Authors response to referee 2:
Thank you very much for your constructive feedback. We appreciate your valuable suggestions, and incorporating them into our work will undoubtedly contribute to the improved quality and comprehensibility of our content.

1. Referee comment and authors answer

Referee:
“There is a clear shift in the quality of argumentation between the chapters 2-4 und chapter 5. In the chapters 2-4 the information given often left me puzzled why the authors choose to do it that way (unusual choice of the reference period, reference data with clear deficiencies for the purpose, partly somewhat arbitrary metrics,…). Finally, in chapter 5 some of it is put into a context. I would suggest to give some of the context earlier at the appropriate places and consider it in the result section. This might change some of the sparsely given explanations in the result section and put the descriptions there in some context earlier on.“

Authors:
“In the revised version the chapters will be modified according to your comments and incorporate important explanations in appropriate places beforehand. For example, in Section 2 "Data," we will explain why we have chosen a non-typical time reference and the reference data and in Section 3, where we explain the methodology, we will add an explanation for why we chose this specific metric. By providing this explanation early on, readers will gain a clear understanding of our approach from the outset.”

2. Referee comment and authors answer

Referee:
“The four indices given lack a systematic approach and proper discussion. The authors distinguish between “heavy” and “extreme” precipitation. However, there seems to be no systematic approach to that, for instance covering the range of return periods. Furthermore, for “heavy” precipitation, they use a) linear trends and b) the difference between two time slices. The manuscript does not explore thoroughly how the differences in methodology affect the results. Is the trend always linear, for all ensemble members? Does the trend depend on the climate sensitivity of the GCM/RCM chain? Is there an effect of long-term climate variability? A regional shift in the sign of the change over time? Going into some of these topics would make the paper more valuable.“

Authors:
”Heavy events are generated directly by the simulations in any case, occur on average 3 times per year or half year and can be handled by a direct statistical analysis (counting of occurrences per year). Heavy events represent return periods of 4 respectively 2 months. Extreme events are much rarer and their range is not sufficiently covered by the simulation for a direct statistical analysis over 30-year periods but is statistically estimated by extreme value theory. The idea behind this distinction is to investigate if all types of intensities show a comparable development. The effect of the different methodologies is explained and discussed in chapter 4.3 lines 316-328 and in chapter 5 lines 434-441. In the revised version we will emphasize more how the differences in methodology affect the results.
We did not test the linearity of the trend but we applied with Sen’s (1968) test a not strictly linear but monotonic trend analysis. We derived a mean temporal gradient from this analysis

to enable a comparison with the gradient calculated by the difference method over a 120-year interval.
The dependence of the trends on the climate sensitivity is an interesting aspect. We have analyzed if specific GCM/RCM combinations lead to a systematically stronger increase in heavy precipitation events than others, and we will add a corresponding paragraph in the revised version.
To analyze the effect of long-term climate variability and temporal shifts for the whole ensemble is a complex task, as this variability is given by the GCM's individual variability and may differ in phase and period between the models. We have additionally analyzed with method 2 (difference method) for selected GCM/RCM combinations how stable the calculated tendencies are with respect to an increasing time interval. The results show a widely monotonous increase on decadal scales of threshold exceedance amounts with the strongest increase towards the end of the 21st century for GCM/RCM combinations with a high and low climate sensitivity."

3. Referee comment and authors answer

Referee:
"Lines 116ff, Table 1: The EURO-CORDEX ensemble used here is more an ensemble of opportunity than a balanced one (cf. Sobolowski et al., 2021, DOI: 10.5281/zenodo.7673400). There are for instance about twice as much rcp8.5 than rcp4.5 simulation included. In addition, some GCMs and RCMs are more often represented than others. Who does this affect the robustness of the results? Does the lack of consistency in some regions arise from more the GCM or the RCM spread?"

Authors:
"The statement of Sobolwski et al. is correct. The EURO-CORDEX ensemble does not represent a balanced but a very consistent ensemble of regional climate simulations. But we checked the influence of different ensemble sizes on our results.
We employed a bootstrap test to determine the minimum number of members required for a grid-point to exhibit robustness under the RCP8.5 scenario. The robust area fraction becomes stable from 10-15 ensemble members depending on region. We also repeated our analysis for the RCP8.5 scenario with a reduced ensemble, where we considered only the same GCM-RCM combinations as used for the RCP4.5 scenario. The relative deviation between the climate change from the reduced and full ensemble is less than 5% for most parts of the model domain apart IP and MD, where the robustness is weak at all.
Our finding aligns with previous studies that have also observed stronger and more robust characteristics associated with the RCP8.5 scenario compared to RCP4.5.
Furthermore, the study by Dosio and Fischer (2018) acknowledge the potential impact of ensemble size on robustness. However, they note that the results for different warming levels, such as 2°C warming computed using only the RCP.8.5 runs, are similar to those obtained using the entire ensemble. This suggests that the influence of ensemble size on the results can be neglected above a certain number of members (about 10-15). Nonetheless, we will explicitly address the issue of different ensemble sizes in our revised version to ensure more transparency in our analysis."

Citation: Dosio, A., & Fischer, E. M. (2018). ["The robustness as defined in our methodology is dependent on the models' ensemble size; the results for the 3°C warming are computed with a smaller ensemble compared to those for 1.5°C and 2°C warmings, which may affect the results. However, from a sensitivity analysis, it turns out that the results, for example., for 2°C warming computed using only the RCP8.5 runs are similar to those using the whole ensemble.”]

