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 1, we will make the following changes—including clarifications to our methodology—to make our findings more robust:

- 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 MDB drought 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

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 study focuses on deepening the understanding of the natural ranges of Australian droughts, which can be used to better assess intrinsic and externally forced drought risks in future planning. They address this by comparing droughts in the 20th century (1900-2000) based on observations and models, with simulated droughts during the pre-industrial millennium (850-1849). They seek to assess if drought characteristics (mean duration, maximum length, intensity, etc.) have changed during the last century, compared to the past millennium.

One of the main conclusions is that multi-year droughts have been longer on average during the 20th century over part of Australia (including the MDB), compared to the last millennium, and that anthropogenic forcing is the likely cause of this change.

The authors also conclude that having a larger sample (pre-industrial millennium) allows for a better characterization of natural drought variability, reaching out extreme events (longest droughts) that have not been observed over the last century. Based on this, the authors conclude that such extreme events are part of the natural range of droughts in Australia, and thus they can be expected in the future. This, superimposed with projected drying trends, pose critical challenges for adaptation planning.

The article is well written and the motivation and research question is clear. However, there are some methodological aspects that should be addressed before drawing robust conclusions:

> 1) There are models that perform better than others during the historical period, and this is quantified as part of the analysis (Sect. 2.2.1; 2.3.3; Supporting Fig. 19). In this line, the interpretation of results should also account for these different performances. We should trust more those models that better represent the observations in the historical period, right?
>
> For example, if the physical mechanisms represented by a particular model structure leads to lower interannual precipitation variability compared to observations, it is expected that its simulations during the last millennium

reflect the same bias, and vice versa for the case of higher interannual variability. However, the conclusions of the paper are based on the average of all models, independently of their performances during the historical period.

Author response: *We did not weight the multi-model averages as there is evidence from studies analysing future projections that weighting for model performance does not necessarily produce better projections (e.g., Abramowitz et al., 2019). Such weighting approaches can be problematic as they can lead to results that are heavily weighted towards a small number of highly dependent models and there is no universally agreed way to do such weighting (Eyring et al., 2019). There are additional considerations in our case, including:*

1. *there are many non-significant correlations, meaning the values are not necessarily meaningful, and*
2. *as we state in the paper (and show in Supp. Fig. 20), in the short 101-year time period of comparison, there is a high random element to the multi-year drought metrics.*

This means that the weighting may not necessarily be a fair representation of each model's long-term skill in simulating multi-year drought characteristics.

Given this context, ***at the Editor's discretion*** *we will re-calculate the ensemble-mean values (shown in Figs. 5-7), weighting according to each model's correlation with observations of that same metric. If not, we will add a statement to the end of Section 2.3.2 stating:*

*"*We calculated arithmetic multi-model means rather than weighting the models according to the spatial correlations. Evidence from future projections suggests that weighting for model performance does not necessarily produce better projections (e.g., Abramowitz et al., 2019). Additionally, weighting can lead to results that are heavily weighted towards a small number of highly dependent models and there is no universally agreed way to do such weighting (Eyring et al., 2019).*"*

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dynam., 10, 91–105, https://doi.org/10.5194/esd-10-91-2019, 2019.

Eyring, V., Cox, P.M., Flato, G.M. et al. Taking climate model evaluation to the next level. Nature Clim Change 9, 102–110 (2019). https://doi.org/10.1038/s41558-018-0355-y

Given the large differences between models (spatial precipitation patterns, MAP values and performance against observations), I don't think mean ensemble values can be directly used for interpretation of results. For example, the assessment of the "Possible anthropogenic influence on Australian multi-year droughts" (Section 3.4) relies on the ensemble mean of models, however, we know that there are models that perform better than others.

Author response: *Analysis of ensemble mean values is extremely common, particularly when using CMIP/PMIP ensembles; it is a straightforward way of identifying common (and therefore climatically meaningful) signal whilst smoothing over individual model biases. Additionally, we do not purely rely on the multi-model mean, but also show all individual model results in the Supplementary Information. If we weight the means according to the Reviewer's request, the ensemble-mean values will more closely reflect values of the models with better apparent skill in this historical period. However, this is not necessarily a guarantee of more accurate results, for the reasons outlined above.*

A way to account for these different model performances could be to apply a statistical correction to the models before analyzing droughts, similarly than those applied to GCMs in the historical period before analyzing their future projections (e.g., Cannon, 2018 and references therein). This data-process involves that each model is

> corrected according to their own performances in the historical period, and then results can be interpreted similarly across models.

