Review of "A novel method to identify sub-seasonal clustering episodes of extreme precipitation events and their contributions to large accumulation periods" by Jérôme Kopp1, Pauline Rivoire1, S. Mubashshir Ali1, Yannick Barton1, and Olivia Martius1

Main comment

The authors study the clustering of precipitation extremes and their relevance for accumulated precipitation extremes at the global scale. They use ERA5 data and aggregate precipitation over river catchments, which is the basis for an interesting study. They introduce metrics for investigating the above from a novel perspective.

I read the paper with high interest. I appreciate the effort done by the authors in providing graphics for explaining the procedure. However, unfortunately, I found the methodology very difficult to understand. In my view, the presentation of the methods, which is - together with the results - the fundamental aspect of the paper, requires a thorough revision. In fact, it is unclear to me from many points of views. In this context, I find it difficult to judge how well the metric captures the investigated physical processes and whether a more straightforward (easy to interpret) metric could have been designed.

After an improvement of the presentation, which should make everything clear to the reader (see specific comments below), I think that the following crucial aspects should be discussed thoroughly.

The authors propose a novel metric, hence high attention is required to the physical interpretation of (1) the defined metric (i.e., explain the reasoning beyond the choice of the metric based on simple physical arguments to the reader) and (2) the associated results. This is fundamental to allow the reader to well understand metric and results (and ultimately to maximise the impact of the work). On the same topic, as also states by the authors in the discussion, "a shortcoming of the method is the lack of a simple assessment of the significance of the clustering". In fact, this shortcoming, combined with a non-clear (according to me) presentation/explanation of the metric, makes it difficult to interpret physically the spatial distribution of the clustering and its relevance for accumulated precipitation. I fully understand that the results are novel and, for this reason, it can be sometimes difficult to compare with previous literature, however, the authors should try to explain whether the results are consistent with some physical understanding/expectation. (I do provide some possible ways to go in this direction below.) This would help to make the work more robust.

I hope that my comments can help the authors to improve the manuscript.

Specific comments

L25, I agree, but isn't the third point a consequence of the two above, so should not this presented in a non-parallel fashion?

- L31-40
  "In these studies, clustering in time was assessed using the index of dispersion (variance-to-mean ratio) of a one-dimensional homogeneous Poisson process model i.e., a Poisson process with a constant rate of occurrence (Cox and Isham, 1980)."
  "All studies discussed above used statistical models to identify significant serial clustering of extreme events. However, none of those methods are able to directly identify individual clustering episodes."
  "To our knowledge, no procedure exists that (1) automatically identifies individual serial clustering episodes of extreme (precipitation) events, and (2) subsequently uses the identified episodes to evaluate the clustering properties of a region."

Aren't Bevacqua et al. doing so (for precipitation from storms), i.e. introducing a counting-based procedure to identify individual clusters and avoid issues with the Poisson-process methods? Their approach does not rely on parametric distributions (related to your L275). If so, this should

be acknowledged and the text fixed accordingly where necessary. Similarly, are Dacre and Pinto presenting counting based procedures as well?
(The two references are those in the original manuscript.)

L56 “Precipitation in ERA5 is a prognostic variable.”
I understand the sentence, however, I suggest to expand the text by mentioning the implication and what does that mean for a non-specialist (in a few words).

L64. Can you explain better to the reader why you do this choice, i.e. using level 6? Thanks

L70, “We retained only catchments containing at least five ERA5 grid points for our analyses.” Does this mean that you consider only catchments with a catchment's area above about 5*25*25km? (I am assuming a resolution of 25km for the grid points.) If so, this means that you are considering relatively large catchments, where the clustering may be more important as they are responding slower to rainfall. If you agree (supported by a reference), this could be mentioned to reinforce your approach.

L86, "After applying the declustering approach, a series of independent extreme daily precipitation events was defined”. I understand that you end up with a time series of binary events (fig 3b). Specifying that would help the reader.

Depending on the local autocorrelation of the precipitation time series, after applying the high-frequency declustering, you will end up having a different number of extreme events at different locations. Does this affect your final results, which may differ at different locations simply because of that? Please clarify/discuss.

Could not Figure 3 and 4 be merged, i.e. keep only 4? The first two panels are *about* identical to Fig. 3. (They are not exactly identical as stated in the caption of Fig 4 given that there are no lines in panel 4b).

Figure 4, Can be adding 14 days after the last day in the panel help to read the panels? (Such to be able to well understand why n14 is 0 in the last days in panel c.)

L100, when you talk of extreme events in this section, I assume you refer to extreme events identified though the high frequency decluttering defined in the section above. Please make this clear/explicit.

L104, at the end, are windows centred or not? In Fig 4d, there is a centred window.

L105-106. You refer to Figure 4d. n14 is computed over the next 14 days, while acc14 is computed over a centred window. You explain why later, but it is confusing for the reader to find this in the Figure at this stage (as you refer to Figure 4d).

