Review of EGUsphere-2024-729

We thank the reviewer for their valuable input, suggestions, and recommendations. We describe below in blue font how we implemented all the points that were raised.

This is a review of "A comprehensive verification of the weather radar-based hail metrics POH and MESHS and a recalibration of POH using dense crowdsourced observations from Switzerland" by Kopp et al. The study presents some useful new findings on hail detection from radar, taking advantage of Switzerland's world-leading observational network for both radar and crowdsourced reports of hail. The important findings include verifications of the often used metrics probability of hail (POH) and maximum expected severe hailstone size (MESHS). A report clustering method that uses only report data and can be used instead of radar-based filtering shows promising results for quality control of dense hail report data. MESHS shows high false alarm rates. An analysis of POH shows that the original version of POH does not correspond with a probability in Switzerland, and the authors show a new calibration of POH.

This is an excellent paper. The results are sound and will be of use to the community who use MESHS, POH and related radar products to infer hail occurrence and/or hail size at the ground. These radar-based metrics are empirical and without well-defined uncertainties. This work, which helps to quantify uncertainties on these measurements in the European domain, is therefore valuable. The paper is highly relevant to AMT. I suggest that only minor changes and some further explanation of thresholds is required before the paper will be ready for publication.

General comments

1. Threshold choices: there are thresholds used in the paper (e.g. "populated areas" are those with a population of at least 100 people per km2, reports from a user submitting more than four reports per day are removed, five reports are required for a cluster to be defined, etc) that are given without explanation or justification. I understand the difficulty in choosing a reasonable threshold but it would be helpful to include a discussion of how these thresholds were chosen and whether or not results in the article are sensitive to these choices. I note that the authors have done a great job of justifying and testing other thresholds (such as for EpsD, EpsT, and POH and MESHS thresholds in the verification, for example), so what is required for these other more minor thresholds is simply some reasoning for why the specified values were chosen.

We now explain in more detail the choice of our thresholds in the manuscript. The justifications for the four reports per day and the five reports to form a cluster are given as answers to specific questions from RC2.

Regarding the population threshold of 100 inhabitants per square km, it assumes that if a person can see up to 100 m around, then a hundred people can cover a square km, provided that they are equally distributed in space. Considering a higher threshold would strongly reduce the area and make it too discontinuous.

Section 2.3 has been modified to better explain the choice of the Swiss100 and ZRH regions.

Specific comments

1. Line 33: “because of the nature of hail” - the authors could explicitly name here what the difficult nature is (hailstorms are relatively small and rare, essentially).

We made the sentence more explicit as suggested: “Ground-based observations are challenging to gather **because hailstorms are scarce and most of them have a small spatial extension (see, e.g.: Brimelow, 2018)**

2. Line 48: Regarding advection of falling hailstones, a relevant reference may be Ackermann et al 2024 (DOI: 10.5194/amt-17-407-2024).

We thank the referee for making us aware of this recent reference. Following the comments of the second referee, we present a more detailed review of the wind drift effect in the introduction, which includes the findings of Ackermann et al. (2024):

L48: Looking at 12 severe hailstorms that have occurred in Switzerland, Hohl et al. (2002) found that the best correlation between radar-derived hail kinetic energy and hail damage claims was achieved for wind drifts of 3 to 4.3 km. Analyzing data from seven severe storms in Australia's metropolitan Sydney and Brisbane areas, Schuster et al. (2006) found a wind drift ranging from 2 to 2.8 km. More recently, Ackermann et al. (2024) applied a virtual advection algorithm to 30 hail events that happened between 2010 and 2022 in Australia and found that most events had an estimated wind drift of less than 2 km and that none of them had a wind drift above 4 km. Such values are lower than most distance buffers used in the literature to evaluate the performance of POH.

3. Line 104-105: It would be helpful to mention whether the COSMO runs used here were done for this study or whether operational MeteoSwiss runs were used.

H0 was retrieved from operational COSMO runs done by MeteoSwiss. We added this information on L104-105.

4. Line 117: The definition of maximum column reflectivity (CZC) is missing the fact that the CZC is defined for a given time period (e.g. one radar set of elevation scans or one scan cycle).

