Review of Di Capua et al.

The manuscript analyses tropics-NH teleconnection patterns on weekly time scales in boreal summer in ERA5 and the SEAS5 seasonal forecast system and mainly attempts to understand biases in the latter. It uses causal effect analysis, a method that is potentially stronger at identifying biases related to cause and effect between parts of the climate system than standard correlation analysis.

I am positive about the attempt to use causal effect analysis in model evaluation, which I think could be very beneficial in general. The paper seems quite thorough in a number of respects compared to others in the literature (e.g. I like the use of two different SEAS5 datasets based on different start times and consideration of multiple testing) and the writing style is mostly clear.

However, I think the way the analysis has been done in this paper makes interpretation difficult, and I think major revisions are necessary to produce useful results. I also think the analysis of the role of mean state biases and ENSO doesn't produce clear results and these sections are overly long – cutting this down would improve the quality by making the paper more succinct I think (many of the best papers are brief!).

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. 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.

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.

L439 It seems like this part could do with references regarding waveguides.

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?

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 summarised quite briefly too.

L479-80 "If a dependence is found…" – I don't follow this.

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 initialised on May 1.

L499-502, L514-9 When talking about similarities over such small regions, it's not clear that this isn't noise.

Other comments

1. Abstract - it would be good to have a brief summary of the quantitative size of the most important results.
2. L71-3 This final statement about the superiority of dynamical forecasts doesn't seem clearly justified.
3. L143 What's the process of removing the interannual variability, seasonal cycle and any long-term trend?
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.
5. L188 It could be helpful to define the ERA-S and ERA-L periods earlier where the data is described.
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.
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.
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?
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.
10. L272-4 There seems to be lower OLR over India in MCA2 for both ERA5 and SEAS5 to me.
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.)
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?
13. L290 It could do with saying here what lag these maps are for.
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.
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.
16. L297 and North Asia?
17. L301 fig. 3c rather than 3e?
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.

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.
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.
21. L423 I suggest “bias” -> “mean state biases” in the section title.
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.
23. L558 Is “negative bias” referring to the CEN coefficients with respect to that mode?
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.
25. L569-71 Again, this doesn’t clearly follow.
26. L574 “quite satisfying” is quite subjective – if you mean the signs of the coefficients seemed similar in the datasets, then please clarify.
27. L599-600 I suggest "if EC-Earth behaves similarly to ERA5".
28. L602 “fairly well represented” – magnitude does matter!
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.
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.
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)
32. L623-4 Where was this result made clear?
33. Fig.3 – the caption should explain how the causal maps should be interpreted, or refer to the text.
34. Fig.3 - Some explanation should be given for the boxes drawn on the maps.
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?
36. Figs.5,6 - The "mean" and "std" values should be explained.