Ukkola et al. analyse an important topic, the impact of climate change on three drought types in Australia, including their seasonal changes. They use an ensemble of 32 members (4 x 4 x 2) based on the Australian Landscape Water Balance model (AWRA-L), forced with downscaled and bias-corrected 4 CMIP5 models. They used three statistical bias correction approaches and one combined downscaling and bias correction approach for two RCPs (4.5 and 8.5). Their results suggest an overall increase in all three drought types (meteorological, hydrological, and agricultural), particularly in winter and spring. They also attempt to quantify associated uncertainties. The topic of this manuscript is relevant to HESS readership and it is a nice contribution to the community. I find the current manuscript is very suitable for publication in HESS after addressing some minor comments listed below. The paper is mostly very clearly written and well-referenced.

We would like to thank the reviewer for taking the time to review the manuscript and for their positive assessment.

Here are my minor comments:

- The abstract does not clearly address, which drought characteristics you quantify. You mention only one aspect in the abstract, L24: "time spent under drought", which could be better referred to as "drought duration"? From the abstract should be clear whether you also considered other characteristics, such as severity or spatial extent. If not, then the title of the manuscript should be adjusted accordingly. Also, L30 " future increases in drought" doesn't specify which aspect of drought is analysed. Also, in L98 in the Introduction, they say "across different indicators of drought", but they don't explicitly mention them. You mention them for the first time on L231ff. Results on drought intensity should be also mentioned in the abstract.

  Thank you for pointing this out. We analyse three drought metrics (time under drought, duration and intensity) and will modify the abstract to make this clearer as well as mention the results on drought intensity and the different drought indicators used (L98).

  We have opted to not change L30 as this statement applies to all drought types and metrics analysed here. The details are provided earlier in the abstract and this statement is to provide an overview of the results.

  We note that "time spend under drought" and "drought duration" are not identical metrics. Time under drought is the proportion of months in drought over a time period, whereas drought duration is the length of individual drought events (as detailed in section 2.2.3). Drought duration describes the sequencing of drought months and could be very different for the same number of drought months depending on how they cluster (e.g. lots of short droughts vs fewer longer droughts). Hence we consider both in the manuscript for completeness but mainly concentrate on time under drought as the key metric to keep the manuscript a manageable length.

- Line 109: "model that is calibrated towards observed river streamflow, satellite soil moisture and evapotranspiration across the continent." It is not clear, whether it was done in this study, or they refer to any other previous work, This was done as part of previous work (as we state on in the Methods, the projections were obtained from the National Hydrological Projections dataset and not developed by us). We will reword the sentence and add a reference.

- Lines 121-124: did you do any of these evaluations, which you could possibly include in the supporting information file?

These have also been done as part of previous work with the relevant reference provided (Frost and Wright, 2018). We have provided an evaluation specifically for drought metrics (Figures 5-7) and will also add an evaluation against observed streamflow droughts (see comment below).

- Section 2.2.1. on historical observations should include some of the model's evaluations. I fully understand that "gridded runoff and soil moisture observations are not available", but still, you could compare the routed runoff against observed streamflow observations to assess the credibility of your simulations in the historical period.

We will add an evaluation against observed streamflow using a newly-developed CAMELS-AUS v2 streamflow dataset (Fowler et al., 2024). This is the most comprehensive streamflow dataset of unimpaired catchments currently available for Australia. We will use data over the period 1970-2020 for the evaluation as was done elsewhere in the paper (Figures 5-7).

- Lines 161-179: was this done by the authors, or taken from other authors, i.e., Peters et al.?

As stated on L149, we obtained the projections from the readily available NHP dataset. We will reword L161-179 for clarity.

- Line 197: are you sure that you are able to obtain steady-state conditions of your model's states in 10 years, given the very arid region? I guess in the dry regions, it can take longer… Did you check the time series of selected pixels? I can imagine the disagreement in results (grey in Fin 3) in central Australia could be driven by this factor of too short initialization.

While the GCM- forced runs analysed in this study begin in 1960, the initial conditions come from a much longer reference run forced with observed climatology (beginning 1911, effectively giving an extra ~50 years for spin-up). The GMC forcing has been bias-corrected to the same observations and as such should not result in large jumps in the forcing meteorology. As such we have only removed the first 10 years from the projections. Removing more years would result in a too short a time series to analyse historical drought metrics reliably. A key factor in model disagreement in central Australia is likely the highly stochastic nature of rainfall in these arid desert regions which would also be reflected in the runoff and soil moisture simulations (as we discuss on L638 onwards).

- Section 2.2.3 You run the model at a daily time step, and then drought analysis is done at a monthly time scale. I guess, you need to state this somewhere explicitly, possibly in this section. And then you apply 3 months averaging. This sequence needs to be stated clearly here in the section. Then, I would suggest moving L206-214 elsewhere because they are a bit distracting where they are. I would start directly with L203-205 and then move directly to L223 onwards. L205-208 could go to discussion.

