General comments

This paper presents a new count-based method to identify episodes of clustered extreme precipitation events and quantify their contribution to large precipitation accumulations. There are a number of potential benefits of this approach relative to existing approaches including i) the lack of a need to make assumptions about the underlying statistical distribution, ii) an ability to identify individual clustered episodes, iii) a framework that allows for quantifying contribution of clustering to total precipitation, iv) its global applicability, and v) ready extension to other extreme phenomena. The work is therefore scientifically significant and should be well read by the community.

The methods are valid, though I do have requests to elaborate further on the data and methods (see specific comments below).

The presentation quality is good. The manuscript is well written and the figures are clear. The abstract can be understood without reading the main paper. The work is well motivated with strong reference to prior studies concerning the clustering of climate extremes. I particularly appreciate the section comparing this new method to the more traditional dispersion metric. I also appreciate that the code is publicly available and easily accessible.

The subject matter is appropriate for HESS and is worth being published after my comments below have been addressed.

Specific comments

How adequate are the daily ERA5 precipitation data in capturing the extremes for this study? This is mentioned in passing in the main text, but I think the use of ERA5 needs further justification, citing of the literature, and a statement about whether the authors expect results to change using gridded observations or surface station data.

Response: We agree with the referee that this is a relevant point which deserves further justification. A study by Donat et al. (2014) assessed the consistency of precipitation extremes in various gridded observational and reanalysis datasets. They found a reasonable agreement between the observational precipitation datasets in extreme precipitation patterns and time series, while the reanalysis datasets showed lower agreement but generally still correlated significantly. However, this study predates the release of ERA-5 and used its predecessor ERA-Interim.

Yang and Villarini (2019) examined the capability of four reanalysis products (MERRA, MERRA2, ERA-Interim, JRA- 55) in capturing the temporal clustering of heavy precipitation in the observations. The results from all the four reanalysis products tend to agree in terms of spatial extent, even though there are small-scale differences.

More recently, Rivoire et al. (2021) compared moderate to extreme daily precipitation from ERA-5 against two observational gridded data sets, EOBS (stations-based) and CMORPH (satellite-based). Using the hit rate as a measure of co-occurrence, they found that for days exceeding the local 90th percentile, the mean hit rate is 65% between ERA-5 and EOBS (over Europe) and 60% between ERA-5 and CMORPH (globally). They also found that the differences between ERA-5 and CMORPH are largest over NW America, Central Asia, and land areas between 15°S and 15°N (the Tropics). They also compared the full precipitation distributions between ERA-5 and CMORPH and found that precipitation intensity agrees well over the midlatitudes and disagrees over the tropics.

Another recent study by Tuel and Martius (2021, preprint) on sub-seasonal clustering compared ERA5 with three satellite-based datasets (TRMM, CMORPH and GPCP), as well as output from 25 CMIP6 Global Climate Models (GCMs). They found a good agreement on the spatio-temporal clustering patterns across datasets. We are happy to share the preprint with the reviewer.

Based on those studies, we don't expect our results to significantly change using a different observational or reanalysis dataset. The use of ERA5 was motivated by its global coverage, its regular spatial and temporal resolution and its consistency with the large-scale circulation (see e.g., Rivoire et al., 2021). We also emphasize that our method can be applied to any kind of datasets, independently of their spatial configuration and temporal resolution.

New references:

Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V., Lippmann, T., & Zwiers, F. W. (2014). Consistency of Temperature and Precipitation Extremes across Various Global Gridded In Situ and Reanalysis Datasets, *Journal of Climate*, *27*(13), 5019-5035. Retrieved May 26, 2021, from https://journals.ametsoc.org/view/journals/clim/27/13/jcli-d-13-00405.1.xml

Rivoire, P., Martius, O., & Naveau, P. (2021). A comparison of moderate and extreme ERA-5 daily precipitation with two observational data sets. Earth and Space Science, 8, e2020EA001633. https://doi.org/10.1029/2020EA001633

Reference: Tuel A. and Martius O. (2021), A global perspective on the sub-seasonal clustering of precipitation extremes, submitted to Weather and Climate Extremes, in review. This manuscript is confidential but if reviewers want, we are happy to share it.

Change: L77 The choice of ERA5 was motivated by its global coverage, its regular spatial and temporal resolution and its consistency with the large-scale circulation (Rivoire et al., 2021). While our method can be applied to any kind of datasets, independently of their spatial configuration and temporal resolution, we don't expect our results to change significantly using other gridded datasets, surface station data or satellite observations. Indeed, previous studies have shown that precipitation extremes in gridded observational and reanalysis datasets correlated significantly (Donat et al, 2014), and that reanalysis products tended to agree in capturing the temporal clustering of heavy precipitation (Yang and Villarini, 2019). These studies used ERA-Interim, the predecessor of ERA-5. More recently, Rivoire et al. (2021) compared moderate to extreme daily precipitation from ERA-5 against two observational gridded data sets, EOBS (stations-based) and CMORPH (satellite-based). Using the hit rate as a measure of co-occurrence, they found that for days exceeding the local 90th percentile, the mean hit rate is 65% between ERA-5 and EOBS (over Europe) and 60% between ERA-5 and CMORPH (globally). They also found that the differences between ERA-5 and CMORPH are largest over NW America, Central Asia, and land areas between 15°S and 15°N (the Tropics). Another recent study by Tuel and Martius (2021, preprint) on sub-seasonal clustering compared ERA5 with three satellite-based datasets (TRMM, CMORPH and GPCP), as well as output from 25 CMIP6 Global Climate Models (GCMs). They found a good agreement on the spatio-temporal clustering patterns across datasets.

Line 74: Please explain what you mean by ‘timing’. Are you referring to the time of day, or time of the year? I think we need more explanation about why timing errors are so critical for this study to justify excluding the tropics.

When referring to the timing of precipitation, we refer to the days when extreme precipitation occurs (time of the year). As our method is based on counting how many extreme events happen in a certain time window, differences in the timing of the extreme events could result in different counts. We now discuss this point in more detail in section 2.1 (see previous answer, particularly the paragraph on the study by Rivoire et al. (2021)).

Changes: L74 The timing of extreme precipitation (**time of the year**) is important for the present study **because our method is based on counting how many extreme events happen in a certain time window (see section 2.3).** Rivoire et al. (2021) showed that **this** timing of extreme precipitation is well captured by ERA5 in the extratropics but less so in the tropics.

The runs declustering step needs more justification. I can understand its purpose for the case of slow-moving synoptic cyclones. But for the case of a multi-day sequence of afternoon severe convective storms, these are multiple events that are clustered rather than a single event.

As the first referee also raised a question regarding the runs declustering, we copied its question and our answer below in *italic* for completeness.

We agree with the referee that by applying a runs declustering, our methodology will not pick up this specific scenario of a multi-day sequence of afternoon severe convective storms at the same grid-point. The spatial (0.25° lat/lon) and temporal (daily) resolutions of ERA-5 is too coarse to properly target convective scale precipitation, and we would miss many convective extremes. The present research is more targeted at the larger scale structures, such as mid-latitudes cyclones and cut-off lows. The runs declustering removes the short-term temporal dependence of extremes so as to focus exclusively on clustering at longer timescales (weekly and above).

That being said, it would be interesting to apply our approach to shorter time scales by using input data with a higher temporal and spatial resolution.

Change: L87 The runs declustering successively removes the short-term temporal dependence of extremes so as to focus exclusively on clustering at longer timescales (weekly and above). In this framework, a multi-day sequence of afternoon severe convective storms at the same grid-point would be reduced to a single event, while being composed of multiple independent events. This is not an issue because the present research is more targeted at the larger scale structures, such as mid-latitudes cyclones and cut-off lows. More importantly, the spatial (0.25° lat/lon) and temporal (daily) resolutions of ERA-5 are too coarse to properly target convective scale precipitation, and many convective extremes would be missed. Input data with a higher temporal and spatial resolution should be used to apply our approach to shorter time scales.

1st referee question: 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.

Response: we agree with the referee that the declustering reduces the number of extreme events differently at different locations. In a catchment where extreme precipitation is on average more persistent, the number of independent events retained after the declustering is smaller than in a catchment where extreme events have a short duration. The goal is to identify independent extreme events (ideally these are related to independent triggering weather systems). Note however that these differences in number are not relevant for our analysis as we focus on the clustering of independent events. We further limit our analysis to the top 50 clustering episodes for each catchment so the same number of episodes is used for all catchments (we checked that all catchments had 50 episodes with at least one (declustered) extreme event).

The sensitivity analysis presented in Fig. 9b also reveals that a change in the run length parameter r from 2 to 1 days resulted in the smallest differences in Sr. Not applying the declustering is equivalent to setting r = 0 days, and consequently this should have a very limited impact on our results and wouldn't affect our conclusions.

We also emphasize that the precipitation accumulations are not affected by the declustering, only the event counts.

I appreciate the discussion on lines 277 to 287 of whether seasonality affects your Sf metric. But I still don't understand how seasonality is not a problem with your method. Using an annual percentile means that in cases with strong seasonality your episodes will mostly occur in the wet season. Does this mean that the method can't say anything about the role of clustering in drier seasons? Why can't a seasonally varying percentile be used?

Response: We thank the referee for this question. Here we chose to work with annual percentiles as those percentiles are most impact relevant when considering flooding. However, our method can be applied using seasonally varying percentiles, taking certain precautions in the identification of episodes to avoid edge effects at each season transition (consider each year separately to avoid artificial clustering across years; add days at the beginning and end of each season to account for episodes starting in one season and ending in another (Barton et al., 2016)). Such an analysis could be conducted in a further study.

