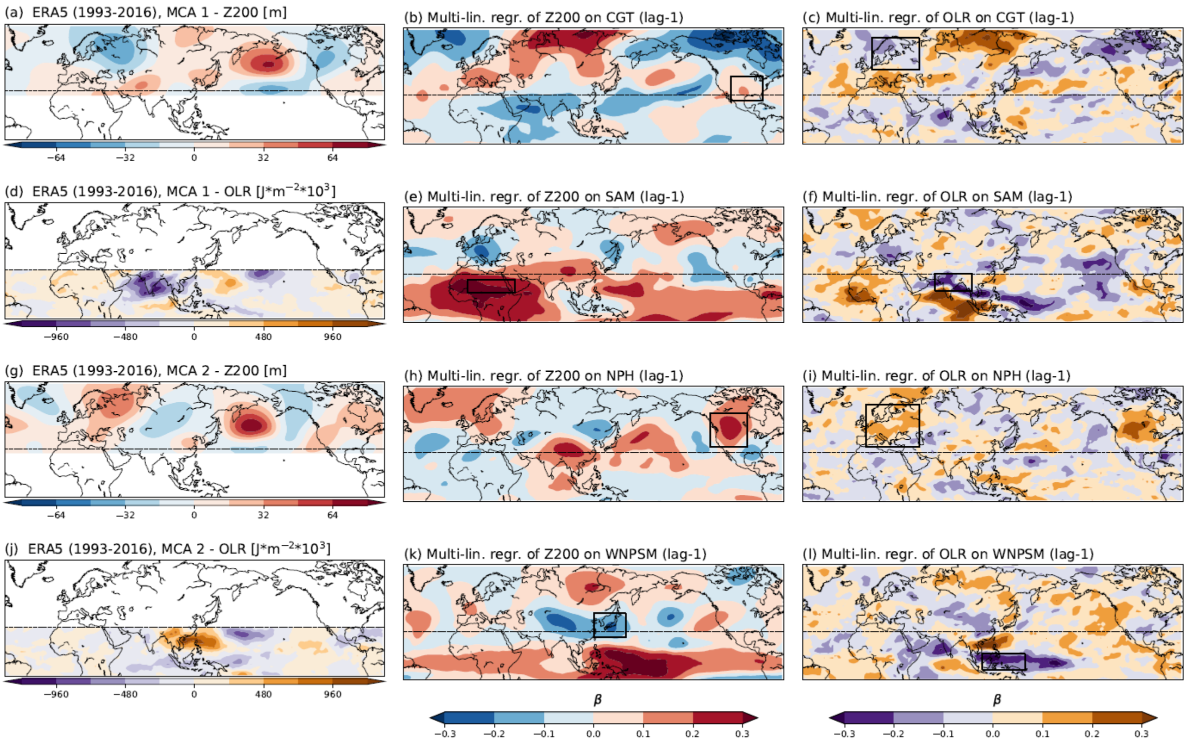

Fig. R2.1. Same as for Figure 3 but showing the multilinear regression coefficient of Z200 on CGT (Panel b) and OLR on CGT (panel c) with the regression coefficients calculated assuming that the grid point time series for Z200 (or OLR) depend on Z200 itself, CGT at lag -1 and SAM at lag -1. Panel e and f: same as for panels b and c but for the effect of SAM on Z200 and OLR. Panels h,i and k, l: same as for panels b,c and e,d but for the effect of NPH and WNPSM on Z200 and OLR fields. Plot obtained using the 600 years as a unique time series.