L107-118, In my view, the explanation of the procedure needs major improvement. The statements below can help the reader to understand points where the text needs improvements.

L107 Add a sentence at the beginning of the paragraph explaining that through your procedure you aim at reducing the number of clustering episodes up to a number N_ep, to avoid having overlapped clusters. The reader is then able to read the step with this in mind and things will be easier to understand.

L107 “highest count of extreme events”. What does “highest” mean? “Largest precipitation” The same with “largest”. It seems that there are two different thresholds involved in the selection, in addition to the constrain on N_ep and other thresholds. Please clarify.
Does changing these thresholds affect the results (in terms of matching between Cl_n and Cl_acc? (This is related to line 116)

L113, do you mean you sort by the number of counts in extreme events, and if that is equal among clusters you then sort by precipitation?

L115. To me, it is unclear how Cl_acc is obtained. You state: “This is done by applying steps (ii) to (iv) of our automated identification algorithm to the original precipitation time series.”
Hence, I would assume that you only apply steps ii to iv. Is this correct?
If so, this would imply that there is no association between Cl_n and Cl_acc, in the sense that Cl_n and Cl_acc can be associated with different dates as the two procedure are carried out independently (this seems in line with L164). In this context, I think that the sentence at line L122-124 is not necessarily obvious, and should be explained better to the reader.

L114, “The episodes picked out by the clustering episode identification and the extreme precipitation accumulation identification can be partly or completely identical. Examples of Cln and Clacc for the time series of Fig. 4 are shown in Table 1.”
- Is the example in the table one where they are identical or not? It seems they are in terms of dates (which I assume is not always the case - please clarify), but not in terms of rank. Please clarify.
- If selecting episodes associated with different dates is possible (as I understand), I strongly suggest creating an example where this also occurs. This would help to avoid any confusion in this regard.

L117. You refer to the table where Sr Sf S’ is discussed but it has not been presented to the reader yet. This can be confusing.

L120, this sentence is not precise. I guess you mean that the clustering is present if the variance of the number of extreme events across Cl_n is above a certain threshold.

L125 start a new paragraph before “We would like”. (“We would like” is too colloquial in my personal view.)

After clarified things about the weights (see below), consider whether having their description in an appendix would help the reader. This could allow focusing directly on the metrics S. You should provide at around L 125 a general explanation on the way you are going to build the metrics S and why you need weights there. This should be before going into the details of the weights, which is a more technical aspect.

L130, clarify the difference between “points” and “weights”.

L132. Aren’t the results therefore strongly sensitive to your choice of the weights? I mean, the condition “the difference between the ith place and the (i+1)th place should be larger than the difference between the (i+1)th place and the (i+2)th place”? This seems to be a very relevant point to discuss. For example, why isn’t the difference between adjacent points always the same?

L140, what is lambda?

About L150, You do not state explicitly whether qi is different in the two classifications.

L150-155. Explain better to the reader why: “it measures how often sub-seasonal clustering episodes happen and how many extreme events these episodes contain”. (I appreciate the link to the metric phi in the next section, and I can somehow see why this happen. However, the reasoning beyond the choice of the metric should be provided clearly to the reader).

Does Sf depend on the high-frequency decluttering procedure, which - depending on the serial correlation of the precipitation - can lead to a different number of extremes at different catchments? If so, is it possible then to compare different catchments via Sf? In figure 8 you implicitly do such a comparison via selecting locations based on a global unique threshold for Sf.

L160 Would the mean number of extreme events in the windows selected in Clacc divided by the total number of events provide information on the role of clustering for precipitation in a simpler fashion?
- Please present Sf, and explain it physically. Then S’f and explain what information it conveys from a physical point of view. Then present the ratio Sr.

- Especially, explain Sr in the context of the fact that Sf and Sf' may represent events associated with different dates (see comment above).

- A suggestion is to use subscripts or superscripts "acc" and "n" for S such to clarify instantaneously when this is related to Cl_n and Cl_acc. This could help the reader.

L205, Section 3.1. At the moment this section provides a description of the spatial pattern of the maps. Is it possible to provide some physical insights into the interpretation of the maps?

L205, Section 3.1, feel free to consider whether the following can be interesting questions/aspects to investigate or not. It is up to the authors.
- are results dependent on the catchment size?
- are results dependent on the (i) mean precipitation spatial variability or (ii) precipitation temporal variability?
- Focussing on some catchments (through showing precipitation time series) where you do find opposite behaviours based on the S metrics could help the reader to better visualise the differences and see what the metric captures. This would also allow for describing some physical aspects leading/not leading to clustering (precipitation relevance) in the direction of Figure 11.

L 220, (I see that you discuss this also in the final discussion). Can using an arbitrary percentile provide a good understanding of the spatial patterns?
For example, in the context of the metric phi, studies have looked at values significantly higher than zero, given that this implies clustering.
If based on theory it is not possible to define reference thresholds, is it possible based bootstrap procedures to define some thresholds for a "null case" to be used as a benchmark?