Change: L287 **Finally, we note that our method can be applied using seasonally varying percentiles, by taking certain precautions in the identification of episodes to avoid edge effects at each season transition (Barton et al., 2016).**

Figure 8, showing the intersections between clustering and large precipitation accumulations, is perhaps the key results figure of the paper. The regional differences are intriguing and the reasons for this regional variability likely depends on the regional climate processes. Can you suggest a few?

We agree with the referee that this is a particularly interesting and relevant aspect. A detailed analysis of the drivers of subseasonal clustering is beyond the scope of this paper, whose focus is on introducing a new methodology. However, we now discuss the underlying structure of the precipitation time series for representative catchments (new Appendix A with examples) and added references to existing literature.

Change: new section 3.1 and Appendix A.

L243: The physical drivers of the sub-seasonal clustering of extreme precipitation are numerous and a detailed analysis of the identified clustering patterns is beyond the scope of the present research. Generally speaking, sub-seasonal clustering of extremes requires either very stationary or recurrent conditions that locally provide the ingredients for heavy precipitation (lifting and moisture) (Doswell et al. 1996). In some areas, large-scale patterns of variability have found to be relevant, such as the North Atlantic Oscillation (e.g., Villarini et al., 2011; Yang and Villarini, 2019; Barton et al., in preparation), the El Niño Southern Oscillation (Tuel and Martius, 2021) or the variability of the extratropical storm-tracks (Bevacqua et al., 2020). However, in other areas the circulation patterns associated with clustering differ from the patterns of variability (Tuel and Martius, in preparation). We direct the interested readers to the above-mentioned publications.

Did you see any 40-year trends in extreme event counts or large precipitation accumulations? Do the dates of the episodes mostly fall in the latter half of the 40-year period? It could be interesting to map the ratio of the numbers of episodes in the first 20 years vs. the final 20 years, and whether the contribution of clustering changes across the two periods. This is a suggestion for additional analysis and is not required in the revision.

Response: We thank the referee for this suggestion and agree that analysing the presence of trends is a relevant point. We didn't perform any trend analysis neither in the extreme event counts nor in the accumulations for the present research but that would be an interesting aspect to study.

Change: see below.

My understanding is that the method converts the precipitation data to binary, and therefore loses information on the magnitude of the individual extreme precipitation events. If this is correct, then I don't think it would be too much additional data processing to additionally retain magnitude information. In doing so, many other scientific questions could be pursued. For example, you could look at sequencing and explore statistically significant differences in the magnitudes of the 1st, 2nd , 3rd events within an episode, and how this varies regionally. I'm not suggesting you add this to the paper, but maybe note this as a potential extension in the Discussion.

Response: we agree with the referee that this is a potentially interesting question to explore. We indeed convert the precipitation data to binary events/non-events in our study, and don't retain the magnitude of each event. However, information on the magnitude of each extreme event within each episode could easily be added using the daily precipitation data.

Change: see below.

I think the paper would be stronger with a more in-depth discussion of how this method can aid physical process understanding of the clustering mechanisms. You go some way down this route in Fig. 11 but there is a lot more that could be done. For example, you could look at scalings between clustering and temperature or other environment variables. Or you could map out the time window length of the strongest clustering to get clues about contributing processes. Again, I'm not suggesting you do these analyses for this paper, but some further discussion about the ways the method aids process understanding is needed.

Response: We thank the referee for these very interesting suggestions. We will expand the discussion on how the method can be applied to get further insights on process understanding.

Change:

L306 The objective of the present paper was to introduce a new methodology and to demonstrate its application to the study of sub-seasonal clustering of extreme precipitation. It paves the way for further research on several aspects. First, potential extensions of the method itself could be explored, such as integrating the magnitude of each extreme event within an episode and sequencing its variability. Second, possible trends in the contribution of clustering to accumulations could be studied by comparing values of Scl and Scont in the first half and the second half of the investigated period. Third, the method could provide insights into the physical drivers of clustering by looking at scalings between the two metrics and other environmental variables (such as temperature or pressure) during selected clustering episodes or globally.

<u>Technical Corrections</u>

Fig3 and Fig4 could be merged into a single figure.

I didn't see Table 3 referenced anywhere in the main text.

We thank the referee for identifying those two points. Fig.3 will be merged with Fig.4 and a reference to Table 3 is now made at line 52, in the end of the introduction:

Change: L50: The paper is organised as follows: the data and methods are introduced in section 2. The results are presented and discussed in section 3. Finally, general conclusions and future research avenues are presented in section 4. **All important quantities used in this study are listed in Table 1.**

Table 3 is now Table 1.

Revised section 2.3 (Identification of sub-seasonal clustering episodes):

L100-118 **The identification of sub-seasonal clustering episodes is equivalent to searching for time periods (here 2 to 4 weeks) that contain several extreme precipitation events**. The first step is to count the number of independent extreme precipitation events (n_w) in a **running (leading)** time window of w days, after the runs declustering has been applied to the time series. This count is computed for each day of the time series over the next w − 1 days (not w, as the starting day is included in the time window length). In parallel, we calculate the running sum of daily precipitation (acc_w) over **the same leading** time window w. Time windows of w = 14, 21 and 28 days were investigated. **Fig. 3c and 3d show the values of n_21 and acc_21, corresponding to the time series of Fig. 3a**.

We then run an automated clustering episode identification algorithm that consists of the following steps: **(i) isolate the days with the largest value of n_w (highlighted in red in Fig. 3c). (ii) Among these days, retain the one with the largest accumulation acc_w (the purple bar in Fig. 3d). This selects a clustering episode which starts at the retained day and ends w-1 days later (shown by the red rectangle in Fig. 3a). The clustering episode identified in Fig. 3 contains four extreme events (n_21 = 4) and the related accumulation acc_21 is 275 [mm]. (iii) reduce the time series by removing all days within w − 1 days before and after the starting day of the selected episode (the purple window in Fig. 3d)**, to avoid further selected episodes from overlapping. **(iv) repeat steps (ii) and (iii) on the reduced time series to successively select the next episodes with the largest values of n_w and acc_w** until a predetermined number of episodes N_ep **= 50** is reached. The choice of N_ep is discussed below in greater detail, and **at this stage we emphasize that limiting the selection to 50 episodes is sufficient for our method**. This iterative selection results in the identification of 50 non-overlapping clustering episodes **sorted by the number of extreme events (n_w) and then by accumulations (acc_w)**. We denote this classification as Cl_n. **The left panel of Table 1 shows the Cl_n classification obtained for a subcatchment of the Tagus river in the Iberian Peninsula (HydroBASINS ID: 2060654920)**. **The Cl_n classification contains information about the frequency of sub-seasonal clustering. In a catchment where sub-seasonal clustering scarcely happens, Cl_n would typically be composed of a majority of episodes having a small number of extremes (e.g. n_w <=2). Whereas for a catchment where sub-seasonal happens frequently, Cl_n would be composed of several episodes with more extreme events (e.g. 2 <= n_w <= 6). Additional examples of catchments can be found in appendix A.**

In addition, we identify and classify the episodes with the **largest** precipitation accumulations as follows: we apply steps (ii) to (iv) of the automated identification algorithm to the **accumulation** time series. **This is equivalent to selecting episodes using the sole criteria of maximising acc_w (the 21-days accumulations) at each iteration**. **This**

second selection results in the identification of 50 non-overlapping episodes sorted by accumulations (acc_w). We denote this classification as Cl_acc. **The right panel of Table 1 shows the Cl_acc classification obtained for the same catchment as the left panel. All episodes listed in Table 1 are represented on the yearly timeline of Fig. 4 (in orange for Cl_n, in blue for Cl_acc and in grey when they overlap), along with the timing of all extreme events (black dots). We note that the choice of a centred or lagged window, instead of a leading window, does not change the values of n_w and acc_w**, except for the first and last w days of the time series. **This has no significant impact on the results**.

The degree of similarity between Cl_n and Cl_acc is the key point in our method to evaluate the contribution of clustering to large accumulations. This degree of similarity can be evaluated by doing a rank-by-rank comparison of the number of extreme events (n_w) in the episodes of Cl_n with the episodes of Cl_acc. If the episodes composing Cl_acc and Cl_n have the same n_w at each rank, then it means that the episodes with the largest number of extreme events are also leading to the largest accumulations. In this particular case, the contribution of clustering to accumulations is maximised. If an episode of Cl_acc has fewer extreme events than the episode with the same rank in Cl_n, then the contribution of clustering to accumulations is below the maximum contribution. The episodes selected in Cl_n and Cl_acc can be the same and ordered similarly or differently (they appear in grey in Fig. 4), but they can also differ (they appear in orange or blue in Fig. 4). The fifth columns of the left and right panel in Table 1 illustrate such a comparison, where the corresponding rank of each episode in the other classification is displayed. If the column is empty, it means that the episode is not present in the other classification. In this example, both classifications share the same first episode (n_w = 5), but their second and third episodes have different n_w. We also note the episodes without extreme events in Cl_acc (at ranks 11, 24, 30,...). The additional examples in appendix A illustrate cases with different degrees of similarity between Cl_n and Cl_acc.

