Review of revised manuscript "Evaluation of convection-permitting extreme precipitation simulations for the south of France" by Luu et al. (2021)

The revised manuscript represents a faithful response to most of the comments. The issue of comparing model and observational data at different spatial scales remains unresolved.

To recap, my previous criticism was that the authors were directly comparing results from the 12 km and 3 km resolution models with station data and 1 km resolution observations, before judging which model performs better. I argued that this is not a fair way to judge added value as the 12 km model is designed to represent grid box means at the 12 km scale, not values at the point (station data) or 1 km (gridded observations) scale. Closer agreement of the 3 km model with the aforementioned observations does not, therefore, necessarily mean that the 3 km model is "better" than the 12 km model, but rather likely simply reflects that the observations' resolution are closer to that of the 3 km model. My argument was that to identify if the 3 km model "adds value" to the 12 km model, one must first upscale the model and observational data to the resolution of the coarsest data (in this case, the 12 km model).

Response: We thank the reviewer for his/her enthusiasm in reviewing our revised manuscript and providing further discussions on the "change of support" issue.

Main comments.

The authors disagreed with my above criticism. Their arguments are summarized below in "Authors C1-4". My responses follow in "Reviewer C1-4".

*Authors C1. Model users regularly use coarse-resolution data (e.g. 5 to 50 km) for local climate studies. The 3 km model's higher spatial variability and improved precipitation at small scales thus represent added value for these users. The authors only focus on to what extent the 3 km model improves the extreme precipitation at local scale.

**Reviewer C1. It is true that some users directly use low-resolution climate data for point- or local-scale studies. This, however, does not mean that those users are correct to do so and is, anyway, of secondary importance to my criticism. If the stated aim of the research is

35 "evaluation" and to "investigate the added value" (see title/abstract), then the fact remains that it is not appropriate to do this at a spatial scale that the 12 km model is not intended to represent. It is trivial that the 3 km model exhibits higher spatial variability (simply because it has more grid cells); added detail is not added value.

40 The further the model resolution is away from the observation's resolution, then the less appropriate the comparison. Hence, if the 3 km and 12 km models were to be perfect at their own spatial scales, then the 3 km model must be in better agreement with the point- and kilometre-scale observations, compared to the 12 km model. This does not mean that the 3 km model "adds value"; it simply reflects the different scales the models are intended to represent.

45

In short, it is not possible to make conclusions on added value if the two models are being compared at different spatial scales.

Response: In order to broaden the discussion and include the reviewer's viewpoint, in this
50 revision, we also upscaled (with some modification in upscaling procedure compared to our previous revision) our convection permitting simulations to 12 km as in EUR-11 simulations and present both comparisons. The results allow an interesting comparison which is presented in the response beneath.

55 *Authors C2. Their goal is to "assess the overall improvement against observed station data", not to disentangle the causes of 3 km model improvement, i.e. resolution or physics. Comparing the 12 km and 3 km models at the same resolution (i.e. 12 km) would only answer whether (or why) the fine-scale resolution (3 km) can improve the larger scale (12 km).

60 **Reviewer C2. I accept that disentangling the contributions of different resolution and physics to any added value is not the aim of the study, so no problems there. I also agree that comparing the 3 km and 12 km models at the same (12 km) resolution will "only" answer whether the fine-scale resolution adds value at the 12 km scale. For the reasons outlined above, this (12 km) is however the minimum scale at which you can assess the added value.

65

Response: We now include both comparisons so the discussion can be broadened.

*Authors C3. There is no "standard" way of evaluating model added value: the appropriate method depends on the scientific question.

70

**Reviewer C3. I agree, but the scientific question also has to be appropriate. Asking whether 3 km simulations add value over 12 km simulations for representing point- or kilometre-scale observations (without upscaling) is, in my view, not an appropriate scientific question.

Response: We understand the viewpoint, and include the proposed comparison with upscaled results, along with a direct comparison.

