Replies to Reviewers Comments on

Modelling crop hail damage footprints with single-polarization radar: The roles of spatial resolution, hail intensity, and cropland density

Reference:
Portmann, R., Schmid, T., Villiger, L., Bresch, D. N., and Calanca, P.: Modelling crop hail damage footprints with single-polarization radar: The roles of spatial resolution, hail intensity, and cropland density, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2023-2598 , 2023.

We thank the Reviewers for their valuable comments and feedback and appreciate the time they took to read the manuscript. In the following we reply to each of their comments in blue. Text from the *original manuscript is in blue italic*, and *new or rephrased text is in blue italic with yellow background color*.

In response to the reviewers comments we will implement (alongside other minor changes) the following main changes

- Add an additional Figure (new Figure 1) showing a map of the study region and descriptions of the region names appearing in the text.
- More detailed explanations of the radar product, including the physical background
- Add more references regarding the Heidke Skill Score of radar-derived hail products.
- Add some more discussion of the changes in skill as a result of changes in resolution
- Extend Tab. 3 with the frequency bias and additional model setups and discuss the selection more in the text

Reviewer 1 (Rob Warren)

This study uses single-polarisation radar observations and high-resolution insurance data to develop a model relating the Maximum Expected Severe Hail Size (MESHS) diagnostic to crop hail damage. A detailed investigation is performed into the sensitivity of the model performance to data resolution, highlighting important differences between field crops and grapevine, which are linked to their differing spatial distribution. The impact of variations in MESHS threshold and the minimum crop field density used to define exposure are also explored.

I was really impressed with this paper. It is well written and scientifically rigorous, with high-quality figures and appropriate discussion of relevant literature. I particularly appreciated the detailed analysis of how spatial resolution and cropland density impact the different verification metrics. I think this will be an invaluable reference for future studies of hail-related damage to crops and other assets. As such I have no hesitation in recommending it for publication, subject to some minor revisions. There are a couple of issues related to the use of radar data for detecting hail that I would like to see briefly discussed in the paper. These are noted below, together with a few other substantive comments. Beyond this, I have numerous suggests for minor textual and grammatical changes, as well as adjustments to some of the figures. Rather than listing these all out, I have provided an annotated PDF (see attached). I thank the authors for their efforts in putting this work together and encourage them to reach out to me via email should they have any questions regarding my review.

We thank the reviewer for his positive evaluation of the manuscript and marking the many detailed suggestions for textual and grammatical changes in the attached PDF. For some of the PDF comments we added direct replies at the bottom of the replies to Reviewer 1. The remaining small comments that mainly suggest reformulations or different arrangements in Figures will be largely followed as suggested.

Substantive comments:

- Several factors can impact the quality of radar-based hail retrievals. Among these are (1) the reflectivity calibration, (2) the methods used to map observations from spherical polar coordinates to a Cartesian grid, and (3) attenuation and resonance scattering effects. The first of these is discussed in my 2020 paper (Warren et al. 2020), the second has been explored by my colleague Jordan Brook (Brook et al. 2022), while the third is addressed in several previous studies (Battan, 1971; Kaltenboeck and Ryzhkov, 2013; Ryzhkov et al., 2013). I think it would be useful to provide some brief discussion on these topics in section 2.1 of your paper. You should note what procedures (if any) are used in Switzerland to ensure an accurate and spatially consistent radar calibration. If no such procedures exist you should consider what impact this might have on your results. Likewise, you should briefly describe the method used to map data from the five individual radars to your common 1km grid and the potential impacts this may have on your hail retrievals. Given that your radars are all C-band, attenuation and resonance scattering effects could both be significant. The former can be corrected for given dual-polarisation measurements (e.g., Gu et al., 2011), although the corrections may be underestimated attenuation due to hail (Borowska et al., 2011). Again, you should note what (if any) attenuation correction is employed in the Swiss radar network.

  Thanks for this comment. We will state the following additional sentences in the corresponding section:

  The two products used here, the Maximum Expected Severe Hail Size (MESHS), and the Probability of Hail (POH) are single-polarization radar products provided by MeteoSwiss (Betschart and Hering, 2012; Trefalt et al., 2022; Germann et al., 2022) and are computed operationally. The underlying reflectivity data are mapped to a regular 1x1km grid using a pre-calculated projection table relating polar to Cartesian coordinates. Reflectivity is calibrated with multiple independent sources of information, including various types of weather echoes, ground clutter, signals from the sun, and others (Germann et al., 2015). Radar signal attenuation can lead to a bias in individual hail cells, but we expect this bias to be small compared to the known inherent uncertainty of hail detection from radar data, which is due to the indirect estimation of hail size. Note that direct hail size estimation is limited by resonance scattering effects in large hail stones (e.g. Kaltenboeck and Ryzhkov, 2013).