L121: ~~As a preliminary remark, we note that if the Cln classification of a given catchment has many clustering episodes that contain several extreme events, then sub-seasonal clustering is occurring frequently in that catchment. Similarly, if the two classifications Cln and Clacc have episodes with the same number of extreme events at identical ranks, this implies that the episodes with the largest number of extreme events correspond to the episodes with the largest precipitation accumulations. In this case, the contribution of sub-seasonal clustering to large precipitation accumulations is maximised.~~

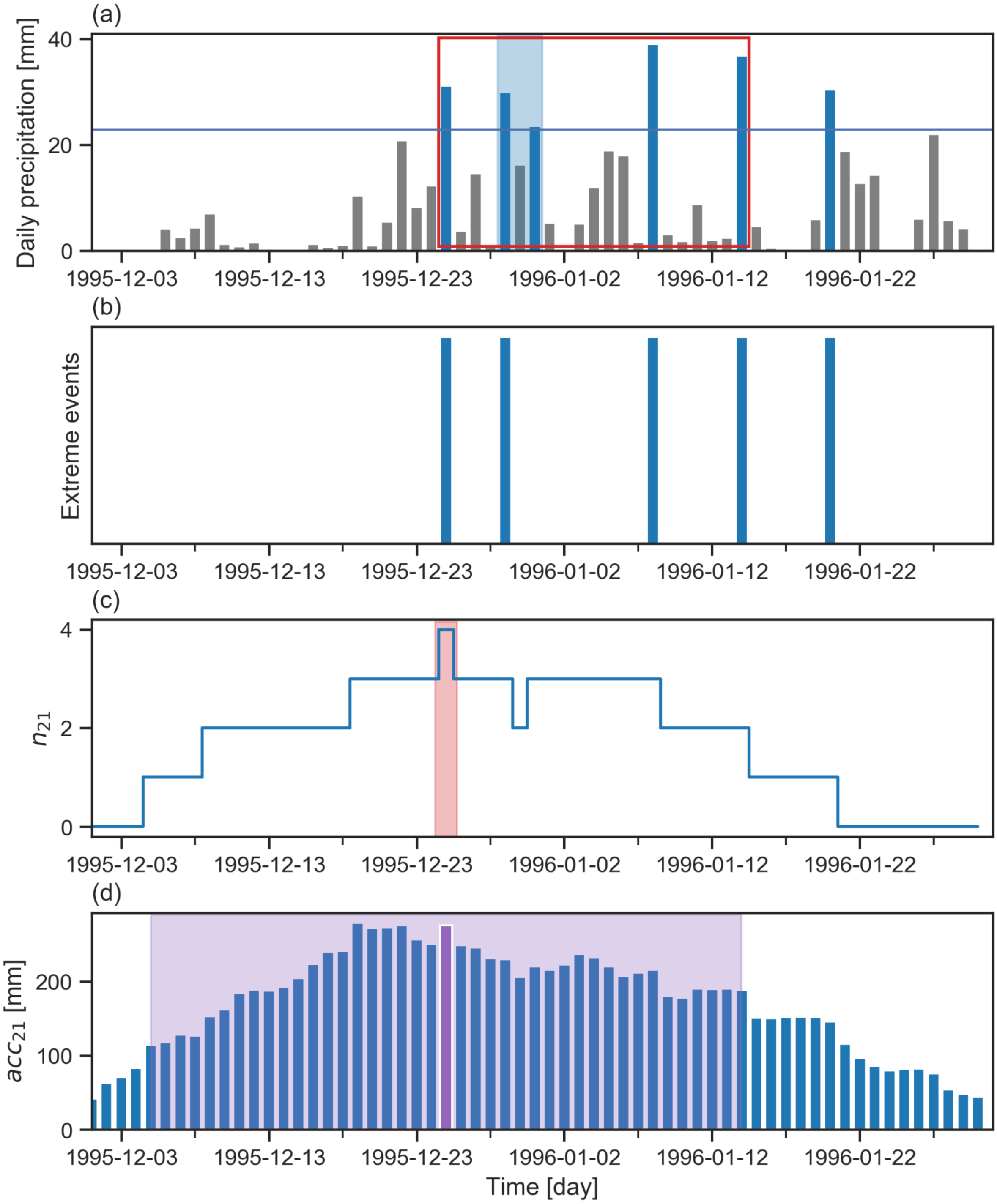


Fig. 3. Schematic illustration of the identification of a sub-seasonal clustering episode with w = 21 days. (a) Time series of daily precipitation with extreme precipitation days marked by blue bars; the horizontal blue line represents the threshold t (e.g. the 99th percentile) defining the extreme events; the light blue shading highlights a high-frequency cluster (r = 2 days) and the red rectangle denotes the clustering episode identified using the information of panel c and (d). (b) Series of binary events of extreme precipitation obtained after applying the declustering approach to the daily precipitation. (c) Number of extreme precipitation events in a running (leading) time window of 21 days (n_21) based on the time series in panel (b); the light red shading indicates the day with the largest n21. (d) Precipitation

accumulation in a running (leading) time window of 21 days (acc_21) derived from the time series of panel (a); the purple bar denotes the day with the largest acc21 among the days with highest n21; this day is the starting day of the selected clustering episode; all days within the light purple shading are removed from the initial time series in the next step of the selection algorithm.

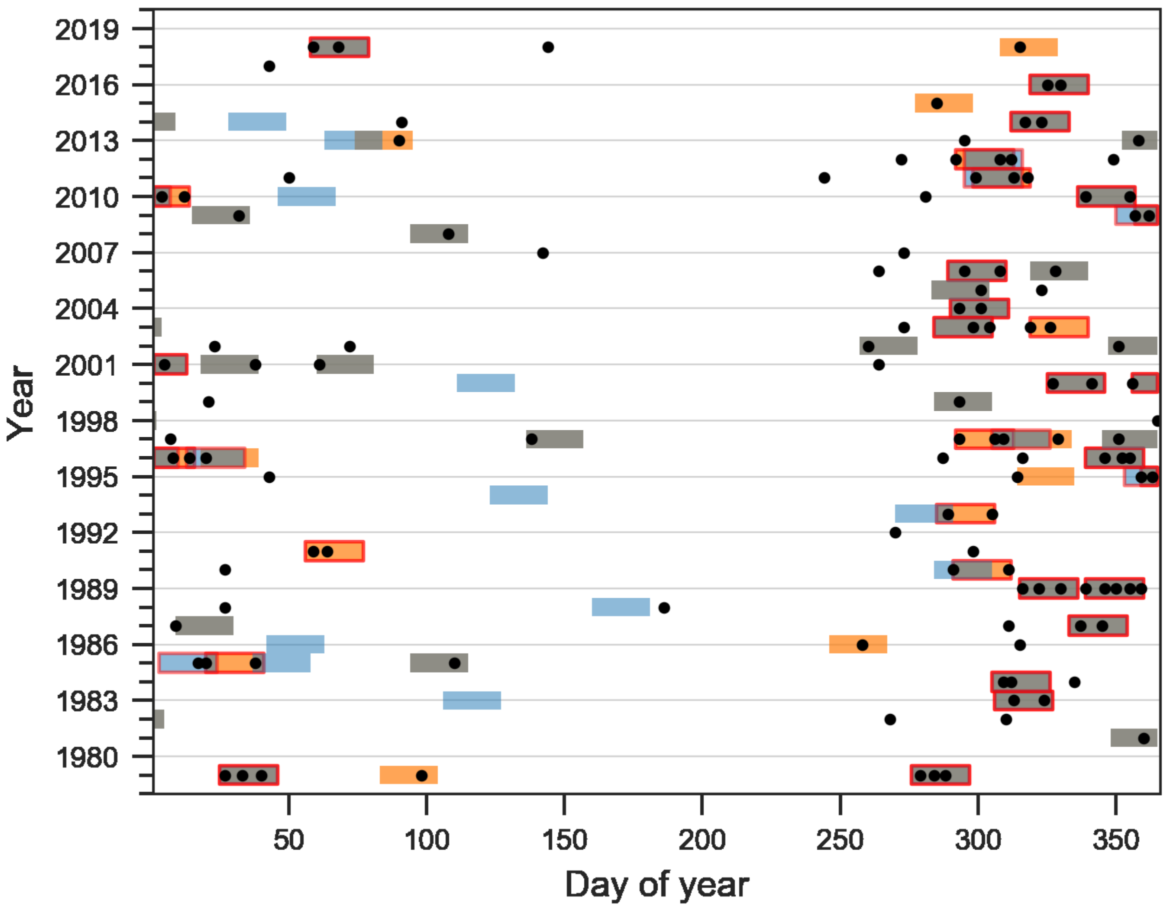


Fig. 4. For the catchment 2060654920, all extreme events are shown as black dots and 21-day episodes are highlighted by the colored rectangles. Episodes appearing in both classifications are shown in grey and those appearing only in the Cl_n classification are shown in orange whereas those only in the Cl_acc classification are shown in blue Episodes containing two or more extreme events (n_w >= 2) are highlighted with a red edge.