Author response: *There are multiple different ways to bias-correct and the choice of correction method is subjective and can strongly influence the results. Using CMIP5 simulations over Australia, Vogel et al. (2023) for example recently showed that different bias correction methods can lead to large differences in simulated rainfall climatology, variability and extremes. Furthermore, no single bias correction method was able to outperform the others when evaluated for multiple metrics. Choosing one bias correction method for our study would thus be highly subjective and a comparison of multiple methods is out of the scope of this study. We also note that our method accounts for some of the differences in the simulated MAP by applying the drought metrics separately to each model's own climatology (such that all drought metrics are calculated relative to the model's own mean).*

*To clarify this, we will add the following statement at Line 150: "*Note that we do not use the results of this verification to bias-correct the models. Vogel et al. (2023) used CMIP5 simulations of Australian precipitation to demonstrate that different bias correction methods can lead to large differences in simulated rainfall climatology, variability and extremes, and that no single bias correction method outperforms the others when evaluated for multiple metrics."

Vogel, E., F. Johnson, L. Marshall, U. Bende-Michl, L. Wilson, J. R. Peter, C. Wasko, S. Srikanthan, W. Sharples, A. Dowdy, P. Hope, Z. Khan, R. Mehrotra, A. Sharma, V. Matic, A. Oke, M. Turner, S. Thomas, C. Donnelly, V. C. Duong, An evaluation framework for downscaling and bias correction in climate change impact studies, Journal of Hydrology, Volume 622, Part A, 2023. https://doi.org/10.1016/j.jhydrol.2023.129693

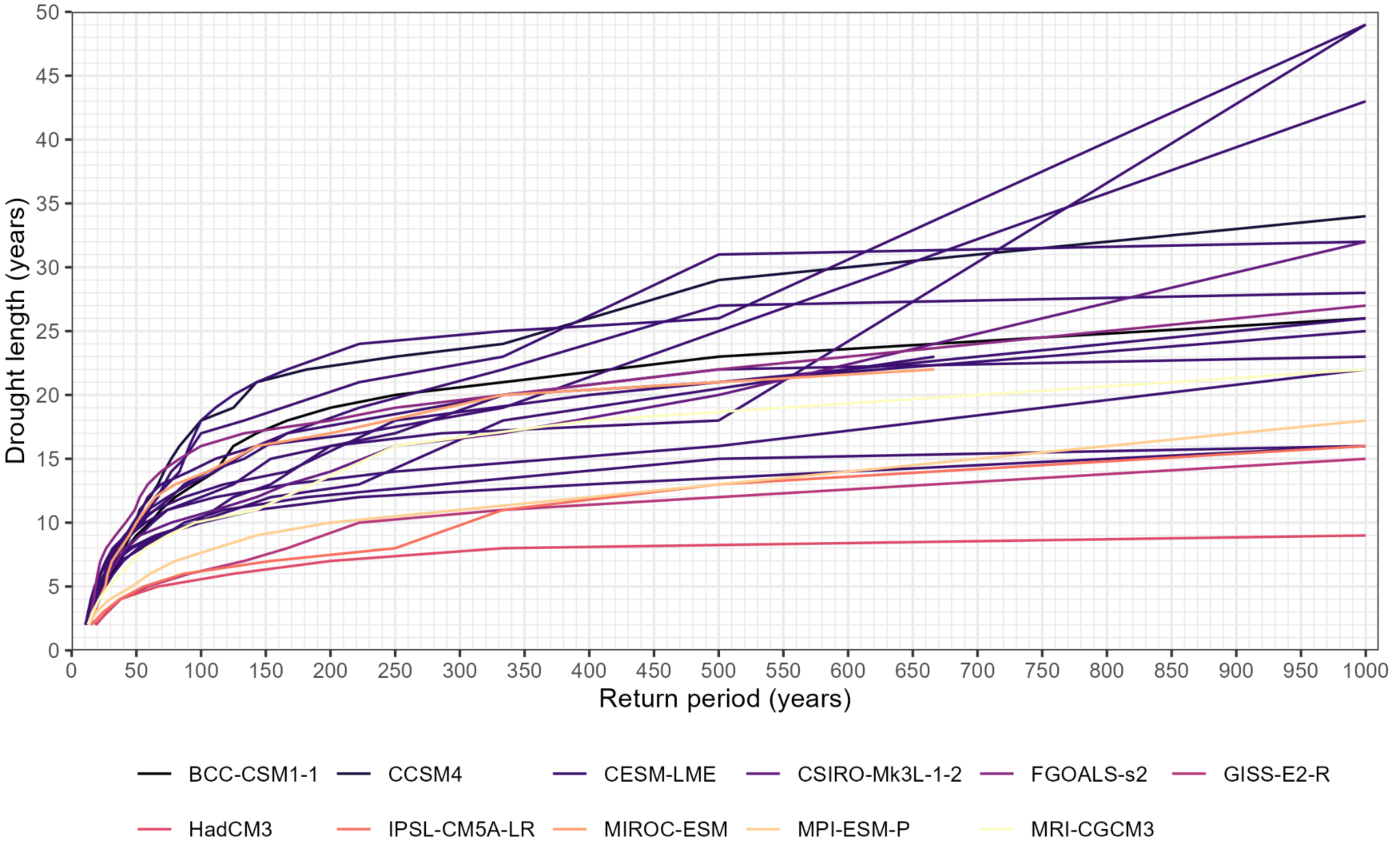
> From Sect. 2.3.2, it is inferred that droughts are defined as deviations from the climatology of each model (right?) If the models are bias corrected, the same climatological mean (that from AWAP) could be used for drought definition. And direct comparison between models could be applied, instead of %. This is easier for interpretation than "For example, 0% represents the climatological mean precipitation, and 100% represents zero precipitation". Same for severity, it would be much easier to compare directly mm across models, instead of % ("For example, a value of 200% represents a total deficit equal to two years of mean precipitation.")
>
> Comparing deviation metrics as % is influenced by the native MAP of each model (deviations from a low absolute MAP values represent larger % than when the MAP is larger). By comparing Fig 2.a and Fig. 3, it can be seen that some models have MAP biases up to 100%, with similar absolute biases that observed MAP in Fig. 2a.

