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

The authors perform an evaluation study of a convection-permitting model (CPM) at 3 km resolution. The simulation domain covers the south-east of France and part of the Mediterranean Sea. The CPM downscales a 0.11° model, which was run over the EURO-CORDEX domain. A nice aspect of the study is that there are three realisations of the CPM simulations, each based on a different GCM; this aspect could potentially be given mode attention, as it is unusual in the literature. The authors then evaluate the performance of the CPM and 0.11° models and conclude that the CPM produces more realistic precipitation.

I think the study is a reasonable contribution to the literature and could in principal be published. However, at present the study has some limitations which must first be addressed before publication. These are detailed below in the main comments section.

Main Comments

1. Novelty and relation to similar literature. In their abstract, the authors state of their climate-length convection-permitting simulations that "… *this approach has never been used in a climate simulation for the Mediterranean coastal region*" (L4-5). There's a similar statement in the Introduction (L69-70): "… *such long simulations, to the best of our knowledge, have never been done for coastal area in the Mediterranean region*".

This is not correct. I can think of at least five studies which perform convection-permitting simulations at climate timescales over the north-western Mediterranean, which the authors don't cite. These studies all cover the area of the CPM domain used by the authors, as opposed to the studies of e.g. Armon et al. (2020) and Zittis et al. (2017) cited by the authors, which are for other parts of the Mediterranean and aren't on climate timescales. The studies I have in mind are (there may be more):

[1] Berthou et al.: https://doi.org/10.1007/s00382-018-4114-6
[2] Vergara-Temprado et al.: https://doi.org/10.1029/2020GL089506
[3] Meredith et al.: https://doi.org/10.1088/1748-9326/ab6787
[4] Adinolfi et al.: https://doi.org/10.3390/atmos12010054
[5] Caillaud et al.: https://doi.org/10.1007/s00382-020-05558-y

Ref. [1] has a specific section on heavy precipitation events in SE France. Refs. [4] and [5] also assess intense hourly and daily precipitation events in CPMs over France using similar observation sets to the present authors. Ref. [3] uses the same annual re-initialization technique as the authors and also focuses on the Autumn months in the NW Mediterranean, just as the present authors do.

Around lines 55-67 it would also be good to cite these climate-scale studies, as most of those presently cited are for case studies or selected events.

The authors need to cite and discuss the relevant literature, not just in the Introduction, but also where appropriate in the Results and Discussion. The results of the present authors should be presented in the context of the pre-existing relevant literature. That means, wherever appropriate, compare your results with those in the pre-existing literature. This is particularly important if your results are different, in which case possible explanations would be helpful.

2. Comparison of model data and observations at different spatial scales.

A major issue with the evaluation is that model data and observations on different spatial scales are being compared directly. While it's arguable that model data on a 3 km grid could be compared directly with station data, what do the authors hope to learn by comparing data on a 12 km grid (that means grid box averages over an area of 12 x 12 = 144 km^2) with station data (point values)? Or even with the 1 km COMEPHORE product?

It is not surprising that Figures 4, 5 and 6 show the lowest intensities in the 12 km model, followed by the 3 km model, followed by the point observations. This simply reflects the fact that the extremes are being averaged over ever greater areas as the grid spacing increases, thus the intensities are "smoothed out"; the same applies to the "mean 14/23 stations" in Figs. 2 and 3. Indeed this also applies to the box means in Figs. 2 and 3, because the area mean of high-resolution extremes must by higher than the area-mean of low-resolution extremes. These comparisons don't tell us whether or not the 12 km model is worse than the 3 km model, or vice versa. Suppose your 12 km model was perfect at the 12 km scale: the extreme intensities would still be much lower than those at the 3 km or point scale. Or imagine you aggregated your 1 km COMEPHORE data to the 12 km grid and then compared it against the 1 km data at some point: the 12 km data would have a strong negative bias, even though it's the same dataset. For further discussion of this topic, I suggest the study of Göber et al. (2008, https://doi.org/10.1002/met.78).

In the case of Fig. 4 (temperature scaling), what's important is that the models have similar scaling curves to observations, the intensities don't need to match to validate the models.

As pointed out in Göber et al. (2008), the standard/appropriate way to compare observations and model data is by upscaling the observations to the coarsest model grid (EUR-11 in your case). The CPS publications the authors cite all upscale their observations to the coarsest model grid: Kendon et al. (2012), Fosser et al. (2015), Knist et al. (2018), Chan et al. (2013, 2014). Also Refs. [1], [4] and [5] above.

In the cases of the gridded observations (SAFRAN, COMEPHORE), it is certainly possible to compare models and observations at the same spatial scale (i.e. that of EUR-11) through conservative remapping. In the case of the station data, there's no simple solution. As stated above, comparing the 3 km intensities with stations could be defensible. I don't see much value in comparing the 12 km intensities with stations; but if the authors really want to do this then they need to give a very strong warning to the reader that this has limitations, and these limitations should be communicated in the text.