- When you first introduce MESHS (on L43) it is worth adding a footnote clarifying that MESHS is different from MESH, the Maximum Expected Size of Hail (MESH) originally introduced (as MEHS; maximum expected hail size) by Witt et al. (1998). This is important as MESH is much more widely used (at least outside of Switzerland).

  Thanks for the remark. We will clarify this aspect by adding the following text
  The climatology is based on two hail products that are computed operationally: The Maximum Expected Severe Hail Size (MESHS), which is different from the widely used maximum expected

size of hail (MESH, Witt et al. 1998), and the Probability of Hail (POH; Betschart and Hering, 2012).

- On L200-201 you note that my study (Warren et al., 2020) achieved an HSS of ~0.5. It might be worth mentioning a couple of other previous studies that have quantified the skill of radar-based hail detection (e.g., Cintineo et al., 2012; Ortega, 2018; Skripniková and Řezáčová, 2014; Kunz and Kugel, 2015).

  Thanks for the comment. We will add references to some other studies and the corresponding section will be adjusted to

  For comparison, prior studies achieved a HSS of radar-based hail detection around 0.3-0.5 (e.g. Kunz and Kugel, 2015, Ortega, 2018, Warren et al. 2020). Warren et al. (2020) regarded their values around 0.5 as 'moderate skill'.

- When introducing the inflation factor (L260), it would be useful to provide a proper definition of this quantity.

  Agreed. We will define the inflation factor more clearly and also adjust the legend and axis description of Figure A2 to be more aligned with this definition

  The reason for these differences can also be expressed in terms of an areal inflation factor. That is, the area covered by all exposure grid cells at a given spatial resolution divided by the area covered by all exposure grid cells at 1km resolution. By this definition, the inflation factor is 1 at 1 km resolution and increases for coarser resolutions.

