Point by point response – reviewer #2

We thank the anonymous reviewer for their further suggestions of how to better compare the strength of the causal effect between the two datasets. We have followed the reviewer's suggestion to impose the same causal parents detected in ERA5 also in the SEAS5 ensemble. Our new results confirm that SEAS5 has a pronounced tendency to underestimate the strength of the beta values (details can be found in our response below). Moreover, we have partly altered the structure of the paper, both to accommodate the new results and also to better separate information belonging to the Method section from the results. We would like to highlight that now only Fig. 4 presents results obtained using all 600 years available in SEAS5 at once. These results are only meant to give the reader a first intuitive idea of what the overall spatial pattern and sign of the causal links would look like. However, the strength of the beta values is now compared strictly by using causal maps obtained from a set of 24 years. This is true both for Figs. 5 -6 and Figs. 7 – 8. We also want to highlight that the two subsampling experiments, which we now term experiment A and B in section 2.4 of the revised manuscript, present two complementary approaches for comparing ERA5 and SEAS5 causal maps (see point -1- in our response below). Moreover, we further checked all concerns regarding the MCA patterns (see point -5- in our response below).
We are confident that we have implemented all the suggestions of the reviewer to the best of our understanding and that the proposed changes have helped increase our confidence in the original message. We hope that the reviewer will find that the revised paper satisfies their concerns. A point-by-point response is provided below.

-1-
From the authors' response and the revised manuscript, it seems that they have not been able to address the first major concern of my previous review, that trying to use their method to compare datasets of different sizes seems liable to give erroneous results.
This follows from using statistical significance thresholds to select predictors in the analysis, which will tend to isolate larger magnitude signals in the smaller dataset (ERA5 here), which looks to me like it will tend to produce a bias in the differences between datasets. This is a critical part of the work and it is necessary to show that the presented method of diagnosing model biases works reliably. Once it is shown that the method being applied is reliable, it could be quite valuable for better understanding model biases, so I do think it's worth getting this right.
One problem that comes to mind is that it sounds like predictor variables are only being included in the statistical models if their regression coefficients are found to be statistically significant, and this is differing between datasets. Then it is difficult to compare the coefficients i.e. if in one dataset the model used is z(t) = A*x(t)+B*y(t) and in the other dataset the model used is z(t) = A'*x(t), the coefficients A and A' are not clearly comparable in general (if x and y are correlated, for example). It may be necessary to decide to use a common set of predictors for the regression for each dataset for each predictand, in a way that isn't biased towards either dataset.
The multiple linear regression results shown in the response do not clearly address the point about potential bias because, again, for datasets of different sizes, MLR coefficients will tend to be noisier and larger in amplitude for a smaller dataset (if the noise is large enough to change the sign sometimes, as it probably is here). It's not clear to me from the maps how the distributions of ratios of coefficients look. It also doesn't address the problem that the main numerical results in the paper comparing ERA and SEAS5 may be biased.

We have now explicitly addressed this major comment by directly implementing the suggestion of the anonymous reviewer. We save the information on which set of causal parents is identified for each grid point in each causal map produced with ERA5 and then run a subsampling experiment for

SEAS5 (using 24 years for each subsample) and calculating the causal effect (beta coefficient) by imposing exactly the same set of causal parents as identified in ERA5. Thus, each sample in the subsampling experiment has exactly the same number of years as the ERA5 dataset (1993-2016) and the causal links are not detected in SEAS5 but imposed from our previous knowledge of those detected in ERA5. As a result, we obtain a set of 1000 causal maps for each of the 8 analyzed links. In this new set, the same grid points that were significant in the ERA5 dataset are shown (Fig. 3) thus allowing a comparison to be made of the strength of the ERA5 beta values with the distribution of 1000 SEAS5 beta values for each grid point. The new results are shown in Fig. 5 and 6 of the revised manuscript. In the left column of both figures, the average beta values coefficients are shown, i.e., the beta values coefficient for each grid point is obtained by averaging on the 1000 beta values available from the subsampling ensemble. Qualitatively, the average strength of the beta values coefficients is very similar to that shown in Fig. 4 (where all 600 years are used together). However, the average value does not take into account the spread of the ensemble, to which we want to compare the ERA5 beta values coefficient. See our next comments to see how we now tackle this aspect. We describe these changes in the revised manuscript in lines 254-273:

"We perform two sub-sampling experiments: experiment A aims to better understand differences in the strength of causal links between ERA-S and SEAS5-R, while experiment B evaluates the spread inside the SEAS5 ensemble. For each subsampling experiment, we select 1000 samples of 24 years each (for each year, one ensemble member is randomly selected out of the 25 available 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). Reducing the length of the time series in this way increases the variability and hence lowers the significance of the obtained β values. However, this should not by itself lower the strength of the β values themselves. Thus, a priori, we might expect fewer regions to show a significant β value in a smaller dataset than in a larger one, but not a difference in the strength of the β values. Hence, this 1000-ensemble member subsampling experiment allows us to evaluate the distribution of β values around their mean value and to compare it to the ERA-S values of reference. 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 < 0.1 are retained.

