Author response: We thank the Reviewer for their kind comments, and for their constructive review of our manuscript. We have addressed all suggestions (details below). In response to both these comments, and the suggestions of Reviewer 2, we will make the following changes to improve the clarity of our manuscript:

- Add two new Supporting Figures:
  - One showing the return period of multi-year droughts in the Murray-Darling Basin, in each PMIP3 model's pre-industrial last millennium simulation
  - One showing the relative severity of the longest droughts in each simulation from each model, as well as observations
- Provide increased clarity around the interpretation of Supporting Figure 20, and also slightly modify this figure for ease of interpretation (details below)
- Provide more detail of calculation of the spatial correlations, and - at the Editor's discretion - either use these correlations to weight the calculation of multi-model means, or add a statement as to why we did not do this
- Add a statement as to why we did not bias-correct the models

We also provide two new analyses in this response:

- A comparison of multi-year drought characteristics in the observational dataset used in this paper with observations that extend to 2021, thereby encompassing the Millennium and Tinderbox droughts.
- For each model, a timeseries of the *maximum* relative intensity of each drought across 850-2000 CE for the MDB, based on the Reviewer's suggestion.

Additionally, we will address all general and specific comments from the Reviewer as outlined below. We consider that these changes will result in a stronger paper with clearer, more robust findings.

Review of 'Emerging anthropogenic influence on Australian multi-year droughts with potential for historically unprecedented megadroughts'

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:

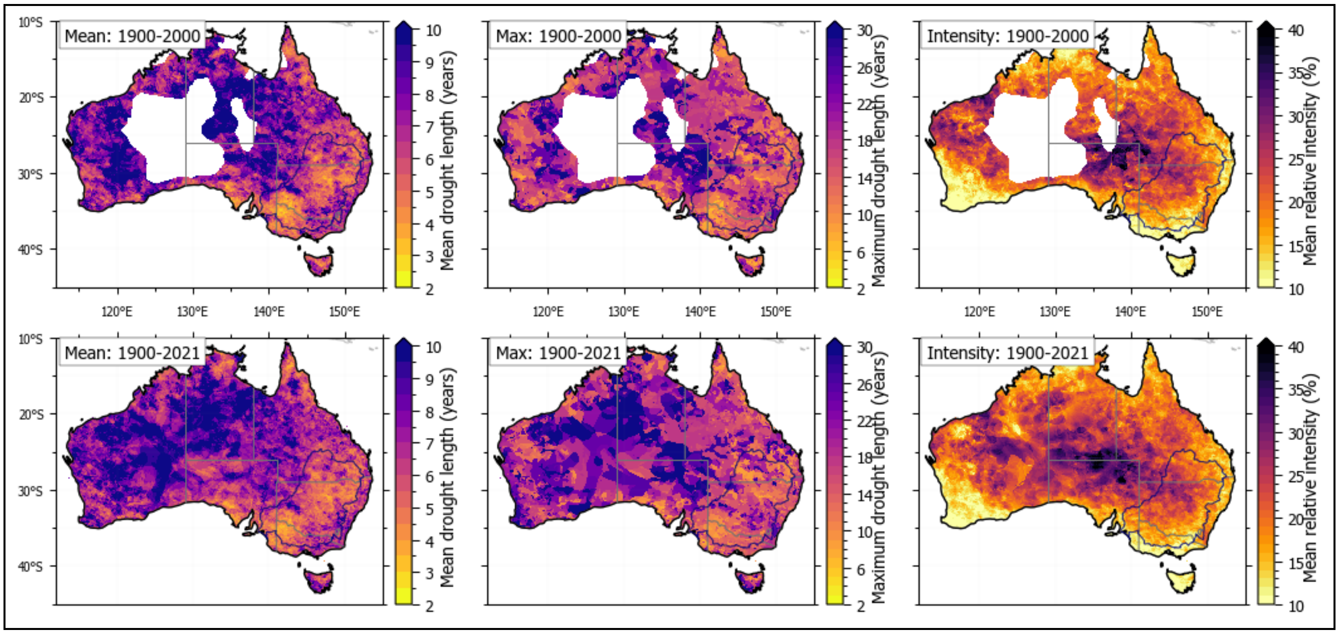
- My most pressing concern is ensuring that the validation is sound and appropriately quantified;
- At present, I do not believe that the presentation of figures intended for communicating the impact of different sample sizes is sufficient; and
- I feel that a measure of drought intensity that is comparable across events of different length is missing
- More detail for these three dot points are provided under the general comments.

