We would like to thank the reviewer for their valuable comments, which we address point by point below. The reviewer's original comments are shown in black and our responses are shown in blue. Line numbers in our responses refer to the revised version of the manuscript submitted herewith.

General comments

1. According to the Authors the main aims of the manuscript are:

   (1) to provide a benchmark for future generations of GCMs with grid spacings on the order of 10 km, as well as for km-scale RCM simulations with respect to the UPSCALE project atmopsheric GCM (AGCM) simulations in terms of seasonal mean and extreme precipitation;

   (2) to provide a methodology (a combination of two previous applied methodologies) to evaluate extreme (daily) precipitation by fitting extreme value distributions and evaluating the model outputs over large (>50000 skm) river basins in Europe for future applications as input to impacts models.

   (3) to determine to what extent the resolution sensitivity in precipitation is due to the sensitivity to resolution of the simulated North Atlantic storm track and [...] to contrast that with the role of local forcing from the orography at different resolutions;

   With regard to aims 1 and 2, correctly, the Authors point outs that "is an obvious requirement for climate models to simulate precipitation in a realistic way if the models are to be applied, for example, in process and impact studies of the hydrological cycle [...] Due to the wide range of applications, the required realism concerns all aspects of the precipitation distribution in space and time including the probability distribution function of the precipitation time series and its extremes.".

   Unfortunately, extreme value analysis results seems to refer to averaged values over selected river basins losing the spatial distribution information that is one of the most relevant for impact studies (together with temperature analysis). The analysis of spatial distribution of extreme precipitation (and their timing) is a point to be investigated/clarified in the manuscript.

   The choice of river basins of an area of 50000 km^2 or more is adequate in view of the fact that the resolution of the GCMs evaluated here is between about 135 km and 25 km. For some kinds of impacts, these scales are relevant as demonstrated for example by flood events affecting much or all of a river basin of this size (e.g. [Grams et al., 2014]). We do, however, fully agree with the reviewer that for other applications smaller scales are important. For such smaller scales, higher-resolution RCM output may be preferable, but the quality of the RCM output may still largely depend on the performance of the driving GCM [Graham et al., 2007]. We have made this point clearer in the introduction (Page 2, Line 33 – Page 3, Line 3). As suggested by Reviewer 1, we have also added an additional

Table 1: Maximum, mean, and standard deviation of orography over European land (-14–50°E) in the control and sensitivity experiments (m).

	max	mean	std. deviation
N480	2844	366	453
$N480_{N96}$	1977	337	370

paragraph summarising results obtained with the EURO-CORDEX RCM ensemble (Page 2, Lines 14–22).

2. A second point to be addressed is the sensitivity of the analysis results to the E-OBS dataset horizontal resolution (not reported in the manuscript), is it possible that the more the AGCM horizontal resolution is close to E-OBS horizontal resolution the better the results are, just because the data spatial resolution is more similar and values are not averaged in space ?

   This is an interesting comment and the reason why we have conducted the scale-dependent evaluation (Fig. 3 of the main manuscript) showing that this is not the case, our results are robust throughout a wide range of spatial scale. As far as extreme precipitation is concerned, this is one of the reasons to aggregate all data sets over river basins larger than 50000 km^2, i.e. above the grid scale of any of the models and at a scale where E-OBS is representative (see also previous comment).

   With regard to aim 3, the Authors test two alternative hypothesis the sensitivity to the large-scale circulation, specifically the North Atlantic stormtrack and to orography finding that orography effect is dominant with respect to North Atlantic stormtrack in improving precipitation description using a slightly different AGCM.

Specific comments

3. Page 10 Line 3-4 Authors write For most cases, the statistical model fits the observed maxima well, but there are a few discrepancies for larger return periods of more than about 20 years. This reasonable considering the sample size of 26 years for each ensemble member.

   We agree, we cannot sample internal variability on longer timescales than the individual simulations, even with an ensemble.

4. Page 17-Section 5.2 It will be of interest to report how orography varies within the sensitivity experiment, i.e. change in maximum, mean and standard deviation.

   These values are shown in Table 1.

5. Table 1. According to values reported in Tab.1, N216 simulation is the "worst" one, but it is also the same with only 3 ensemble members, of

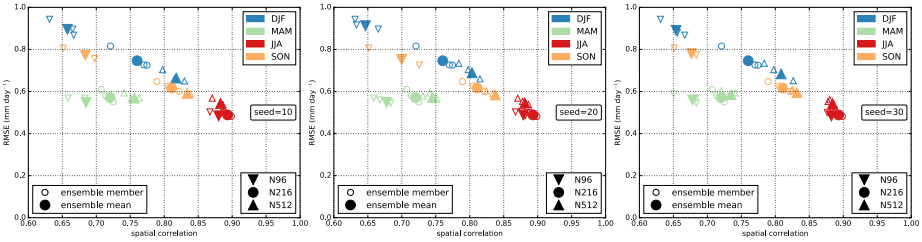

Figure 1: Alternative versions of Figure 2 in the main manuscript using only three randomly sampled ensemble members for the N96 and N512 models. Three versions of the figure with a different random number generator seed are shown.

which one is quite different from the other two in winter (Figure 2). Which will be the statistics of N96 and N512 if only 3 member are considered, or, how much does the ensemble size (for a given resolution) affect the statistics of the results? In Figure 2, it is quite evident that N216 winter values are more scattered than N96 and N512 values for the same season.

As far as Table 1 of the main manuscript is concerned, the 95% confidence intervals of European mean precipitation tend to be larger for N216 than for N96 and N512. This is indeed due to the fewer ensemble members (3) in N216. The mean values are however very similar for all three resolutions and the confidence intervals strongly overlap, so that there is no clear "worst" model.

The fact that the ensemble size differs (5 members for N96 and N512, and 3 members for N216) does not impact the conclusions drawn from Fig. 2 as can be seen from the results for individual ensemble members included in that figure. To illustrate this further, we show versions of Fig. 2 of the main manuscript but using only three randomly sampled members for the N96 and N512 models (Fig. 1). These figures are very similar to one another and to Fig. 2 in the main manuscript showing that our conclusions are robust.

Technical comments

6. Page 1 Line 5 the model resolution is indicated as 135km but in other part of the manuscript is 130km, please fix it across the manuscript

   Thank you for spotting this, we now use 135 km throughout.

7. Page 2 Line 15 introduce here the meaning of UPSCALE acronym instead of Page 3 Line 6

   This has been changed, thank you.

References

[Graham et al., 2007] Graham, L. P., Hagemann, S., Jaun, S., and Beniston, M. (2007). On interpreting hydrological change from regional climate models. *Clim. Change*, 81(SUPPL. 1):97–122.

[Grams et al., 2014] Grams, C. M., Binder, H., Pfahl, S., Piaget, N., and Wernli, H. (2014). Atmospheric processes triggering the central European floods in June 2013. *Nat. Hazards Earth Syst. Sci.*, 14(7):1691–1702.