*Authors C4. The authors provide an additional analysis in their response where they upscale the 3 km data to the 12 km grid (observations are not upscaled) and re-compare the seasonal maxima (3 h, 1 day) against observations (1 km and stations). Based on this comparison, the 3 km model is deemed to still outperform the 12 km data.

**Reviewer C4. I would like to know how the authors performed the upscaling for the results shown in Figure 4 of the response (this was unfortunately not mentioned). I ask this because the boxmean values for CPS (Figures 2/3, manuscript) and CPS-11 (Figure 4, response) are identical within a rounding error of 0.1 mm, which seems highly implausible.

The correct way to do the upscaling would be to upscale all the 3 hourly (daily) data to the EUR-11 grid, and then *after* that compute the Rx3hour (Rx1day) values. In Figure 4 it looks like the upscaling has simply been performed on the final Rx3hour (Rx1day) results of the original CPS grid. What else could explain the identical boxmean values between CPS and CPS-11?

Response: In our previous revision, we indeed upscaled the final results of Rx1day and Rx3hour using conservative remap method provided in CDO. So, the identical box mean values between CPS and CPS-11 is understandable because the conservative remap is expected to retain the flux of a variable over the domain. This could be the explanation for the approximation of box means between the CPS and CPS-11 (i.e., the CPS upscaled to 0.11 degree). However, as proposed by the reviewer, we applied this upscaling procedure to every single field of daily rainfall and daily maximum 3-hourly rainfall from the CPSs, then we calculated Rx3hour and Rx1day again. We also upscaled the results from 11 years of COMEPHORE radar from 1 km to 12 km for Rx3hour as a reference to allow comparison. For Rx1day, we used SAFRAN (1961-1990) at its original resolution (8 km), which is, in our perspective, comparable to simulations at 12 km. The results are shown in Fig. 1 (for Rx3hour) and Fig. 2 (for Rx1day) below. Generally, the results from CPS-11 of this revision are 2% to 5% lower than the CPS-11 from our previous revision. For Rx3hour, the mean of the Cévennes box from CPS-11 ranges from 1% to 15% lower than the result of upscaling COMEPHORE (Fig. 1). Meanwhile, the EUR-11 simulations (Figure 3a-c in

our previous revised manuscript) underestimate by 50% the mean of the Cévennes box in comparison with the upscaling COMEPHORE. For Rx1day (Fig. 2), the results from CPS-11 show biases of mean of Cévennes box ranging from -20% to 14% compared to SAFRAN (8 km). While the EUR-11 simulations underestimate the mean of the Cévennes box by 20% to 40%. In summary, we find here that the upscaling procedure (to 12km) barely alters the results (5% max), while the differences between simulations are between 15% and 50%. Therefore the convection permitting model improves extreme precipitation simulation over the south of France. And as we stated above, we included both CPS and CPS-11 simulations and updated the text in our revised manuscript.