In experiment A, we impose the set of causal parents which have been detected as significant for ERA5 in the SEAS5 ensemble and then calculated the corresponding causal effect in SEAS5. In this way, we provide a fair comparison between the strength of β_{ERA5} and β_{SEAS5}. In other words, SEAS5-R causal maps obtained from experiment A will show significant causal links for the same grid points as in ERA5 causal maps, but the sign and the strength of the β coefficient will vary following the physical representation on these teleconnections in the SEAS5 dataset. In contrast, in experiment B we let the PCMCI algorithm identify the causal parents in SEAS5 and then estimate the causal effect. Thus, new causal links that were not detected in ERA5 may appear while others, that were significant in the reanalysis dataset, may disappear. Analysing the results for these two subsampling experiments will enable us to compare the strength, sign and location of tropical – extratropical teleconnections between SEAS5 and ERA5."

and lines 379-391:

"

To assess the difference in strength between SEAS5-R and ERA-S β values, we use the causal maps obtained from subsampling experiment A, as described in Section 2.4. To make sure that we properly assess changes in the strength of the causal effect between the two datasets, we (a) use the same number of years as in ERA5-S (24 years) and (b) fix the sign and significance of the causal links for each grid point to be the same as those detected for the ERA5 dataset (Fig. 3). As an example, if we analyse the effect of SAM and CGT on the Z200 field in the ERA5 dataset for the grid point at 20°N, 16°E, we find a significant positive β coefficient for the link SAM → Z200|CGT and a significant negative β for link CGT → Z200|SAM (plus a certain self-influence of the Z200 time series on itself). Then, we calculate the corresponding β coefficients for each of the 1000 SEAS5 subsamples by imposing the same causal

parents as those found in ERA5. This way, we obtain a set of 1000 β coefficients, for each analysed causal link and for each grid point, which would be visualized in 1000 causal maps (see Fig. S4a in the Supplementary Material). The causal maps obtained by averaging these 1000 causal maps are shown in Figs. 5a,d,g,j for links CGT→ Z200|SAM, SAM → Z200|CGT, CGT→ OLR|SAM and SAM → OLR|CGT, respectively. In general, the strengths of the β values obtained in Fig. 4 tends to be of the same magnitude as the mean β values obtained in the 1000 subsamples, further suggesting that SEAS5 underestimates the strength of the β coefficients.

"

-2-

It also occurs to me that the key quantity Delta = beta_SEAS5/beta_ERA5 has expected value <Delta> that is not clearly equal to <beta_SEAS5>/<beta_ERA5> – indeed, <Delta> may not be well-defined. So <beta_SEAS5>=<beta_ERA5> does not clearly imply <Delta>=1. This is another reason why the presented results are not clear evidence of SEAS5 underestimating the coefficient magnitudes. A metric needs to be used where its expected value is clear in the case that <beta_SEAS5>=<beta_ERA5>.

In the new Figs. 5 and 6 we now use a different way to compare SEAS5 and ERA5 beta coefficients than the Delta_beta method used in the previous version of the manuscript. We now have available a distribution of a 1000 beta coefficients for each grid point and we calculate the 0^{th}, 20^{th}, 40^{th}, 60^{th}, 80^{th}, and 100^{th} percentile. Then we categorize the ERA5 beta coefficient as falling in a certain category and plot the results (Fig. S4 illustrates this in the revised Supplementary Material). The results are shown in the middle column of Figs. 5 and 6. The majority of the grid points show colors ranging from orange to dark red, meaning that for those grid points, the ERA5 beta coefficients falls above the 80^{th} percentile of the SEAS5 beta value distribution. The percentage of grid points falling in each category is shown in the histograms in the right column of both Figs. 5 and 6. We describe these changes in the revised manuscript in lines 392-413:

"Next, we compare the β_{ERA5} values to the absolute values of the 1000 β_{SEAS5} obtained in the subsampling experiment A. We use the absolute values as we intend to compare the strength of the β coefficients and not the sign. Notably, the average percentage of β values in the subsampling experiment for which the sign does not agree with that of βERA5 is ~20%. Thus, for each grid point we calculate the probability density function describing the distribution of the 1000 β_{SEAS5} and estimate the 0th, 20th, 40th, 60th, 80th and 100th percentiles (Fig. S4b in the Supplementary Material). Then, we categorize the β_{ERA5} as falling in one of the selected quantile ranges or above/below the maximum/minimum of the distribution. For example, in Fig. S4 the β coefficients for the grid point at 20°N and 16°E are analyzed and the β_{ERA5} value is shown to fall beyond the 100th percentile, meaning that for that grid point, all of the 1000 β_{SEAS5} coefficients are smaller in strength than the observed βERA5 value. The overall results for the analysed region are shown in Figs. 5b,e,h,k, where yellow shaded grid points indicate that β_{ERA5} coefficients fall in the middle of the β_{SEAS5} distribution, while blue/red shaded grid points indicate β_{ERA5} coefficients falling in the lowest/uppermost tails of the βSEAS5 distribution.