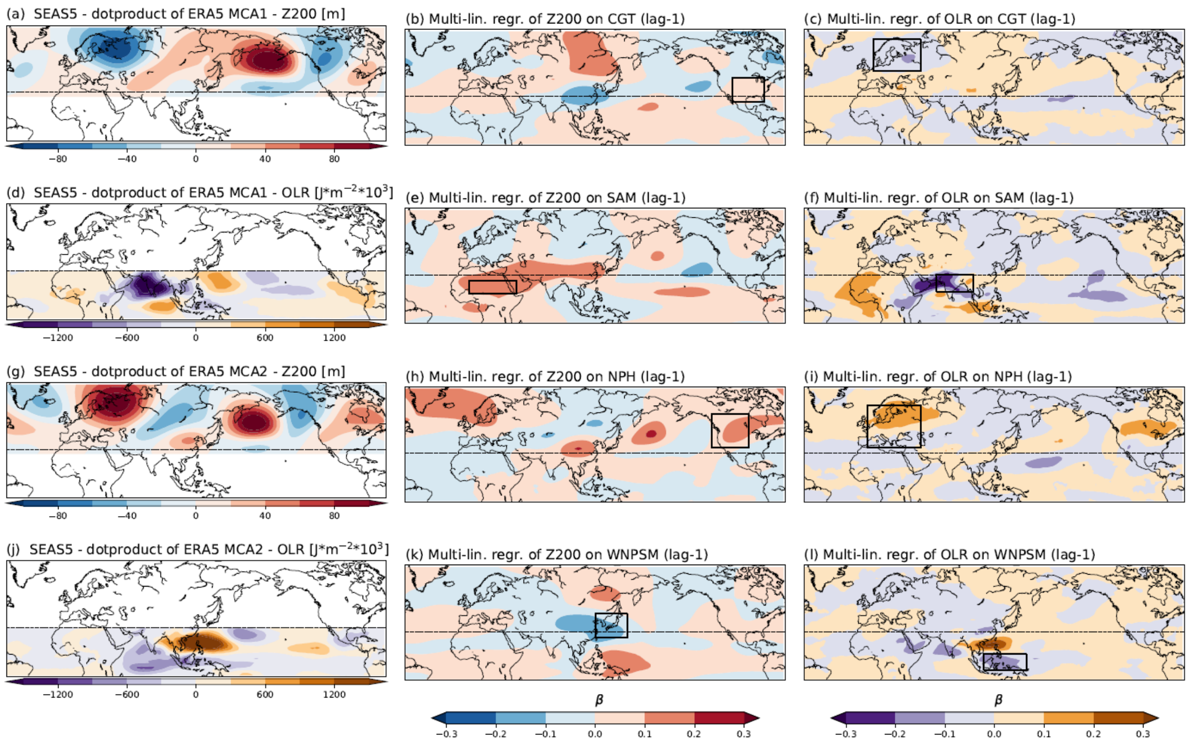

Fig. R2.2 Same as for Fig. R2.1 but for SEAS5 data.

However, the method to use 60 year samples from SEAS5 (L374) also does not allow a clear comparison with ERA5, as the longer time series would give lower sampling variability. I think it would work to create 24-year pseudo-time series of SEAS5 by randomly selecting one member from each forecast year and concatenating them – this would give time series of equal length to ERA5-S and would remove any variation due to different sampling of sea surface temperatures etc. (Random selection of years, as done in the current manuscript, would generally mean not all SST states are represented in a given sample, which could cause sampling variability to be underestimated.) The causal coefficients could be calculated for each pseudo-time series. Then the statistical significance of the difference at each grid point could be assessed based on whether fewer than 5% of the coefficients across the pseudo time-series are as far away from the mean as are those for ERA5-S (similar to the sort of reasoning behind figs.7 and 8). (Applying a multiple testing criterion here would be good as well. It would also be good to do a cross-check that the mean coefficients across the pseudo-time series are similar to those calculated from SEAS5 all together, to check that there isn't a bias dependent on the time series length.) The authors could use another method that produces a clearly unbiased comparison if they have one.

