Detection and characterization of precipitation extremes and geohydrological hazards over a transboundary Alpine area based on different methods and climate datasets

GENERAL COMMENT

Crespi, Enigl et al. study different rainfall datasets and test their potential for predicting geohazards. The test is conducted over a comparatively large are in parts of Italy and Austria. Results include how well the tested datasets and statistical descriptions of rainfall extremes identify storms and recommendations on how the which dataset should be used.

The strength I see in this study is more on the comparison of the datasets than on the testing of statistical thresholds to identify storms. Some of the datasets compared in this study are often used, also in other data-sparser regions of the world, making such a comparison useful. Which rainfall statistic is most powerful in predicting geohazards is a widely studied topic and I don't think the authors do this in much depth in this study. Furthermore, the results and conclusions are not presented in a very accessible way. I therefore mainly recommend streamlining and restructuring to frame the research in the right context and make it more accessible (i.e. higher impact). Nevertheless, I congratulate the authors on the work they've done so far, which I find useful and with practical impacts.

We thank the reviewer for taking the time to read the manuscript and providing useful comments and suggestions. We need to clarify that the objective of our study is not to introduce novel statistics for describing triggering precipitation to use for hazard predictions. Instead, our aim is to systematically evaluate how well, through simple statistics or common definitions and through different datasets, we can identify precipitation events that are associated with the occurrence of hazards. Importantly, this assessment is carried out without imposing any assumptions on the physical or temporal dynamics of the hazard processes themselves, such as the role of antecedent conditions or specific triggering precipitation mechanisms.

Based on the reviewer's comment and the feedback provided by the other reviewer, the main changes applied to the manuscript are:

- *We restructured Data and Methodology sections by moving part of the contents to the Results. In particular, the comparison among the datasets and the overview of collected hazard records are now in the new subsection 3.1 ("Precipitation statistics from meteorological datasets and hazard record overview").*
- *Based on the feedback received from the other reviewer, we performed the hit rate analysis for each hazard category separately (flood and mass movement), although results are used only for discussion purposes and we kept the hit rate based on the full hazard database (floods and mass movements together) as main analysis in the manuscript.*
- *We updated the hazard dataset by integrating the new version of WLV and GERIOS, which slightly increased the number of hazard records in our set. We updated all numbers and results based on the updated version.*

- *We revised Discussion and Conclusions by shortening them and making key messages more prominent and better related to the research questions of the study.*

Specific comments are addressed below.

I list my main comments below and line-by-line comments further down.

The paper cannot be read very fluently, and one often has to guess the intention of the authors with certain paragraphs/figures. For example, the methods around L180 on the rainfall datasets are a mix of methods, results and discussion. Likewise for the hazard catalogues, where trends are calculated and discussed in the methods (~L250). Also the discussion and conclusions could be better structured to better convey the key messages by adding subsections to the 3-page discussion that explicitly address the goals of the paper (testing the methods for extreme rainfall definition, testing different datasets, implications for practitioners).
We thoroughly restructured the manuscript to better separate methods, results, and discussion. Specifically, we moved the rainfall dataset comparison and hazard catalogue overview to the new subsection 3.1 under Results. In this way, the methodological descriptions remain clearly distinguished from results and their interpretation.

In addition, we followed reviewer's suggestion and reorganized Discussion and Conclusions by explicitly address the main objectives of the study. In particular we split Discussion into four main paragraphs reflecting the structure of the analysis workflow (4.1 Temporal patterns of precipitation statistics and hazard records, 4.2 Methodological choices for comparing extreme precipitation events and hazard records, 4.3 Temporal match between extreme precipitation events and hazard occurrences, 4.4 Spatial coherence between extreme precipitation intensities and hazard records) and we shortened the Conclusions by focusing on key messages only.

