Response to Reviewer #2

We thank the anonymous reviewer for their positive evaluation of the paper and for their final inputs to further improve the manuscript. We have addressed all the comments and modified the manuscript following the reviewer's suggestions. A detailed point-by-point response to the comments is found below.

Overall, **my recommendations based on this are**:

- Remove quantitative comparisons of magnitudes of causal links in ERA and SEAS5, since the possible range of differences between them seems very large e.g. at L48-49 in the abstract, and other places. Referring to quantitative results of particular analyses is still encouraged – it just shouldn't be implied that any particular result shows the true bias in SEAS5.

We have modified our claim in the abstract of the manuscript, which now reads "By performing a subsampling experiment (over time), we analyse the strengths of causal links in SEAS5 and show that they are qualitatively weaker than those in ERA5." (lines 47-48). We have checked the reminder of the manuscript and, when necessary, modified our claims in accordance with the reviewer's suggestion. We avoid quantitative comparison, but instead still refer to qualitative results showing that beta values in SEAS5 seem to be consistently underestimated when compared to those found in ERA5.

- Include some of the analysis in fig. S10 in the main manuscript, including the p-val=1.0 case, to illustrate dependence on the p-value threshold used e.g. one panel showing middling results such as (d) or (f). Also include clear discussion of the potential role of selection bias in the results – this will be helpful for anyone else wishing to use this method. It would also be interesting to quantify the average underestimation of link strength in the fig. S10 p-val=1.0 case e.g. in percentage terms – this may not be the same as the underestimation of causal link strengths, but it seems like a useful diagnostic.

We have modified the revised manuscript following the suggestion of the anonymous reviewer. We have now included former Fig. S10 as Fig. 7 in the revised manuscript. The figure is described in lines 431-440: "Finally, in Fig. 7 we check that the underestimation effect does not depend on the chosen p-value threshold (see also Supplementary Material, Text T1 and Fig. S9). We demonstrate that even using non-causal β values, i.e., those β values calculated for grid points where the p-value > 0.1 and therefore not representing causal links in Figs. 3, 5 and 6, the underestimation by SEAS5, though less strong, still dominates all causal maps (Fig. 7). We would like to highlight that using a p-value =1.0 (i.e. no threshold), up to 80% of the obtained β values do not represent causal links. Nevertheless, a peak between the 80th and 100th percentiles is still clearly visible also for very high p-values and the number of β values that fall above the 80th percentile (which should be 20% in case an underestimation effect of the β_{ERA5} is not present) is still found between 30% and 49% even in the most extreme case (p-value=1.0). Thus, we can confidently state that qualitatively, the underestimation of β_{ERA5} by SEAS5 does not depend qualitatively on the chosen p-value. However, the exact magnitude of the underestimation effect, i.e. the obtained β values for each p-value thresholds, can vary."
Note that small changes in the percentages reported in the newly added Fig. 7 (with respect to those shown in Fig. S10 in the previous version of the revised Supplementary Material) are due to the correction of a mistake in the code. Specifically, the definition of the quantiles was biased towards higher values do to the assumption that all grid points had the same number of significant beta values (1000). This is however

not true in some areas. This correction changes the numbers reported in the previous paragraph from "24% to 42%" to "30% to 49%", thus further sustaining our thesis. This correction also reflects on Figs. 5 and 6, where minor quantitative changes are visible in some areas. The correction does not affect the qualitative interpretation on any of the reported results.

Other comments

1. L189-194 It would be useful to clarify here if the time series for the two MCA modes are independent, or if not what their correlation is (c.f. the above discussion of sampling variability in diagnosing biases in the regression relationships).

We have checked the correlation coefficients between different combinations of the ERA-S MCA modes, which are reported in Table S1 in the revised version of the Supplementary Material. The results show that, as expected, by construction the Z200 and OLR time series of the same mode are significantly correlated (r ~ 0.5-0.6, p < 0.01), while when the time series of the two modes are mixed (e.g. MCA1 Z200 with MCA2 Z200) the correlation coefficients is small (r~0.03) and not significant. These results are to be expected as by construction, the MCA analysis identifies pairs of patterns which share the maximum covariance, while different modes tend to be independent from each other. By applying causal maps, however, we can filter out these dependences and check which of the two paired time series is the causal driver of the chosen atmospheric field in a certain grid point. This is now described in the text in the revised version of the main manuscript in lines 337-339 "Note that both the CGT – SAM time series pair and the NPH – WNPSM pair share a significant correlation of r ~ 0.5-0.6 at lag 0 (Table S1 in the Supplementary Material). In contrast, the SAM – WNPSM and CGT – NPH pairs show a non-significant correlation close to r ~ 0. Thus, we analyse two MCA pairs separately.".