CIn					CIacc				
starting day	acc_21 [mm]	n_21	Rank CIn	Rank CIacc	starting day	acc_21 [mm]	n_21	Rank CIacc	Rank CIn
05.12.1989	281	5	1	1	05.12.1989	281	5	1	1
25.12.1995	275	4	2		19.12.1995	279	3	2	
23.12.2009	213	4	3		16.10.2006	275	2	3	11
25.01.1979	247	3	4	5	27.02.2018	255	2	4	12
11.11.1989	242	3	5	6	25.01.1979	247	3	5	4
04.12.1996	229	3	6	7	11.11.1989	242	3	6	5
03.10.1979	188	3	7	16	04.12.1996	229	3	7	6
19.10.1997	188	3	8		16.12.2009	220	3	8	
18.10.2012	161	3	9		21.12.2000	214	2	9	13
25.10.2011	141	3	10		02.11.1983	212	2	10	14
16.10.2006	275	2	11	3	15.02.2010	202	0	11	
27.02.2018	255	2	12	4	14.12.1981	196	1	12	28
21.12.2000	214	2	13	9	01.11.1997	191	2	13	
02.11.1983	212	2	14	10	20.11.2000	191	2	14	15
20.11.2000	191	2	15	14	13.01.1996	190	2	15	
20.01.1985	160	2	16		03.10.1979	188	3	16	7
02.12.2010	159	2	17	20	21.10.2012	169	2	17	
29.11.1987	154	2	18	23	06.02.1985	163	1	18	
16.10.2004	144	2	19	28	27.09.1993	161	1	19	
31.10.1984	142	2	20	29	02.12.2010	159	2	20	17
08.11.2014	141	2	21	31	11.10.1999	156	1	21	29
12.10.1993	139	2	22		13.12.2002	155	1	22	30
11.10.2003	135	2	23	36	29.11.1987	154	2	23	18
14.11.2016	118	2	24	48	04.03.2013	151	0	24	
25.02.1991	117	2	25		18.12.2013	150	1	25	32
18.10.1990	112	2	26		22.10.2011	147	2	26	
15.11.2003	85	2	27		01.03.2001	144	1	27	33
14.12.1981	196	1	28	12	16.10.2004	144	2	28	19
11.10.1999	156	1	29	21	31.10.1984	142	2	29	20
13.12.2002	155	1	30	22	11.02.1986	142	0	30	
18.01.1996	152	1	31		08.11.2014	141	2	31	21
18.12.2013	150	1	32	25	15.11.2006	138	1	32	35
01.03.2001	144	1	33	27	14.09.2002	137	1	33	36
15.03.2013	142	1	34		03.01.1985	137	2	34	
15.11.2006	138	1	35	32	10.10.2005	136	1	35	37
14.09.2002	137	1	36	33	11.10.2003	135	2	36	23
10.10.2005	136	1	37	35	11.10.1990	135	1	37	
03.04.2008	134	1	38	38	03.04.2008	134	1	38	38
09.01.1987	126	1	39	43	28.01.2014	133	0	39	
09.11.1997	122	1	40		03.05.1994	129	0	40	
15.01.2009	121	1	41	44	16.04.1983	128	0	41	
11.12.1997	120	1	42	45	20.04.2000	127	0	42	
18.01.2001	119	1	43	46	09.01.1987	126	1	43	39
04.04.1985	119	1	44	47	15.01.2009	121	1	44	41
16.05.1997	118	1	45	49	11.12.1997	120	1	45	42
04.10.2015	116	1	46		18.01.2001	119	1	46	43
04.11.2018	116	1	47		04.04.1985	119	1	47	44
03.09.1986	114	1	48		14.11.2016	118	2	48	24
24.03.1979	113	1	49		16.05.1997	118	1	49	45
10.11.1995	112	1	50		08.06.1988	117	0	50	

Table 1. Left panel: Episodes with the largest number of extreme events (n_21) retained in the CI_n classification for catchment with HydroBASINS ID: 2060654920 (corresponding to a subcatchment of the Tagus river in the Iberian Peninsula). Columns are (from left to right): starting day of the episode, accumulation during the episode (acc_21), number of extreme events during the episode (n_21), rank of the episode (Rank CI_n), rank of the episode in the CIacc (Rank CI_acc), an empty Rank CI_acc column means that the episode is not

present in this classification. Right panel: Same as left panel but for episodes with the largest accumulations (acc_21) retained in the Cl_acc classification.

Revised section 2.4 (Metrics for sub-seasonal clustering):

Next we define metrics that synthesize the properties of the two classifications to compare catchments. An intuitive choice for the metrics would be to average the number of extreme events, however such a would result in a loss of information (see Appendix D for a more detailed discussion on this). We take a different approach, equivalent to defining a scoring system, where each episode is given a weight q_i depending on its rank in the classification, and this weight is used as a proportion factor for the number of extreme events in the episode. **We have many options for defining the weights. For example, taking the average over the N_ep episodes (as discussed in Appendix D) is the same as setting all weights equal to 1/N_ep.** Sitarz (2013) discusses a mathematical approach for defining a scoring system in sports, with two intuitively appealing properties. First, the first place should be rewarded more points than the second, and the second more than the third, and so on. In our case, rewarding more points is equivalent to giving a larger weight. Second, 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. **The second property means that someone gaining a place (or a rank) should be rewarded more if the initial rank is higher, as improving at upper ranks is more challenging than improving at lower ranks**. We then follow the method of the incenter of a convex cone (Sitarz, 2013) to construct our weighting scheme (see Appendix B for a detailed description). **The same weight q_i is assigned to the ith episode of each classification (Cl_n and Cl_acc)**. **We have tried two other weighting schemes, also satisfying the two required properties: the inverse of the rank (q_i = 1/i) and the inverse of the square root of the rank (q_i = 1/sqrt(i)). The former gave slightly too much weight to the very first episodes of the classification and the latter gave almost identical results to the incenter method. Our results are hence only slightly sensitive to the choice of the weighting scheme, as long as it satisfies the two desired properties.**

We can now use each weight q_i as a proportion factor for the corresponding number of extreme events in the ith episode for both classifications **and derive the three following metrics:**

Clustering Metric: Scl = $\sum_{i \in Cl_n} n_w(i) \cdot q_i$

Accumulation Metric: Sacc = $\sum_{i \in Cl_{acc}} n_w(i) \cdot q_i$

Contribution Metric Scont = Scl / Sacc

The first metric S_{cl}, **called the clustering metri**c, is the weighted (q_i) sum of the number of extreme events ($n_w(i)$) over all episodes (i = 1 to 50) in the Cl_n classification. **S_{cl} is proportional to the number of extreme events in the clustering episodes. It is most sensitive to the number of extreme events in the first clustering episodes, which are given the largest weight. In section 2.5, we show that S_{cl} correlates well with the index of dispersion -- a widely used measure of clustering. Appendix A provides examples of catchments with high and low values of S_{cl} for illustration.**

The second metric S_{acc}, called the accumulation metric, is computed similar to S_{cl}, but using the episodes of the Cl_{acc} classification, where episodes were ranked according to their accumulations. As S_{cl} and S_{acc} are computed using the same weights, their ratio S_{cont} can be used to make a rank-by-rank comparison. S_{cont} is equal to 1 when S_{acc} = S_{cl}, i.e. when the two classifications have episodes with the same number of extreme events at identical ranks. S_{cont} is equal to 0 when S_{acc} = 0, i.e. when all episodes in the S_{acc} classification contain no extreme events (n_w(i) = for all i in [1,N_ep]). In this particular case, **subseasonal clustering does not contribute to large accumulation and there is even no contribution of single extremes to large accumulations.** In other cases, **a proper assessment of the contribution of clustering to large accumulations is done by considering both S_{cl} and S_{cont}. S_{cont} alone evaluates the similarity of the two classifications and catchments can have low values of S_{cl} (limited sub-seasonal clustering) and high values of S_{cont} at the same time**. The exact interpretation of intermediary values of S_{cont} requires looking at both classifications (Cl_n and Cl_{acc}) in detail to see where they differ from each other. **For example, if S_{cont}= 0.8, both classifications have a high degree of similarity, but it does not necessarily imply that 80% of the episodes are ranked equally. Appendix A provides examples of catchments having high and low values of S_{cont} as an illustration. We normalize S_{cont} to compare different catchments and to assess their sensitivity to the choice of the parameters.**

We now briefly address some technical points related to the definition of the metrics. We note that performing a regression between Cl_n and Cl_acc would be a more conservative approach in assessing their degree of similarity because it would require giving a unique identifier to each episode according to its starting day. In that case, the strength of the regression would be lowered when two episodes containing the same number of extreme events just swap their ranks in the two classifications. Such a change does not affect S_cont.

S_cl and S_acc both increase with the number of extreme events per episode so any parameter change which increases this number will also lead to an increase in Scl and Sacc. Appendix C shows boxplots of S_cl for all parameter combinations. We see that a lower threshold t, a shorter run length r, and a larger window w lead to an increase in the values of S_cl. However, the sensitivity of S_cl and S_acc to the parameters does not affect our general conclusions. First, a change of parameters impacts all catchments, so while the scale of S_cl (or S_acc) is changed, the comparison of two catchments will result in the same conclusion in almost all cases. That is, a catchment with a relatively low value of S_cl compared to other catchments for one parameter combination will also have a relatively low value for other

combinations and similarly for high values. Second, the sensitivity of S_cont to the parameters (which depends on both S_cl and S_acc) is explicitly assessed in section 3.2 and accounted for in our results.

[Continue at L177]

Appendix B: Calculation of the weights (new) - composed of L129 to L150.

Revised section 2.5: (Correlations with index of dispersion and significance test):

L185-197: unchanged

L197: ~~This is illustrated in Fig. 6, which shows Sf and phi for the initial parameter combination and where it can be seen that regions of high (low) Sf correspond to regions of high (low) phi. Figure 6a is further discussed in the results section.~~

L197: This is further illustrated in **Fig. 6a ad Fig. D1** in Appendix D, which respectively show a map of S_cl and a map of phi_hat for the initial parameters combination. A visual comparison of the two maps reveal that regions of high (low) S_cl correspond to regions of high (low) phi_hat.

