Review of Di Capua et al. revised manuscript

The authors have provided substantial extra analysis that helps in assessing the main issue in the previous review of the concern that the methods are potentially biased, particularly in figs. S9 and S10, showing sensitivity tests of histograms of the percentile rank of ERA-S results amongst the SEAS5 resampled dataset.

Overall, I think the analysis indicates that SEAS5 does probably underestimate the regression relationships between the variables being considered, though not with a very high degree of confidence. The magnitude of the underestimation seems very uncertain due to selection bias effects and sampling variability. I think the results can now potentially be published if the conclusions are appropriately modified to reflect this, the results presented in the main manuscript are made more complete and other concerns are addressed below.

(This is contingent on the time series of the two MCA modes being quite independent of each other, something I'm not sure about as I'm not that familiar with MCA, else sampling variability may explain the key results. I suggest that this is shown so the editor can be satisfied of this.)

Analysis relating to potential selection bias in causal relationships based on ERA5
If I understand correctly, in fig.S9 the causal relationships identified in ERA5 using the method in the main manuscript are used, and so the results here are still potentially affected by a selection bias. In fig.S10, it looks like the statistical significance threshold for identifying causal relationships is varied, and more regressions between pairs of variables are included as the p-value threshold is raised. I think the only result for which it is possible to be confident that there is no substantial selection bias is that with p-val=1.0 as the threshold in fig.S10.

The authors argue that "the underestimation effect does not depend on the chosen p-value threshold". However, this is very different from my impression, since from fig.S10 it looks like the attained histograms vary a lot depending on the choice of p-value threshold, with apparent systematic underestimation of the causal coefficients in SEAS5 reducing strongly as the threshold is relaxed. This could be because SEAS5 underestimates the most strongly causal relationships more than other regression relationships, which I think is the authors' view, but it could also be because the difference across histograms is caused by the selection bias, which would become stronger with stricter thresholds. I don't see a way to distinguish between these possibilities based on the presented analysis. It was not clear previously that this is how the results would come out – it could have been that the p-value threshold made less difference, indicating the selection bias was limited in size.

For the minimally biased (I think) p-val=1.0 curve in fig. S10, the histograms do still show magnitudes of ERA-S values tending to be at the higher end of the percentile distribution, falling into the top 20% of SEAS5 values 24-42% of the time across the different causal relationships examined. So this gives an indication of support for the idea that SEAS5 generally underestimates regression relationships as opposed to the opposite.

However, to make a confident conclusion requires analysis of whether sampling variability could have produced this result. This is not completely clear to me from the given analysis. The consistency across different pairs of causal relationships does give an indication that the results are not only due

to sampling variability, but the different pairs share common variables and are not fully independent of each other. Presuming that the time series of the two MCA modes are quite independent (as said above, I'm not actually sure if this is the case), then there are perhaps ~3-4 "equivalent independent" relationships depending on how independent are Z200 and OLR (at a guess based on the relationships shown in fig.3). Getting this degree of consistency across ~3-4 independent results is quite unlikely to happen by chance, though it may not be at the typical 95% confidence level threshold.

So overall, I'd say the results show a good case for believing that SEAS5 systematically underestimates regression relationships between the variables being considered, and it seems more likely than not that this is also true for the variables picked out by a causal discovery algorithm. The SAM → Sahel Z200 link, which is one of the links analysed in more depth, does also look likely to be underestimated. The magnitude of the underestimations is hard to identify, however, due to the potential role of selection bias when p-value thresholds are applied. If the authors wanted to justify making a stronger conclusion, this could potentially be done using something like a bootstrap test based on replacing ERA-S with samples from the SEAS5 distribution, repeating the analysis many times and checking that deviations from uniform distributions like those shown in S10 for p-val=1.0 are rare enough. But it's not necessary if the conclusions are stated with the appropriate degree of uncertainty.

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

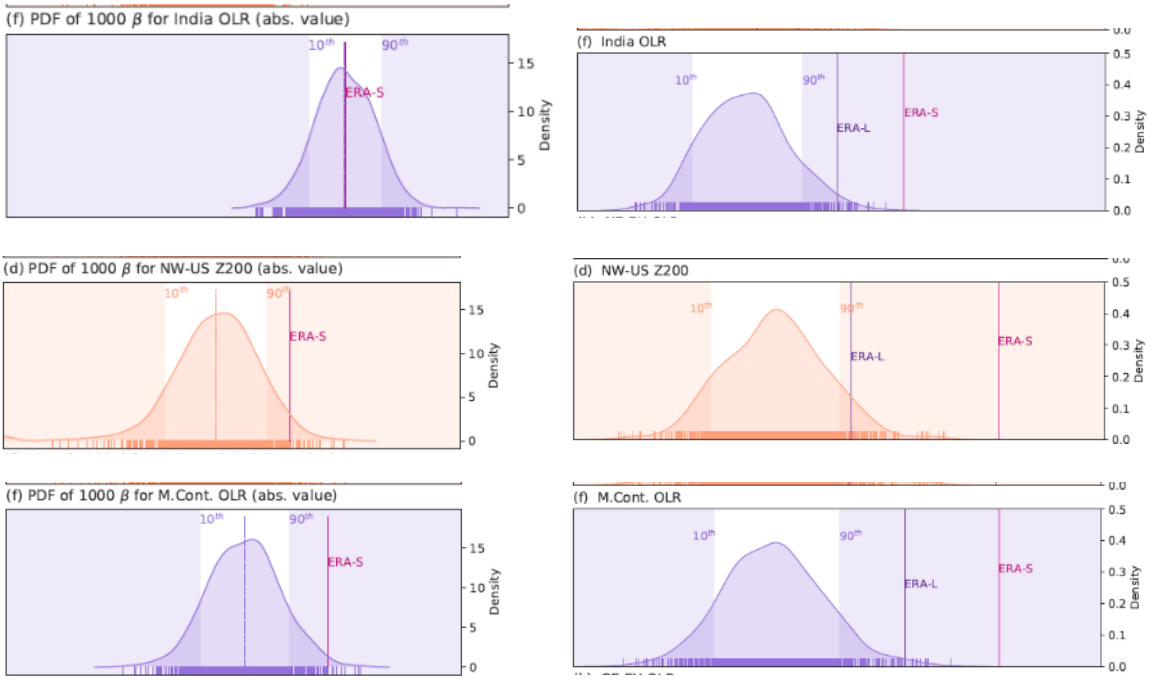
- Remove quantitative comparisons of magnitudes of causal links in ERA and SEAS5, since the possible range of differences between them seems very large e.g. at L48-49 in the abstract, and other places.
    - Referring to quantitative results of particular analyses is still encouraged – it just shouldn't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

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).
2. L262 The comparison is not "fair" for all the reasons discussed in the last couple of reviews.
3. L267 It would also be assuming that the effect of sampling variability in ERA5 on the selection of causal links is zero.
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.)
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).

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.
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:



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?)
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.
10. L380 I think "changes" -> "differences" would capture what is meant.
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?
12. L397 "Thus" implies the following sentence is somehow implied by the previous sentence, but I don't see the link.

13. L482 I think S11 should be S13.
14. L506-15 I think this is using SEAS5 data - it should be said clearly.
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).
16. L564 I suggest “biases” -> "time-mean biases" (to distinguish from causal coefficient biases)
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.
18. L643 I think putting “qualitatively” before “correctly” would better reflect what is meant here.