Another indicator that the results might be being affected by the comparison of different spatial scales is the added value you find for daily precipitation. Studies show that CPMs generally don't add value for daily mean or extreme precipitation, e.g. Refs. [1] and [4] above, Chan et al. (2013), Ban et al. (2014). It's likely that a lot of the added value you find for daily precipitation statistics is simply due to the different spatial scales you're comparing against observations. Having said that, Berthou et al. (2018, Ref. [1]) did find added value at the daily scale for CPMs in the case of autumnal precipitation extremes in the Mediterranean.

<u>Other Comments</u>

1. Ideally this study would have been performed using reanalysis as boundary forcing. Since you are using free-running GCMs, you therefore need to inform the reader early on (i.e. in the methods) that the regional models will inherit biases from the GCMs, and that any biases you find therefore

result from a combination of both the GCM and RCM biases. Later on in your results, we see quite different results depending on what the GCM is, so the role of the GCM is clearly not trivial.

2. The CPM simulations cover the Autumn months because this is the time when the most intense events occur in SE France. Maybe not all readers will be aware of this or know why, as many expect the most intense short-duration events to be in the summer. I think a few sentences in the Introduction and/or Methods explaining why the strongest events are in Autumn would be useful. E.g. warmer Mediterranean SSTs, low pressure systems advecting warm moist air at lower levels from the Mediterranean into southern France and then orographic lifting, etc. Maybe the studies of Labeaupin et al. (2006, https://doi.org/10.1029/2005JD006541) and Toreti et al. (2010, https://doi.org/10.5194/nhess-10-1037-2010) would be of interest to you.

3. Temperature scaling of extreme precipitation (L118-126). What steps have you taken in order to avoid effects from under-sampling? Do you require some minimum value of data points to be in a bin before you compute the percentile? If so, what? Boessenkool et al. (2017, https://doi.org/10.5194/nhess-17-1623-2017) show that the downturn at higher temperatures can simply be a statistical artefact if the bins are not sufficiently populated. In your Figure 4, the deviations away from CC or 2xCC scaling occur at low and high temperatures, exactly the range where there are less events. This could be due to insufficient data points in the bins.

Also, in Figure 4, do the numbers in the inset table represent the mean scaling rates? If so, how do you compute them? Over the entire range of data? Or is it an average across all stations?

4. There are lots of different data sets used: Gridded data, 14 stations, 23 stations, etc. When the biases are presented in the text (Section 3.2), it is sometimes not clear with respect to which data the bias is for. It might help the reader if you state this more explicitly in the text.

Minor Comments

L15-16: "*because of the limitation in computer resources, deep convection processes have rarely been solved explicitly in long climate simulations*". This is again a bit of an exaggeration with respect to the existing literature. There are really quite a lot of CPM studies on climate timescales. For example, there are the studies which you already cite: Ban et al. (2014, 2015), Fosser et al. (2015), Hodnebrog et al. (2019), Kendon et al. (2014), Knist et al. (2018), Vanden Broucke et al. (2019). Then there are the five I've listed under Main Comment 1. There are a lot more if you take a look on Google Scholar, and not just for Europe like those already listed.

L28-44: Please remember to also cite literature relevant to your study region.

L41: "added value" is always singular, i.e. not "added values". Also in other parts of the manuscript.

L82: Could you please also give the resolution of the CPM in degrees?

L103: Could you please state what the model top is? With only 32 levels, the spacing between layers could be quite high. You should avoid having a vertical spacing which is greater than your horizontal spacing, which may be a risk here for your CPM simulations. It's too late to change this now, but it's useful to keep in mind for the future.

L105-109: Is the shallow convection parametrized in the CPM? If so, what scheme?

L124: Maybe you mean "same" instead of "similar"? "Similar" doesn't mean "identical", but "same" does.

L136: Unit of g is m s^{-2}.

L148-163: The authors could consider making these lines into a separate Section 2.3 for the data sets? If they don't want to, that's also OK.

L178-180: Do these biases refer to the bias over the whole box against SAFRAN? If so, the numbers don't agree with my calculations based on the insets in the panels of Fig. 2. please check.

L203-205: Are these 23 stations for the time period in 3 (h) or 3(j)?

L220: Instead of "we model the Clausius-Clapeyron relation …", it would be more correct to say "we investigate if the temperature-precipitation scaling follows the Clausius-Clapeyron relation in observations and models", or similar.

L230: The EUR-11 model can't be expected to have similar intensities as the point-scale observations, simply because you're comparing at different scales here (see main comment 2). What's important is whether the EUR-11 and CPM have the same scaling rate. Same goes for L243.

L235-240: Maybe your super-CC scaling results from the combination of strong moisture convergence in autumn precipitation extremes in SE France (due to onshore moisture advection) and deep convection. These ingredients aren't present simultaneously at other times of the year.

-L249 (Section 3.4): My understanding is that the analysis in this section is based on wet-events, i.e. days without precipitation are excluded. If this is the case, it would be useful for the reader to know what fraction of days contain wet events and if this differs much between the different simulations.

-L252: Change "either … or" to "both … and".

-Figure 3: There's no panel (i) after (h), so I think you need to change (j) to (i).

-Figure 3: What does the yellow colour over Italy represent? If this is simply an area of no data, then it would be good to mask it in white like in Figure 2 (g).