While I think the detection thresholds calculated from "areal mean", local p99" and "anomaly" I think generally are meaningful statistics to use. But I don't see much reasoning on why exactly these were chosen and there are not many references either in this part. Given that rainfall thresholds for geohazards has been a research topic for a long time, I miss the novelty compared to other studies or even just the justification for using exactly these statistics, while so many other statistics could be computed too (antecedent rainfall, multi-day cum. rainfall, …).
It is important to clarify that the goal of our study is not to develop novel statistical methods for characterizing triggering precipitation for hazard prediction. Rather, our objective is to systematically evaluate how effectively simple statistics or widely used definitions for extreme characterization, applied across different datasets, can identify precipitation events associated with hazard occurrences. Crucially, this evaluation is performed without making any

assumptions about the physical or temporal dynamics of the hazard processes themselves, including the influence of antecedent conditions or specific precipitation-triggering mechanisms.
We chose these three statistics to measure the extremality in different features of rainfall: spatial extent (areal mean), local intensity (local p99) and magnitude, i.e., the combination of spatial extent and level of above-normal intensity (anomaly). We better specified the aim of the study in the Introduction:

"In this framework, the study aims to i) evaluate how metrics for precipitation intensity, not a-priori tailored to a specific hazardous process, enable to capture extreme events with triggering potential for geohydrological hazards over complex topography; ii) assess the suitability of precipitation datasets of different types and spatial resolution to describe extremes; iii) investigate the optimal combinations of metrics and datasets for characterizing extreme precipitation events and their spatio-temporal relation with hazard records. To answer these questions, three metrics measuring different aspects of rainfall extremes are calculated from 1-day precipitation fields of four meteorological datasets over a transboundary Alpine area between Italy and Austria and used to identify precipitation events over 2003-2020. Subsequently, they are compared with a harmonized archive of geohydrological hazard records to quantify the spatio-temporal match between identified events and observed records."

We also provided a motivation for the choice of the three metrics in the Methodology section: "The metrics adopted for event detection are chosen to consider three different aspects of extreme conditions, i.e., the spatial extent of intensities, the local intensity peak, and the combination of anomalies and their spatial extent."

The dataset comparison is conducted by comparing the hit rate, eg in Table 2, at an artificially set threshold of top 5% rainfall events. However, from my experience it is more common and interesting to compare the predictive power of these datasets to separate hazardous from non-hazardous dates at a range of thresholds. For landslide early-warning, it is almost standard to report receiver operating characteristics. You will easily find references on this and the statistics can be calculated from the data you have with the eg the scikit learn library in python (eg ROC-curve).
We initially computed receiver operating characteristic (ROC) metrics as part of our analysis. However, our case differs from the standard ROC application. Because our study focuses exclusively on the top 5% most extreme events—identified by applying the respective methodological approaches to each dataset—the vast majority of days within the study period (2003-2020) are classified as non-events. This leads to a large number of "misses," which biases the ROC curve and limits its interpretability in our context.
A more meaningful ROC-based evaluation would require considering all days in the period together with the full set of reported hazards, which would result in a dataset-independent

analysis. Such an approach, however, is beyond the scope of this study. Importantly, our objective is not to develop or evaluate an early warning system, but to assess the ability of three different statistics to identify extreme days that led to hazards when applied across different datasets.

Specific comments:
L17: can you say more about the three definitions? Abstract readers will want to know the temporal scales of your analysis.
We added the temporal scale, i.e., we specified that we used 1-day precipitation totals in our study, and we explicitly reported the extreme aspect measured by each metric instead of listing metric names.

L21-24: Please specify in the abstract which data products you are testing. Now it only becomes clear that ERA5-Land is bad. But what is good? What do you mean by «high-resolution observation»?
We have updated the abstract to specify which data products are being tested. It now clearly indicates not only that ERA5-Land performs poorly, but also which datasets (INCA and secondarily SPARTACUS-TST) show better performance.

L79-80: also, the cited papers all seem hydro/flood related but not landslides
We revised the Introduction by citing more studies linking precipitation and geohydrological hazards, covering both floods and mass movements (e.g., Peruccacci et al., 2017; Steger et al., 2023; Vaz et al., 2018; Araújo et al., 2022; Banfi and De Michele 2024). We also revised the previous paragraph about extreme definition by reporting more hazard-related studies for both floods and mass movements (e.g., Barton et al., 2022; Meyer et al., 2022).