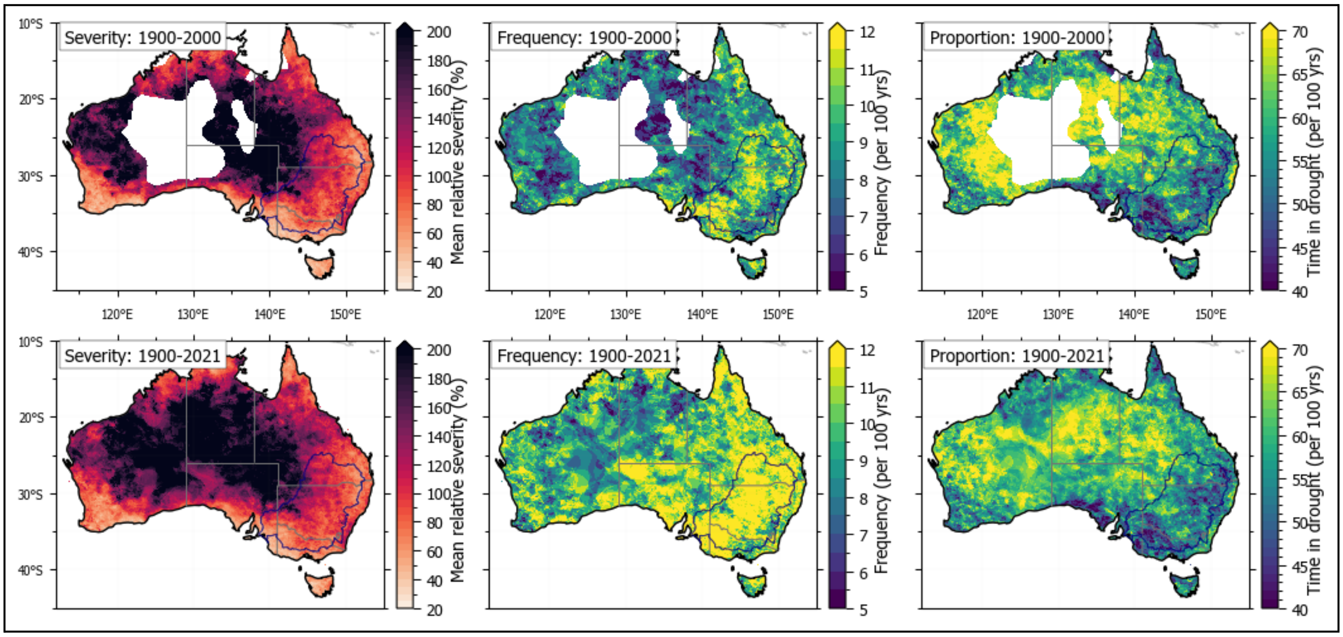
- And finally, I would like to see an acknowledgement of the limitations in making comparisons with historical droughts, particularly since the analysis omits most of the millennium and all of the Tinderbox droughts.

Author response: *Regarding the last dot point, which is not addressed under 'General comments': we re-calculated all multi-year drought metrics, but using a version of the observational dataset that extends to the year 2021 (thereby including the Millennium and Tinderbox droughts). We provide that comparison on the following page. For mean and maximum drought length, relative severity and intensity, and proportion of time spent in drought, differences between the two datasets are negligible. Drought frequency increases slightly with the addition of the extra 21 years - particularly in eastern Australia.*