An evident drawback of S_cl compared to phi_hat is the lack of a reference value above (below) which there is (no) clustering (phi_hat = 1). While we cannot derive such a reference value, we can still use a bootstrap based approach to assess how significant the value of Sn is for each catchment. More precisely, we tested the following hypothesis:

H0: The clustering episodes contain a number of extreme precipitation events (n_w) which is not higher than for a distribution of those extremes without temporal structure (random).

H1: The clustering episodes contain a number of extreme precipitation events (n_w) which is significantly higher than for a distribution of those extremes without temporal structure (random).

and we reject H0 if the observed value of S_cl is significantly greater than a given threshold. A rejection of H0 at a certain level of significance will be further noted as "significant sub-seasonal clustering" for simplicity. To this end, 1000 random samples were generated by doing permutations of the precipitation time series (i.e. each daily value is drawn only one time in each sample, without repetition, this way the distribution quantiles remain identical.). S_cl was calculated for each sample, using the initial parameters combination, and leading to an empirical distribution of S_cl values. An empirical cumulative distribution function (ECDF) was calculated from the S_cl empirical distribution, and an empirical p-value was obtained by evaluating the ECDF at the observed S_cl value: 1-ECDF(S_cl(obs)). At a 1% level, approx. 42% of

the catchments (2729 out of 6466) show significant sub-seasonal clustering (Fig. 6b, catchments in red).

Interestingly, the whole S_cl empirical distribution based on the random samples is almost identical for all catchments, with a mean value around 31.42. This means that a selection of catchments based on a given level of significance can be well approximated by a selection based on relatively high observed S_cl values. In section 3, we select catchments which are either below the 25th percentile or above the 75th percentile of the observed S_cl distribution for all catchments. It allows for a quick selection of catchments with rare or prevalent sub-seasonal clustering for each parameters combination, whereas the permutation/resampling approach would have required more computational time. We compared the two selection methods for the initial parameters combination and found only limited differences.

Many catchments have a very low p-value because we take an annual percentile for defining the extreme precipitation events. With this definition, catchments with strong seasonality in the precipitation (e.g. with extremes occurring during a "wet" season) will have their extreme events occurring only during a few months. A random permutation of the daily precipitation will redistribute the extremes equally during the year in most cases, corresponding to much lower values of S_cl. Taking seasonal percentiles would most likely result in fewer catchments having very low p-values. The implications of seasonality and the choice of an annual percentile are further discussed in section 4.

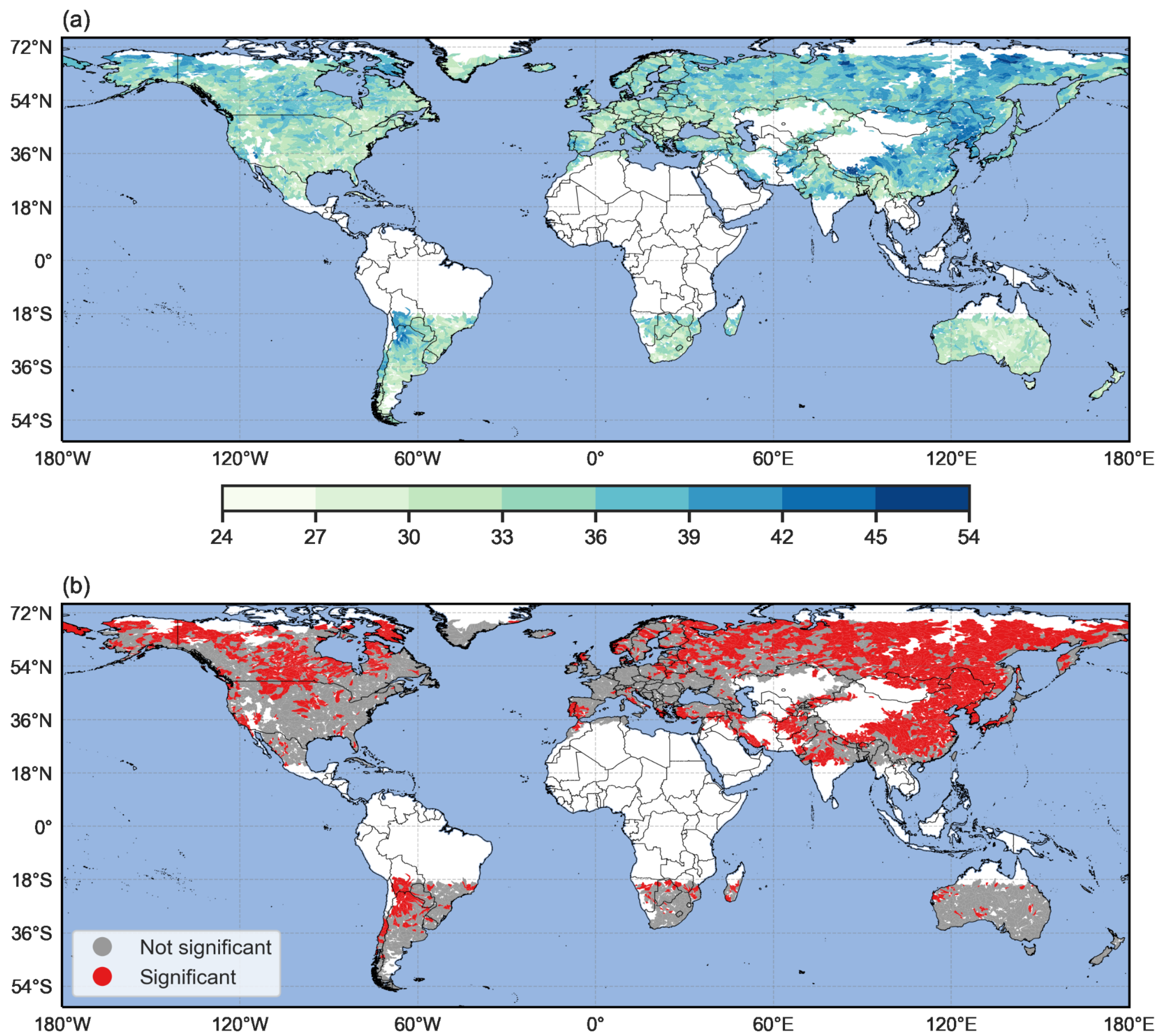


Fig. 6. Metric S_cl (a) and sub-seasonal clustering significance (b) by catchment, for r = 2 days, t = 99p, w = 21 days. In (a), high values of S_cl denote catchments where sub-seasonal clustering is prevalent. In (b), catchments where S_cl is significantly higher than for a distribution of extremes events without temporal structure are shown in red at the 1% level.

Revised section 3 (Results):

~~*First, world maps of the **clustering (Scl) and contribution (Scont) metrics** for all selected catchments are shown using the initial combination of parameters (r = 2 days, t = 99p, w = 21 days). These maps indicate regions where sub-seasonal clustering is prevalent. Then, the sensitivity of the sub-seasonal clustering to the parameter choice is assessed by testing 12 different parameter combinations: w = 14,21,28 days; u = 98p,99p; r = 1,2 days.*~~

3.1 Sub-seasonal clustering and its contribution to accumulations:

Sub-seasonal clustering is prevalent in catchments having high values of S_cl (see section 2.5). Such catchments are located in the east and northeast of the Asian continent (northeast of Siberia, northeast of China, Korean Peninsula, south of Tibet); between the northwest of Argentina and the southwest of Bolivia; in the northeast and northwest of Canada as well as in Alaska; and in the southwestern part of the Iberian Peninsula (Fig. 6a and 6b). Regions with low values of **S_cl** are located on the east coast of North America, on the east coast of Brazil, in central Europe, in South Africa, in central Australia, in New Zealand and in the north of Myanmar (Fig. 6a and 6b). Catchments with strongly contrasting values of **S_cl** are rarely found in close proximity, except for a group of catchments located northeast of the Himalayas (south of Tibet), and another group located southeast of the Himalayas (Bangladesh and Myanmar). The catchments to the north have high values of **S_cl**, whereas the neighbouring catchments to the south exhibit low values of **S_cl**.