2. L262 The comparison is not "fair" for all the reasons discussed in the last couple of reviews.

We have modified this statement by substituting the term "fair" with the term "fairer" (line 262).

3. L267 It would also be assuming that the effect of sampling variability in ERA5 on the selection of causal links is zero.

We have modified this sentence including the suggestion of the anonymous reviewer: "…(which would be the equivalent of assuming that the effect of sampling variability in ERA5 on the selection of causal links is zero and that the model does a perfect job in reproducing all observed teleconnections, …" (lines 267-268).

4. L406-7, 415-6, 641 etc. I don't think the magnitude of apparent underestimation of causal links in SEAS5 can clearly be fairly compared between spatial regions because each will have different variability characteristics, which will vary the size of any selection bias arising from the use of a p-value threshold. (It's also not clear to me from the figures that the differences in the diagnostic in the tropics are larger anyway.)

We would like to point out here, that comparison of ERA-S β coefficients to the distribution of SEAS5-R β coefficients does intrinsically take into account the different variability of each grid point. Nevertheless, we have deleted the statement regarding the overestimation of β coefficients being prevalent in tropical regions (former lines 419-422).

5. L509-10 Using different sample sizes for estimating causal link strengths is likely to introduce biases due to selection effects as discussed in previous reviews. This should at least be corrected for e.g. subsample La Nina years to equalise numbers – it should also be made that the mean Nino3.4 magnitude is close to being equal in each sample (and the values should be given).

See our answer in point 6, below.

6. L527-8 It is not shown that the differences between ENSO phases are larger than can be explained from sampling variability, so the conclusion that ENSO modifies the strength of the causal links isn't justified. E.g. show that differences in magnitudes are quite similar in independent subsamples, or use a suitable bootstrap test.

We have modified the ENSO analysis in accordance to the reviewer's suggestion. The revised version of the manuscript now contains an updated Fig. 10, in which the same number of years is used to define both El Niño and La Niña years (90 years in each case). We now also perform a bootstrapping test by calculating 1000 subsamples of 90 years each (same length as the El Niño and La Niña samples) and use these to assess whether the strength of the beta values in the El Niño and La Niña samples changes significantly from the rest of the distribution. The new results are shown in Figs. 10 and S16-S20 in the revised Supplementary Material and are described in lines 522-548 in the revised version of the manuscript:

"Finally, we investigate the effect of ENSO states on the sign and strength of the tropical-extratropical causal interactions shown in Fig. 4 and find that the effect of ENSO positive and negative phases on the SEAS5 dataset is mostly marginal with a few exceptions. We define Niño3.4 positive and negative 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 effects of SEAS5-R MCA modes 1 and 2 on Z200 field separately for 90 Niño3.4 positive and 90 Niño3.4 negative years. Note that this set corresponds to those years that fall above/below the 85^{th}/15^{th} quantiles of SST anomalies in the Nino3.4 region. These thresholds correspond to [-0.62K, +0.73K] for SEAS5 initialized on the 1^{st} of March and to [-0.69K, +0.63K] for SEAS5 initialized on the 1^{st} of May. Both pairs of thresholds are more stringent than the [-0.5K, +0.5K] threshold generally used in observational data. This definition allows us to have the same number of El Niño and La Niña years, and thus avoid differences in the magnitude of the β values due to different size of each sample. Moreover, this definition is also consistent with the inherent skewness of ENSO time series, which have larger positive peaks though for the majority of the time the index is negative.

The results for SEAS5-R MCA mode 1 are shown in Fig. 10 for both Niño3.4 positive (left column) and Niño3.4 negative years (right column), as well as for different initialization dates (1^{st} of March and 1^{st} of May). Comparing the causal maps in Fig. 10 left and right column with those in Fig. 4 shows that, in general, the spatial patterns and the sign of the causal links are not affected markedly by the sign of the ENSO anomalies. A comparison of the $\beta_{niño}$ and $\beta_{niña}$ values with the β_{SEAS5} distribution obtained from 1000 subsamples of 90 years each (consistent with the length of the El Niño and La Niña samples) is presented in

Fig. S16 (see Supplementary Material) and shows that $\beta_{niño}$ and $\beta_{niña}$ values tend to fall in the middle of the β_{SEAS5} distribution. One noteworthy exception is presented by the $\beta_{niño}$ and $\beta_{niña}$ values for the SAM → Z200|CGT link in the tropical central Pacific (red cross in Figs. 10a-d). While for El Niño years, the causal effect of SAM on the tropical central Pacific is positive and markedly strong ($\beta_{niño}$ ~ 0.2-0.3, Figs. 10a,c), during La Niña years the causal effect is almost absent or weaker in strength ($\beta_{niña}$ ~ 0.1, Figs. 10b,d). In contrast, $\beta_{niño}$ and $\beta_{niña}$ values for the SAM → Z200|CGT link over the Maritime Continent regions (blue cross in Figs. 10a-d) show the opposite result: a strong negative causal link is found during La Niña years ($\beta_{niña}$ ~ -0.2, Figs. 10b,d), while during El Niño years this causal link is completely absent (Figs. 10a,c).