*We currently state in the Discussion: “*Major droughts in the MDB during the first two decades of the 21st century (i.e. the Millennium and 2017-2019 droughts) are not included in our analysis, but strengthen this finding of an apparent change towards longer droughts and more years spent in drought during the 20th century simulations compared with natural variability during the pre-industrial last millennium*” (L431). We will re-phrase this as follows, to state that our conclusions—particularly regarding anthropogenic impacts on Australian droughts—may be influenced by the fact that the PMIP3 simulations do not span the Millennium or Tinderbox droughts.*

“We note that the PMIP3 *past1000* simulations do not cover two of the Murray-Darling Basin’s most impactful droughts of the historical period: the Millennium and 2017-2019 droughts. However, the occurrence of two major droughts in the first two decades of the 21st century provides additional support for our finding that the MDB is spending more time in drought during the historical period compared with natural variability during the pre-industrial last millennium.”





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.

Author response: *We will add the following brief description of possible implications of using a January-December year across the broad range of climate zones in Australia, at Line 110.*

"Precipitation seasonality varies across Australia. For example, in tropical northern Australia, 'years' may be better represented by 'tropical years', where each year starts e.g. in May of calendar year 1 and finishes in April of calendar year 2. However, this is not applicable across the entire continent, where some regions have winter-dominated precipitation, and others have no distinct seasonality. Given this study focuses on multi-year events, our choice of a calendar year for calculation of annual totals should not have a major influence on results."

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.

Author response: *For Supp. Figs. 1-6, we show a pattern correlation. We will add the following, more detailed, description to the Methods (at Line 150): "*Spatial correlations between pairs of two-dimensional grids were calculated by flattening each grid, resulting in two directly-comparable vectors, with each index position of each vector representing the values at a particular latitude-longitude pair. We calculated the Pearson correlation coefficient, and provide an estimate of the significance of that coefficient (reported 'significant' if $p < 0.05$).*"*

For Fig. 9 where we calculate significance per pixel, we did not account for spatial dependence or False Discovery Rate. We will adjust the p-values as suggested by the Reviewer, and update our findings if necessary.

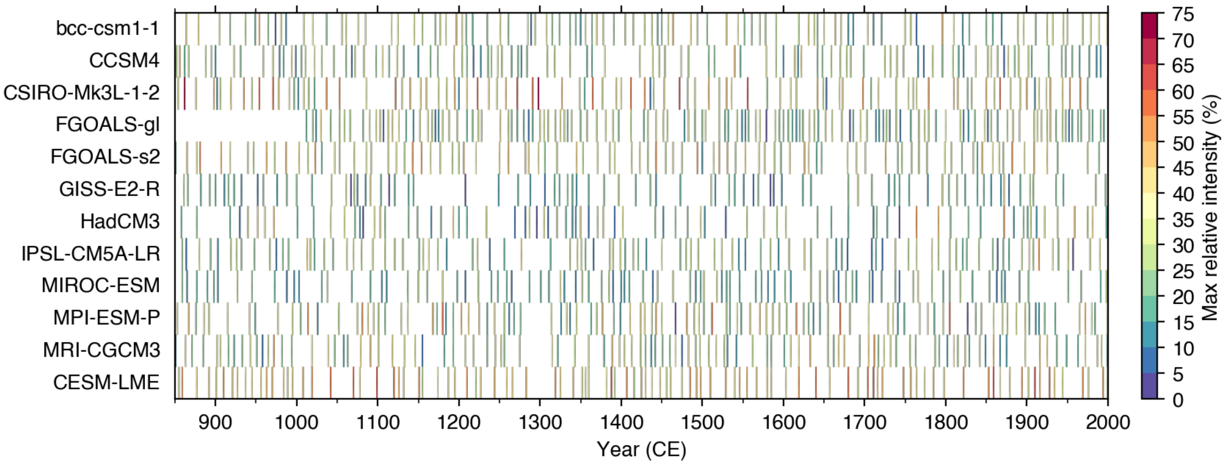
The measures of both relative drought intensity and severity appear to be functions of the average deviation from climatology across the event. It appears to me that relative drought severity is a superfluous metric since drought length is also presented (although the figures show the mean over time and sometimes across models, so I realise it is showing something different then taking the product of, say, fig 5b and fig 6b).