L113: a short intro to this section and the reasoning on how you chose the datasets would be helpful here. Also, a table with key facts about the different datasets would be very helpful
We added an introductory paragraph in Section 2.2.1 and added a summary table with the main dataset features in the Supplementary Material (Table S1):
"Four climate datasets covering the study area are selected to assess their ability to detect and characterise extreme precipitation events over the transboundary region. The selection aims to evaluate precipitation fields from different types of products i.e., observation-based grids against reanalyses, and across different spatial resolutions. Two regional products are considered as km-scale datasets, one based purely on the interpolation of in situ observations and one incorporating multiple sources including observations and weather radar fields. The state-of-the-art European reanalysis CERRA-Land at 5.5 km and the global reanalysis ERA5-Land at 9 km are chosen to account for two widely used large-scale products and to evaluate to what extent their precipitation fields are comparable with those resolved by regional datasets. Each dataset is described in detail in the following, while key facts of each product

are summarized in Table S1. To enable the comparison, all analyses were based on the congruent period 2003 to 2020, while each product was used in its native spatial resolution."

L181-196: These paragraphs are a mix of methods, results and discussion. Furthermore, I miss the link to your study. Could you say why showing these monthly means is important for your study on extremes?

The monthly means were intended to provide a preliminary description of the dataset features, which in turn might be reflected also in the representation of extremes, e.g., spatial patterns, resolved scales, and seasonality. However, we recognized that showing a different statistic is more appropriate given the objectives of our study. We have restructured the relevant paragraphs and displayed the monthly 99th percentile both in Figure 2 and Figure 3 of the revised manuscript (we reported them below). The previous version of Figure 2 displaying the monthly means is now in the Supplementary Material (Figure S2) and used to support this preliminary dataset comparison.

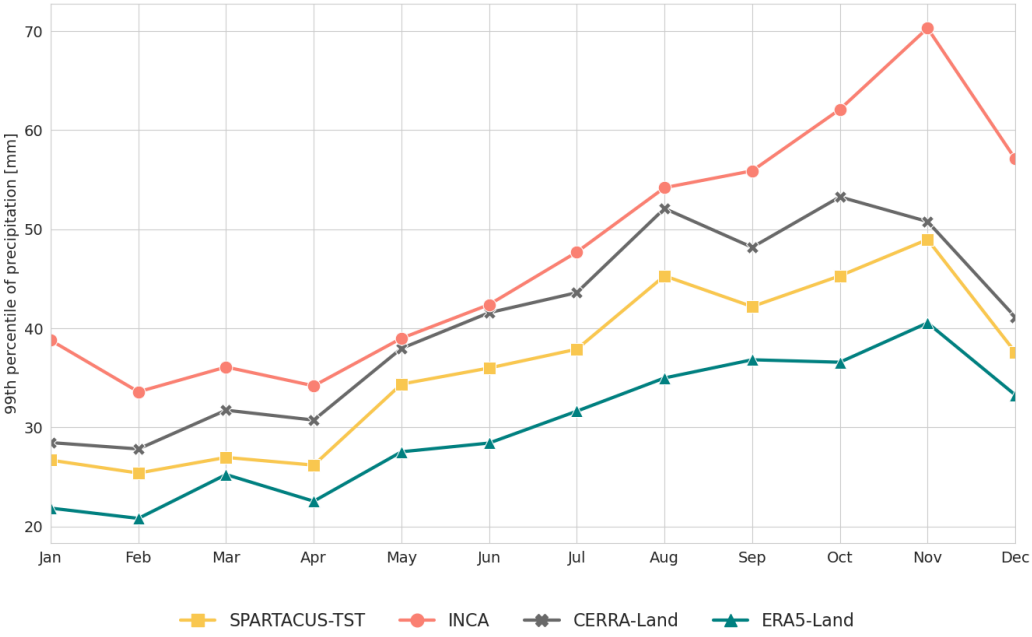

Figure 2: Monthly 99th percentile of daily precipitation calculated over all days in 2003-2020 and all grid points in the study area for the four gridded datasets considered.