4. Referee comment and authors answer

Referee:
“Lines 127-128: For the Mediterranean extreme precipitation season the summer to winter separation might be disadvantageous, since it often starts in September (e.g. Grazzini et al., 2019). Could shifts in the heavy precipitation period between the ensemble members affect the lack of “robustness” for that area?“

Authors:
“It is possible that the robustness of changes in the Mediterranean could increase in the winter-half year if we shift this period by one month earlier to September. This is supported by the annual analysis, which is dominated by the most extreme events from September to December and which shows a slightly stronger robust area fraction than in winter half-year but only for the MD (not IP) region and the 10- and 100 year extremes (Method 3) (see Figure 12). However, this would imply a region- and model-specific definition of half-year periods.”

5. Referee comment and authors answer

Referee:
“Chapter 2/4.1/5.: The problems with the quality of the reference data are discussed in chapter 5. You might consider if you use some of the statements from chapter 5 for your analysis, which might exclude some areas from a comparison at least with E-Obs. Furthermore, how adequate is ERA5 as a reference? Since - as stated in chapter 5 – ERA5 precipitation is a forecast product and does not include assimilated observed precipitation directly. It would be good to explain your choice of references already in chapter 2. Furthermore, in chapter 4.2 the manuscript shows, that the reference data are outside the range of the observations over large areas. These areas partly differ for both reference data sets. E.g. or E-OBS it seems, that for Germany, where the dataset is based on many stations, the result is quite different compared to Eastern Europe, where it is based fewer station data. In addition, the comparison with KOSTRA-DWD for Germany is not that bad. It seems necessary to put these findings into a perspective. Is the ensemble not suitable or are the references not suitable for the specific requirements? Or do you deem it as not crucial?“

Authors:
”The analysis is most reliable for Central Europe. The additional comparison with ERA5 should demonstrate the substantial uncertainty of the reference data in large parts of the domain where observational reference data are based on fewer stations. We think that the references are not suitable for an analysis of extreme amounts over large parts of the domain, particularly if daily amounts from sparse stations are smoothed by flat interpolation algorithms over large areas. The KOSTRA-DWD dataset should have the highest reliability as it is particularly created for extreme events. A fair comparison would require a Europe-wide data set of the same quality, but this is currently not available. We express this more clearly in the conclusion of the revised version.“

6. Referee comment and authors answer

Referee:

“Chapter 3.3, Line 199ff: The text is a bit confusing. Consider reformulation. Since the 100-year return values are far outside the range of the 30-year input data, confidence intervals should be given.“

Authors:

”The basic idea of extreme value statistics is to derive long-terme return values from shorter time periods with generally much smaller extreme events. Because these estimates have a high uncertainty, we have introduced a double check procedure to guarantee that the calculated GPD is actually represented by the simulated precipitation events of the corresponding time period and if the simulated extreme events of the future period do not fit the GPD of the historical period (and vice versa). We rephrase the paragraph as suggested. The potential confidence interval can be estimated from the range of return values in the full ensemble. We modifed Figures 4 and 7 so that the bars represent the area median of the confidence interval for the median”

7. Referee comment and authors answer

Referee:

“The authors give criteria for the non-applicability of the methods or the exclusion of certain grid points. How does this affect the “robustness” of the results in such areas?“

Authors:

”If methods are not applicable, for example because precipitation amounts are generally too low (see white areas in Fig 3, 5 and 6 over North Africa), or the calculated GPD does not pass the goodness of fit test (contributes to grey areas in Fig. 9 and 10), a grid point can be completely excluded from the analysis (this is the case if the non-applicability applies for more than 33% of the simulations) and the area ratio of robust changes for a particular sub-region is reduced, which means that we can identify robust changes only for a reduced or minor part of the particular region.”

8. Referee comment and authors answer

Referee:

“Chapter 4 structure: Chapter 4.1 and 4.2 are mostly about evaluation, whereas chapter 4.3 is about the climate change signals. You could consider separating them.“

Authors:

”Chapter 4 represents three different aspects of the analysis of results, and we would like to keep this structure.”

9. Referee comment and authors answer

Referee:

Line 40: “the people died due the flooding caused by extreme precipitation.”

Authors:

“Corrected as suggested.”

10. Referee comment and authors answer

Referee:
Line 95: “Sørland et al. (2021) is not about CPM simulations”

Authors:
“Corrected as suggested.”

11. Referee comment and authors answer

Referee:
Line 151ff: “With a threshold of 3 events per 6 month, you consider a 2-monthly return period in method 1 and 2. Consider stating that to get an easier distinction between your terms “heavy” and “extreme” precipitation”

Authors:
“We clarify this in the revised version. See our answer to the 2nd referee comment above.”

12. Referee comment and authors answer

Referee:
Line 240: “..extreme events are lowest in SC…” Consider reformulation like e.g. “..least intense..”

Authors:
”Corrected as suggested.”

13. Referee comment and authors answer

Referee:
“Line 248: “…median s of extreme events are greater..” should be changed to “higher”

Authors:
”Corrected as suggested.”

14. Referee comment and authors answer

Referee:
”Line 316-318: Long somewhat confusing sentence. Consider reformulation.”

Authors:
”Corrected as suggested.”