Review of Di Capua et al. revised manuscript

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.

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>.

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. 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.

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.

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?

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.