We have opted to not move L206-214 as these detail the drought threshold used including why it was chosen and are an important part of the methods. This information also directly follows our choice to use percentiles to identify droughts and is thus appropriate in its current section. However, we will shorten this section for clarity.

- L 231: you could have also analysed spatial drought extent? Do you see distinct results for behaviours in duration and time under drought? If not, then I would suggest keeping just one.

As discussed above, time under drought and duration are not identical metrics and provide different information. Hence we have opted to present both. We do not think changes in spatial drought extent would provide substantially different information compared to the metrics already presented as it would largely follow the country-wide results presented in Figure 2 and the NRM region specific results in Figures 6-7 (we assume the reviewer means the commonly used "area under drought" metric). As the paper is already fairly long, we have opted to not add additional metrics.

- In results, the results quantify the drought types. It would be interesting, which aspects lead to the runoff droughts, which seem to be by 20% longer, it's not only because of precipitation deficits, but surely from evaporative increases due to increased temperature? Also, in Fig.2, there are the reference values missing, to better relate the percentage increase to a reference. The 20% would be different from 2 months or from 4 months … ?

Figure 2 shows that the robustness of runoff droughts is in fact similar to precipitation (e.g. 17% of land area shows robust runoff drought changes vs 20% for rainfall under RCP4.5). However, it is true that the magnitude of changes tend to be larger for runoff.

There are several possible explanations for this. One likely factor for the stronger runoff increases is the amplification of precipitation changes in runoff. Because runoff is a smaller component of the water cycle compared to P and ET over most of Australia, any given relative change in P or ET is amplified in runoff, resulting in larger relative changes in runoff. There are several observed examples of this in Australia e.g. from southwestern Australia where 15-20% declines in rainfall led to >40% declines in dam inflows (Petrone et al., 2010). Ukkola et al. (2016) also found that fairly small changes in ET of ~6% led to much larger changes of 25-30% in streamflow across Australian catchments. A third possible factor is changing evaporation. We will further quantify this aspect using evaporation outputs from NHP in the revised manuscript.

As for Figure 2 reference values, this figure shows the ensemble mean future change in time under drought relative to the historical baseline (i.e. the fraction of time under drought). As we state on

L284, the reference value is 15% as per our definition of droughts as months below the 15th percentile:

> "*Fig. 2 shows the ensemble mean future change in time under drought relative to the historical baseline (during which ~15% of the time is under drought as per our definition).*"

Reference values for drought duration and intensity are presented together with the future changes in Figures S1-S2.

- It might be useful to rearrange the sequence of the results. How about starting with 3.4, where observations are compared, and show basic characteristics of individual realization (Fig 5), then showing aggregated characteristics of drought for the full ensemble (Figs 2 and 3)…

We prefer keeping the current order as Figures 2 and 3 present the overall continent-wide results. The subsequent results then explore the specific sources of uncertainty (e.g. GCMS vs bias correction/downscaling methods) and the reliability of these projections (through a comparison to observations). We feel presenting the aggregated results first gives necessary context to the subsequent sections.

- Why does the GFDL model stand so much apart? Is it because of precipitation or temperature differences?

Yes GFDL tends to project hotter and drier conditions in the future compared to the other GCMs. We discuss this in section 4.2:

> L565: "[..] *we found that ensemble members using GFDL-ESM2M as the forcing model were particularly anomalous compared to the rest of the NHP ensemble, indicating stronger increases in most regions. [..] The GFDL-ESM2M model projects greater future warming and drying over Australia than the other GCMs used here (Peter et al., 2023); our finding of larger drought increases in GFDL-ESM2M are consistent with this tendency*."

- Nicely written discussion section, but could the conclusions be taken apart into one paragraph section at the end?

We will add a brief conclusions section to the revised manuscript.

**References**

Fowler, K. J. A., Zhang, Z., and Hou, X.: CAMELS-AUS v2: updated hydrometeorological timeseries and landscape attributes for an enlarged set of catchments in Australia, Earth System Science Data Discussions, 1–21, https://doi.org/10.5194/essd-2024-263, 2024.

Petrone, K. C., Hughes, J. D., Van Niel, T. G., and Silberstein, R. P.: Streamflow decline in southwestern Australia, 1950-2008, Geophys. Res. Lett., 37, L11401, https://doi.org/10.1029/2010GL043102, 2010.

Ukkola, A. M., Prentice, I. C., Keenan, T. F., van Dijk, A. I. J. M., Viney, N. R., Myneni, R. B., and Bi, J.: Reduced streamflow in water-stressed climates consistent with CO2 effects on vegetation, Nature Clim Change, 6, 75–78, https://doi.org/10.1038/nclimate2831, 2016.