Comments on: Weather extremes over Europe under 1.5◦C and 2.0◦C global warming from HAPPI regional climate ensemble simulations by Sieck et al. (esd-2020-4)

This paper presents results from a regional dynamical downscaling of two GCM ensemble from the HAPPI project and compares the model output to 4 extreme event indices.

The large number of ensemble members and important can potentially make a good contribution in how to analyse and estimate the difference between relatively close climate change “targets”.

One major discrepancy however is that there are no evaluation of the quality of results with respect to observations or reanalysis. If the model results deviates to much from the “real world” one can not trust the conclusions given for the warm periods. The discussion does not have to be very advanced but I think you should include a point on this.

In general both the description of methods and results are not always very clearly defined. Sometimes it is just a matter of language. The sentences are sometimes so long that the reader loose the thread before getting to the end, so please try to be more focused.

A problem is that it is sometimes unclear whether the assumptions used are your / RCM limitations or inherited from the Happi protocol. Even though I know the Happi protocol it was sometimes hard to separate. I see that many of the questions from other reviewers on critical assumptions e.g. how does pre-industrial come into play, when is the future time slice .. is often related to the protocol. I think the authors should include a summary of this information, in particular as you rightly point out that the Happi protocol is quite different from the traditional CMIP settings.

The defintions/ presentations of the indices and how they are used can be improved (Section 2.2 and 2.3) It is particularly hard to follow the set-up of significance tests.
“RX5 days are computed ...similar to ATG28, however …

CDD is similar to ATG28 but with the method used for RX5 but now however?
There are no significance test for RI50yr ? There does not have to be one but I would like to know.

With only 4 indices I think you should avoid “similar” as much as possible and just describe the methods for each.

The discussion on being able to reproduce the more noisy results of the smaller ensemble by picking a smaller number of the largest ensemble was interesting. I wonder if it is any way of presenting this in a cumulative manner, with respect to number of ensemble members. The change rate may also indicate whether even the largest ensemble is too small. I realize however that this is likely beyond the scope of the article.

To the point on data availability: Is it the data used specifically for this study, or is it general RCM output.

Some minor suggestions /corrections

Line 20: “differ” not needed
Line 25: Include a line on scenario definition in CMIP6?
Line 33:skip the word “indeed”
line 70: Green-house-gas -→greenhouse gas
line 81: Sea-Ice → sea-ice (usually not capital letters)
line 121: Missing intervall (“100” years on both sides)
Figure 1. Is the outer map equal to the model domain?

Line 135-136 The explanation is fine, but then you do not need to mask it out either. It would show up as insignificant?

Line 141 NOResm → NorESM

Figure 6. The difference between the two downscaling sets are quite large. Any comments.

line 260 -265 use significance instead of unlike / similar