The authors have presented a well-motivated and, largely, clearly executed analysis of changes in large droughts in Australia using validated modelled drought outputs for the historical period and over the last millennium as derived from paleoclimate data. I believe that this manuscript is a worthy contribution to the scientific discourse, but I believe some revisions and additional analysis is required to make this a robust study as follows:

- My most pressing concern is ensuring that the validation is sound and appropriately quantified;
- At present, I do not believe that the presentation of figures intended for communicating the impact of different sample sizes is sufficient; and
- I feel that a measure of drought intensity that is comparable across events of different length is missing

More detail for these three dot points are provided under the general comments.

And finally, I would like to see an acknowledgement of the limitations in making comparisons with historical droughts, particularly since the analysis omits most of the millennium and all of the Tinderbox droughts.

Otherwise, there are a small number of clarifications, particularly of the caveats, and these are detailed in the minor comments.

General comments

A brief discussion is needed to acknowledge the limitation of using one definition of a water year for all of Australia and locations where the water year definition used is most/least relevant.

I need to see more detail about how the spatial correlations were calculated. Was this a calculation of correlations between matching locations that were then averaged across the region, or was the significance at each location assessed independently, or was a field significance considered and if so, which method was used (e.g. false discovery rate, walker's test, counting test)? The latter (a measure of field significance) is what is required. An averaging of correlation results is not appropriate, and assuming spatial independence is also inappropriate. The results of quantifying the field significance of the similarities between the observed and modelled droughts will impact on the credibility and interpretability of the remaining results. If fiend significance results markedly alter the validity of the modelled results, the interpretation of the pre-industrial millennia results will need to be re-interpreted accordingly.

The measures of both relative drought intensity and severity appear to be functions of the average deviation from climatology across the event. It appears to me that relative drought severity is a superfluous metric since drought length is also presented (although the figures show the mean over time and sometimes across models, so I realise it is showing something different then taking the product of, say, fig 5b and fig 6b). What seems to be missing is a measure that reflects the most severe drought year (or years) such as the most intense two consecutive years of drought within an event and to see if the maximum (annual or consecutive multi-annual) intensity is changing in different time periods. A measure like this would prevent the metric from being influenced by the definition of the event duration,

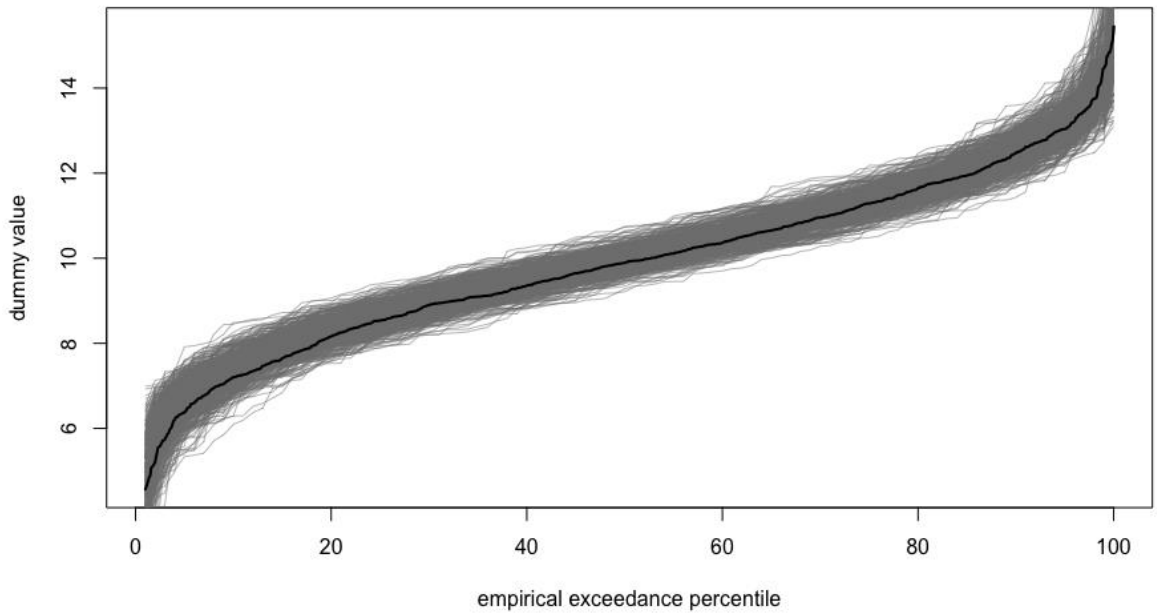
which is the case for drought intensity and severity metrics, particularly since event length would be sensitive to the definition of drought in determining onset and termination.

Section 3.2.3. Results of Fig 6 b-d and f-h are presented in the text, but references to the figures need to be made. I recognise that reference is made to supp. Figs 1-2 and 5-6 that reflect the points made in the text in more detail, but Fig 6 b-d and f-h are still relevant and need to be referenced in the text.

