Comment on wcd-2022-54
Anonymous Referee #1

Referee comment on "What distinguishes 100-year precipitation extremes over Central European river catchments from more moderate extreme events?" by Florian Ruff and Stephan Pfahl, Weather Clim. Dynam. Discuss., https://doi.org/10.5194/wcd-2022-54-RC1, 2022

Review of Ruff and Pfahl (2022): What distinguishes 100-year precipitation extremes over Central European river catchments from moderate extreme events.

In this paper, the authors study very extreme precipitation events over Central Europe river catchments, and the associated atmospheric dynamical conditions, in a robust way, thanks to the use of a very large ensemble of operational weather predictions. This clever and original approach allows them to study a much larger sample of extreme events than it would be possible with standard approaches based on observations or reanalyses.

The paper is generally well written, and discusses interesting results. But there are also some potential methodological issues, the methodological choices are not always well justified, and the implications of these choices are not always discussed. Therefore, I think that major revisions are needed.

General comments
The authors could have done the same analyses with climate models outputs: large Single Model Initial-condition Large Ensembles (SMILEs) exist and could be used to obtain large sample sizes to do robust analyses of extreme precipitation events. The implicit hypothesis of the authors is, I think, that a weather prediction system provides more accurate representation of extreme precipitation events than climate models. I think this implicit hypothesis should be stated, and discussed, using references.

Precipitation observations are not assimilated in weather prediction systems, I think (is it the case here? It should be discussed), and, in the end precipitation is strongly the result of the atmospheric model, especially at day 10. So, what is really the advantage of a weather predictions system compared to a climate model? Only resolution?

As discussed below, there are also a few important drawbacks to the approach they follow compared to using SMILEs, and therefore I think it is important to discuss these points.

The authors use a clever and original approach to obtain a very large sample of extreme precipitation events to study, thanks to the use of the results of a large ensemble of weather predictions. But there are some unacknowledged limitations. Even if the ensemble of weather predictions is large, it spans a very short time period for climatological studies (2008-2019) and therefore samples a small sample of sea surface temperature (SST) conditions, for example. It means that it does not sample correctly interannual and low-frequency climate variability e.g. ENSO variations, decadal variations in NAO, AMV etc. These modes of variability or others may impact precipitation extremes. For example, what if the link between atmospheric circulation and extreme precipitation events is different during El Nino and La Nina events? Also, the 2008-2019 period is strongly impacted by anthropogenic forcings. This issue, and how it may impact the results of the study, should be discussed.

More generally (as noted in the specific comments), the authors should sometimes better describe their analyses, better explain why they do them, how they reach their conclusions based on these analyses, and discuss their limitations. A part in the conclusion section should be dedicated to the discussion of the limitations of the analyses.
The physical analysis of extreme precipitation events could have been more developed. For example, the cyclone tracking algorithm is only based on SLP, which might not allow to capture the potentially complex atmospheric circulations associated with extreme precipitation. Also, the authors do not look at atmospheric stability, convective precursors etc., which are likely to play a very important role regarding these events.

Other comments

Section 2.1

Very little is said on the weather prediction system. There are almost no references on the model, assimilation system, on the skill of the prediction system etc. Please add some information and references.

I suppose that the model has evolved during the period studied by the authors, with also changes in the assimilation system and observation networks etc. Am I right? We really need to have information on the evolutions of the system during the period studied by the authors.

Given these evolutions, there could be potential issues with the temporal homogeneity in the dataset studied by the authors and the analysis in Fig S1 is far from sufficient to show that it is not the case. At the very least, a discussion is needed on this point.
The authors cite Breivik (2013) to support the hypothesis that precipitation on day 10 of forecasts is independent. But this was with a previous version of the weather prediction system, I suppose. The skill of weather prediction systems increases with time and maybe it is not the case anymore?