Author response: *In the case of our analysis, all drought metrics are calculated relative to the models' own climatologies. Therefore, correcting for MAP biases would have a minimal impact on the results as most of the drought metrics presented are based on years above/below the models' own climatological means. The Reviewer is correct that bias-correcting the mean would allow a direct comparison in mm but the relative intensity metric used here in fact already allows for direct comparison across models. Changing this to mm would not change the relative differences across models, and hence would not affect our results or conclusions.*

> *Cannon, A.J. Multivariate quantile mapping bias correction: an N-dimensional probability density function transform for climate model simulations of multiple variables. Clim Dyn50, 31–49 (2018).* https://doi.org/10.1007/s00382-017-3580-6

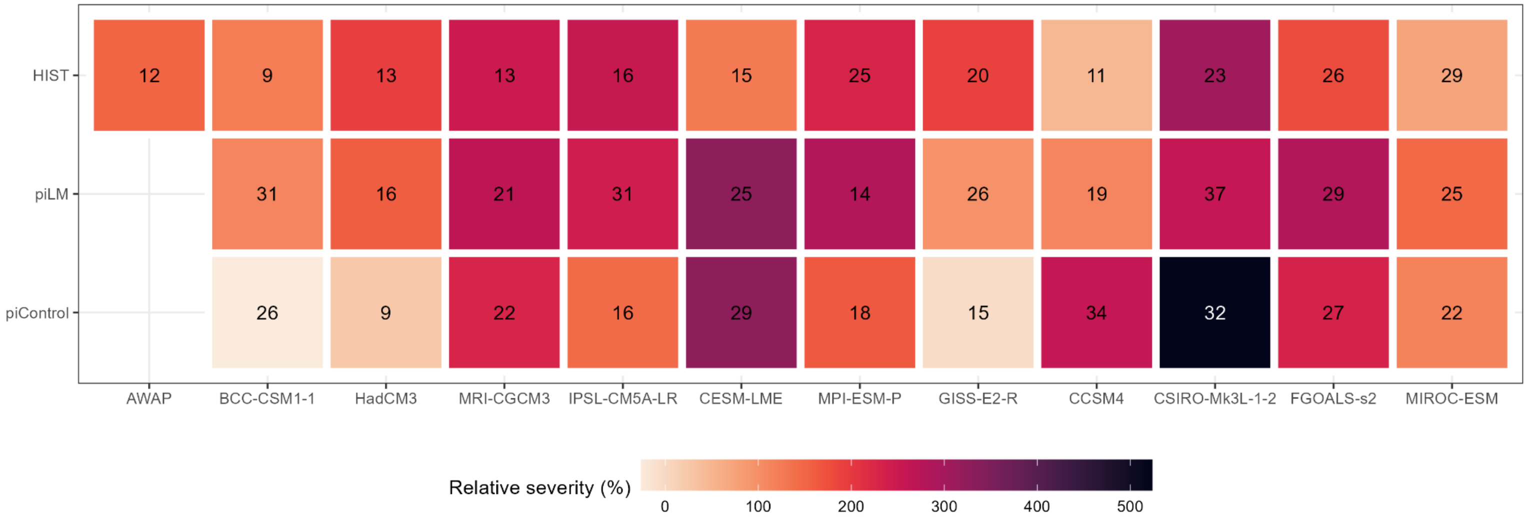
2) Having a more extreme event in a large sample can be somehow expected, but I am missing an assessment of the return period of such events. The longest droughts simulated over the pre-industrial millennium, can be expected to happen over the next century, couple of centuries, thousand years?