Causal maps for MCA mode 2 and for the ENSO years versus neutral years (i.e. analysing El Niño and La Niña years together and separating them from neutral years) are displayed in Figs. S17-S20 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 qualitatively, with the only exception being the ENSO region itself, where El Niño years see a more prominent causal effect over the tropical central Pacific and La Niña years over the Maritime Continent."

7. Figs.7,8: Regarding the response to my previous comment that the ERA-S values in these plots seem to have shifted since the first submission and it's not clear why, the plots the authors say are from their first submission are not the same as those in my copy. Perhaps it's due to a change in the analysis method that I've lost track of. Here are the plots in the current manuscript (left) and in my copy of the first submission (right) where the relative position of the ERA-S values looks to have shifted considerably.

In our response to the third round of reviews, we state that "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." This is correct, as we showed in the figures created appositely in point 9 of our response. However, the reviewer is here referring to the first (original) submission, where indeed fig. 7 and 8 differ from those shown after the first round of reviews. This is due to the fact that, while in our initial submission we used 60-year long subsamples, after addressing the first round of reviews we then use 24-year long subsamples, as suggested by the anonymous reviewer. This change in the number of years used in the subsamples thus explains the differences in the shape of the PDFs between the original submission and all following versions.

More minor comments

8. L248 The false discovery rate used here needs to be stated. I think this is still leftover from my first review. (The authors have added a statement explaining what the term means, but not what the value of the FDR of their approach actually is. Or is it meant to be the same as the "significance threshold", which usually means something different?)

The false discovery rate value mentioned in lines 245-248 is the significance threshold value mentioned in the same paragraph.

9. L331 I think the SEAS5 MCA modes being used from here are those calculated by projecting the ERA-S MCA patterns onto SEAS5 fields - it could do with being stated clearly.

We have clarified that the abbreviation "SEAS5-R" refers to the MCA modes obtained by projecting the ERA-S MCA patterns onto SEAS5 fields in lines 324 "Then, to visualise the equivalent SEAS5-R MCA modes (where SEAS5-R depicts the MCA modes calculated by projecting ERA-S MCA patterns onto SEAS5 fields), …" and 331-332 "In the reminder of this paper, we will analyse the causal effect of SEAS5-R MCA modes shown in Fig. 4.".

10. L380 I think "changes" -> "differences" would capture what is meant.

We have included this suggestion in the revised manuscript (line 386).

11. L396-7 I'm confused here because it says the sign of beta sometimes doesn't agree, but in L384 it says the sign is constrained to match that in ERA-S. Did it mean the direction of the link is fixed?

We thank the anonymous reviewer for giving us the opportunity to clarify this point. We have now included the explanation as to why this may seem an incongruency, "Note that fixing the set of parents implies that the β coefficients are calculated by regressing on the same set of parents used for ERA-S, however in SEAS5 the sign and strength of the β coefficient can still vary." (lines 403-405).

12. L397 "Thus" implies the following sentence is somehow implied by the previous sentence, but I don't see the link.

We have included this suggestion in the revised manuscript (line 405).

13. L482 I think S11 should be S13.

We thank the anonymous reviewer for spotting this mistake, we have corrected the Figure number in the revised version of the manuscript (line 497).

14. L506-15 I think this is using SEAS5 data - it should be said clearly.

We have included this suggestion in the revised manuscript (line 522-524).

15. L545-6 I suggest putting "effects of" before "actual physical mechanisms", to distinguish from saying that the causal maps themselves necessarily have a simple physical explanation (which would take more work to show than is presented here).

We have included this suggestion in the revised manuscript (line 565-566).

16. L564 I suggest "biases" -> "time-mean biases" (to distinguish from causal coefficient biases)

We have included this suggestion in the revised manuscript (line 584).

17. L642 I still don't think “meaningful” is the best choice of word as I think it implies that there is a simple physical interpretation of the patterns. I think it would make more sense just to say the results indicate SEAS5 represents the diagnosed links qualitatively well.

We have modified this sentence accordingly with the suggestion of the anonymous reviewer “Thus, our confidence in the ability of the SEAS5 forecasting system to qualitatively represent those causal links identified in ERA5 reanalysis is increased.” (lines 662-663).

18. L643 I think putting “qualitatively” before “correctly” would better reflect what is meant here.

We have included this suggestion in the revised manuscript (line 662).