The authors provide a Complex Networks perspective on climate model evaluation. The analysis is based on the d-MAPS methodology. The d-MAPS method has been used in climate science as a powerful way to investigate local and non-local connectivity patterns in spatiotemporal climate fields. And such method has been adopted for climate model evaluation and to study climate regime shifts in paleoclimate simulations.
The authors did not just adopt the d-MAPS framework but nicely contributed to it by proposing, adding and considering different new metrics. The first goal of this paper is then to improve the d-MAPS framework and to broaden its applications. The new metrics considered can quantify nonlinear relationships and to investigate linkages among multivariate fields. Moreover, several metrics for network comparison are also proposed, adding to the metrics initially proposed by Falasca et al. (2019). At a second step, the authors adopt such newly proposed metrics to investigate sources of differences and biases in CMIP6 models.

I believe that the paper can be suitable for publication in ESD following some revisions. As it is right now, I find it difficult to follow. The tools proposed are very interesting and powerful and there is lots of potential in their application. Novelty of such approach for climate model evaluation should be discussed in depth. What are we learning in terms of CMIP6 ensemble? What are the new tools showing us? This should be better discussed starting from the abstract. Part of this may be improved already by being clearer and more concise in few sections (such as Section 4.3.2). I provide comments below, ordered by sections.

- Abstract

  The novelty of this work is proposing and implementing a series of different tools and metrics in the d-Maps framework. This is powerful as it further broadens the tools available in the d-MAPS methodology. This should be clearly stated in the abstract, currently is not. Please, revise the abstract by specifying that this is not only an application of d-MAPS, but it is an actual contribution to the overall d-MAPS framework. In fact, some modifications, such as the Spearman's Rank correlation in the domain identification step may result in very different results from the usual Pearson correlation in case of strongly nonlinear associations. So, I suggest to (a) rephrase the abstract in terms of the true novelty of the paper and (b) add some sentences on the results obtained with such new tools. Right now, there is only one line…what are the main results in the context of model evaluation? Which are the best models in terms of their network connectivity? Where do the models tend to fail?

- Introduction
  - Line 38: El Niõ -> El Niño
  - Line 46: I suggest adding the paper of Tantet and Dijkstra to the references https://esd.copernicus.org/articles/5/1/2014/esd-5-1-2014.pdf

- Data
  - What is the temporal resolution? I don't see it written (apologies in case I missed it)
  - The paper uses two reanalyses products: CERA20-C and 20CRv3.
    In the data section can we see that CERA20-C is abbreviation for Coupled Reanalysis for the 20th century. Maybe I missed it, but I do not see anywhere that 20CRv3 is abbreviation for NOAA-CIRES-DOE Twentieth Century Reanalysis version 3 (also is not obvious). Please add the "20CRv3" abbreviation.
  - Line 88: was there a reason to remap the two fields to two different resolutions? Why not both at 2.25 degrees?

- Methods
  - Domain identification

    Line 184-191: there is no need in this case to adopt the scheme proposed by Falasca et al. (2020). That scheme has proven to be useful to identify abrupt (and non-abrupt) shifts in climate variability at paleo-scales, which is not the goal of the submitted paper.

  - Network of domains

    In Line 208 the author say that every possible lag is analyzed from -*L* to *L*. What is *L?* Is it the length of the time series? Please clarify.

    - 3.3 Distance covariance/correlation.

    In line 264: "The correction of autocorrelation is a rather unrobust statistical technique…statistically advantageous". Is there a reference for this claim? If yes, please add the reference. If not elaborate on why that is the case.

    - 3.3.1 Distance multivariate/multicorrelation

    This is a nice and interesting addition to the d-MAPS framework.
    Results are discussed both in the case of distance covariance and distance multivariate. It is my understanding that in the case of Distance covariance/correlation, the authors considered all possible lags; while in the distance multivariate/multicorrelation they only consider *instantaneous* connections. Is this correct? If this is the case, can the author please add this point in the discussion of the metrics. At the risk of being repetitive I think it is useful to underline that (a) results with Distance covariance/correlation are computed using lags and that (b) results with distance multivariate/multicorrelation are instaneous.

Adding to this, is there a reason for this choice? I understand the choice, as the problem may become easily intractable when considering *lagged* higher order dependence. This seems to be said in line 474: I suggest to be clear on this point in Section 3.3.1.