~~Regions with large values of the relevance metric (Scont, see Fig. 7) are in the east and northeast of the Asian continent, west of India, central Australia and central North America. Areas with low values of Scont are located in central China, on the east coast of North America, in the south of Brazil and in France.~~

The contribution of sub-seasonal clustering to precipitation accumulations is analysed with both S_{cl} and S_{cont}. Catchments with high values of S_{cl} and S_{cont} are of special interest, because in these catchments, sub-seasonal clustering is prevalent and contributes substantially to large 21-days precipitation accumulations. We identify such catchments by considering those whose values of S_{cl} and S_{cont} are greater than the 75th percentile of their respective distribution for all catchments. The choice of the 75th percentile makes it possible to focus on the highest values, without being too restrictive, and follows the quick selection method mentioned in section 2.5. Catchments where sub-seasonal clustering is prevalent and contribute substantially to large accumulations are mainly concentrated over eastern and northeastern Asia (Fig. 7a). The largest continuous area of such catchments is located in northeastern China, in North and South Korea, Siberia and east of Mongolia. Other areas with several catchments of interest are central Canada, south California, Afghanistan, Pakistan, the southwest of the Iberian Peninsula, the north of Argentina and the south of Bolivia. Every continent includes groups of two to three or isolated catchements. **Appendix A1 contains detailed information for an example catchment with a strong seasonality located in northeastern China (S_{cl} = 41.14, S_{cont} = 0.93). Almost all extreme events happen between June and August, which make clustering episodes and periods of large accumulations more likely to overlap.**

We also identify catchments with values of S_cl below the 25th percentile and values of S_cont above the 75th percentile (Fig. 7b). Low values of S_cl mean that the clustering episodes identified by our algorithm contain a small number or even no extreme events, and high values of S_cont mean that those episodes lead to the largest accumulations. Such regions that exhibit rare clustering and where this rare clustering contributes substantially to large accumulations are the following: Taiwan, most of Australia, central Argentina, South Africa, south of Botswana and south of Greenland. **Again, every continent includes groups of two to three or isolated catchements**. Interestingly, the identified catchments are almost all located in the Southern hemisphere. **An example located in Australia is presented in detail in Appendix A2 (S_cl = 26.79, S_cont = 0.90). The extreme events are distributed throughout the whole year and only a limited number of episodes contain two or more extreme events.**

Finally, we identify regions with values of S_cl above the 75th percentile and values of S_cont below the 25th percentile (Fig. 7c). The high values of S_cl mean that the clustering episodes identified by our algorithm contain a relatively large number of extreme events, whereas the low values of S_cont mean that episodes leading to the largest accumulations contain a low number or even no extreme events. Such regions that exhibit prevalent clustering with a limited contribution to large accumulations are the following: the south of Tibet, the south of the Qinghai and west of the Sichuan Chinese provinces and central Bolivia. **Again, every continent includes groups of two to three or isolated catchements**. Only a few catchments exhibit this combination of high S_cl and low S_cont values, highlighting the importance of the clustering of extreme events for generating the largest accumulations for the majority of the catchments. **An example located in central China is presented in detail in Appendix A3 (Scl = 43.23, Scont = 0.59). The seasonality is present but less pronounced than in example A1: almost all extreme events happen between mid-May and September. However, in this case, clustering episodes and periods of large accumulations tend not to overlap as much as in Example A1. This is a particularly interesting feature, especially because the two different patterns exemplified by Appendix A1 and A3 happen in neighbouring regions.**

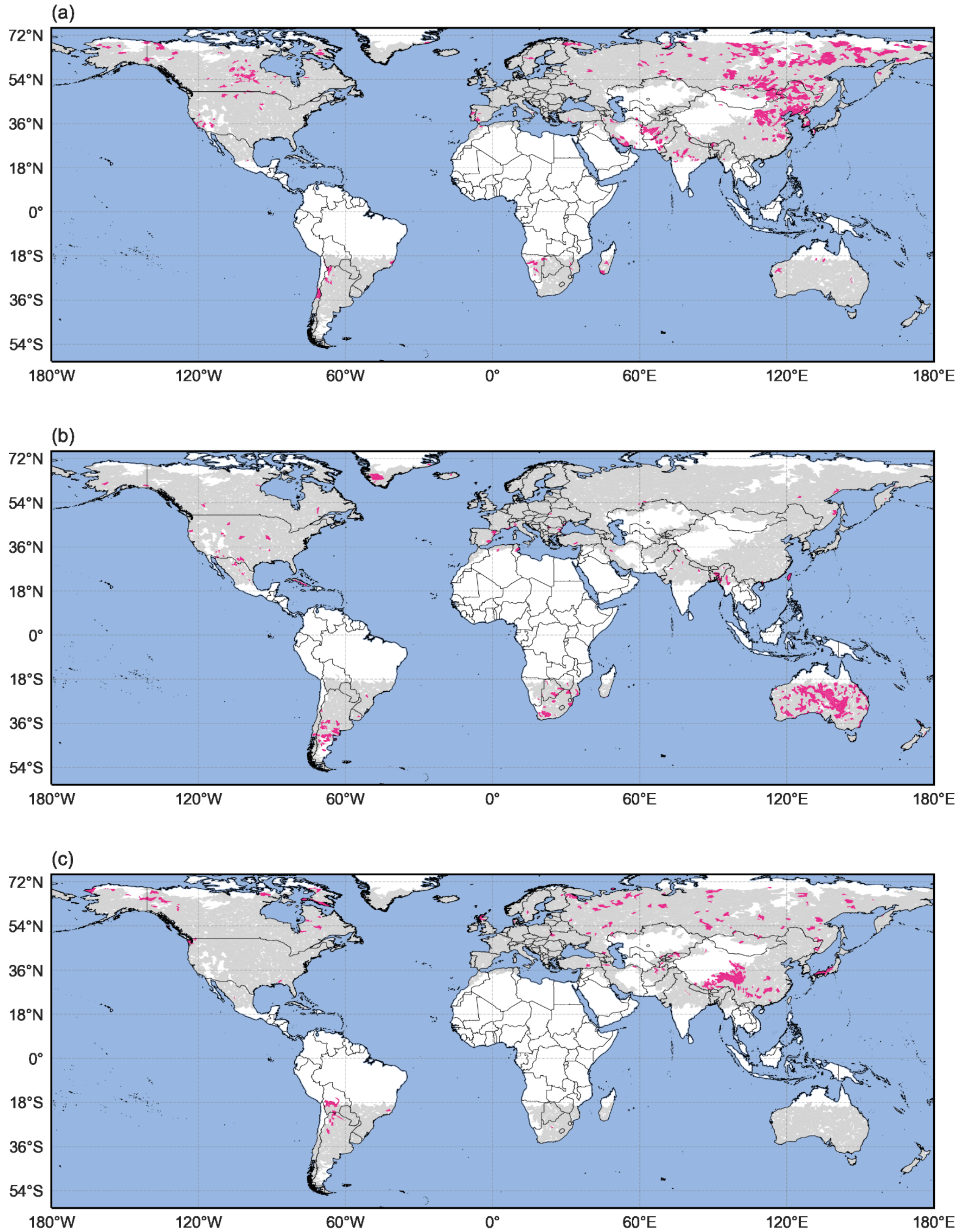


Fig. 7. (a) Catchments where S_{cl} and S_{cont} are both above the 75th percentile of their respective distribution (pink areas); (b) Catchments where S_{cl} < 25p and S_{cont} > 75p (pink areas); (c) Catchments where S_{cl} **>** 75p and S_{cont} **<** 25p (pink areas). In all panels, catchments in grey do not satisfy the respective conditions, whereas catchments in white were excluded from the analysis according to the criteria defined in section 2.1.

We investigated a potential link between the catchment size (in km2) and both the clustering (S_cl) and the contribution metric (S_cont), by computing their Spearman rank correlation coefficient, but found no significant correlations (not shown).

The physical drivers of the sub-seasonal clustering of extreme precipitation are numerous and a detailed analysis of the identified clustering patterns is beyond the scope of the present research. Generally speaking, sub-seasonal clustering of extremes requires either very stationary or recurrent conditions that locally provide the ingredients for heavy precipitation (lifting and moisture) (Doswell et al. 1996). In some areas, large-scale patterns of variability have found to be relevant, such as the North Atlantic Oscillation (e.g., Villarini et al., 2011; Yang and Villarini, 2019; Barton et al., in preparation), the El Niño Southern Oscillation (Tuel and Martius, 2021) or the variability of the extratropical storm-tracks (Bevacqua et al., 2020). However, in other areas the circulation patterns associated with clustering differ from the patterns of variability (Tuel and Martius, in preparation). We direct the interested readers to the above-mentioned publications.

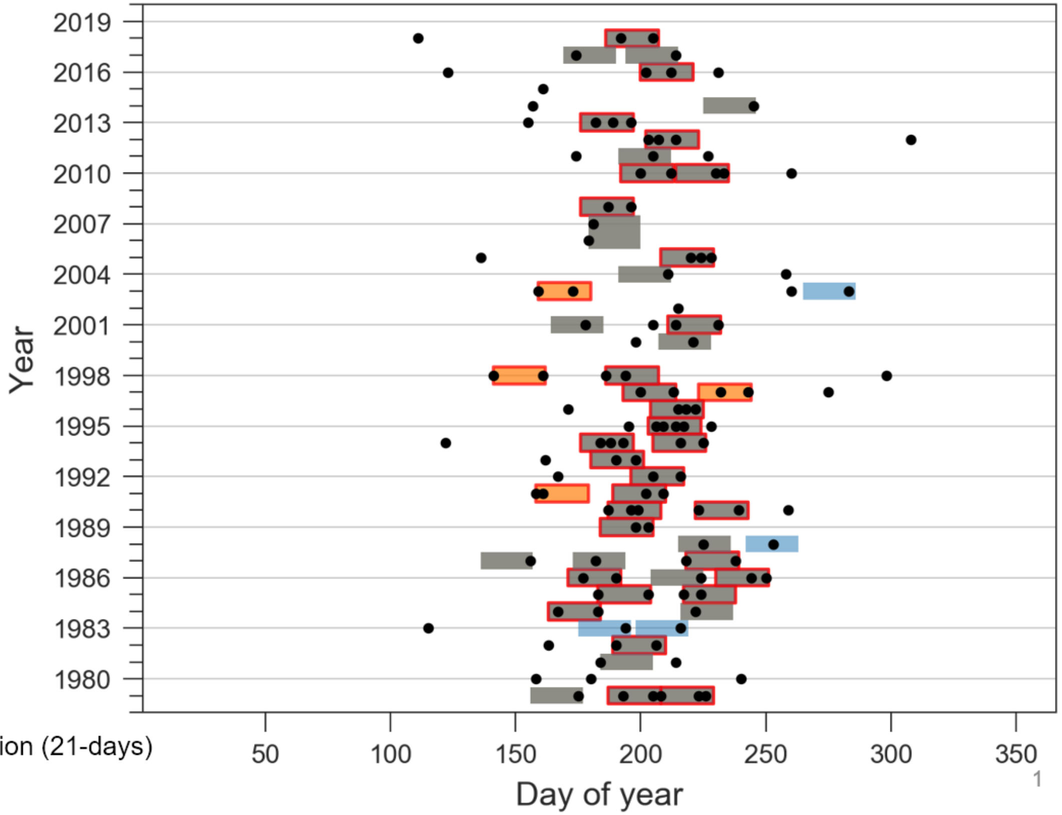
Appendix A: Examples of catchments

A1: Catchment with frequent subseasonal clustering contributing substantially to large accumulations (S_f = 41.14; S_r = 0.93)

