This paper motivates that the assessment of decadal climate predictions is essential to providing reliable information on hot-dry compound extremes because of their potential impacts on environments and societies. There is mention that previous research has focused mostly on climate variables or univariate climate extreme prediction at this timescale and this study aims to fill that gap for hot-dry compound extremes. It evaluates the ability of the CMIP6 multi-model decadal climate hindcasts in predicting hot-dry climate extremes as well as hot and dry univariate counterparts for forecast years 2-5, investigates the added value of model initialization by comparing the forecasts to historical simulations, and compares the modeled correlations between compound and univariate extremes with observation-based datasets, ERA5 and GCPP-BEST (as referred to in the paper).

While the paper addresses a relevant gap and performs a worthwhile correlational analysis, much revision is required to make this study well-explained and well-interpreted.

We thank the reviewer for taking the time to review the paper and post the public comment. We did our best to take your comments into account and improve our manuscript. Please see below our answers.

Major Points – Scientific Methods
Strengths:
Two different CMIP6 experiment types (DCPP MME and Hist MME) are a meaningful way to explore the skill contributions for initialization and external forcing.

Good choice to use multiple reference datasets (GPCC-BEST and ERA5).

Critiques:
Model initialization details should be discussed (e.g. initialized over land/sea). - Model selection and some information/justification concerning the models should be mentioned instead of simply referring to the appendix table.
Model initialisation in the DCPP project is not standardized. For this reason, we refer to the reference in the appendix table for each single model, as each one of them applies different methods and data. We have added a sentence in section 2 (line 63) referring to this issue.

More discussion about the potential oversimplification due to all of the ensemble averaging should be discussed.
Please note that the extremes are calculated for each ensemble member individually, without averaging. In climate prediction, the ensemble mean is typically interpreted as the predictable component of the signal. We follow the same approach in this study, and to make predictions, we average the member-specific extremes to derive the predictions. There are well-known issues that the magnitude of variations in the ensemble mean is smaller than the observed variations, a direct consequence of averaging the noise component across individual members.

"Observational uncertainty" is only quoted however, discussion is required to compare the general higher agreement with GPCC-BEST in comparison to ERA5 at least.
Thank you for pointing this out. We have added a paragraph in section 4 (line 326) comparing the results between the observation and the reanalysis reference dataset.

The compound extremes definition is somewhat rigid and may miss more complex co-occurrences. There should be consideration given to the simultaneity of event dynamics.
We understand and agree that we computed the compound extremes based only on statistical properties, which are meaningful in this specific study. Computing the compound extremes based on dynamical characteristics may also be very relevant. However, in the context of this paper, this is out of scope, but could be investigated in future studies.

A 3 month accumulation window is chosen and is appropriate for meteorological drought but may not capture many other drought timescales. The sensitivity to accumulation period should be discussed.
We have performed a sensitivity analysis for a 6-month and 12-month accumulation period (as well as different extreme thresholds). The results have been added to the supplementary material (Figures S7-S12). 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 on the overall sensitivity analyses.

Major Points – Explanations
Strengths:
Model-dependent variability is mentioned.

Critiques:
Explanation of indices calculation relies on references and is otherwise unclear/not justified (e.g. why use the analysis period as reference period for TX90p and are observation years used for the comparisons aligned with forecast years 2-5?). Should include some explanation similar to references (e.g. for SPI/SPEI calculation and standardization processes).
In section 2.2 (line 124), we have added a sentence explaining how the reference datasets are aligned with the average forecast years 2-5 for the correlation analysis. In addition, we refer to the calculation of the extremes to studies which are widely known and established in the literature. More specifically, TX90p is part of the ETCCDI suite of indices (we refer in the manuscript to the study of Zhang et al. (2011)). On the other hand, for SPI and SPEI, we refer to the studies McKee et al. (1993) and Vincente-Serrano et al. (2010). Finally, we use the entire time series as a reference period for TX90p to ensure consistency with the reference periods of SPI and SPEI.

Methodological details feel condensed or fragmented:
Mention that PET is calculated using the Hargreaves's method but no justification is provided.
In section 2.1 (line 99), we have added a brief explanation of why the Hargreaves method was reasonable for this specific study.

Description of percentile estimation using a 5-day running window is noted but how seasons are handled should be discussed.
The running window is short enough not to incur in the seasonality biases that are common in the longer window periods. We have added a sentence in section 2 (line 84) explaining this concept, and also added a reference (Brunner et al., 2021).

Missing data mentioned in figure captions but not discussed.

We have added a sentence in section 2 (line 75), indicating how missing data are part of observational datasets.

A statement about why hot-wet compound extremes are not discussed would strengthen the focus of this analysis.
We appreciate the suggestion, and indeed an investigation of different types of compound extremes would be interesting to analyse. However, in this study we focused on this specific type. We believe that analysing also hot-wet compounds would broaden too much the spectre of this specific research (including the need to define indices that meaningfully reflect e.g. the health risks of hot and humid conditions), and suggest this would be best to do in future research.

There is mention in the introduction that dependence among univariate variables of a compound extreme can decrease the return period of such events but this is not discussed or mentioned again any further.
The mention of the return period has been removed.

Major Points – Conclusions/Interpretation of Results
A general comment about the statement of results: overall it reads a bit as a list, with a lack of sufficient meaningful connections made or real reasons explored for greater "skill" in particular regions compared to others.