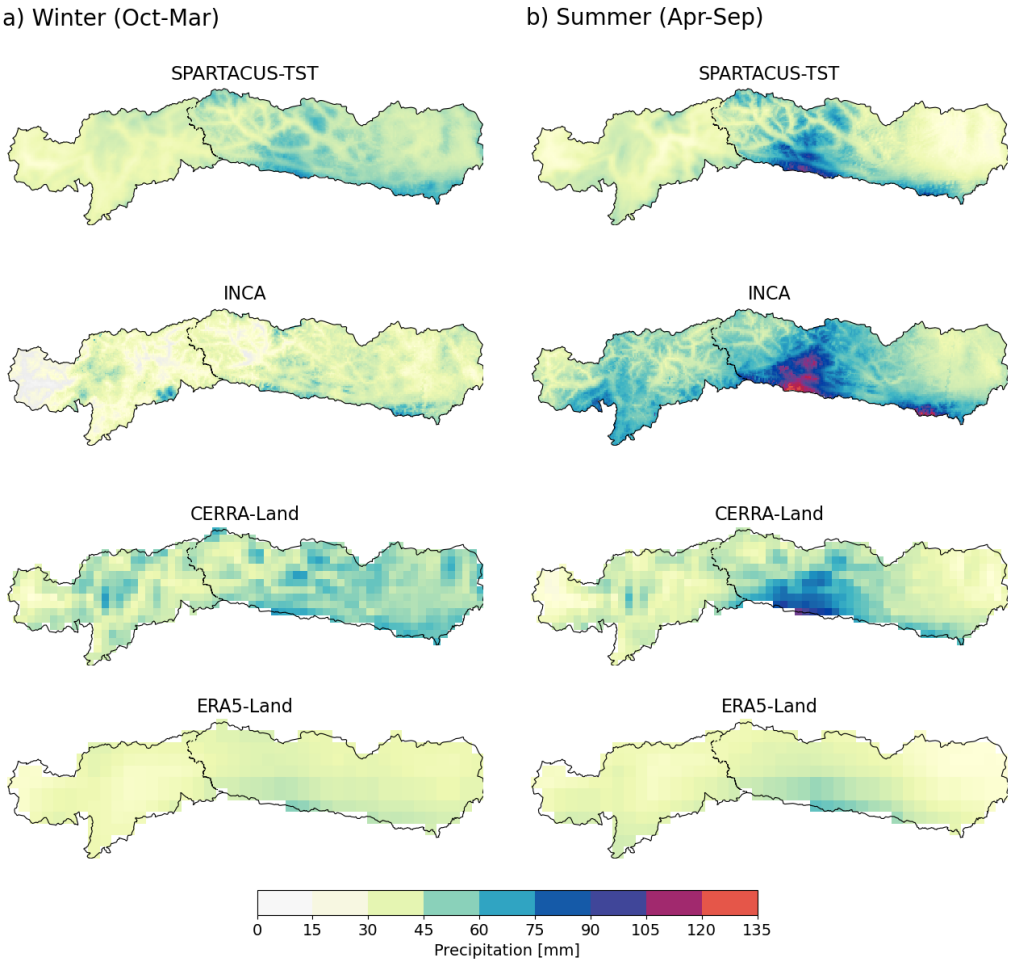

Figure 3: a) Winter half year (October to March) and b) summer half year (April to September) 99th percentile of daily precipitation totals over 2003-2020 in the study area based on SPARTACUS-TST, INCA, CERRA-Land and ERA5-Land. Each dataset is shown in its native spatial resolution.

L201: Please specify how you define “gravitational mass movement” as this is a very broad term. Does it include rock glaciers? deep-seated landslides or only shallow? Debris flows? Rockfall?

Gravitational mass movements comprise the following process types: shallow landslides, debris avalanches, debris and mud flows, and rotational and translational slides. Rockfall events are not included in the analysis. The detailed list of hazard processes considered is reported at the beginning of Section 2.2.2.

L250: these lines again seem like results to me. Unless they were taken from other studies, but then a citation should be enough.

These considerations are based on the preliminary assessment we performed on collected hazards from the different catalogues and their distribution over the analysed period. To make it clear that they are results of our analyses, we moved this part to the Results section in the subsection 3.1, together with the comparison of precipitation fields from different datasets.