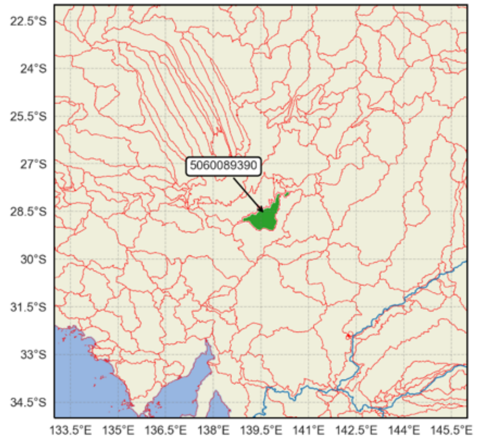
Green area: catchment location (northeastern China).
Annotation: HydroBASINS number.
Red lines: HydroBASINS catchment boundaries.

Area: 9893.4 [km]
99th quantile: 30.16 [mm]
# of episodes with ≥ 2 extreme events : 34

- Extreme events (> 99th percentile)
- Subseasonal clustering episode (21-days)
- Episode of large accumulation (21-days)
- Clustering episode contributing to large accumulation (21-days)
- Episode with ≥ 2 extreme events



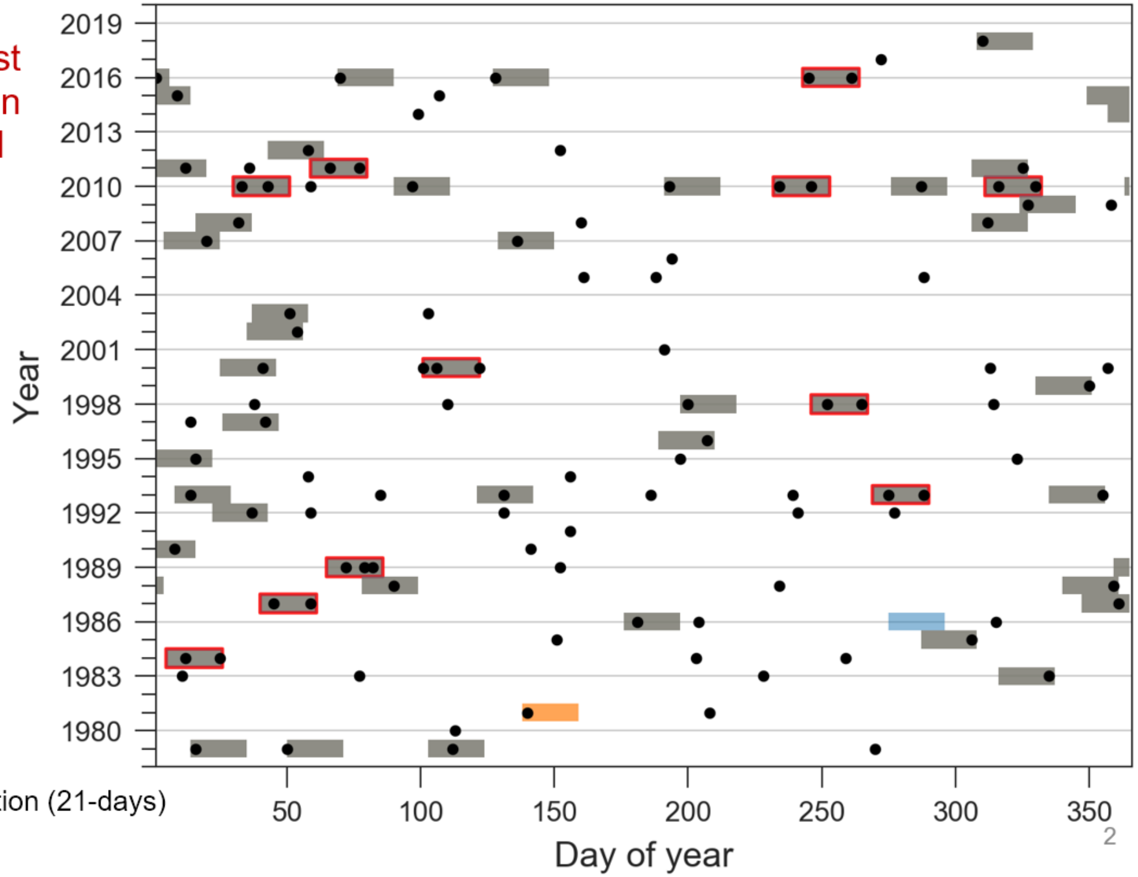
A2: Catchment with rare subseasonal clustering contributing substantially to large accumulations* (S_f = 26.79; S_r = 0.90)



*in that case most of the contribution is due to isolated extreme events

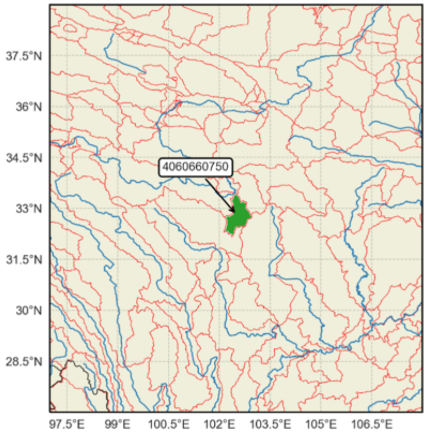
Green area: catchment location (Australia).
Annotation: HydroBASINS number.
Red lines: HydroBASINS catchment boundaries.

Area: 7769.8 [km]
99th quantile: 12.10 [mm]
# of episodes with ≥ 2 extreme events : 11



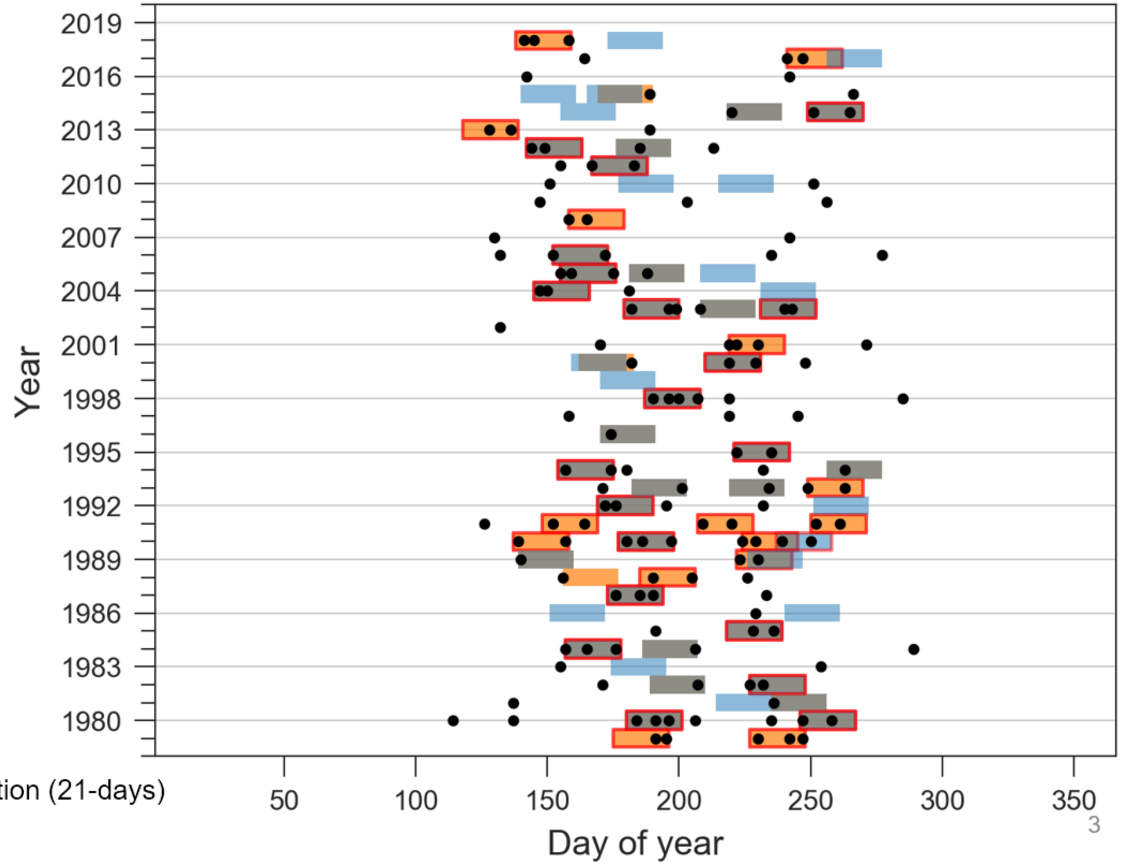
- Extreme events (> 99th percentile)
- Subseasonal clustering episode (21-days)
- Episode of large accumulation (21-days)
- Clustering episode contributing to large accumulation (21-days)
- Episode with ≥ 2 extreme events