Author response: *That is correct - the 'drought severity' metric combines the information from the drought length and drought intensity metrics; drought intensity is independent from duration. The severity metric provides an estimate of the* total *precipitation deficit over the drought, which is not easily apparent from inspecting the length and intensity metrics separately. We consider this to be an important aspect of the drought characterisation processes, as it provides insight into the total water stress placed on the region over the span of a particular drought. We also chose this particular metric of drought severity as it is consistent with our other metrics (e.g. drought mean and maximum length), which are expressed for the whole event.*

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.

Drought severity is indeed dependent on the metric used to identify droughts. In the '2S2E' method used in this paper, multi-year droughts tend to be longer and more severe (but less intense) compared with, for example, a method where 'multi-year droughts' are two or more years below the 20th percentile of climatological precipitation. Given the already very large number of figures in this paper, we do not provide a test of the sensitivity of our results to different drought identification methods. However, we will add a statement to this effect.

Whilst we agree that the new analysis proposed by the Reviewer is certainly an interesting one, assessing temporal variability in maximum drought intensity is not practical for a continent-wide analysis. Here we have performed this analysis for the Murray-Darling Basin: the below figure shows, for each model, the maximum *relative intensity of each drought, for 850-2000 - i.e. the single most intense year within the drought (as proposed by the Reviewer). Across the PMIP3 models, there is no consistent temporal variability in drought maximum intensity through time.*



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.

Author response: *We will make the following changes to ensure specific reference to Fig 6b-d and f-h:*

- *At Line 286, we will reference Fig. 6b and f at the end of the line.*
- *At Line 290, we will edit the figure reference to state* 'Figs. 5-7a,e compared with Figs 5-7c,d,g,h'. *We will also add the word* 'broadly' *before 'resemble' in Line 289*

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

Author response: *We will state that we used* α *= 0.05 as our significance threshold.*

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.

Author response: *We did not originally include this text, as the models' skill in simulation precipitation variability over the MDB is very similar to the skill over the entire continent. We will add the following text to a new sub-section of Section 3.1. Accordingly, we will remove the current sentence about CV in the MDB at Line L254.*

"**Section 3.1.1 Evaluation of climate models' precipitation variability: Murray-Darling Basin**
Model skill in capturing observed precipitation variability over the Murray-Darling Basin (MDB) is very similar to model skill over the entire continent. Most models have a positive overall MAP bias (Fig. 3); exceptions to this are CSIRO-Mk3l-1-2 and IPSL-CM5A-LR (negative overall bias). Although most models slightly underestimate precipitation CV in the MDB, the spatial CV patterns are reproduced fairly well (Fig. 4)."

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.

Author response: *Major and random spatial variability, rather than well-defined spatial patterns, implies a large random element to the patterns. In this particular context, this suggests that there are insufficient multi-year droughts for a climatological pattern to emerge. This points to an important role of model internal variability at these time scales as the different ensemble members discussed here only differ in their internal variability (rather than other factors such as model physics). To clarify, we will add the following text at Line 348:*

"That is, the presence of random spatial variability, rather than well-defined spatial patterns, implies a large random element. In this context, this suggests that there are insufficient multi-year droughts in a 101-year sample for a climatological pattern to emerge."

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.

Author response: *Supporting Figure 20 demonstrates that no single 101-year segment can capture the full range of variability present in a 1000-year sample. Hence, any single 101-year segment (such as the historical period) likely gives a skewed representation of long-term variability in that particular drought metric. This is true for all metrics, not just maximum possible drought length.*

We recognise that by showing the overall mean values from the piLM simulations as blue dots on Supp. Fig. 20, we are not making this point particularly well. We will therefore replace these with dots showing the equivalent values from each model's ***HIST*** *simulation. We will also add the following text to Section 3.3.2 to clarify this (replacing the current sentence at Line 351-352):*

"That is, individual 101-year segments do not capture the full range of variability represented by the 1000-year piLM period, with the magnitude of drought durations, intensity, severity, frequency, and proportion of time spent in drought varying markedly from one 101-year period to another. This means that selecting any single 101-year period (such as the historical period) is not representative of the full variability in the models' simulated precipitation."