The specification of significance level needs to be stated in terms of what significance level has been chosen for the evaluation i.e. α = 0.01, rather than reporting overall p-values.

Section 3.3. I would like some text around how the precipitation mean and variance compare between modelled and observed specifically in MDB (rather than relying on the reader to interpret the figures themselves) as this could help explain the difference in drought results.

Section 3.3.2. and supporting fig 20. It does not follow that large spatial variability implies anything about the adequacy of record length. Another justification is needed here. Also, I can see the intent of what the authors are aiming to communicate here: that shorter simulations fail to fully explore how variable drought can be given the range of drought conditions that can be explained over a longer period. I think Supp Fig 20 b is sufficient because it is clear that the maximum drought length obtained from a shorter 101-year sample will likely underestimate the maximum plausible drought length. However, I don't think the remaining plots in this figure demonstrate that the drought characteristics sampled in a 101 year-long sample are not representative of what could be reasonably expected in our climate and an alternative way of presenting this data is needed.
One suggestion I have would be to plot a cumulative density line or scatter for each of the 500 samples on a single plot and then overlayed would be the cumulative density line for the 1000-year long simulation. As a dummy example, I've done this for 500 samples of n=100 for a normal distribution of mean=10 and sd=2 (these values have no meaning or significance, it's just for an example) with an additional sample of n=1000 shown in bold. The historical or HIST ecdf could also be added and discussed with respect to over/underestimating flood characteristics at different magnitudes with respect to the longer record.



I'd be more than happy for the authors to either adopt this or develop an alternative for presenting their findings that would provide a figure that supports the argument they are making in the text of section 3.3.2.

Minor comments:
L48: A reference is needed for the Tinderbox drought being a "major" drought.
L102: Full stop at the end of this sentence
L105: Use "0.05° × 0.05° latitude/longitude resolution" for consistency with later model resolution descriptions.
L146: clarify that the bias relative to observations is shown for each member as well as the overall ensemble mean.
L193 and elsewhere: specify that the resolution is for latitude/longitude
L209: is the percentage bias also reported for the rest of Australia? If not, why not?
L222: could you clarify in L126 how many members are run in natural or fully forced or single forcing so it is clear what this "30" is based on?
L225: to improve clarity, extend this sentence with "...(>60%) ensemble members were not in drought at the same time *as this would indicate*….."
L233: replace "," with "…(101 years) differ *and* affect any disparity…."
L235: for clarity, it would be worth reconfirming the number of distributions that are generated (i.e. 500)
L255: do you mean "observed variability" instead of "MAP"?
L264: in addition to overall bias in mean MAP across the continent, the models also largely generate precipitation with reduced variability (with the exception of CSIRO-Mk31-1-2 and IPSL as previously stated).
L280: to improve clarity, insert "across ensemble members" prior to referencing the supp figs.
L285: does the statement "suggest similar spatial patterns" apply to all members, or just some? If this is across the ensemble, please state this up front in the paragraph as this would help demonstrate that the simulations are an adequate representation of the observations, which I believe is the intent of this paragraph, and it is key to providing a basis on which comparisons of HIST, piLM, and pi Control can be assessed. The message at present is a little

lost because the shortcoming are presented first and the purpose of the 20th century simulations is not clearly stated (i.e. for validating the model runs and providing credibility for the piLM and pi control runs).
L311: particularly *in* southern and eastern Australia
L323: for consistency, include "ranging from" or "range of" prior to "5.1-8"
L325: Is this supposed to be "mean maximum drought length" for both metrics on this line?
L328: I'm not sure what is meant by "continent-spanning grids". Is this just "all locations"? Or "grids covering the mainland"?
L330: could you clarify which model simulations? I believe it would be the HIST model simulations
L332: add "%" to these numbers
L335-336: Does this statement not apply across the ensemble results too? Or is it just confined to the three best performing models?
L336: "worse" is subjective. Use "more severe" or similar
L385: "The exception is volcanic forcing, where ~~CESM LME~~ most ensemble members *in the CESM LME* run with volcanic forcing are not in drought….". Also, it seems like a discussion of the agreement between ensemble members under LULC forcing is missing. It would also be good to comment on the variability of the forcing as it's very easy to see when volcanic forcing imposes a large change in the radiative forcing, but the variations in solar and LULC are less easy to identify.
Line 391: specify that this is in reference to results that are averaged across Australia.
L415 to 417: Given the findings that 100-year samples result in different summary statistics compared to a single 1000 year record, these comparisons really should be made in the context of the distribution of 100 year samples taken from the longer record as opposed to comparing a 100 year long record with a 1000 year long record (i.e. fig 9a-d).
L427: MDB (rather than MBD)
L438: Can text be added to make this finding a bit more explicit? Such as: "The co-occurrence of volcanic eruptions and supressed drought conditions over the MDB appear to contradict existing understandings of the impacts of volcanic eruptions on El Niño-like conditions and subsequent impacts on rainfall in the MDB".
L 443-L445 needs to be clarified. At present it appears to contradict the first sentence of the conclusion.