Reviewer #1 (Robert Jnglin Wills)

I find this to be an interesting paper and the conclusions can largely be supported by the work presented. Overall this is a substantial contribution and the needed revisions are minor. Nice work.

We thank the reviewer for the careful revision of the manuscript and appreciate the positive feedback and the helpful comments and suggestions.

My only major concern is with the discussion of the links to the AMO. The AMO is generally associated with a center of action in the North Atlantic subpolar gyre (e.g., O'Reilly et al. 2016, Wills et al. 2019, and references therein), which shows no clear anomaly in either of the resented composites. To the extent that the Atlantic temperature anomalies are there at all, they (Fig. 4c) look more like NAO-coupled variability of the ocean gyre circulation (i.e., warming in the Gulf Stream and GIN seas, but cooling in between; Curry and McCartney 2001, Eden and Jung 2001, Sun et al. 2015, Wills et al. 2019). Since these anomalies are weak anyway, I would limit your discussion of connections to the AMO, instead saying something like "there appear to be differences in Atlantic temperatures between the two drought types, and that this could be related to modes of Atlantic multi-decadal variability such as the AMO or the NAO-coupled variability of the gyre circulation, as discussed in the literature". Note that I've included a lot of Atlantic multidecadal variability literature here because of my own interest in that part of the story, and in case it is useful, but I don't actually think it is necessary to go into/reference all of it in this manuscript.

We agree with the reviewer and gladly include the suggested phrase in the revised manuscript: "*There appear to be differences in Atlantic temperatures between the two drought types, which could be related to modes of Atlantic multi-decadal variability such as the AMO or the NAO-coupled variability of the gyre circulation, as discussed in the literature".*

Scientific questions/issues:

17 droughts is a small number of degrees of freedom to be computing clusters from. Could you quantify what you mean by "most conclusive clustering result" or give some metric of how this clustering depends upon sampling? Furthermore, you then explain a principal component analysis based approach and this left me confused as to which method you were using. Are you using two separate methods to characterize the droughts? Do they get the same answer?

A sample size of 17 droughts is admittedly small and quite sensitive to the number of clusters or generally the sampled area. We thus tried many different settings of the clustering by changing the cluster numbers, the chosen spatial and temporal domain as well as the clustering approach itself. We found that limiting the clustering to two instead of three clusters resulted in the more evident classification of the 17 droughts and is furthermore consistent with the literature (*e.g.*, Fye et al., 2003). In that process, we also decided to exclude the turn-of-the-century drought from the clustering because of its inherently different spatial signal compared to the other 16 droughts in our sample.
In terms of the clustering method: We tested two different clustering approaches: k-means clustering and ward clustering, both of which have their strengths and weaknesses. While both methods resulted in an almost identical classification of the droughts, they disagreed on the class affiliation of one drought. Therefore we chose to combine the two clustering methods and make use of the method's strengths to more accurately define the right cluster for the remaining drought. (page 6, l. 14-15) "Ward hierarchical clustering was used to determine the cluster centers, which were then used as a starting point for the k-means clustering.")

With regards to your methodology of "each drought period was first expressed relative to a reference period that comprised 5 years before and 5 years after the drought period", have you compared this to the simpler approach of using anomalies from the long-term mean? It seems that this would be a simple check and I would hope it doesn't make a huge difference.

We did compare the ±5 years composites approach to anomalies with respect to the different climatologies. We found that it is dependant on the long-term mean chosen. E.g.: When anomalies from 1901-2000 are compared with the ±5 years composites then the resulting patterns look very similar for the two droughts during this period (1931-1939 and 1952-1665) but the accordance gets worse the further back in time a drought period is. An analogical outcome is obtained when performing the analysis with different reference periods. This points towards the fact that spurious trends are biasing the results. Since we are looking at a 400 year long time period, it is difficult to find an appropriate long-term mean that would capture the differences between the drought and non-drought periods in a satisfactory manner. We believe that the ±5yr composite approach is much more suitable for our analysis.

Do you have an explanation why the SLP anomalies tend to be weaker / less significant than the GPH anomalies? Physically this would arise if the circulation anomalies were baroclinic (consistent with a shift of the subtropical jet in the longitude band of Pacific/North America), but I am not sure the EKF400 reanalysis can be trusted to that great of a degree. Could it possible be reconstructing less of the SLP variance than it does the GPH variance? Are the differences in anomaly amplitude actually quantitatively different? It may be helpful to rescale the SLP colorbar and to consider my following comment.

That is an interesting comment. We suspect that this might be due to the fact that the 500hPa field includes more of the temperature signal and thus performs statistically slightly better. We will make sure to increase the visibility of the differences in the anomaly amplitude in the revised manuscript.

