Crespi and Enigl et al test the suitability of four different meteorological datasets and three different extreme event metrics for the detection of observed geohydrodological impacts, namely floods and mass movements. They study the temporal consistency of extreme precipitation events extracted from the four datasets with the impact events, and further try to show spatial consistency as well. The analysis is done for a transboundary domain in the European Alps.

I think that the question on which datasets can be used for these types of events as well as which extreme precipitation metric can detect these events is relevant.

I think one limitation in the analysis is that floods and mass movement events are pooled together in the analysis. While I understand that both events can be triggered by extreme precipitation, I do think that it would be worth while also analyzing the dependence of your results on both of these events independently. The reason for this is, that the low "hit rate" of maximum 50% could be related to that only one type of event is well represented, and the other is not. I think there would be merit in giving recommendations on whether the dataset and methods work for floods or mass movements, or both or both equally good/bad.

The methods section generally needs more clarification of the individual steps and should also include some justification of the method choices. Some of the contents should be moved to the results section.

Generally, it is not always easy to follow, and I would recommend working a bit more on the flow of the text and potentially simplifying the sentences. You partly use very long sentences with a lot of information. Further, I think in the results part some sections are quite long. I would suggest splitting these into multiple parts. This will improve the flow of the text and will make the results more accessible to the readers.

Lastly, I think the conclusions can be condensed quite a bit and suggest to focus on the main results of your analyze.

Abstract:

While the abstract is generally well written, the sentences are very long containing a lot of information. Try to rework the sentences in a way that you maintain good flow but reduce the sentence length to be more accessible to non-experts. I would maybe mention which four meteorological datasets you used, but at least mention the different types, e.g. gridded observations, reanalysis, radar based.

Introduction:

P1 L 35: "hydrometeorological events" -> can you mention which ones?

P2 L 50-68: Three remarks: 1) I think the first part on the definition of extremes is quite generic. You have two very clear impacts that you are trying to link to precipitation extremes. Therefore, I would suggest that you try to be more specific and focus on what event definitions have been used in the context of floods and mass movements. Make this first part its own paragraph. 2) I would merge second half of the paragraph (starting in L 58 "Moreover, …") with the next paragraph. 3) "Seneviratne et al 2021" is not included in the reference list.

P2 L 77: Wood et al (2024) -> replace with published version

P 2 L81f: Be more specific here. Define clear research questions and briefly describe how you will answer these questions (e.g. To answer these questions we compare four datasets of different complexity and three different extreme event metrics ....)

Material and Methods:

Study area: Why did you limit your analysis to this study region? From the data description it sounds like you could have extended your analysis also to other Austrian regions. Maybe just include one short sentence on this.

P 4 L 130-32: “The temporal dimension ...” -> How was this achieved? Was the TST dataset available in higher temporal resolution and then daily totals have been re-calculated?

INCA: Mention in the description that only very few stations were included in South Tyrol and that there are no radars on the Italian side.

P5 L163: “primary advantage” of CERRA-Land -> I would mention here that CERRA-Land is also one of the only reanalyses that assimilate precipitation gauge data. Except for ERA5 over the US.

P5 L 169: “lapse rate correction” -> I think that the lapse-rates correction was only applied to temperature and not to precipitation. Precipitation is only linearly interpolated to the higher resolution.

P6 L181f: This and the following paragraphs are part of the results section. You could add a new first result section titled "Comparison of general extreme precipitation statistics". Further, since your study is mainly motivated by extreme precipitation, wouldn't it be more meaningful to show some metric that represents extreme precipitation in Figures 2 and 3? In Figure 2, for example monthly 99th percentile or monthly max precipitation. You can place the current figure on mean seasonality in the supplement and briefly mention it in one sentence. The same would then apply to Figure 3.

P6 L 190: “SPARTACUS-TST and INCA are comparable” -> I would maybe mention that INCA is considerably and systematically lower than SPARTACTUS-TST over South Tyrol, which is likely due to the limited number of stations used and no radars. Also west of Lienz, the precipitation is lower in INCA. I would suspect that the radars are blocked in the north and east by topography.

P 7 L202: “processes are closely linked to extreme precipitation events” -> For floods this is quite obvious, but could you include some studies that show the connection between extreme precipitation and mass movements.

Figure 4 & 5: I would suggest merging both figures and slightly change the contents. Have Fig 5 as new panel a, and then figures 4 below as panels b and c. Would it be possible to have the Yearly and monthly distributions as stacked bar plots consisting of the two types of hazards (floods and mass movements).? Also adapt the new figure caption giving a descriptive title first before describing the contents of the panels.

Methodology:

This section is currently difficult to follow, and the methods described need a bit more clarification and a justification of the methods choices. It is very important that everyone understands the event definition and selection.

P10 L 284: “the available precipitation data...” -> Why this assumption? With most of your datasets you could test this assumption.

P11 L 296: “The ranked values …” -> I would remove this sentence since you are not doing this.

P11 L 304: “is expected to capture …” -> Doesn't this bare the risk of only sampling events that are located in the "high precipitation" areas?

P11 L314: “all wet-day values” -> Have wet days also been excluded from the other two methods? Or is it uncommon that the entire region has zero precipitation?

P11 L317: “The product of the two…” -> Please be more explicit how you combined these two metrics.

P11 L322-23: “To ensure that … is retained.” -> Does it make a difference whether you do the event filtering (i.e. clustering events within a 5-day window) before or after the ranking? When you remove the clustered dates from the ranked list, do you adjust the rank of the remaining list? Do you remove these dates entirely from all analysis?

P11 L 324: “top 5% of sorted dates” -> Why the top 5%? And this 5% applies to the ranked list where dates belonging to the same event "within 5-days" have been removed?

