Review of Di Capua et al. revised manuscript (Jan '23)

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 focussed 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 50^{th} 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.

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.

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.
3. L327 It's good to see the MCA-1 results looking consistent. I think the same should be shown for MCA-2.
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.
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.
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.
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.
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.
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.
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".
11. L618 As before, I don't see clearly larger biases in the tropics.
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?
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.