Why do your GPH figures seem to have a mean over the plotted domain that is less than zero? This could be due to variability in the Southern Hemisphere that is not relevant here. Could you remove this so that the plots are easier to parse?

Well spotted. N-S asymmetries can play a role, but one has to keep in mind, that our approach is not mass conserving (applicable for SLP).

How is the 95% significance level computed for the figures? In particular, how are you computing the number of temporal degrees of freedom? It would be helpful to state this in the caption.

This is mentioned in the method section but we will add it to the captions as well. All significant tests are based on a non-parametric Wilcoxon-Mann-Whitney test.

Please check that there are no major differences between a composite of SST and the T2M composite shown. No need to show it, but it would be good to check this and state whether there are any significant differences.

We are currently looking into this comparison and will provide some statement/validation in the revised version addressing this matter. Furthermore, we plan to include a new supplementary figure in the revised manuscript showing the comparison of SSTs and T2m.

I don't fully agree with your interpretation of Fig. 4. There are not particularly stronger or more significant ocean T2M anomalies in the North Atlantic than the North Pacific. Given the larger influence of tropical SST anomalies on the atmospheric circulation (e.g., Kushnir et al. 2001), the different atmospheric anomalies are just as likely to result from the tropical Pacific or tropical Atlantic temperature anomalies, even though those anomalies are smaller and not significant. You state multiple times in the discussion that the warmer North Atlantic (while not significant) could explain this or that atmospheric change, but I don't think these results make a strong case for that, especially not for any role of the AMO, which should have larger-scale coherent warm anomalies focused in the subpolar gyre (such as those seen in Fig. 5).

It may be helpful to consider Ruprich-Robert et al. 2017, which looks at the differing impacts between the tropical and extratropical component of "AMO" anomalies in a climate model.

In the revised version we will limit our discussion of connections to the AMO and instead point towards the differences in Atlantic temperatures between the two drought types potentially related to different modes of Atlantic multi-decadal variability.

Could you extend the latitude range of your T2M plot over the equator? Any SST anomalies in the 0-20° S latitude range could still have a large impact on the atmospheric circulation in the Northern Hemisphere.

We will extend the T2m to 20°South in the revised manuscript to get a more comprehensive view.

Technical corrections:

Thank you for pointing out typos/wording problems. We will correct them in the revised version.

- *Page 1, Line 14 typo: "show" should be "shows"*
- *Page 2, Line 13: typo, extraneous "of" after behind*
- *Page 3, Line 18 typo: "or" instead of "of"*
- *Page 4, Line 25-25: I think "opposed to decadal variability" is not the correct word choice for what you are saying. Should be "compared to decadal variability" instead.*
- *Page 6, Line 8: missing word(s) between La Niña and El Niño*
- *Page 6, Line 17 typo: "at in"*
- *May not Mai*
- *Mid-19th not mit-19th*
- *Page 6, Line 17: former and latter are both singular, and you should use "exhibits" with them, not "exhibit"*
- *Page 8, Line 27: "turn-of-the-century drought" not "turn of the century"*
- *Page 1, Line 16-17: positive and negative anomalies in what index?*

  We will specify: positive and negative GPH anomalies

- *Page 2, Line 4-6: the words "most relevant" are not very precise, consider rephrasing*

  Agreed, we will use "alarming" instead.

- *Page 2, Line 11: Is "moisture interpretation" a vocabulary word I am not aware of, or is this simply a wording problem where you should have said "are mostly restricted to interpretations as temperature and moisture"?*

  This is a wording problem, we will use your suggestion instead.

- *Consider referencing Enfield et al. 2001 as well for the Atlantic SST influence on multi-decadal drought.*

  Good idea, we will add Enfield et al. 2001 in the revised version.

- *Your abstract had me wondering why only summer SST/SLP/GPH is relevant. If you say you are looking at summer drought, then it would become clear why, and you then don't even need to say that it is summer SST/SLP/GPH.*

  Elegant, we will add "summer" in the abstract to clarify that we are focussing on multi-annual droughts during summer.

- *Page 4, Line 6: please state how/why the ensemble members differ*

We will add a sentence in the revised manuscript on how the 30 ensemble members in EKF400 differ.

- *Page 6, Line 9/10: twice you say "three" where I think you mean "two"*

  Yes, that is a mistake, we will correct this to "two" in the revised version.

- *Page 8, Lines 19-20: the second half of this sentence needs to be reworded, this word order (especially with contribute at the end) does not work in English.*

  Agreed, we will reword this sentence in the revised manuscript.

- *First sentence of conclusions: please add that this is the first time this has been studied in a climate reconstruction, because there have of course been model-based studies*

  Yes, that's a good point. We will add this specification in our revised conclusion.