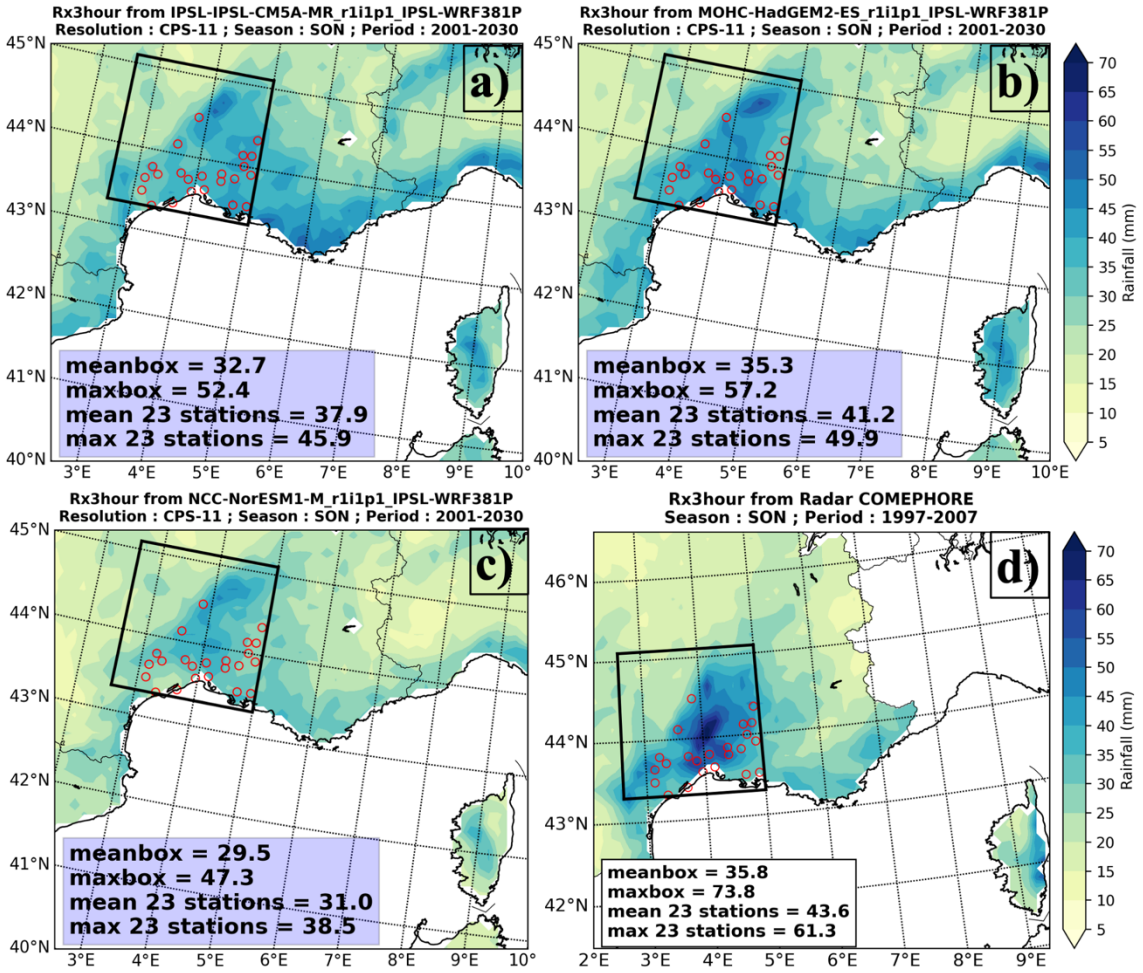
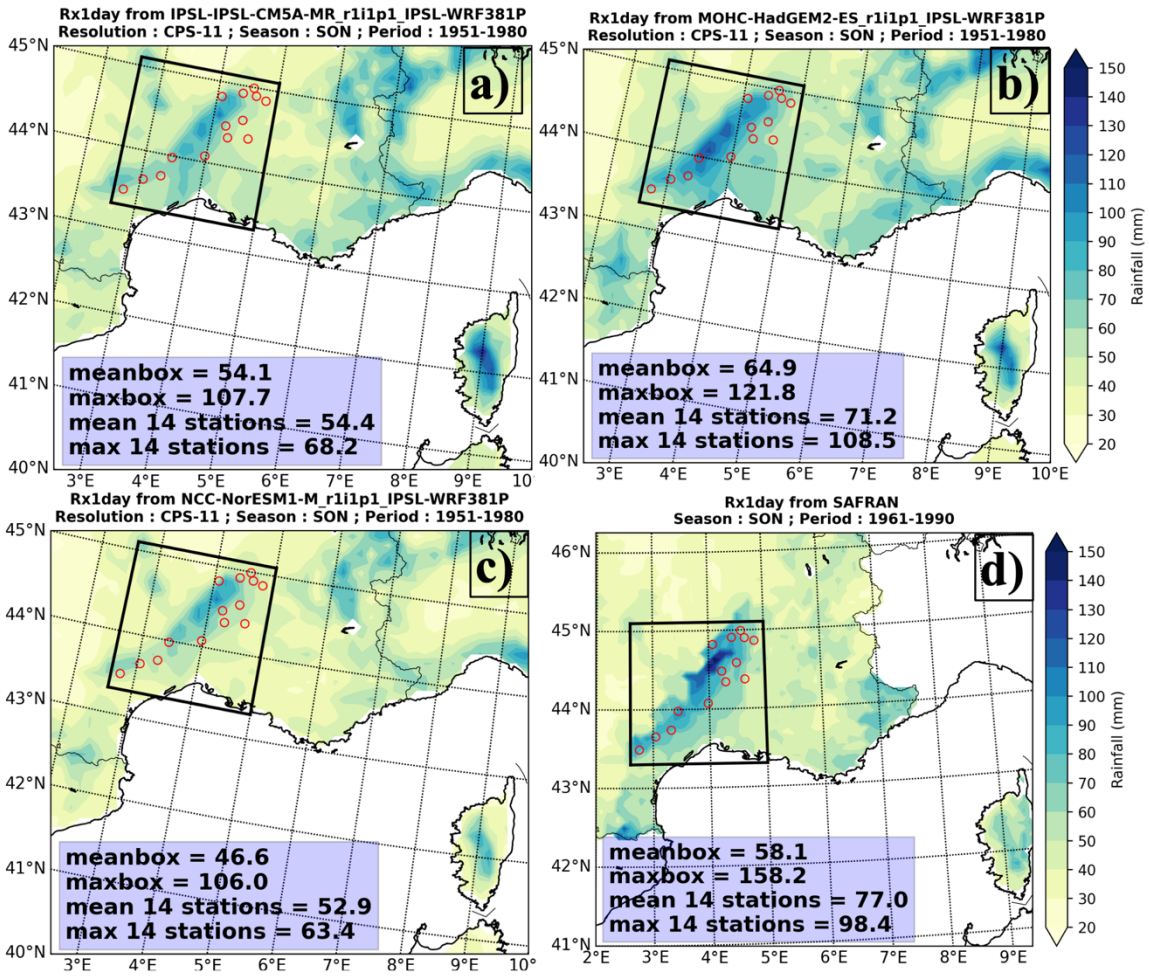