We thank the anonymous reviewer for this insightful suggestion. We have re-calculated the subsampling experiment as suggested: we now provide 1000 subsamples each containing the same 24 years as in ERA5 (where for each year one member is randomly selected from the 25 ensemble members available). Figures 7 and 8 in the revised version of the manuscript now show these new figures. We have updated the text accordingly in the methods section and in the results section. Results show that reducing the number of years increases the mean value and the standard deviation of the distribution. Thus, in general, ERA5 β values fall better inside the distribution of β values from the 1000 subsamples. We now also plot on the pdf a vertical line showing the mean β value of the causal maps shown in Fig. 4. Notably, β values obtained using all 600 years are in general lower than the mean values obtained from the subsampling experiment, further supporting the idea that proving the subsampling experiment with the same number of years used in ERA5 provides better and more insightful results. Finally, two of the four regions analyzed for MCA mode 1 have been substituted: the south-eastern US has been substituted with Southeast Asia and western-north Europe has been substituted with the Mediterranean. This change is motivated by the lower p-values available after reducing the length of the samples from 60 to 24 years. The new figures are provided in the revised manuscript and the revised text is shown below:

Lines 387-389: "We select 1000 samples of 24 years each (for each year one ensemble member is randomly selected out of the 25 available ensemble members), and for each sample we provide the corresponding causal map. In this way, the number of years used in each subsampling experiment (24 years) is the same as those available from ERA-S (24 years)."
Lines 393-398: "For each causal map, the p-values are corrected by applying the Benjamini-Hochberg false discovery rate correction and only β values with a corrected p-value < 0.05 are retained. The resulting 1000 causal maps are averaged and shown in the left column in Figs. 7 and 8. For each grid point, the mean β value is calculated only if at least 100 β value results are significant at the $\alpha = 0.05$ threshold, however, non-significant values do not enter the mean. Applying this double threshold (which is not done in Fig. 3) shrinks notably the area of the spatial patterns if compared to those in Figs. 3 and 4, however here we concentrate only on the β values contained in the regions highlighted by black boxes in Figs. 7 and 8."
Lines 415-417: "The mean β values obtained by the SEAS5 causal maps calculated using all 600 years (Fig. 4) are represented by solid vertical lines (orange for Z200 and light purple for OLR PDFs) together with reference β values calculated in ERA-L and ERA-S (shown as a purple and magenta vertical lines) in each PDF. All three β values are standardized by dividing by the standard deviation and mean value of the SEAS5-ERA5 subsample distribution"

Lines 422-424: "In general, mean β values obtained in Fig. 4 tend to be lower than the mean β values obtained in the 1000 subsamples, indicating that taking all the 600 years together, despite spatial patterns showing good agreement, effectively increases the underestimation effect of the strength of the β values when compared to the average of the 1000 subsample."

Lines 428-431: "By contrast, the causal effect of SAM on OLR over India (SAM → India OLR | CGT), of CGT on OLR over the Mediterranean (CGT → Medit. OLR | SAM) and of CGT towards Southeast Asia Z200 (CGT → SE-Asia Z200 | SAM) all fall in the range of the possibilities of the 1000 subsamples with both ERA-L and ERA-S β values falling between the 10th and the 90th quantile (Fig. 7d,f,h). Thus, in this region the β values show a good agreement with reanalysis data when the spread of SEAS5 β values is considered."

2.
The discussion of the potential role of mean state biases (sec. 3.4) seems rather speculative and it's not clear to me what value this is providing, as it doesn't really seem to narrow down the cause of any biases. The end result is to say that the analysis is inconclusive – then it seems like it could all be briefly discussed in a paragraph or so.

We agree with the anonymous reviewer #2 that the results of this section can be moved to the Supplementary Material and discussed more briefly. We have adapted the manuscript following this suggestion. Former Figures 9 and 10 are now Fig. S8 and S9 in the Supplementary Material. The text in the main manuscript has been adapted accordingly:

Lines 445-457: "We investigate how the bias in convective activity between ERA5 and SEAS5 may affect the monsoon-desert mechanism and find inconclusive results. SEAS5 shows enhanced convective activity with respect to ERA5 around the equator (negative OLR anomalies in Fig. S8m,o) and a drier tendency over central India and the Arabian Sea (positive OLR anomalies in Fig. S8m,o). Rodwell and Hoskins (1996) has shown that the heat source provided by the convective activity in the Indian Ocean/Bay of Bengal region generates Rossby waves that reach the Sahara Desert. However, the latitudinal position of the heat source is critical: a heat source located in the south (10°N) does not act as a source of Rossby waves capable of reaching the Sahara Desert, while a heat source located around 25°N does. Thus, we investigate whether the dry bias over central India may explain low causal effect values over the Sahel and North African region and calculate causal maps for years with enhanced convective activity over central India and for those with enhanced convective activity over the tropical Indian Ocean. Despite a slight tendency towards higher β values over North Africa being detected during years with enhanced convection over central India in SEAS5 initialized at 1st May (40% higher compared to years with enhanced convection over the Indian Ocean, Fig. S9e), this result is not found in SEAS5 initialized at 1st March and thus remains inconclusive (Fig. S9j)."

L439 It seems like this part could do with references regarding waveguides.
Due the shortening of this section, the sentence has been removed.

L445-8 It's not clear why a bias in mean convection would affect the strength of the link to North Africa – this would probably affect the mean state in North Africa, but why would it be particularly important for the regression coefficient?
We agree with anonymous reviewer #2 that in general, a bias in the mean state does not necessarily affect the strength of the β value but just the total effect (as discussed in lines 528-529). However, when the monsoon-desert coupling is examined, if convective activity were situated too far south over India, then no signal would be seen over Northern Africa, and for this reason we wanted to check whether this bias in the tropical Indian Ocean would potentially affect the β values over North Africa. Nevertheless, we agree that this is not the case and in the revised version of the manuscript this comment has been removed (see lines 445-457 reported above).

3.
Similarly, the discussion about ENSO influence is quite lengthy and doesn't really produce clear results, so it seems to me that this could be summarized quite briefly too.

Lines 458-479: Similar to the previous comment, we agree with anonymous reviewer #2 that the results of this section can be moved to the Supplementary Material and discussed more briefly. We have adapted the manuscript following this suggestion. Former Figures 11 and 12 are now Fig. 9 (revised paper) and S10 in the Supplementary Material. The text in the main manuscript has been adapted accordingly:
"Finally, we investigate the effect of ENSO states on the sign and strength of tropical-extratropical causal interactions shown in Fig. 4 and show that the effect of ENSO positive and negative phases is mostly marginal with a few exceptions. We define El Niño and La Niña years based on seasonal (JJAS) SST anomalies averaged over the Niño3.4 region (5°S-5°N, 190°-240°E) and calculate causal maps for the effect of MCA mode 1 and 2 on Z200 field separately for the 102 Niño3.4 positive and 142 Niño3.4 negative years (those years that exceed the +0.5°C/-0.5°C thresholds are defined as El Niño/La Niña years, respectively). The results for MCA mode 1 are shown in Fig. 9 for both Niño3.4 positive (left column) and Niño3.4 negative years (middle column) separately and for the difference

Niño3.4 positive minus Niño3.4 negative (right column) and for different initialization dates (1st of March and 1st of May). Comparing the casual maps in Fig. 9 left and middle column with those in Fig. 4 shows that, in general, the spatial patterns and the sign of the causal links are not affected by the sign of the ENSO anomalies.

Changes in the strength of the links are shown in the right column in Fig. 9. For each grid point, the difference Δ_{β} between the β value for Niño3.4 positive years ($\beta_{niño}$) and the β value for Niño3.4 negative years ($\beta_{niña}$) is calculated and then divided by $\beta_{niño}$, following Eq. (3):

$$\Delta_{\beta} = \frac{\beta_{niño} - \beta_{niña}}{\beta_{niño}} x100\% \quad (3)$$

