Second review of ACP-2020-565

Recommendation: Reject

General comments: While the authors did make strides to improve the explanation of the methods they employed, and their efforts did make their methods more clear, I obtained additional clarity in their approach and found several glaring concerns that need to be addressed. These concerns fall into three primary areas:

1) The blended use of teleconnections for specific spatial regions as a global product, even though many of these indices were not derived globally when they were created
2) The use of the ERA model data versus raw satellite data, and the subsequent use of their TCWV approach as effectively a verification measure for the ERA
3) Their empirical "reversed index" approach, which effectively is just a basic non-parametric hypothesis test.

The third of these issues impacted the interpretation of all results as it caused the authors to identify results as significant without proper context. Fixing this issue would require the authors to completely redo the analysis with more appropriate methods. As a result, I must still recommend rejection of this manuscript.

Major comments:

The authors did not really address my concern with the barotropic conditions in the tropics and why these regions are not used. In my experience deriving teleconnections, including barotropic regions generates large areas of high correlation due to the minimal height gradients in the tropics. Often these large areas wash out teleconnection features that would otherwise be present when using PCA (i.e. the first PC almost always exclusively identifies the tropics instead of a hemispheric teleconnection). As many teleconnections the authors considered (the NAO, PNA, etc.) were derived without including the tropics or the Southern Hemisphere, it is unclear how employing these teleconnections on a global study even makes sense, or even employing the teleconnections in the tropical latitudes which were not used in deriving the indices. How did you address this issue?

It seems like a strange methodological approach to verify the ERA model representations of TCVW using these teleconnection renderings (lines 87-90). Why not just directly verify the TCVW model data with the observation dataset? It seems like a very roundabout way to do model verification. Maybe a bigger question is why you are including the ERA data. Why not just use the satellite data directly? The differences do not seem that dramatic and there is no effort to explain why you did this except to compare model data against satellite, which does not help remove the "forecast" confusion in this study.

The use of the delta RMS quantity is strange. The authors even note (lines 209-210) that this quantity is basically the same as the correlation coefficient between the fit and the teleconnection. This makes sense as essentially you are doing a multivariate linear regression with an extra time term and you are just computing the variance explained by each teleconnection. This should scale almost exactly to just the correlation squared between the predicted value and the teleconnection. Why use this delta RMS instead of something simpler like R^2 to quantify the relationship between the teleconnection and the TCVW? Are there studies that employed a similar methodology?

I'm not sure your interpretation of the fit coefficient is correct. Are the time coefficients always the same for all teleconnections? If so, differences in magnitude in the teleconnection could be the reason for the change in the fit coefficient, not the actual amount of fit. You even show an example of this in Fig. A7, where the normalized ENSO index has a range from roughly +/-3 while the WHWP index has a range from +/-5. An almost 100% increase in magnitude in the index would affect the coefficient magnitude dramatically yet not explain any more variability.

With the volume of indices considered (Fig. A6), many of these have notable longer-term trends (MGII, Sunspots, PDO, AMO, etc.) where you do not even get an entire phase shift in your 20 year study period. How can you say with any certainty that there is a relationship without getting more than a single period of these quantities being in a "high" or "low" phase?

I would guess the small RMS values in the tropics are almost entirely a function of the barotropic conditions in the tropical regions (see my first comment above), not related to the impacts of clouds on the observations (as suggested in lines 222-224). Polar regions have similar issues as their conditions tend to be quasi-stationary. This alludes back to my earlier comment regarding the use of the tropics in this study.

While I technically agree with your statement on line 255, this is the nature of hypothesis testing. Most studies select a level (typically 95% or 99%) and go with it. I think the more important issue is trying to establish significance of these results using a hypothesis testing approach, since these tests are sensitive to sample sizes (e.g. you can get significance with a large sample size that could still have a poor relationship between the variables).

Are there citations of other studies that have used the "reversed index" approach you employed in this study? I have several concerns about it. First, reversing the time series does not ensure this is a completely uncorrelated relationship; random number generation would do a better job of that. As an example, I used a 70 year ONI time series and simply correlated the time series against its reverse and found a correlation of 0.16, which certainly is higher than what a random dataset should yield. This is even clear with several of your indices where the RMS values were fairly large considering this is supposed to emulate a "random" comparison. Second, the black dotted line is based on the mean and standard deviation of the reversed time series RMS 99th percentiles, yet when I look at the plot I see several points that would be "significantly" better than the reversed time series threshold. These points would also drive that mean upward as they are outliers (S107, MG11, etc.). This calls into question the validity of this approach since using a different statistics (e.g. the maximum) would cause almost all of your "significant" points to shift to non-significant. Maybe most importantly, this approach does not really show statistical significance. If you are treating this as a multiple-comparisons problem (which it appears you are), you need a Bonferroni correction on the cutoff threshold to ensure you are not committing type 1 errors (which are basically guaranteed with the 57 comparisons being done here). This would further shift the cutoff threshold upward and make more of your results non-significant. Why are you not using more traditional methods, such as bootstrapping, permutation testing, etc., to quantify this significance?

The latitudinal results in Fig. 11 make sense to me since most of the teleconnections related back to ENSO. Why are there so many teleconnections near the International Date Line? You never even discussed the longitudinal plot in the paper from what I could tell and that is a more interesting plot to me (and more difficult to explain).

The authors state that an advantage of their empirical approach is that it "avoids problems of existing algorithms for the determination of significance, because no assumptions on the significance level or the measurement uncertainties have to be made." However, by selecting the 99th percentile you have effectively created an $\alpha = 0.01$ significance level as you are comparing your observation against an the 99th percentile of an empirical distribution. In effect you just did a hypothesis test, just with a slightly different appearance. If it is different than a hypothesis test it needs to be explained more effectively.

Minor comments:

Remove all of the uses of the word "like" in the e.g. statements (lines 46-47 and any others in the manuscript).

Line 98: "In section 2 the global datasets used in this study" is not a complete thought.

The y-axes in Fig. 7 should be consistent for both indices.

What do you mean by "reduced number" of data available? How small? What are the differences? (lines 228-229)

Would the result regarding zonal winds in the tropics not just be a consequence of the relationship between geopotential height and wind (lines 315-316)?

Why would a high surface albedo be a systematic measurement bias? Is the satellite instrument the one with the bias or the ERA data? (Lines 341-342).

There is strange comma use and formatting issues in section 7 of this paper.

Maybe I missed it, but why are there massive data gaps over Siberia and into India in the satellite data (Fig. 10)?