We specified that CZC is defined for a given 5-minute time step (i.e. a full 20-elevation volume scan) on L117.

5. Figure 4 caption: the Natural Earth populated areas dataset requires a reference.

According to the Terms of Use of Natural Earth (https://www.naturalearthdata.com/about/terms-of-use/), only the mention “Made with Natural Earth” is necessary. However, we still added a proper reference to their website.

6. Line 150: I read this as meaning that for a given day and a given point, if the daily maximum radar reflectivity occurred outside 6-22 UTC then the entire day is considered hail free at that point. If this is the case could some valid hailstorms be removed (if for example there was hail during the day but a larger reflectivity was recorded overnight). The authors should consider stating how often grid cells were removed in this way to show any possible impact on the results.

We thank the referee for pointing this out. In fact, this refers to a previous approach that we changed specifically because it could remove valid hailstorms as mentioned by the referee.

What we do in the present study is simply compute the daily maximum of the radar metric between 6:00:00 UTC and 21:00:00 UTC.

We removed the following part "we consider only grid cells where the peak of the hail (defined as the daily maximum of the radar metric) occurred between 6:00:00 UTC and 21:00:00 UTC to ensure enough users are awake to make reports. If the daily maximum is not reached in this interval, the corresponding value for that grid cell is set to 0, as if no hail was predicted."

7. Line 161-162 and 167: "hail < 5 mm" - since the WMO classifies hail as being at least 5 mm by definition I suggest changing this to "ice particles < 5 mm in diameter".

We changed as suggested by the referee.

8. Line 262: What the authors refer to here as the POH is also called the success ratio (see e.g. Roebber 2019, DOI: 10.1175/2008WAF2222159.1.)

We added that **1 - FAR is also called the Success Ratio (Roebber, 2009).**

9. Line 273: CHF is used without a definition here and should be replaced by "Swiss franc".

We replaced CHF with Swiss franc throughout the manuscript.

10. Line 281: It is not really explained why the authors focus on the Zurich region and then present the results for all Switzerland in the appendix. The authors should explain this choice.

We agree that it should be better explained and included this paragraph at the beginning of section 3.1:

The results for the Swiss100 and ZRH region are comparable and most of the conclusions are identical for both. Therefore, we discuss in details the results for the ZRH region in this section, while the figures for the Swiss100 region are shown in appendix A. We chose to discuss the ZRH region because we think that it has a higher and smoother population density compared to the Swiss100 region, which likely leads to a better estimate of the False alarm ratio.

11. Line 415: To be picky, "this probability" here refers to a matching distance of 2 km as defined by Equations 2 and 3. I suggest replacing "This probability" with "The probability of observing hail in the neighbourhood of a given point", or similar.

We changed as suggested by the referee.

12. Figures 14, 15 and 16: It was not initially clear to me how the probability on the y axis of these plots was calculated. Reading back I see it is 1-FARprob. The authors could consider making this more clear, for example by including 1-FARprob in the y axis labels.

We changed as suggested by the referee.

Technical corrections

1. In general I think “weather radar-based” should be written “weather-radar-based”.

We changed accordingly.

2. Lines 108-110: The units do not need to be italicised here. This is true elsewhere in the text also, such as on line 119. There is some inconsistency in the writing of units. Units should at all times be non-italicised and with a space between the numeral and the unit.

The typography was checked and all physical units were written in non-italicised format with a space between the numeral and the unit.

3. Some in-line citations still have brackets (e.g. on line 21 and in the caption for Figure 2).

We corrected the citations.

4. Line 180: “HRC” is introduced here without definition (the definition comes a few lines later). Acronyms should be defined when first used.

We defined the acronym when first used at Line 180.

5. Table 2: There is a mix of capitalisations in the column of seasons. Also, “year” is not a season so perhaps the title should be “time period” or similar.

We changed the title of the column to “Time period” and capitalized the period names.

6. Line 384: “restraint” should be “restrained”.

We corrected it accordingly.

7. Figure 14: ’red’ and ’green’ curves appear as yellow and blue on my display (also affects the text around Line 394).

The curve colors are blue (Swiss100) and orange (ZRH). The text was corrected accordingly.