Hence, Δ_{β} is expressed as a percentage, where a zero value represents perfect agreement between $\beta_{niño}$ and $\beta_{niña}$, a positive value of, e.g., 50% means that $\beta_{niño}$ is 50% larger than $\beta_{niña}$ and vice versa. In general, β values in the tropical Pacific and over eastern Africa and the western tropical Indian Ocean are 40 to 80% larger during El Niño years and this result is consistent for both initialization dates. However, in other regions, results differ depending on the initialization date, e.g. North-western Africa or the North Pacific region for the effect of SAM on the Z200 field. ENSO causal maps for MCA mode 2 and for the ENSO versus neutral years are shown in Figs. S10-S12 in the Supplementary Material and show similar results. Thus, we conclude that in general ENSO does not alter the sign and spatial patterns of tropical extratropical teleconnection but can however modify the strength of these connection in some specific areas, especially close to the equator."

L479-80 "If a dependence is found…" – I don't follow this.
This sentence has been removed in the revised version of the manuscript.

L495-9 The similarities here that I tried to check do not look very substantial e.g. for the western central Africa and tropical central Pacific SAM->Z200 connections, there is hardly any signal in the runs initialized on May 1.
In the revised version of the paper, we now refer only to the most clear and robust difference between different ENSO phases, i.e. the enhanced β values over the tropical Pacific and over eastern Africa and the western tropical Indian Ocean (see lines 458-479 or the text reported above).

L499-502, L514-9 When talking about similarities over such small regions, it's not clear that this isn't noise.
The description of these regions has been removed in the revised version of the manuscript (see lines 458-479 or the text reported above).

Other comments

1. Abstract - it would be good to have a brief summary of the quantitative size of the most important results.
We now specify the strength of the causal links in line 47: "However, the strength of causal links in SEAS5 (β values ~ 0.1-0.3) is often too weak (about two thirds of those in ERA5, β values ~ 0.2-0.4)."

2. L71-3 This final statement about the superiority of dynamical forecasts doesn't seem clearly justified.
We have improved the explanation of this concept in the revised manuscript. See lines 69-74 "However, when the focus is on the representation of physical processes rather than the forecast skill, dynamical forecasts, generated by general circulation models (GCMs), provide a more complete representation of the atmospheric physics that governs weather and climate behaviour (Shukla et al., 2000). Dynamical seasonal forecasts explicitly resolve dynamic and thermodynamic equations and are better suited for representing the dynamic and thermodynamic processes and emerging dynamical teleconnections within the climate system."

3. L143 What's the process of removing the interannual variability, seasonal cycle and any long-term trend?
We have improved the explanation of this pre-processing step in the revised manuscript:
Lines 143-147: "The interannual variability, seasonal cycle and any long-term trend are removed. To do so, first the interannual variability, i.e. the average value of each May-to-September period is subtracted from the corresponding year (thus ensuring that the weekly signal does not include the interannual variability). Then, for each of the 21 time steps considered in each season (e.g., the first time steps for May for each year) the trend

over the 24 (or 600) analyzed years is removed and anomalies around zero are calculated, thus removing both the trend and seasonal cycle."

4. L166 It said above that May is used in the analysis, so the spin-up time is less than 1 or 3 months for each SEAS5 case.

The reviewer is correct, the month of May, though outside the target season, is used when time steps previous to lag 0 are considered for the month of June. That is the reason why we use two sets of seasonal forecasts to make sure that the initialization date has no or little influence. We have adjusted the text so that it is not misleading for the reader:

Lines 171-173: "This way, the model has up to three (and at least one) months of spin-up to reduce the influence of the initial conditions. However, for the 1st of May initialized forecasts, although the month of May is outside the target season, May time steps enter the set of precursors for June time steps. Thus, we provide a sensitivity analysis to show which results depend on (or are independent of) the chosen initialization date."

5. L188 It could be helpful to define the ERA-S and ERA-L periods earlier where the data is described.
We now define the ERA-S and ERA-L periods in the Data section (line 137).

6. L223 Ah, so separate time series are created for the OLR and Z200 for each MCA mode. It could be useful to clarify this above.

We have clarified this in Section 2.2.