Fig. 1: Rx3hour (2001-2030, panels from a to c) from CPSs and COMEPHORE radar (1997-2007, panel d) upscaled to 12 km resolution.



120 Fig. 2: Rx1day (1951-1980, panels from a to c) from CPSs upscaled to 12 km resolution and SAFRAN (1961-1990, panel d) at its original resolution (8 km).

As a final point, I'd like to add that in my last review I listed a large number of CPM evaluation studies which all, before assessing the added value, upscale their observations and higher-
125 resolution simulations to the scale of the lowest-resolution model. I would be interested if the authors can point to any published CPM evaluation papers where what they are proposing has been done, i.e. no precipitation upscaling prior to evaluation of added value.

Response: We appreciate that the reviewer pointed out the importance of upscaling steps in
130 added value assessment. Indeed, all recent studies upscaled their finer simulations and gridded observations to the coarser resolution, or vice versa.

Minor comments.

- Section 2.2 / scaling. In your description of how the binning works you should also add that you set a minimum of 300 observations per bin in order to avoid undersampling, as stated in your response. This is important for the reproducibility and interpretation of your results.

Response: We added some clarifications to our revised manuscript around line 141: "... calculate the 99th percentile for rainfall and the mean temperature for each bin. *We use a threshold of having at least 300 points of precipitation to take a bin into consideration. This is to avoid the under-sampling effect on the final scaling results*".

- L301-303: I suggest adding some of this information to the caption of Figure 7, so that the figure can be understood on its own.

Response: We changed the caption of Figure 7 into: "*Mean moisture transport of the 12 heaviest daily rainfall events from all simulations (2001-2030, from a to c for EUR-11 simulations, and from d to f for CPSs) and ERA5 (1989-2018, panel g). Note that the domain of CPS is far smaller to meet the requirement of these analyses. Therefore, we embedded each CPS moisture transport inside its corresponding driving EUR-11 for the same 12 events of that CPS. This means that results from EUR-11 in these cases (panel d to f) may differ from those from panel a to c*".