Critiques:
The minimal added skill from initialization is important, if a real artifact, but under-discussed – what does this imply for the use of decadal forecasts in operational contexts? In section 3.2 (line 183), we have added a sentence explaining the usefulness of residual skill. In addition, at the end of the section, we have added a sentence explaining how that section underlines the limits of the current predictive potential of multi-annual predictions for hot-dry compound extremes. In addition, we have added in the discussion (section 4, lines 310-316) a paragraph underlining the distinction between forcing-derived and initialisation-derived skill, and its impacts on the usability of these forecasts.
The regional and seasonal variations are mentioned as future work but some exploration into this would give more content/meaning to the study. A preliminary investigation would be interesting and relevant given the spatial nature of the data and existing literature.
We agree that exploring seasonal and regional variations would be a significant contribution to the field in which this study is situated. However, we believe that this would broaden the focus of this study; therefore, we prefer to reserve it for a future investigation.

Statements about better/under-performing regions are made but connections between them and between areas for similar analyses should be explored. These really should be accompanied by considerations of the frequency and intensity of these extreme events in those areas (e.g. high skill in California important for multi-day droughts in the region).
At this point, we can only speculate regarding the sources of better/worse skills in specific regions. However, a more conclusive and detailed analysis of the sources of this skill in such regions should be conducted in follow-up studies.

Recommendations for Further Discussion/Improvement
Aforementioned justifications/clarifications should be included. Thank you for the suggestions provided. We have tried to include them as best as we can.

Schematic for event selection (compound index calculation) would enhance clarity.
We appreciate the suggestion. However, we believe the compound index to be already quite schematic. In addition, the study of De Luca and Donat (2023) already offers a quite comprehensive explanation of this specific method.

Consideration of other compound event types such as humid heatwaves, lagged dependences (or explanation for why specifically these that were chosen) would expand the framework's applicability.
As stated above, we believe that considering other types of compound extremes and dynamics could be a valuable contribution to the topic, but is out of the scope of the current study. We have added a comment in Section 4 (line 345) where this framework should be applied also to other compound indicators.

Mention and discussion of the socio-economic impacts since the motivation centers on "high-impact" extremes but is never discussed again. Linked the compound events to impact datasets could strengthen the relevance and make the closing statement make more sense.
We agree that an exemplification of the impacts would strengthen the manuscript's motivation. We have added a sentence in Section 1 (line 24), where we put a specific example of the effects of the hot-dry summer in Russia in 2010.

Specific mentions of events should be made rather than "compound extremes during 2003, 2010, 2015 and 2018 in Europe stand as an example".
We have extended the sentence in Section 1 (line 25-26), talking about the examples of compound events. In addition, we refer to the references added at the end of the sentence in Section 1.

Sensitivity analysis on SPI/SPEI accumulation periods could reveal the robustness of event definitions.
See comment above regarding the SPI/SPEI accumulation periods.

Skillful areas should be connected to real examples to emphasize the importance of the system's performance.
We appreciate the suggestion. However, the scope of this study is to assess the skill of the hindcast over the whole time period. Exploring the hindcast's ability to predict single events would broaden the scope of this study too far.

Similarly, discussion of how the lack of skill in extreme-prone areas is a major limitation (e.g. North Africa, where certain types of hot-dry events are more common).
We discuss the limits of the skill, particularly from the perspectives of residual correlation and interannual variability. Specifically, we have added short paragraphs in Section 3.2 (line 250) and Section 4 (line 314) to highlight this limitation. However, exploring the sources of such a limitation would be, at this point, a speculation. We could conduct a more specific analysis on the topic in future studies.

Minor Points:
Figures are should be made larger and less crowded.
Thanks for the comment. However, we use a maximum of six panels per figure. We are also utilising the largest available space and the optimal panel configuration to display the figures. The plots for the correlation, residual correlation, and correlation-trend analysis are created to highlight the hot-dry compound extremes.

This is small but there are too many references to the appendix mostly in the discussion for the ERA5 results but not presented in the body.
We use ERA5 as a secondary dataset. However, we decided to show only one dataset in the main paper, especially because the two datasets (GPCC-BEST and ERA5) are mostly in agreement.

Bias correction and model drift should be noted. This is mentioned in one of the authors' other papers referred to in the Data and Methods section.
Decadal predictions can be affected by model drift, which is typically corrected for by calculating anomalies relative to a lead-time dependent climatology. Here, we use a lead-time dependent percentile to calculate the temperature extremes. Specifically, we build a distribution for every lead-day of the predictions (with its 5-day window). On the other hand, for the standardisation of SPI and SPEI, the distribution is built for every lead-month. These specific steps for calculating the extremes, being lead-time dependent, implicitly correct for the drift in the decadal predictions. To better clarify this point in the manuscript, we have added, in Section 2.1, two brief sentences explaining these concepts, specifically at line 85 for the hot extremes, and line 105 for the dry extremes.

Grammatical/Written Structure:
Generally clear, scientific tone and use of domain-specific terminology.

Good integration of citations.

Minor/spelling/grammatical errors.
We have proofread the manuscript, correcting all the typos, grammatical errors we encountered.
Incorrect figure references (e.g. "Figure 2d" instead of correct "Figure 2c").
The reference has been corrected, and also the other references have been checked for other potential mistakes.
Incorrect figure references (e.g. "Figure 2d" instead of correct "Figure 2c").
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).
Paragraph transitions need to be made smoother and often miss key points – topic sentences should be prominent with a clear message while more concluding sentences need to be added to complete the idea of the paragraph.
We have tried to make the transitions between the paragraphs, and also the different sections of the manuscript, more fluent.