Author’s response to the anonymous referee comments on nhess-2023-91

Dear anonymous reviewer,

we sincerely thank you for the time and effort you took to review our manuscript and for providing constructive feedback and comments that will help improve the quality. The manuscript has been revised according to your suggestions. Please find below our response to each of your comments.

NOTE: Reviewer’s comments are in black, our responses to the comments are given in blue below.

Response to RC3 on nhess-2023-91

After reading the manuscript, I found the experimental setup and results convincing but, in my opinion, the discussion section is hard to follow and needs to be more concise. In the revised manuscript, a lot on new text has been added into this section probably based on reviewers' suggestions (I have not participated on the first round of revisions). In my opinion, the authors focus too much on explaining individual models' biases in heat waves' simulation and these explanations are often not very conclusive and sometimes even contradicting each other. Unfortunately, it is difficult to provide specific comments how to improve this section but the length of discussion (nearly 3,500 words not counting conclusions) is in my opinion not justifiable. Maybe focusing more on WRF@5km (main message of the manuscript) could provide a more concise discussion.

We can see the point that the discussion appears to be too comprehensive. We have shortened this section and moved part of it to the results section (see below). As a result, the section is now less than 2,500 words long.

Other comments:
Line 22: 'Maximum temperature was reproduced reasonably well by all models' – then why calculating the 90th percentiles individually for each data set to circumvent the biases (Line 144)? It also contradicts the sentence in Line 236. I suggest revisiting this sentence.

We agree and can see the reason for confusion. This is why we adjusted the sentence in Line 22 to: "Maximum temperature was only reproduced satisfactorily by some models".

Line 49: Lhotka et al. (2017) → Lhotka et al. (2018a)

Correct, we have adjusted this.

Line 215: 'RACMO is the best performing EURO-CORDEX RCM, ALADIN the worst.' please add in what metric for better clarity.

We have adjusted the sentence to "RACMO is the best performing EURO-CORDEX RCM **in all of the three categories (correlation, CRMSE and standard deviation match)**, ALADIN the worst" for more clarity.

Table 1: Authors may consider adding Tmax bias values for 90th, 95th, and 99th percentiles, which may help interpreting the inter-model differences in heat waves' simulations.

We do not consider this as very useful, since these percentile bias values would refer to the overall time series, while in this context it is about the percentile values for the respective days based on the whole time period.

Line 528: 'an overestimation of which metrics? I suppose it should be 'persistence' (?) Please clarify.

Correct, we have added this for more clarification.

Lines 530–552: In my opinion, this newly added information belong to the results section.

We agree on this. Since this information refers to the bias in general, we have added an extra section in the results for that. This also shortens the discussion.

Line 615: 'overall best performances' should be defined, otherwise it is a subjective metric.

We agree. We have changed the sentence to “The three RCMs with the overall best performances **especially regarding the reproduction of heat wave characteristics** are ALADIN, REMO and WRF@15 km”.

Line 680: Authors should mention both 5 an 15km WRF runs.

We agree. We have adjusted the sentence to: “For this purpose, we employ an ensemble of six ERA-Interim-driven EURO-CORDEX RCMs of 12.5 km horizontal grid resolution **as well as outputs of a target area tailored, ERA-Interim-driven WRF simulation at 5 and 15 km resolution”**.