Review of "Network connectivity between the winter Arctic Oscillation and summer sea ice in CMIP6 models and observations" by Gregory et al.

General comments

This study by Gregory et al. investigates the spatio-temporal covariability of the winter Arctic Oscillation (AO) and summer Arctic sea ice and its representation in CMIP6 models with a complex network approach. This is interesting for the predictability of summer sea ice in the light of a suggested physical mechanism between the winter AO to summer sea ice, supported by the observations. The authors point out discrepancies between the covariability of winter AO and summer sea ice in observations/reanalyses and CMIP6 models, particularly also in the transition from earlier to more recent years, and hypothesize that these could be due to regional biases in sea ice thickness.

This is a carefully performed and well-presented study and a valuable addition to previous literature on mechanisms for sea ice predictability. The paper is clearly structured and well written. The methods are well explained, so that also someone who is not familiar with complex networks can easily follow and understand the results. Previous literature is thoroughly cited. Overall, I recommend the paper for publication after consideration of the comments outlined below.

We would like to thank the reviewer for their kind words and for the time they have given to review this work. Their comments have been very insightful and have certainly strengthened this manuscript. Please see our responses below.

Specific comments

- As you write in L66 ff., climate network analysis is an useful addition to more conventional statistical approaches. I find the climate network analysis very intriguing, but (being new to the method) it is not entirely clear to me why you chose it here and what is the added value to more conventional statistics. Would you expect to arrive at the same conclusions with e.g. an EOF analysis?
  This is a great question. In some specific test cases we would indeed find that the networks approach and EOF analysis produce very similar results. For example, the dominant spatial patterns of variability seen in the network strength map of one particular model ensemble member are likely to also be present in the first few leading EOF maps of that same ensemble member. The purpose of using the networks method in this paper was two-fold. On the one hand, we wanted to use this paper to bring attention to the method itself, as complex networks have become increasingly used in climate science over the last decade, although perhaps not so much in cryospheric/polar domains. Furthermore, it provides a nice initial framework which can be modified for deeper analysis of climatological interactions, such as causal inference techniques. On the other hand, the networks approach does provide some advantages over EOF analysis. For example, EOF analysis is variance greedy, so each of the (EOF) derived modes reflect the direction along which the data exhibit the largest variance, and so patterns of lower variance are inherently masked. Furthermore, it is also perhaps cumbersome to convey how each of these EOF modes are interrelated, while in the networks framework we do not mask patterns of lower variance, and we can also neatly summarise the interconnected nature between nodes(modes) with the strength and link maps. For a more in-depth response to this discussion point, we have provided additional information in the response to reviewer 2's first comment.
- To L124 ff.: Is the number of network nodes N determined by the clustering algorithm or is it prescribed? What would be a typical number for N?
  The number of network nodes N is determined by the clustering algorithm. This is specific to our chosen algorithm, which is a grid-based clustering approach whereby each node tries to continuously grow in spatial extent, so long as the correlations between all points within the node are above a specific threshold. Once the node cannot grow anymore, the next available point is selected, and the process repeated (available point meaning a grid cell that does not already belong to a network node). Alternative clustering algorithms such as K-means require N to be

defined a-priori. The sea-level pressure networks typically see N in the range of 20 to 30, while the number of sea ice concentration network nodes can be typically on the order of 80 to 100.

- Equation 2: I understand that the grid cells need to be weighted by their grid cell area. However, I find it a bit puzzling that for the polar grid the weights should be the square-root of the cell area, while for a regular grid the weights are a non-dimensional number between 0 and 1. Would it be possible to introduce normalized area-weights, such that there is no [km] in the unit and the anomaly time series (and with that also the covariance and node strengths) are not dependent on the node size? This could then maybe also resolve the issue that a high node strength does not necessarily coincides with large explained variance, as you explain in L252 ff.?

  If we understand correctly, you are happy with the weighting of the lat-lon gridded (sea-level pressure) data -- which is a form of normalised area weighting, however your concerns lie primarily with the weighting of the polar stereographic (sea ice concentration) data, and you would like to see a similar kind of normalised area weighting to these data?

  We believe that the current approach is justified because in either the lat-lon, or polar stereo case, the weighting is simply a means to ensure that data points contribute proportionally to the eventual network node anomaly time series, and hence temporal covariance between network nodes. Normalising the area weights for the polar stereo case would produce the same network node anomaly time series as the current approach, but there would just be a difference of scaling. To remove the relationship between covariance and node size we could simply standardise each network node anomaly time series (at which point the links between nodes become the correlation coefficient, as opposed to covariance). It is worth pointing out that both the sea-level pressure and sea ice concentration data show examples of where high node strength does not necessarily coincide with large explained variance and we feel that this result in itself is interesting as it suggests that CMIP6 models produce sea ice concentration and/or sea-level pressure fields which are more homogeneous than the observations (recall that an individual network node represents a geographic region which has behaved homogeneously over the length of the time series record, so the larger the node, the larger the region of homogeneity).