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.

Author response: *Thank you for the suggestion. In response to this and a similar suggestion from Reviewer 2, we will add a new supporting figure showing the return period of droughts of different lengths in the MDB, from each model's piLM run. However, showing return periods calculated from the 101-year segments on that same plot is slightly misleading. For example, the longest single drought of the piLM simulation will occur only once in the piLM run, giving it a return period of ~1000 years, while in the 101-year segment containing that same drought, it will have a return period of ~101 years. This makes it hard to compare the two analyses on the sample plot and may be misinterpreted by readers. This type of analysis is also slightly complicated by the fact that multi-year droughts are discrete events that do not occur every year.*

In replacing the blue spots showing values for the full piLM simulations with spots showing values for the HIST simulations (as detailed above), we consider that we will more clearly demonstrate that using a too-short time period leads to an inaccurate understanding of multi-year drought characteristics. Combined with our additional text as shown above, this will better support our arguments in section 3.3.2.

Minor comments:
L48: A reference is needed for the Tinderbox drought being a "major" drought.
Author response: *We will add the following reference: Nguyen, H., M. C. Wheeler, H. H. Hendon, E.-P. Lim, J. A. Otkin, The 2019 flash droughts in subtropical eastern Australia and their association with large-scale climate drivers, Weather and Climate Extremes, Volume 32, 2021. https://doi.org/10.1016/j.wace.2021.100321.*

L102: Full stop at the end of this sentence
Author response: *Thank you for catching this. We will add this full stop.*

L105: Use "0.05° × 0.05° latitude/longitude resolution" for consistency with later model resolution descriptions.
Author response: *We will make the suggested edit.*

L146: clarify that the bias relative to observations is shown for each member as well as the overall ensemble mean.
Author response: *We will edit the sentence to "*For each model, we show the results as absolute bias relative to observations.*"*

L193 and elsewhere: specify that the resolution is for latitude/longitude
Author response: *We will edit this and similar sentences to "*...into 2° x 2° (lat x lon) resolution…*"*

L209: is the percentage bias also reported for the rest of Australia? If not, why not?
Author response: *We only calculate the percent bias values for the Murray-Darling Basin. This is because the MDB analyses are performed on a single area-mean precipitation timeseries (rather than each individual grid*

cell), which permits these more detailed analyses. We feel this is less appropriate if calculated on area-mean precipitation over the whole continent due to the many climate zones this would encompass.

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?
Author response: *We will add this information to Lines 127-129 in the following format: "*...well-mixed greenhouse gases (n=3), volcanic aerosols (n=4), orbital parameters (n=3), solar irradiance (n=4), and changes in land surface properties resulting from land use (n=3).*"*

L225: to improve clarity, extend this sentence with "...(>60%) ensemble members were not in drought at the same time as this would indicate….."
Author response: *We will add a new sentence at Line 225 stating "*...in drought at the same time. Co-occurrence of droughts across different ensemble members would potentially indicate an externally forced component to drought occurrence.*"*

L233: replace "," with "…(101 years) differ and affect any disparity…."
Author response: *We will make the suggested edit.*

L235: for clarity, it would be worth reconfirming the number of distributions that are generated (i.e. 500)
Author response: *We will edit this sentence as follows: "*...to create 500 distributions of possible values…*"*

L255: do you mean "observed variability" instead of "MAP"?
Author response: *No - the CV metric compares interannual variability with MAP. So in this case, the models' simulated variability is too low compared with their simulated MAP. We will edit the text along the following lines to clarify this: "*...i.e., the model interannual precipitation variability is too low compared with the model 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).
Author response: *We will add the following qualifier at Line 264: "*...despite overall bias in mean MAP across the continent, and generally too-low interannual variability.*"*

L280: to improve clarity, insert "across ensemble members" prior to referencing the supp figs.
Author response: *We will make the suggested edit.*