L283-287: Do you have evidence from other studies to support these assumptions?
Based also on the comments received from the other reviewer, we revised this part of the methodology. We kept only the assumption related to the use of 1-day precipitation fields and removed the other points in this list as they are not essential for the interpretation of our results and might be misleading.

L363: should be "quantile" instead of "percentile"
We corrected that.

Table 3: I'm having troubles understanding this table. I think it shows into which quantiles the 330 events fall. So each row should add up to 100%, which it doesn't, probably due to rounding. What are "intersected hazards"? I couldn't find a definition in the text and I don't get why this number differs among datasets.
For each one of the 330 events, we classified the corresponding 1-day precipitation field into classes based on the quantile ranges of daily precipitation on that date over the domain. The quantile classes are thus relative to the spatial precipitation field of the extreme event. Then for each hazard recorded within the 5-day window of the precipitation event we extracted the precipitation class corresponding with the hazard location. By repeating this analysis over all 330 events, we got the total number of hazard records located in each class and we converted it into a percentage based on the total number of hazard records within all 330 events (last column in Table 3). The fact that the rows did not add up to 100% was a matter of rounding, we reported them with one decimal place in the revised version of the Table (see below). Please note that numbers in Table 3 have been updated based on the updated hazard dataset and the revised method for class assignment.
In the previous version we searched for the maximum precipitation class over the four nearest cells to the hazard record, which corresponds to a different searching radius depending on the resolution of the precipitation product and might penalize the 1-km datasets. We rerun the analysis for all datasets by assigning to the hazard record the maximum precipitation class in a radius of 10 km. The 10-km radius, which is consistent with the coarsest grid of ERA5-Land and the effective resolution expected for the high-resolution datasets, allows for a more robust search as it implies a different number of surrounding cells defined by the grid spacing of each product. The results show more clearly that a higher portion of hazard records (more than 60 %) fall in the highest precipitation class for the events detected and described by the 1-km products, especially for INCA.

		Quantile range						
		[0-0.1)	[0.1-0.3)	[0.3-0.5)	[0.5-0.7)	[0.7-0.9)	[0.9-1]	Total
Areal mean	SPARTACUS-TST	0.1%	3.8%	5.8%	8.4%	19.5%	62.4%	2,364
	INCA	0.3%	2.6%	4.2%	8.5%	17.6%	66.7%	**2,390**
	CERRA-Land	0.3%	3.1%	6.1%	12.3%	23.5%	54.7%	2,286
	ERA5-Land	2.3%	5.9%	11.2%	23.5%	21.8%	35.3%	2,239
Local p99	SPARTACUS-TST	0.2%	3.1%	4.7%	8.4%	18.1%	65.5%	2,521
	INCA	0.6%	2.1%	4.0%	7.2%	15.5%	70.5%	**2,692**
	CERRA-Land	1.7%	2.7%	5.3%	10.9%	31.7%	47.6%	2,688

	ERA5-Land	1.8%	6.7%	12.8%	22.4%	23.3%	32.9%	2,325
Anomaly	SPARTACUS-TST	0.2%	2.8%	5.3%	7.4%	18.9%	65.4%	2,462
	INCA	0.7%	2.1%	3.8%	8.0%	16.4%	69.1%	**2,460**
	CERRA-Land	1.7%	3.1%	6.1%	12.9%	25.1%	51.0%	2,381
	ERA5-Land	2.1%	6.0%	11.1%	24.2%	21.3%	35.4%	2,176

Table 3: Distribution over different precipitation classes of hazards recorded in a 5-day window of the top 330 (5 %) events identified for each dataset-method combination. Precipitation classes are defined as quantile ranges of the gridded precipitation values over the study area. Values are reported as percentage of the total hazard records included in the 5-day windows of the top 5 % precipitation events (in the last column). For each method, the dataset reporting the highest total number of hazards included in the top 5 % precipitation events is in bold.

L695: can you provide an example for an application requiring “accurate description of precip fields”?

We added some examples in the text (hydrological modelling, early warning systems for floods and water-resource-related applications).