Anonymous review for the manuscript "Scale-dependency of extreme precipitation processes in regional climate simulations of the greater Alpine region" by Caldas-Alvarez et al. doi.org/10.5194/wcd-2022-11, 4.5.22

The presented manuscript assesses differences in the statistics of daily ("heavy") precipitation between a 25km RCM simulation and two kilometer-resolution CPM simulations. This analysis has been done before, the results are quite robust across model codes and well described in the literature. The presented analysis is novel nevertheless, but resorts to rather sophisticated methods (PCA, composites) and indices (PSI, FSS, wet-day percentiles), making them rather complex and difficult to interpret. Unfortunately, some of their implementation details are not fully fit for purpose (see major issues below). Also, I am not convinced that the presented results allow supporting the interpretations and conclusions made at the end of the manuscript. In particular, I criticize some of the inferences made from the detected differences (see major issues below), albeit alternative hypothesis are discussed in the literature. Finally, the detected differences are rather small, while the necessary statistical quantification is not provided (E.g., L490 and L303). In my role as a reviewer, I usually abstain from requiring stat. tests, but when an established and robust hypothesis (no differences in daily statistics) is refuted in a manuscript, requiring a robustness assessment seems warranted.

I have the impression that the manuscript title does not reflect the content of the manuscript precisely enough, as neither "scale-dependecy" nor "extreme precipitation processes" are actually assessed in the study. In particular, I don't think that the chosen indices and events qualify as "extreme precipitation", and "extreme precipitation processes" are not considered at all. Meanwhile, the term "scale-dependecy" is usually used differently from the presented use case.

Overall, the simulation configurations are not described in sufficient detail and some important aspects are missing. E.g., for the ALPS-3 simulation, the authors mostly refer to Coppola et al., (2018), which is a MIP overview paper, and thus does provide the necessary detail to ensure scientific reproduceability. For instance, the info that KLIWA-2.8 and ALPS-3 were conducted using two different major releases of the code was not highlighted. I consider this detail is relevant context when suggesting that ALPS-3 is merely a continuation of KLIWA-2.8.

Please find below a list detailing the major and minor issues.

Major Issues

L153: "In spite of these small inconsistencies, we combine both CPM simulations to attain a sufficiently large investigation period for comparison with the RCM simulation and observational datasets."

I don't think that the term "small inconsistencies" is justified here. The KLIWA-2.8 and ALP-3 configurations differ in virtually every aspect, apart from their overlapping computational domains, and their use of the COSMO code (in two different major releases!).

Also, what defines "sufficiently large"? KLIWA-2.8 is 29 years long. I'd consider that sufficient for all of the presented analysis and the qualitative conclusions of the study. Note that KLIWA-2.8 is actually only used in Section 4 (Evaluation of extreme precipitation).

L170: Does the 80th all-days-percentile really characterize "high grid-point intensity"? I think the 80th percentile represents a value of a few mm/day, which is rather typical for a rainy day in Germany.

L193: I don't understand. In Fig. 2 the 11th Jun event is below the 90th percentile of both respective indices, no? This is exactly opposite to the argument made in the text.

L198: I don't think the analysis supports this claim, rather the opposite. As indicated in Fig. 2 the Spearman’s rank correlation between PSI and fldsum is 0.98/0.96. Also it is evident that the solid line mostly tracks the dotted line. That is, at lower amplitude, which does not matter for a ranked index. This means that after applying the second threshold (i.e., L236, L301) almost the exact same events are chosen as would have been when using the fldsum. In other words, Fig. 2 actually demonstrates that PSI is an unnecessarily complex choice for the presented use-case.

L240: Why are the EOFs computed using the RCM and not ERA-Interim directly? My point is that the first few EOFs of the 500 hPa geopotential need to be almost exactly the same in RCM and driving data. After all, if those EOFs of the RCM simulation would become systematically different than those in the driving data, the RCM approach would become somewhat questionable.

L285: The analysis presented in Figure 5 is rather difficult to interpret, since the boxplot parameters (median, quartiles, ...) depict percentiles of a conditional index (i.e., the wet-day precip., P > 1 mm/day). Please consult the following study for a thorough discussion of the problems related to deriving percentiles of conditional indices: Schär et al., (2016)

L327: Why are events chosen that entail bad quality/non-existent observations (since over ocean)? To verify a model, I would intuitively choose case studies that have abundant high-quality observations available. Also, I am not fully convinced about the usefulness of the MSWEP-11km product.

L361: The statement needs to be qualified, also w.r.t. internal variability and accuracy of specific humidity obs with radiosondes. In fact, I was surprised to see only a difference of 0.1 – 0.2 g/kg when comparing profiles of a limited-area climate simulation to more-or-less “instantaneous” soundings. I think, my view is corroborated well when considering Fig. S2, instead of the differences. In contrast to the authors, I think these results are actually rather promising.

L384ff: I am not completely sure if the chosen procedure is appropriate, but honestly, I do not fully understand why it has been chosen in the first place. First of all, why would EOF1 be associated with “heavy precipitation”? I thought that EOF1 portrays the mode with the largest variance, right? That is (by definition) rather unlikely a percentile at the tail of the precip. Distribution, no? Second, EOFs seems rather complicated provided that the results (Fig. 9) exert a very similar pattern as much simpler indices, e.g., the standard deviation (cdo timstd).