Also, what the authors say, i.e. that precipitation from the different members on day 10 is independent implies that there is no predictability of precipitation at 10 days. Is it true? This should be discussed. Is it consistent with what we know about the skill of the weather prediction system? And even if it is true for precipitation, I’m quite sure it is not true for atmospheric circulation and that there is skill at day 10 for SLP, geopotential etc. So maybe precipitation itself from the different members is “independent” at day 10, but it is not the case for the associated atmospheric circulation, which is studied by the authors. What are the implications? I think it may be problematic, for example, to assess the equivalent sample size for the composites of large-scale circulation leading to extreme precipitation events. How is it done? In any cases, this general issue should be discussed.

Figure 2a and b, and near line 205.

It is not totally clear to me how exactly the correlations and auto-correlations are computed. Are they computed at each point and then averaged on the catchment, or is precipitation spatially averaged before computing the correlations and auto-correlations? Is the annual cycle removed before computing the correlations and auto-correlations?

Also, are the correlations and auto-correlations calculated on the complete precipitation series or on the series with only the 10th forecast day? Based on section 2.1, I assume that it is the second possibility, but it is not so clear in section 3.2. Also, for each day there are two values, corresponding to two initializations, right? How is taken into account in the calculation of daily correlations and auto-correlations?

By the way, are the auto-correlations significant?
Around L223. The reasoning behind the analyses in Figure 2a and 2b is not clear, and how exactly these analyses are linked to the hypothesis that precipitation from different members on the 10th forecast day is independent is not very clear. There are no real conclusions regarding this hypothesis.

For example, the authors write “to put the correlation coefficients between the times series into context and also to evaluate the auto-correlation of precipitation time series obtained from one ensemble member” as only justification to the analyses in Figure 2b, with no explicit connection with the previous hypothesis. And in the end, they don’t conclude on the implications of the results for the hypothesis they want to prove.

L252. “can be considered independent”

What are exactly the criteria to consider them as independent? Could you describe the exact reasoning?

L253. “the data is thus suitable for systematic analysis of very extreme, 100-year precipitation events”. Even if we consider precipitation on the 10th forecast day from the different members as independent, all data still only come from a 12-year period, and only sample 12 years of SST variability. As said in general comments, this is really problematic.

Note also that the period studied is strongly impacted by climate changes.
These points should absolutely be discussed, and the potential impacts on the conclusions of the paper clearly stated.

L259. What is the algorithm used to fit the parameters of the GEV distribution? Maximum likelihood?

L260. How do the authors deal with the non-stationarity due to climate change? Over the period of the interest climate trends are very likely to be strong, and therefore the simple GEV model used by the authors, which makes a stationarity assumption, is likely to be quite inaccurate. Some approaches to take into account climate trends with GEV statistical models exist. Why didn't the authors use such an approach? The limitations of their method and its implications should at least should be discussed.

L270-271.

Are all the daily precipitation events greater than the 100-year return level really used for the composite analysis (as it could be understood from the text) or only the events corresponding to block-maxima are used? I.e. if two consecutive days, or days in the same week (or month or semester) are above the 100-year return level, are they all used in the composite analyses? It does not really impact the composites, but it impacts their statistical significance, as it impacts the effective sample size.
The cyclone tracking algorithm is based only on SLP, which is quite basic regarding this kind of algorithm. Is it not problematic to track potentially complex situations leading to extreme precipitation events with such an algorithm? Is SLP not too “smooth” to capture correctly the complex dynamics associated with extreme precipitation events? Could the authors discuss the limitations of such approach or cite studies that show that it is OK to use such tracking algorithm for this kind of events?

Figure 5. It is necessary to add statistical significance in the figure with composites, to demonstrate that the days with extreme precipitation events are really different from the other days. Without significance testing, we don’t really know whether the authors discuss real signals or just statistical noise.

How do the authors know that “most of the individual events develop in a similar way as shown in these composites”? Is this based on a sort of test or analysis (e.g. clustering?) or just by looking at all events? If the second option is right, is it really sufficient?

It is somewhat strange to spend a long paragraph describing these specific events without showing them. They could be shown in SI.
Fig 461. “in an area that is favourable for cyclogenesis”. Why? The authors could cite some papers.

L514. The authors discuss “Rossby wave breaking” at several places in the paper, even in the conclusion, but show no analysis of Rossby wave breaking.

L545. How do the authors explain the differences with Pfahl and Wernli (2012)?