P12 L 344-45: “A lower threshold …” -> Did you also consider days with single hazards but clustered in time and space. Meaning that hazards from the same storm (say 3-days long) may trigger single hazards on each of these days in close spatial proximity. At the moment these hazard events would not be accounted for even though you might account for the 3-day precipitation event.

P12 L 356f: “Since …” -> Create a new subsection. This will break up the methods description and it is easier to follow. This section would only cover the "spatial coherence" analysis. The previous one "hit and miss".

P12 L 365: “To achieve this …” -> On which basis are these percentile classes calculated? based on "all days" in the period 2003-2020, based on "wet days" in the period 2003-2020, or based on "extracted precipitation events" only?

P12 L366: “Each hazard record ….” -> Is the "four nearest grid cells" applied irrespective of the spatial resolution? Isn't this likely penalizing the higher resolution? While the 1km grids suggest a higher accuracy, these datasets still have an effective resolution of 10-15km.

Results

Have you also tested your results independently for the two hazard types (floods vs. mass movements)? I am wondering whether we can say something about whether the hit rate for floods is better than for mass movements. Both can be connected to intense precipitation, but I think it would be valuable if you could say floods are detected in x% of the cases and mass movements in y%. I am simply wondering whether the low hit rate (i.e. 50%) is due to the inability to match mass movement events which are very localized events compared to some of the flood events. For these events likely non of the datasets might be suitable.

I think generally I would deemphasize the trend analysis. Detecting trends from the short time period and the shortcomings of the datasets inhibits any trustworthy trends.

Section 3.2.2: I think the link between temporal hits (based on the areal statistics) and the link to the spatial hits can be strengthened. It would be interesting to know how many of the correctly detected hazard days (up to 50%) also show a correct spatial detection. Meaning that within your temporal search window and your spatial window you have an actual precipitation

value which qualifies as an extreme value. You do analyze the connection already, but you could maybe make this a bit more implicit defining a spatial “hit rate”.

P13 L 387: “INCA is characterized by more pronounced increases in all statistics” -> This might be due to the inclusion of new radars in the recent years….

P14 L415: “72% to 88%” -> Are these numbers irrespective of the ranks, meaning that they agree on common unique days but can show inversed ranks? Did you also check the agreement in the rank locations of the events? Did you also quantify the agreement across the three event metrics?

P14 L418-21: “For all methods ….” -> Where can we see this? How large is the overlap between the different datasets?

P16 L463: “correlation precipitation statistics and hazards ….” -> Are these correlations with all hazard days or with only hazard days with at least two reportings? I thought in your methods section you explained that you remove all days with only a single hazard reported.

Figure 8: Could you maybe add the statistics of the three metrics for this event and their ranks for each of the datasets.

P18 L516-18: “Almost all datasets …” -> Mention SPARTACUS-TST as an exception here. This dataset shows almost equal proportion of events in the second highest class (0.7-0.9), which represents more a moderate event intensity. My hypothesis would be that this is connected to the rather strict spatial rule of 4 closest grid cells and the inherent precipitation smoothing between stations. So, I think if you would extent the search radius to the scale of ERA5-Land (approx. 9 km) then the match to the highest precipitation class might be larger.

P19 L543: “more values in the upper tail” -> but also the lower tail. INCA shows a stronger left skewed distribution, which means that in INCA we have quite a few events with very low intensities.

Figure 9: Can you change the line color of the median to black. The contrast is poor for the salmon color (INCA). You can also remove the unfilled outliers from the boxplots (non filled black circles), since these data points are shown in color anyway. Wouldn't it be better to simple show the precipitation intensities for all hazard events? As I understand you plot the distribution based on the p99 event selection with a hit in hazard, meaning that each distribution is based on a varying number of events. Ranging between 1693 to 2504 hazard events. Or am I interpreting this wrong?

Discussion

P21 L580-83: “The increasing trends of precipitation …” -> I would rephrase this sentence and focus on the temporal consistency rather than mentioning trends. The period is too short to really say something about trends and in two out of the four datasets you have limitations in the consistency of the dataset.

P21 L584: “rise in the number of geohydrological hazards…” -> I think you need to mention here again the limitation of hazard record, which is likely strongly affected by reporting bias. I think you mentioned this is the methods section.

P21 L603-04: “Although arbitrary, …” -> Have you considered to also filter by hazards in proximity, but one day apart?

P22 L629-31: “The 5.5km grid and …” -> You can mention here Wood et al 2025 again to show that this is a consistent finding.

P22 L649-51: “The proportion of hazardous dates …” -> It might also suggest that the extreme event metrics are not capturing the essence of these events.
“For instance, ….” -> From at least one of your datasets (INCA) and potentially TST (if it is an hourly datasets) you could test this hypothesis. You could aggregate the hourly data by taking the daily maximum instead of the daily sums. Then do your analysis accordingly and check whether the detection rate is higher or lower.

P23 L663-64: “This finding suggests…” -> However, as I mentioned before we can see for INCA many instances where precipitation in the vicinity of the hazard is very low (left skewed distribution).

Conclusions:

Shorten the conclusion to the most important take-aways that answer your research questions. You mention several details which you didn't really analyze and which are part of the methods and not a result of your comparison.

P23 L680: “mass waste” -> do you mean "mass movements"?

P23 L681: “a significant increase in daily precipitation intensities” -> not really relevant and not a key finding of your study. I would remove this.

P24 L696-98: “The study also showed …” -> Not relevant, can be removed.

P24 L700-05: “In future studies, ….” -> Move this to the discussion section

P24 L708: “meaningful information about …” -> This statement contradicts a bit your finding of only 50% of hazards being detected by the extracted precipitation events.