L391: Why for the day prior? I thought the authors suggest the detected heavy precipitation events to be primarily associated with synoptic systems.

L398/399: If the first three leading EOFs only explain 39% of the variance. Is that analysis really an appropriate tool? Maybe I do not understand precisely what the authors are trying to achieve.

L415ff: How come? Why not a parameter choice, or a time-step sensitivity in the cloud microphysics parameterization, or the nesting strategy …

L462: No. The statement is conditioned on precipitation being formed the same way in both simulations, given the different surface fields as input. However, the point of the study is that one simulation has the convection scheme active, while the other one does not.

Please find below an alternative (often discussed) hypothesis that would yield similar differences as those presented:

The soil in simulations typically dries out during the summer months. It is very probable that in autumn soil-water content in the RCM and CPM slightly differ, since soil conductivity, soil diffusivity, evaporation and infiltration rate were not calibrated to yield the exact same soil state. The differences in soil-water content will yield differences in the partitioning of the surface latent and sensible heatfluxes (as observed), and ultimately slight differences in surface specific humidity. The signal appears in the composites of all EOFs because it is already there in the mean.

L465: Where does conclusion (c) come from? Maybe a paragraph went missing?

L479: No. (i) scale-dependency is something different. (ii) The considered indices and case studies are not “extreme”. (iii) The manuscript is not assessing how “thermodynamical processes” influence precip. You are looking at composites. (iv) Also consider that “thermodynamical processes” is a rather uncommon term.

L481: I disagree. The analysis also allows other conclusions than portrayed (see above).

L487: It might be worthwhile to note that only daily statistics are considered. In the European summer, a large share of heavy precipitation relates to diurnal convection (see Ban, Kendon). These events only show up when considering hourly precipitation statistics.

L495/L498: I do not agree with the term “explained by” (see above).

L511: Again, there are many other possibilities, e.g., the infamous warm bias related to the reduced turbulent length scale in the CPM simualtions (Baldauf et al., 2011).

Minor Issues:

Introduction: The introduction discusses heavy precipitation and the CPM approach. However, context on the employed approach is not provided. E.g., why are “weather types and PCA” insightful and useful tools for the proposed questions?

L80ff: Indeed there is some connection between moisture, moisture flux, instability and precipitation. After all, these aspects have been under investigation for many decades. I think the authors need to be more specific when making their argument. Slso, what “moisture excess” do you refer to? The term suddenly appears out of thin air (maybe a part of a sentence got lost while editing?).

L100ff: Discussion of precipitation under-catch and spatial representativness, in particular for mountainous regions might be worthwhile.

L188: I am skeptic that this dataset qualifies as an “independent” source for validation. I didn’t know it before, but the description reads like it is mostly based on other models.

L136: “nominal resolution” Do you mean grid spacing?

L143: It might be worthwhile to note that KLIWA-2.8 is embedded into three nests with 0.44 °, 0.0625 °, and 0.025° grid spacing respectivley, as outlined in Hackenbruch et al., (2016).

L146: Could you confirm that ALP-3 was directly nested into ERA-Interim? I am skeptic because (i) the use of intermediate nests wasnt mentioned for the KLIWA-2.8 simulation either, and (ii) the CCLM-5-0-9-KIT contribution outlined in Coppola et al. (2018) specifically mentions an intermediate nest.

L155: Don't you disregard the lateral relaxation zone + additional margin for spinup. Btw: What upper boundary condition is used?

L173: For d=2, PSI considers the sum of 3 days, right?

L300ff: I am not familiar with the analysis presented in Figure 6. I guess its main purpose is a visualization of the spearman rank correlation, and the mentioned "clustering". Could you elaborate on the argument made using that figure? For now, the text mainly refers to the correlation numbers (0.94, 0.41, 0.48).

L304: Please define "hit rate".

L321: Define "percentage of affected area"

L331: Define "ECOs"

L348: Why not simply write specific humidity and temperature in the figure labels? Same for the remaining figures.

L351: How is "spread" calculated?

Figure 5: "The kernel density at each precipitation intensity is shown by the shaded areas." This shaded area is rather confusing, since the corresponding scale is missing. Also, shouldn't the area be the same size across all presented datasets (i.e., sum up to 1), or is it normalized to one dataset?

What does "maximum grid point precipitation" describe? Is that the highest value ever encountered a any grid point in the analysis domain? If yes, does this man that the data is pooled before computing the boxplot?

Table 3: Define "coverage".

Fig. 10: Why does the term "heavy precipitation" suddenly appear again? How is it related to EOF-1? Also, at what time of the day is CAPE computed?

L50: I do not understand what an "energetic low-level" should be.

L54: Why is Khodayar et al., (2021) cited here? I am not convinced it fits the context very well.

References:

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., & Reinhardt, T. (2011). Operational Convective-Scale Numerical Weather Prediction with the COSMO Model: Description and Sensitivities, Monthly Weather Review, 139(12), 3887-3905.

Schär, C., Ban, N., Fischer, E.M. *et al*. Percentile indices for assessing changes in heavy precipitation events. *Climatic Change* **137,** 201–216 (2016). https://doi.org/10.1007/s10584-016-1669-2