Lines 190-191: "Note that separate time series are created for the OLR and Z200 MCA patterns and for each MCA mode, for a total of 4 time series when MCA modes 1 and 2 are analyzed."

7. L228 "lag min" and "lag max" should be defined. The results only seem to consider one lag, equal to one week – if this is what these settings mean, this should be specified.

We clarify this point in lines 234-236: "Note that only results for lag -1 are shown as almost no significant links are found for lag -2 in ERA5 causal maps.".

8. L233 What false discovery criterion is used, more precisely e.g. what is the maximum family-wise error rate, or other metric used to determine whether to accept a coefficient as statistically significant?

We use the false discovery rate described in Benjamin & Hochberg (2001) [quote taken from that paper]: "The false discovery rate (FDR), suggested by Benjamini and Hochberg (1995) is a new and different point of view for how the errors in multiple testing could be considered. The FDR is the expected proportion of erroneous rejections among all rejections. If all tested hypotheses are true, controlling the FDR controls the traditional FWE. But when many of the tested hypotheses are rejected, indicating that many hypotheses are not true, the error from a single erroneous rejection is not always as crucial for drawing conclusions from the family tested, and the proportion of errors is controlled instead."

We have clarified this concept in lines 241-244: "The significance threshold adopted for plotting the results is α = 0.05 and we use corrected p-values by applying a false discovery rate (FDR) correction (Benjamini and Hochberg, 1995) to control for multiple testing among the multiple grid locations in causal maps. The false discovers rate is "the expected proportion of erroneous rejections among all rejections" (Benjamini and Yekutieli, 2001)."

9. L251 How much of the overall variance do these modes explain? This would be useful for justifying why these are important to study, and to help quantify what proportion of the variability this analysis is relevant for.

We have calculated the variance explained by each MCA pattern and the results show that these patterns can explain up to 25% of the variance of Z200 and ORL fields depending on the specific region. We have included this information in the text in lines 272-273: "These four patterns can explain up to 25% of the variance in the Z200 and OLR fields depending on the region (not shown)."

10. L272-4 There seems to be lower OLR over India in MCA2 for both ERA5 and SEAS5 to me.

The OLR signal is indeed negative in both Figs. 3j and S2j, however, while the OLR anomalies in Fig. 3j are very weak and the overall pattern is dominated by the positive anomalies over the South China Sea and negative anomalies over the WNPSM region, Fig. S2j shows more marked negative anomalies over India of the same magnitude of the anomalies over the South China Sea, which are in turn very weak compared to those in Fig. 3j. We have clarified this point in lines 286-291: "Similarly, the convective activity over the Indian peninsula, which in ERA-S represents one of the characteristic features of MCA 1 (Fig. 3d) and is very weak in MCA 2 (Fig. 3j), is found in both MCA 1 and MCA 2 of SEAS5 almost with similar magnitudes, though the negative anomalies are stronger in MCA 1 (Fig. S2d,j)".

11. L275 Though, the Z200 part of MCA1 in SEAS5 looks similar to the negative pattern of the Z200 part of MCA2, so the pattern does seem to be appearing at least somewhat. (But I agree that using a common set of patterns for the rest of the analysis is sensible regardless.)
We have corrected this sentence, which now reads, lines 289-291: "In contrast, the wave pattern over Eurasia characterizing ERA-S MCA 2 showing a high-pressure region over Eastern Europe and a low over Central Asia (Fig. 3g) is very weak in SEAS5 MCA modes 2 (Fig. S2g) and not present in MCA mode 1 (Fig. S2a)."

12. L283 Why not just use a regression of the SEAS5 fields onto the calculated MCA time series based on the ERA5 modes - then wouldn't this be much more comparable to what's shown in the left column of fig.3?
See our answer to general comment 1 earlier in this document.

13. L290 It could do with saying here what lag these maps are for.
This missing information has been added, lines 304-305: "Causal maps calculated for the effect of ERA-S CGT and SAM at lag -1 on Z200 and OLR fields are shown in Fig. 3b,c and 3e,f respectively".