Author response: *We will include the following plot as an additional Supporting Figure, showing the return period of multi-year droughts in the Murray-Darling Basin, in each of the PMIP3 models' piLM simulations. In most models, the longest drought occurs only once, giving an estimated return period of around 1000 years (the length of the simulations). However, most models still simulate very long MDB droughts (10-20 years) with relatively short return periods of ~100-150 years. We will add a brief discussion of these results in Section 3.3*



In the same line, I think that for providing evidence for adaptation planning, the longest droughts should be assessed in conjunction with their deficits: it is not the same to communicate that 20-years of minor droughts (e.g., 0-10% deficits) can be expected that to communicate that 20-years of severe droughts (e.g., >40% deficits) can be expected in the future. This could be done by accounting for relative severity together with maximum length.

Author response: *Given this adaptation-focussed analysis is most likely to be useful for the Murray-Darling Basin, we will add a new Supporting Figure showing the relative severity associated with the longest drought. The figure will be similar to what is shown on the following page, where tile colour corresponds to the relative severity of the longest drought, and the text annotation states the length of that drought, in years. As well as the necessary additions to the Methods and Results, we will state in the Discussion that although the piLM and piControl simulations generally produce longer maximum MDB drought length than the HIST simulations, they are not always more severe in terms of total deficits throughout the drought (in the case of this simple comparison).*



Minors comments:

Supp. Fig. 7: "Mean multi-year drought length in (a) observations (1900-2000) and (b-l) model simulations of the pre-industrial last millennium (850-1849). Showing the CESM LME ensemble mean." It should say, panel Fig. 7l presents the CESM LME ensemble mean. Same for all figures.

Author response: *Thank you for catching this. We will modify the captions for Supporting Figs 1–18 to clarify this.*

Title: it is a complicated title that I don't think is communicating the main messages of the paper. I recommend the authors to consider a simpler one.

Author response: *We agree that it is quite a long title(!) however we consider that it does convey our main takeaway messages. An alternate version could be "*Australian multi-year droughts show emerging anthropogenic change with potential for megadroughts that exceed recent historical experience*", however that is not much less complicated. Alternatively "*Australian multi-year droughts show high variability across centuries and an emerging anthropogenic influence*", however we would like to emphasise the (novel, policy-relevant) finding that natural variability in Australian droughts can produce droughts that exceed historical experience. Of these two alternate titles, the first is our preference. We would be more than happy to work with the Editor in coming up with a shorter title.*

Abstract: "Model simulations suggest future droughts across Australia could be much longer than what has been experienced in the twentieth century, even without any human influence." This can be misunderstood as future projections, please re-phrase. An option could be: Drought simulations over the last millennium suggests that future droughts across Australia could be much longer than what has been experienced in the twentieth century, even without any human influence.

Author response: *We will modify the sentence to "*Model simulations of droughts over the past millennium suggest future droughts across Australia could be much longer than what has been experienced in the twentieth century, even without any human influence*".*