In general, for MCA1 orange/red colours dominate all four causal maps among which tropical and high-latitude regions show the most underestimated β_{SEAS5} if compared to the β_{ERA5} reference point (Figs. 5b,e,h,k). Histograms describing the percentage of grid points falling in each category for each causal map are shown in Figs. 5c,f,i,l. Results show that β_{ERA5} values for the vast majority (~70%) of the grid points fall above the 80th βSEAS5 percentile, and about 5-20% above the 100th percentile, meaning that SEAS5 is never able to reproduce the observed β_{ERA5} value for those specific grid points. Among the few regions where the strength of β_{SEAS5} is overestimated there are the Middle East, Pakistan, Iran and parts of the southern Arabian Peninsula and the Arabian Sea. Similar results are shown for MCA2 (Fig. 6), with β_{ERA5} values exceeding the 80th percentile for 60-70% of the grid points. The percentage of grid points exceeding the 100th percentile is however reduced if compared to MCA1,

with only 1-10% of the grid points showing β_{ERA5} values exceeding the maximum β_{SEAS5} of the distribution. As for MCA1, MCA2 features strongly underestimated β_{SEAS5} values in tropical regions, while in the mid-latitudes, SEAS5 better reproduces of observed strength of the β coefficients."

-3-
Overall, I think it will be difficult to demonstrate the reliability of the method purely through reasoning, given its complexity (with multiple seemingly arbitrary thresholds etc.). I think what is required is a set of tests with data where the error in the causal effect (beta) coefficients is known and a demonstration that the method being applied here gives the correct results. One test would be to compare one member of the SEAS5 ensemble with the rest of the ensemble using exactly the same method, repeated for each ensemble member – there should not be any systematic error, and histograms of the chosen metric like those in figs. 5 and 6 should be centred around a value corresponding to this situation. Another test would be to impose an error of the size claimed to exist between ERA5 and SEAS5 i.e. a factor of 2/3 - this could be done I think by taking one SEAS5 ensemble member, scaling the part of the time series at each location associated with MCA1 and MCA2 by 3/2, then checking that the histograms of the diagnosed metric for the rest of SEAS5 do centre around the anticipated value. Results from these tests should be presented, perhaps in the supplementary information.

We thank the reviewer for providing this suggestion of how to evaluate the error for the beta value. As explained in our answer to point -2-, we have chosen to evaluate the spread of beta value in the SEAS5 ensemble and to use that as a metric (after calculating the quantiles) to compare ERA5 and SEAS5 beta coefficients.

-4-
It would also make sense to make clear the method for comparing the two datasets in section 2, where these tests could be explained – currently, it's hard to follow what is being done with the method being written amongst the results.

We have moved the information regarding the subsampling experiments and the technique used to the method section 2.4. See lines 253-273 in the revised manuscript (also reported in point -1- in this document).

-5-
It also needs to be confirmed that the MCA1 and 2 time series in ERA5 and SEAS5 are comparable. From figs. 3 and 4, it looks like the spatial pattern of the modes for SEAS5 may be generally higher in amplitude (though I know the method for plotting these differs for each dataset – I do not understand the response to my previous comment suggesting to use the same method for each, which says to see the response to "general comment 1", when this is about a different issue). If the modes were larger in SEAS5, this could result in the beta coefficients being smaller.

We thank the reviewer to give us the opportunity to further clarify this point. Indeed, while Fig. 3 (left column) shows the actual MCA patterns as obtained in ERA5-S, Fig. 4 (left column) shows the composites of the time steps for the SEAS5-R time series (i.e., the MCA time series obtained by projecting the ERA5-S MCA onto the SEAS5 dataset) which exceed 1 s.d. First, we would like to point out that both the ERA5-S and SEAS5 MCA time series are obtained in the same way, in that both are the product of calculating the dot product of the ERA5 MCA patterns (shown in Fig. 3a,d,g,l) on the respective fields. Now we show that the mismatch in magnitude of the anomalies between the patterns shown in Fig. 3a,d,g,l and Fig. 4a,d,g,l is greatly diminished if the data are plotted using the same method. Thus, we provide the composites of MCA time steps > 1 .s.d also for the ERA5 MCA and show that the magnitude of these anomalies is very similar to that shown in Fig. 4a,d,g,l. This is