I find this metric very interesting. I am wondering on what are possible some connections/relationships with the following "Tripoles" concept/metric proposed in KDD: https://dl.acm.org/doi/pdf/10.1145/3097983.3098099
And used in this study:
https://journals.ametsoc.org/view/journals/clim/30/1/jcli-d-15-0884.1.xml?tab_body=pdf

- 3.4 Comparison of networks with structural similarity…

Line 313: "Firstly our links are undirectional because distance correlation is much less sensitive to temporal lag than Pearson correlation, such that there is no distinct temporal ordering".
I am confused about this sentence. My understanding was that the network is a *direct* graph, inferred by considering *lagged* relationships. In fact by looking at Figure 1(c-e) the caption confirms "Maximum lagged distance links between…"
However, in the adjacency matrix proposed links are *undirectional.* Why this choice? Why not setting the entries in the adjacency matrix as the maximum link? This is not clear and should be discussed. Links in climate networks can be lagged as they reflect *non-local* connectivity. Such linkages are driven by atmospheric/ocean phenomena such as (for example) Rossby or Kelvin Waves. Therefore, physically many of such links cannot be instantaneous. Moreover, why are *all possible links investigated* (see line 208) if the distance correlation is much less sensitive to lags?
Please discuss this point. To help the reader would it be possible to compute one adjacency matrix using lagged-connections (for example considering the maximum lag) and the other one using instantaneous connections and show robustness? Importantly I suggest to clearly explain the choice of looking at instantaneous links for this part of the analysis, while still looking at lagged links in other sections.

Equation (9). How are the 3 constants c_1 c_2 and c_3 chosen?

- Results
  - Detrending with trend EOF

The detrending procedure proposed here is interesting. However, I find it difficult to follow the discussion when the Figures are all in the Supplemental Information. I suggest moving Figures S1 and S2 in the main paper.

The authors are convincing here in the sense of using trend-EOF to remove a forced trend. How do these spatial patterns of trends (Fiure S1(b)) compare to the simple linear detrending done for each grid cell. It would be useful to show (in the Supplemental Info) maps of slopes fitted from a simple linear regression (as usually done in climate studies) and discuss difference. This could help in making a stronger case for the use of trend-EOF.

- ## Domain Identification

Line 401: *"The map of the domains (Figure 1(a)) resembles the corresponding maps for COBEv2 and HadISST in Falasca et al. (2019) reasonably well, taking into account the different data sets and time period."* The maps in the two studies are very similar, however the similarity metric chosen here is Spearm's Rank, allowing *"for monotone, yet non-linear associations"* (line 179).
I think this comparison can teach us something about the connectivity structure of such fields. A simple conclusion here is that for the temporal resolution and fields considered time series are linearly dependent and linear methods are enough to infer local connectivity (i.e., domains). Additionally, it shows that how the trend is removed makes little impact on the overall result. I suggest to add such a discussion in the text.

- ## Networks of domains

Figure 1(c-e): these metrics encode the "*Maximum lagged distance correlation*". I believe this means that for each couple of domains distance correlations have been computed for all lags and that the maximum (and also significant?) one has been chosen? What is confusing is that, at least by visual inspection, this matrix seems symmetric. If the matrix shows *direct (lagged)* links, it should be asymmetric. Please, can you explain why is this matrix symmetric?

Line 432-435: the S. Tropical Atlantic leading the ENSO domain has been shown to be a time-dependent link in both Falasca et al. (2019) (Section 8, "Climate networks in time") and in Martin Rey paper (see https://link.springer.com/article/10.1007/s00382-014-2305-3).
So the fact that is not picked by the algorithm in the period 1901-2010 is expected.

In both the domain identification step and the network inference step please be sure to say that Figure 1 is for CERA-20C and that results for 20CR3 are reported in the Supplemental Information.

- Section 4.2.3: 3rd order interactions

Results for this section are in a table in the supplemental info. Please, move this table in the main text. In the Table caption please add description of what is dMcor and dCor to make it easier to follow.

Are the interactions reported in Table S1 just the significant ones? If yes, please add it in the caption.

- Section 4.3.1

This section describes results in Figures S4-S7. It is difficult to follow as results are in the SI. I suggest to at least add Figures S4 and S6 and S7 just for one of the two reanalyses and discuss the other reanalysis (plus AMO link in the SI). In this way you would be able to show maps of strengths in the two univariate cases plus the multivariate case. Also, it makes it easier to follow.

Figures S4-S7: are these networks direct graphs (i.e., did you include lags)?

- Section 4.3.2

The authors lost me here: (a) under reanalyses in Figure 2, we have the comparison between the two reanalyses. What are the models compared to? Are they compared to the ensemble mean of *CERA20* or *20CRv3* best estimates? It's difficult to interpret any of these results as it is not clear what they are compared to.

Figures S8 and Figure 2 should be both in the main text (and both in the same Figure).

Figure S8: difficult to understand. I suggest to describe the NQS and point-wise NQS in the case of CERA-20C and leave the comparison with 20CRv3 in the Supplement Info.

Falasca, F., Bracco, A., Nenes, A., and Fountalis, I.: Dimensionality reduction and network inference for climate data us- ing δ-MAPS: Application to the CESM Large Ensemble sea surface temperature, J. Adv. Model. Earth Sy., 11, 1479–1515, https://doi.org/10.1029/2019MS001654, 2019.

Falasca, F., Crétat, J., and Braconnot, P. Bracco, A.: Spatiotemporal complexity and time-dependent networks in sea surface temperature
from mid- to late Holocene, Eur. Phys. J. Plus, 135(5), 392, https://doi.org/10.1140/epjp/s13360-020-00403-x, 2020.