Foremost, I would like to mention, that researching compound extremes in the context of decadal climate prediction is a valuable contribution to the field of inter-annual predictions. The paper does represent the added value of initialized climate predictions well, but also shows the limitation of added value to specific variables and regions.

We thank the reviewer for taking the time to review our manuscript and provide helpful suggestions. Please see below your comments and our answers highlighted in red.

Typos:
4 - „IN this regard“
Amended.

145 - circumscribed - is there a better word possible, or just leave it out?
In response to the reviewer's comment, the sentence "are circumscribed and isolated" has been replaced with "areas are scattered and spatially isolated".

General Remarks:
0 - Is the code with which the results are produced available publicly?
To provide public access to the code, as well as the datasets used in this study, we have added a "Data Availability Statement". Here, we provide links to the public access of the datasets used in this study (CMIP6 models, GPCC, BEST, and ERA5 datasets), as well as a link to a GitLab project where the scripts used to compute the univariate and hot-dry compound extremes can be found (`https://earth.bsc.es/gitlab/aaranyos/hotdry_compounds_dcppa`).

70 - Did you apply any bias correction/calibration to the MME, individual models respectively? If not, why? Wouldn't a non-linear calibration add value to the model outputs?
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.

107 - the authors focus their study on lead times 2-5 which is in my opinion a fair choice. Could you give a reason on why you specifically choose this lead-time? And, if available, please present the findings for other lead times in the way you did for the comparison between GPCC-BEST/ERA5. Decadal predictions extend out to ten years, so I would be curious what happens to the signal of the initialization in this MME context.
The choice for the lead-years 2-5 was taken due to previous consultation of skill assessments of decadal predictions, which indicate the first lead-5 years to have a greater impact from the initialization. To help highlight this point better, we have added a sentence in section 2 (line 67), where we explain how the effect of initialization on this subset of decadal forecasts makes

them a desirable timeframe, especially for the skill assessment of extreme events. We also refer to Section 1 (lines 44-49), where we explicitly state how the first years of decadal forecasts show the most promise in terms of predictive skill. We have not performed the analysis of this study on the whole decadal period, but we would expect a slight degradation of the skill in longer lead times, especially regarding the contribution from initialization. We also refer to the study from Delgado-Torres et al., (2023) "Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes", where in addition of the 2-5 lead years in the main study, in the Supplementary Material figures are shown also for the skill of the whole 10 years. The results show small differences compared to the 2-5 lead-year period, expect a degradation in the residual skill.

135 - In a production scenario, if you only had time to use one reference data set, would you recommend to use gridded observations or reanalysis?
Skill evaluations are always done for hindcasts, which for all decadal prediction systems are produced before any production or (operational) forecasts. This allows to evaluate the hindcasts against all available observational datasets, and in fact, it is important to take potential uncertainty related to the observational reference product into account; we have added a brief discussion on these uncertainties in section 4 (line 326). The forecasts as such would not depend on the reference dataset used for hindcast evaluation. For these reasons, we do not recommend one over the other.

149 - Given the interpretation of skill occurring in certain regions is difficult, I'm curious why Greenland and "Central Asia" stand out as showing increased skill due to initialization. Do you have any idea why that might be?
We agree that investigating the sources of such strong residual skill in hot extremes would be an interesting contribution to the topic. However, at this point we could only speculate about the reasons for the added skill. For example, the strong residual correlations found in Greenland could be related in some way to the decadal predictability of North Atlantic blockings, as stated in the study by Athaniasidis et al. (2020), *"Decadal predictability of North Atlantic blocking and the NAO"*. On the other hand, the residual correlation in Central Asia may be linked to the Siberian High, a relatively stable high-pressure system in the region. However, a more conclusive and detailed analysis of the sources of the skill should be done in follow-up studies. For this reason, we did not add speculations in the manuscript at this point.