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).
Author response: *We will add the statement "*In the multi-model mean, …*" to the start of the paragraph (L277). We will also add that statement in Line 285/286.*

L311: particularly in southern and eastern Australia
Author response: *Thanks - we will make this correction.*

L323: for consistency, include "ranging from" or "range of" prior to "5.1-8"

Author response: *We will make the suggested edit.*

L325: Is this supposed to be “mean maximum drought length” for both metrics on this line?
Author response: *This should be ‘maximum drought length’, except where we are specifically referring to a multi-model mean. To make this clearer, we will change “*length*” in Line 324 to “*lengths*”.*

L328: I’m not sure what is meant by “continent-spanning grids”. Is this just “all locations”? Or “grids covering the mainland”?
Author response: *The former - the grids spanning all of Australia. Fair point - the ‘Australian continent’ technically includes New Guinea as well as mainland Australia and Tasmania. We will replace ‘continent-spanning’ with ‘full’ to differentiate this MDB-focussed analysis from the all-Australian-grid-cells analysis.*

L330: could you clarify which model simulations? I believe it would be the HIST model simulations
Author response: *Sorry for the confusion here. Yes, it is in the HIST simulations - this is stated at the start of the sentence “*The relative intensity of 20th century droughts…*”. To improve clarity, we will add the word “*HIST*” before “*model simulations*” in Line 330.*

L332: add “%” to these numbers
Author response: *We will make this correction.*

L335-336: Does this statement not apply across the ensemble results too? Or is it just confined to the three best performing models?
Author response: *The difference is larger when only looking at the three ‘best-performing’ models, but yes - each model’s HIST simulation generally shows more severe droughts than the same model’s piLM simulations. We will remove the word “*However*” from Line 335.*

L336: “worse” is subjective. Use “more severe” or similar
Author response: *We will replace “worse” with “more intense/severe”.*

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.
Author response: *We see no major influence of LULC on multi-year drought occurrence in the MDB, so in the interests of brevity, do not discuss the forcings individually.*

Line 391: specify that this is in reference to results that are averaged across Australia.
Author response: *We will amend that sentence to state “*Overall our results suggest that across most of the continent, Australian droughts have not changed substantially in the last century compared to model simulations of the last millennium.*”*

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).
Author response: *In this situation, we may consider the 2d grids formed from the statistics of the 1000-year piLM simulations the ‘control’, and the 2d grids formed from the statistics of the 101-year HIST simulations the ‘experiment’. The 1000-year piLM simulations show the climatological patterns we would expect today, without*

the influence of anthropogenic forcing. We are assessing whether there is any detectable *change, following the addition of anthropogenic forcings. If there is yet no detectable change, this could be 1) because there is no anthropogenic influence, or 2) because the anthropogenic influence is not yet large enough/has not had enough time to emerge from the range of internal variability. We state this at line 421: "*Hence, the lack of significant 20th century change across most of Australia in most drought metrics does not imply that there has been no human influence on Australian droughts during the 20th and 21st centuries, but rather that the 101-year HIST simulations are too short for significant differences to emerge.*"*

*To clarify, we will change '*significant differences*' to* 'any significant anthropogenically-driven changes' *at line 424.*

L427: MDB (rather than MBD)
Author response: *Thank you for picking this up - we will make the suggested edit.*

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".
Author response: *Excellent suggestion. We will add the suggested sentence in place of the current sentence at Line 439-440. "*The co-occurrence of volcanic eruptions and suppressed drought conditions over the MDB appears to contradict current understanding of the impacts of eruptions on the El Niño-Southern Oscillation (ENSO), and subsequent impacts on rainfall in the MDB (Gillet et al., 2023)*"*

Gillett, Z. E., Taschetto, A. S., Holgate, C. M., & Santoso, A. (2023). Linking ENSO to synoptic weather systems in eastern Australia. Geophysical Research Letters, 50, e2023GL104814. https://doi.org/10.1029/2023GL104814

L 443-L445 needs to be clarified. At present it appears to contradict the first sentence of the conclusion.
Author response: *We will change "*exceptional*" in Line 444 to "*unprecedented*".*