shown in Fig. R1 in this document: in the first row we report MCA1 patterns as shown in Fig . 3a,d; in the second row we recalculate these pattern as the composites of MCA time steps > 1 .s.d and in the third row we show the same SEAS5-R pattern as shown in Fig. 4a,d. Similar results are obtained for MCA 2. We have added this information in the revised version of the manuscript, see lines 325-328: "Note that the difference in the magnitude of the anomalies shown in Fig. 3a,d,g,j and Fig. 4a,d,g,j is greatly diminished if ERA5 MCA patterns are plotted with the same methods, i.e., plotting composites of time steps with the MCA time series values higher than 1 standard deviation (see Fig. S3 in the Supplementary Material)."

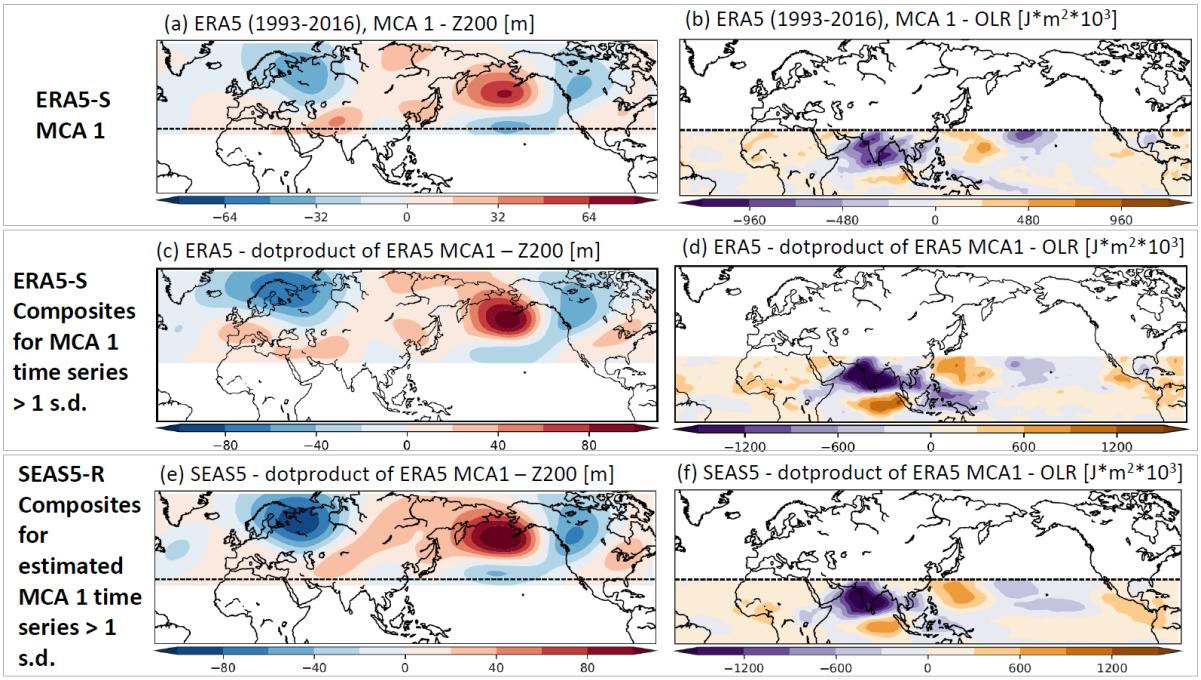

Figure R1. Panel (a) ERA-S MCA mode 1 Z200. Panel (b) ERA-S MCA mode 1 OLR. Panel (c) ERA-S MCA mode 1 Z200 calculated as the composites of ERA5 Z200 for MCA1 time steps > 1 .s.d. Panel (d) same as panel (c) but for OLR field. Panel (e) SEAS5-R MCA mode 1 Z200 calculated as the composites of SEAS5 Z200 for MCA1 time steps > 1 .s.d. Panel (f) same as panel (c) but for OLR field.

-6-

Something else that confused me is that the distributions of beta coefficients shown in figs. 7 and 8 seem to be centred on a value very close to zero – I couldn't make sense of why this would be, when the mean beta values in the regions concerned appear quite substantial. Or are these distributions of differences from the mean?

We thank the reviewer for highlighting the difficulty in interpreting these plots. We now have revised both Figs. 7 and 8 to show the actual value of the beta coefficient instead of the standardized ones. Moreover, we now only overlap the value for ERA5-S beta values, as the values for ERA5-L beta coefficients come from causal maps obtained from a longer time period (42 years) and are therefore not one-to-one comparable.

I've not gone through the rest of the revised manuscript and have not yet made a judgement about that – I can do so when this concern is addressed, if my review is still wanted.