14. L292-315 Some of the signals being discussed, here and below, are only deemed statistically significant over small regions and it doesn't seem clear that they are real - the maximum family-wise error rate of the multiple testing criterion should probably be considered.
In this analysis, we are strongly limited by the length of the time series, which increases the significance level and reduces the spatial extension of the significant patterns. We always calculate the corrected p-values by applying the Benjamini & Hochberg false discovery rate correction. Therefore, comparing Fig. 3 with Fig. S1 or Fig. 3 and 4 in Di Capua et al. (2020a) shows patterns with reduced extent and thus more difficult to interpret. We describe these patterns by using also the information shown in Fig. S1 and referring to Di Capua et al. (2020a), where much stricter significance testing was performed due to the larger number of years available.

15. L293 "are" -> "tend to be"? I think this should be changed to get away from it sounding like these signals will definitely follow - and similarly for other such statements in the following text.
The sentence has been adjusted following the reviewer's suggestion (line 307).

16. L297 and North Asia?
The sentence has been adjusted following the reviewer's suggestion (line 311).

17. L301 fig. 3c rather than 3e?
This mistake has been corrected (line 315).

18. L380 As I said above, I think the size of the dataset will affect the strength of the beta values when there is masking according to statistical significance.
See our answer to general comment 1 earlier in this document.

19. L389 It would be useful to have the boxes marked on figs. 5,6 as well, to be able to see what differences between SEAS5 and ERA5 they correspond to.
Figures 5 and 6 have been modified following the reviewer's suggestion.

20. L393-4 This seems to be effectively saying that the best estimate of the coefficient is zero when it is found to be non-statistically significant. However, it would seem better to me to use the actual result of the statistical procedure, which is probably closer to the truth. It seems like it could introduce complicated effects if some values are set to zero in this part of the analysis.

To clarify this point, we do not set values to zeros, instead we ignore non-significant values and average only on those. See also our response to comment #1 and lines 408-411: "For each region and for each sample in the 1000-ensemble member subsampling experiment, the causal effect is spatially averaged (accounting only for significant values, i.e., zero values are discarded as they are not significant) and the absolute value is taken after averaging. In this way, we obtain a distribution of 1000 β values for each region of interest."

21. L423 I suggest "bias" -> "mean state biases" in the section title.

This suggestion has been implemented, see line 445 in the revised version of the manuscript.

22. L521 The analysis seems statistical to me - "process-based" implies to me that mechanisms explained in terms of fairly fundamental physics were examined, which hasn't been done in general here.

We have modified the text following the reviewer's suggestion "In this work, we provide a process-guided statistical analysis, built on causal discovery…" (line 428).

23. L558 Is "negative bias" referring to the CEN coefficients with respect to that mode?

We have clarified this sentence.

Lines 530-532: "Here, our analysis shows a negative bias in β coefficients over North Africa in the first MCA mode, thus there is not only a negative bias in the precipitation over the Indian peninsula, but also the causal link strength is too weak."

24. L561-3 This doesn't seem to clearly follow logically. Again, see comments above about biases in the mean not clearly explaining biases in a regression coefficient.

We have clarified this statement addressing the reviewer's remark.

Lines 520-531: "In boreal summer, the CGT pattern arises even without the heat source provided by SAM (Ding et al., 2011), as it represents a preferred mode of variability of boreal summer circulation that can be ignited by different forcings (Kornhuber et al., 2020; Teng and Branstator, 2019). Recent work has shown that there is a positive causal link from the SAM to the CGT (Di Capua et al., 2020b, a). In general, climate models struggle to reproduce the climatology of SAM rainfall patterns, both in magnitude and spatial pattern (Menon et al., 2013) and SEAS5 seasonal forecasts underestimate the strength of the SAM convective activity and rainfall over the Indian peninsula and the Bay of Bengal (see Fig. 9 or Chevuturi et al. (2021)). The CGT pattern has been shown to be too weak in SEAS4 (ECMWF's previous operational seasonal forecasting system), likely due to a dry bias in precipitation in SEAS4: weaker convective activity over the Indian continent does not provide the heat source that reinforces the CGT pattern (Beverley et al., 2019; Ding and Wang, 2005; Di Capua et al., 2020b). If the forcing (SAM) is too weak, the response (CGT) will be too weak, but this does not necessarily affect the strength of the causal link. Here, our analysis shows a negative bias in β coefficients over North Africa in the first MCA mode, thus there is not only a negative bias in the precipitation over the Indian peninsula, but also the causal link strength is too weak."