A3: Catchment with frequent subseasonal clustering and limited contribution to large accumulations (S_f = 43.23; S_r = 0.59)



Green area: catchment location (central China).
Annotation: HydroBASINS number.
Red lines: HydroBASINS catchment boundaries.
Area: 5492 [km]
99th quantile: 17.31 [mm]
# of episodes with ≥ 2 extreme events : 35

- ● Extreme events (> 99th percentile)
- Subseasonal clustering episode (21-days)
- Episode of large accumulation (21-days)
- Clustering episode contributing to large accumulation (21-days)
- Episode with ≥ 2 extreme events



Appendix B: Calculation of the weights q_i

Sitarz (2013) assumes two intuitive conditions for a scoring system. First, he assigned more points for the first place than for the second place, and more for the second than for the third, and so on. Second, 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 is equivalent to considering the following set of points:

$$K = \left\{(x_1, x_2, \cdots, x_N) \in \mathbb{R}^N : x_1 \geq x_2 \geq \ldots \geq x_n \geq 0 \text{ and } x_1 - x_2 \geq x_2 - x_3 \geq \cdots \geq x_{N-1} - x_N\right\}$$

where x_1 denotes the points for first place, x_2 the points for second place,. . . , and x_N the points for Nth place. Any choice of points in K would satisfy the two conditions for a scoring system, however we would like to have a unique and representative value. The option chosen by Sitarz (2013) is to look for the equivalent of a mean value: the incenter of K. Formally, the incenter is defined as an optimal solution of the following optimization problem by Henrion and Seeger (2010):

$$\max_{x \in K \cap S_x} dist(x, \partial K)$$

where S_x denotes the unit sphere, dK denotes the boundary of set K and dist denotes the distance in the Euclidean space. By using the calculation presented in the Appendix of Sitarz (2013), and dividing by ~~the parameter lambda and~~ the points of the first place (x_1) to get the weights (q_i), we obtain:

q_i = x_i/x_1 for all i in [1,N]

The weight q_1 is always 1 but the values of weights q_2 to q_N depend on N and in our case N is the number of clustering episodes N_ep.

Appendix C:

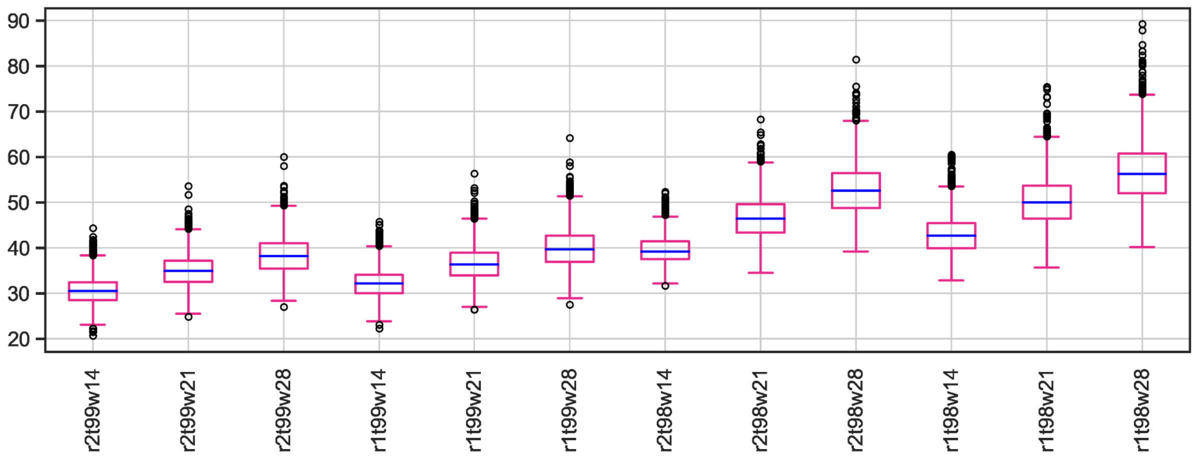


Fig. C1. Boxplots of S_cl for all catchments and parameters combinations. Boxes extend from the first (Q1) to the third (Q3) quartile values of the data, with a blue line at the median. The position of the whiskers is 1.5 * (Q3 - Q1) from the edges of the box. Outlier points past the end of the whiskers are shown with black circles.

Appendix D: Rationale behind the construction of the metrics

An intuitive choice to define the metrics (see section 2.4) is to use the sum or average of the number of extreme events over all (or a subset of) the episodes of CI_n and CI_acc. However, such a choice would result in a loss of relevant information on how the episodes are ranked, and preclude a rank-by-rank comparison between classifications. This can be illustrated with the following theoretical example: let us consider a catchment where CI_n is composed of 5 episodes, each with 3 extreme events, and 5 other episodes, each with 1 extreme event (i.e., N_ep = 10). The average number of extreme events is 2. If CI_acc is composed of the same episodes, then the average remains identical whatever the order of the episodes in CI_acc and we cannot say anything about the contribution of clustering to accumulations by comparing the averages. For example, all episodes with 1 extreme event could have larger accumulations than those with 3 extreme events. There is a low contribution of clustering to accumulations in this case, and metrics based on averages would not be able to capture this feature. A metric based on average would also fail to capture some differences in the same classification between two catchments. This again can be illustrated with a theoretical example: let us consider catchment A where CI_n is composed of 5 episodes: 1 with 5 extreme events, the 4 others without extreme event; and catchment B where CI_n is composed of 5 episodes, each with 1 extreme event. In both cases the average number of extreme events is 1 but the clustering behaviour is different. Consequently, we need a way to properly account for the respective rank of each episode in both classifications.

Appendix E:

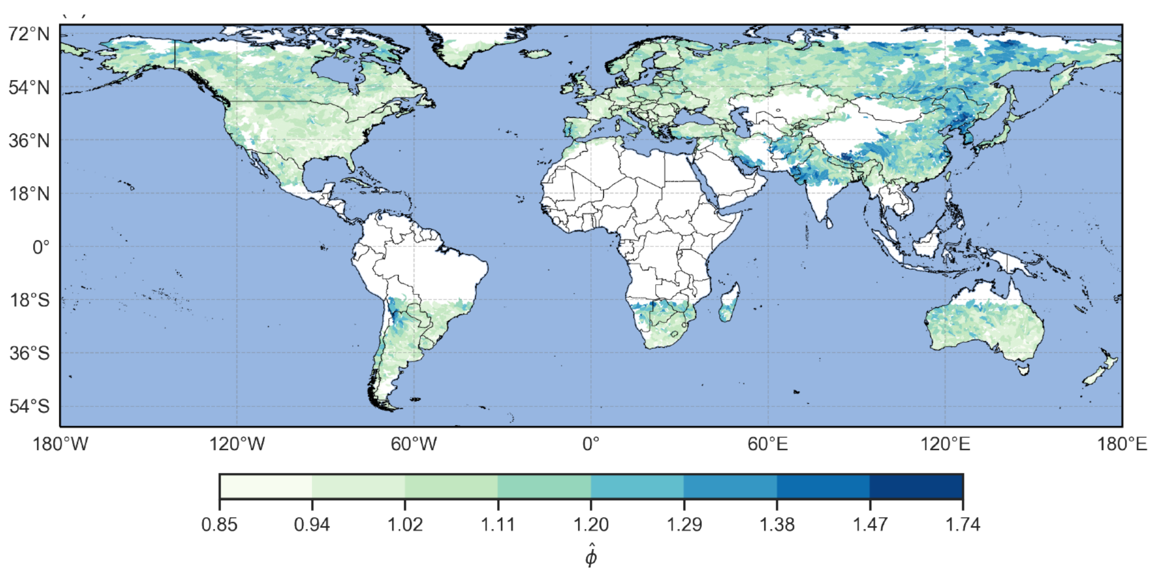


Fig. E1. Index of dispersion phi by catchment, for r = 2 days, t = 99p, w = 21 days. phi > 1 denotes catchments where extreme precipitation events are more clustered than random.

New References:

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., … Roberts, N. M. (2014). Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of Geophysics*, *52*(3), 522–555. https://doi.org/10.1002/2014RG000464

Donat, M. G., Sillmann, J., Wild, S., Alexander, L. V., Lippmann, T., & Zwiers, F. W. (2014). Consistency of Temperature and Precipitation Extremes across Various Global Gridded In Situ and Reanalysis Datasets, *Journal of Climate*, *27*(13), 5019-5035. Retrieved May 26, 2021, from https://journals.ametsoc.org/view/journals/clim/27/13/jcli-d-13-00405.1.xml

Rivoire, P., Martius, O., & Naveau, P. (2021). A comparison of moderate and extreme ERA-5 daily precipitation with two observational data sets. Earth and Space Science, 8, e2020EA001633. https://doi.org/10.1029/2020EA001633

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. Weather and Forecasting, 11, 560-581.

Tuel A. and Martius O. (2021), A global perspective on the sub-seasonal clustering of precipitation extremes, submitted to Weather and Climate Extremes, in review. This manuscript is confidential but if reviewers want, we are happy to share it.