Review: Multi-annual predictions of hot, dry and hot-dry compound extremes
The study by Aranyossy et al. investigates the predictability of hot-dry compound extremes and their univariate components (hot and dry events) on a multi-annual scale (forecast years 2–5), using decadal climate hindcasts from the CMIP6 Decadal Climate Prediction Project (DCPP). The study evaluates the skill of initialized forecasts compared to historical simulations, explores the relative contributions of external forcings versus initial conditions, and assesses whether the model ensemble can reproduce observed relationships between compound and univariate extremes.

This manuscript addresses a timely and scientifically relevant topic—the decadal predictability of compound hot-dry extremes—and provides valuable insights using a multi-model ensemble of CMIP6 decadal forecasts. However, the current version has some substantial shortcomings that limit its scientific clarity and impact. I believe the study requires major revision to strengthen the methodological framework, sharpen the interpretation of key results, and improve its overall scientific accuracy before it can be considered for publication.

We thank the reviewer for taking the time to review our manuscript, and for the constructive feedback. Please see below your comments and our answers highlighted in red.

**Major Notes:**
Unclear Justification for Compound Event Definition: The definition of hot-dry compound extremes is based on overlapping thresholds (TX90p and SPI3dry or SPEI3dry), but the manuscript does not adequately discuss how sensitive the results are to these thresholds or to the chosen accumulation window. The absence of sensitivity analysis raises questions about the robustness of the results.
In response to the reviewer's comment, we have performed a comprehensive sensitivity analysis using different thresholds and included the results as supplementary information. For hot extremes, we use an alternative threshold of the 95th percentile of the maximum temperature (TX95p). For dry extremes, we select accumulated months from SPI3 and SPEI3 with values <= -1.5 (SPI3-1.5 and SPEI3-1.5). Hot-dry compound extremes have also been calculated with the new thresholds. Additionally, we have performed a sensitivity analysis for 6-month and 12-month accumulation periods in the calculation of SPI and SPEI. The results have been added to the Appendix (Figures A7-A12). Based on the results of this analysis, we have observed that a change in the threshold does not yield significant differences. On the other hand, a change in the accumulation period reveals a corresponding change in correlations, particularly in specific regions such as Northern Asia and Southern Africa. We have added a paragraph in Sections 3.1 (line 174) and 3.2 (line 206) comparing the different thresholds and accumulation to the correlation and residual correlation, respectively. In Section 4 (line 329), we have added a paragraph summarising and discussing the overall results of the sensitivity analyses.

Overstatement of Skill Based on Trend Agreement: The study repeatedly refers to "skillful prediction" of compound extremes, but a substantial portion of this skill stems from long-term trend agreement rather than the successful prediction of interannual variability. In many regions, the DCPP ensemble appears to simply capture externally forced warming trends, which correlate with observed trends in hot extremes. However, this does not necessarily equate to predictive skill in a practical, decision-relevant sense. This distinction is mentioned, but not emphasized

sufficiently in the framing of the results or the conclusion. The authors must clearly distinguish between correlation due to trend matching and actual initialized predictive skill.
We agree with the reviewer's comment, and in response, we have attempted to better highlight the implications that a lack of residual correlation brings to overall results. Specifically, in Section 3.2 (line 250), we have added a sentence highlighting how the results of that section underscore the limitations of the current predictive potential of multi-annual predictions for hot-dry compound extremes. In addition, we have added a paragraph in the discussion Section (Section 4, lines 310-316) that underlines the distinction between forcing-derived and initialisation-derived skill, and its impact on the usability of these forecasts.

Presentation and Clarity: The text is dense and often difficult to follow due to inconsistent terminology and lengthy, complex sentences. Key methodological steps are underexplained or relegated to figure captions.
In response to the reviewer's comment, we have proofread the manuscript, with a particular focus on the delivery and clarity of the text. Specifically, we have homogenised the terminology used (Example: referring to hot-dry compound extremes throughout the methodology, results and discussion, instead of using several expressions such as events, extremes... we use only "hot-dry compound extremes"; we use the term "prediction" for the decadal predictions instead of referring to them as forecast, predictions, products...). In addition, we have shortened several otherwise lengthy sentences throughout the manuscript (Example: Section 1, lines 18-20, the sentence "The combination of ... severe underestimation of the risk" was split into two sentences; in Section 2, lines 36-39, the sentence "Unlike climate projections ... model initialisation" was split into two). Finally, we have also added some explanations, which were previously in the captions to the main text (Example: in section 3.2, lines 219-223, we have added a paragraph explaining Figure 3 and how we represented the connection between significant correlations and agreements of the trends' sign.)

**Minor Notes:**
Figure captions could benefit from clearer labeling and direct interpretation; currently, they are overly technical.
To integrate the reviewer's comments into the manuscript, we have reduced the amount of technical information in the captions, especially that which is already provided elsewhere, thus avoiding repetition between the text and captions. For example, in Figure 3 we have removed the explanation of the correlation-trend's sign and added a paragraph in the text. We have removed from the Figures the information stating the two datasets involved. For example, the sentence "Correlations are performed between DCPP MME and GPCC-BEST" has been removed, at least in the main paper, since in the Appendix, where we present the results for ERA5, we believe it is better to specify the dataset. Additionally, we have removed the part in the captions stating the alpha value, as well as the part where we specify whether we used a two-tailed or one-tailed test, since this information is already specified in the methods. Finally, in the Appendix Section, specifically in Figures A4 and A5, we also removed repetitions of the variables shown in the Figures.

l.87 add "as" (... we define months with drought conditions as all months ...)
Added.

Several commas are missing throughout the text.
While proofreading the text, we have added the missing commas.

Geographical areas are not always written in the same way (for example, word “northern“ is sometimes written in lower case and sometimes in upper case).
To homogenise the text, we have checked the text and corrected the typos, following the rule where “northern/southern”, when used as adjectives, are written in lower case (“southern part of... “), while when part of a proper noun, they are written in upper case (Example: North Africa, South America, Northern Hemisphere).

l.238 delete “seen“
Deleted.