25. L569-71 Again, this doesn't clearly follow.

We have clarified this statement addressing the reviewer's remark (see comment above).

26. L574 "quite satisfying" is quite subjective – if you mean the signs of the coefficients seemed similar in the datasets, then please clarify.

We have revised this sentence addressing the reviewer's remark, lines 536-539: "Despite consistent underestimation of causal link strength in certain regions (Figs. 5 and 6), these results imply that 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. 7 and 8)."*

27. L599-600 I suggest "if EC-Earth behaves similarly to ERA5".

We have revised this sentence addressing the reviewer's remark, at lines 562-566: "This information becomes even more relevant in the context of climate change. If EC-Earth (Döscher et al., 2022) (the Earth system model built by the ECMWF which shares the same atmosphere model as SEAS5) behaves similarly to ERA5, we can have some confidence that at least the sign and spatial patterns of these tropical–extratropical teleconnections are well represented, though the strength of the links shows a large spread (Figs. 7 and 8)."

28. L602 "fairly well represented" – magnitude does matter!

We have clarified this statement addressing the reviewer's remark (see comment above).

29. L602 "future projections under global warming scenarios may be fairly reliable" - this requires looking at many more diagnostics and understanding much more about the model.

We have modified the sentence following the reviewer's suggestion.

Lines 565-567: "Future work will analyze how these teleconnections change in future projections under global warming scenarios."

30. L614 It's not clear to me how to do bias-correction based on MCA modes, when these are internal to the atmosphere – I think the claims here need toning down a bit.

We have modified the sentence following the reviewer's suggestion.

Lines 576-578: "By identifying the regions where a certain pattern exerts a significant influence and/or deriving information on which regions have a bias in the model, we provide useful information on the regions where the model representation of these mechanisms should be improved and work towards targeted forecasts".

31. L620-1 needs to be clearer this refers to the sign and not the magnitudes (based on the authors' interpretation of the results, anyway)

We have clarified the sentence following the reviewer's suggestion, lines 583-584: "In summary, this analysis has shown that ECMWF's seasonal forecasts have good ability at reproducing the sign and the spatial patterns of the causal effect of the two main modes".

32. L623-4 Where was this result made clear?

We have clarified the sentence following the reviewer's suggestion "Despite a general underestimation of the causal strength, our subsampling experiment shows that in most of the analysed regions, this negative bias is actually contained in the spread of the SEAS5 seasonal forecasts (Figs. 7 and 8)." (lines 585-587)

33. Fig.3 – the caption should explain how the causal maps should be interpreted, or refer to the text.

Following the suggestion of the anonymous reviewer, we now refer to the text for interpretation of the results in the captions of both in Fig. 3 and Fig 4.

34. Fig.3 - Some explanation should be given for the boxes drawn on the maps.

We have added the definition of each box in the captions of both in Fig. 3 and Fig 4.

35. Figs.5,6 - The colour scale here seems confusing in that more intense red actually means a smaller difference between SEAS5 and ERA5. Maybe shift it so yellow is at 1?

We have modified the colorbar; now dark blue colors represent regions where the beta values are strongly underestimated while yellow colors indicate those with a small bias between ERA5 and SEAS5.

36. Figs.5,6 - The "mean" and "std" values should be explained.

We have included in the caption of Figs. 5 and 6 the explanation of the mean value and standard deviation referred to in the title of each panel.