- Relating to what you write in L175-L187 (which is a very helpful paragraph), I am still wondering how much agreement/disagreement between model simulations (even if they had perfect model physics) and observations you can expect, simply because of internal variability and the time series of 42 years being relatively short, even more so when you look at the first and second half of the time series separately. What could be helpful in this regard is a comparison of the observations to a single-model initial-condition large ensemble, which can give you a good indication of the range of internal variability within the model. The CanESM5 model with its 20 members comes closest to a large ensemble. Looking at the ARI and D values of the CanESM5 ensemble members gives an idea of the expected spread just due to internal variability. I don't have a concrete suggestion for changing the manuscript, but I think this point should be kept in mind when describing and discussing the results.

  This is a great point, and certainly an interesting topic that is worth including in the manuscript. In our case the 20 CanESM5 ensemble members actually comprise two groups of 10 members with different physics (group 1: i1p1f1, group2: i1p2f1). While any analysis of these individual groups while likely produce an under-estimate of the spread in ARI and D from internal variability, we agree that a discussion on this should be included at the least. We have added some content and figures to the revised manuscript to highlight this.

- In all Figures showing covariance and node strengths, please add the units. Moreover, in the Figures for SLP (Figs. 1 and 5), I don't see how the orders of magnitude for covariance (10) and node strength (10^{10}) fit together. As the node strength is the sum of the absolute value of all its associated link weights, it shouldn't be orders of magnitude higher than the individual link weights/covariance, I would assume?

  We have double checked this, and it seems ok. In Fig. 1 for example, the magnitude of the first 5 strongest links belonging to the AO node are as follows: 3.7e9, 2.6e9, 2.4e9, 1.9e9, 0.98e9. Hence the sum of all of these links is only slightly over 1.16e10. We have also added units to each of the plots.

- Figure 8: The two sub-figures are a bit difficult to compare with the different axes ranges/color scales. If you keep the different ranges (which I can see is also useful for comparing Fig. 8a) with the observations), I still think that the range for Fig. 8b) could be capped at lower values, as right now the colors are very light (which is misleading at first glance).
  Thank you for the suggestion, we have edited this figure to cap the range of 8b accordingly.
- In the caption of Figure 12, the values of ARI and D for the individual model simulations don't seem plausible/don't fit to what you write in the main text.
  Thank you for spotting this. The values in the figure caption are incorrect and have now been edited to reflect the main text.
- Figure 14: I don't understand the meaning of the numbers in the legend(?) in the lower right panel of the figure. A proper legend with the names of individual models, the model mean and PIOMAS would be helpful.
  Apologies, this appears to have been a bug in the file upload. The original image contained the individual model names in the legend and no numbers. We have amended this and added the model means to the legend as per your suggestion.

Technical corrections

- L1: Suggest changing "proceeding" to "following", "subsequent" or similar
  Thank you, we have opted for "following"
- L9: Suggest adding comma before "respectively" (here and at other occasions in the manuscript)
  Agreed, we have edited this throughout the manuscript.
- L14: "the north Africa" → "the north of Africa", "north Africa", or "northern Africa"
  Thank you, we have edited this to "north Africa"
- L56: When reading, I was wondering why only 31 of the CMIP6 models, which you explain later in the Data section. Very optional, but you could consider omitting the number of 31 in the introduction to avoid questions at this place.
  Thanks for the tip, we have removed the explicit reference to 31 models in the introduction
- L109: Remove comma: "sea ice concentration and mean sea level pressure"
  Thank you, this has been edited
- L117, Table 1 and other occasions: "ensembles" → "ensemble members". In my understanding, a single simulation is denoted as an ensemble member, while the individual simulations together form a (single-model or multi-model) ensemble. If you change this, please be sure to stay consistent throughout the manuscript.
  Thank you, we have changed this throughout the manuscript.
- Eq. 2: Suggest writing down that E[…] is the expected/mean value.
  Agreed, this is more clear.
- L245: Remove comma: "north Africa and southern Europe"
  Thank you
- L298: Suggest changing heading from "Observations" to "Observations/Reanalyses"
  Agreed
- L430: JC → JS
  Thank you!