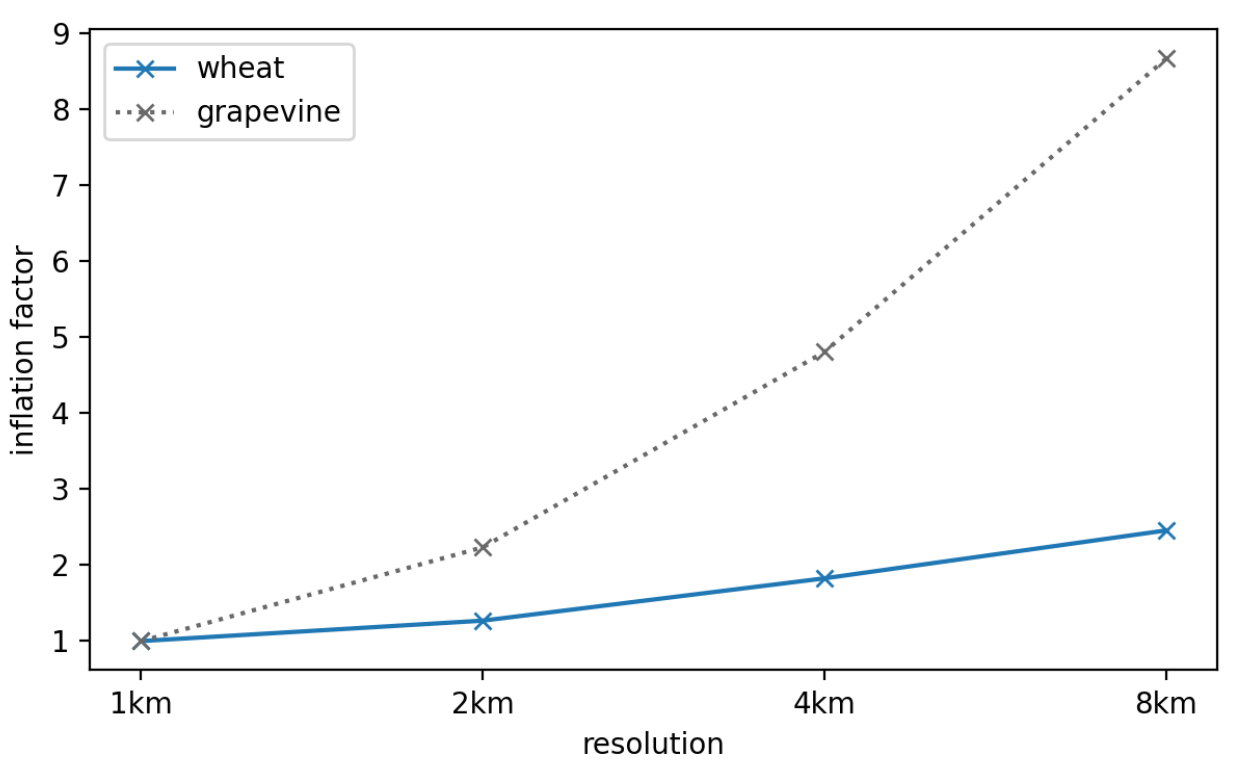

Figure A3: *Total area covered by all exposure grid cells at a given spatial resolution divided by the area covered by all exposure grid cells at 1km resolution (inflation factor) for 1,2,4 and 8 km spatial resolution and for wheat (blue) and grapevine (grey, dashed).*

At the start of section 3.2, you state that CSI and HSS do not exhibit a clear maximum in Fig. 4 (L269-271). I wonder this simply reflects the choice of vertical scale in the figure and the use of 20mm as a lower bound on the MESHS threshold. If you were to plot these metrics separately from POD and FAR, with a maximum y value of 0.4 and a minimum MESHS threshold of zero, I'm sure you would see a more well-defined skill maximum. I was actually curious as to why you didn't consider MESHS values below 20mm. Does this reflect how the metric was defined, perhaps?

Thanks for this comment. Indeed, MESHS is exclusively defined for sizes of 20mm and above (as stated on L98. Note that MESHS represents the largest expected hailstone size within 1 square kilometer; most stones will be smaller than 20mm. If MESHS were to be defined for values below 20mm, the HSS and CSI between 20-40mm might stand out more distinctly as a maximum. This is because we would expect these metrics to have lower values below 20mm. However, it is crucial to emphasize that although the maximum might become more apparent, it would not alter the fact that HSS and CSI remain largely unchanged in the range between 20-40mm, where the majority of events occur. Given these considerations, we have decided not to modify the text at the beginning of section 3.2.

On L273-274 you say "Warren et al. (2020) suggest to additionally constrain the optimal threshold with the condition that B is close to 1 to avoid overforecasting." It is worth noting that Stržinar and Skok (2018) used a similar approach.

Thanks, we will add this reference.

- Regarding your discussion of Fig. 5b on L287-290. I note that there is actually a slight increase in CSI going from 1km to 8km resolution at the 40mm MESHS threshold. This is associated with a more pronounced increase in POD and little change in the FAR. However, for the lower MESHS thresholds of 20 and 30mm, the increase in POD is negated by a commensurate increase in FAR, leading to a reduction in CSI. This difference reflects the fact that the points corresponding to the lower thresholds reside in a part of the phase space where CSI shows little sensitivity to POD but large sensitivity to FAR.

  Thanks for this close observation of the Figure. You are totally right. We will extend the discussion of the figure to:

  Consistent with the results from Sect. 3.1, the main difference is that the FAR is not reduced with reduced resolution but even slightly increased, i.e. the "threshold-resolution web" is squeezed together in the horizontal direction. This results in a tendency for lower skill despite the increases in POD. An exception is the slight increase in CSI for the 40mm threshold from 1 to 8km resolution due to a substantial increase in POD and an almost unchanged FAR. This is because in this region of the phase space, CSI is more sensitive to changes in POD than changes in FAR.

- On L309-301 you say that "the choice of the optimal nthresh heavily depends on the chosen spatial resolution." Can you comment on this a bit more? It might even be worth adding an extra figure to illustrate this sensitivity.

  We will add an extra sentence

  In other words, an n_thresh=20 preserves 99% of field crops at 8km resolution but less than 30% at 1 km resolution.

- Regarding your discussion on using the Peirce Skill Score (PSS) instead of HSS or CSI (L347-353). I wasn't aware of the Ebert and Milne study, so I don't know if they discuss this, but a key problem with the PSS is that it is insensitive to forecast bias for rare events. This can be seen by examining the equation: PSS = H - F, where H = a / (a + c) is the hit rate (POD) and F = b / (b + d) is the false alarm rate (also known as the probability of false detection, POFD). For rare events, d becomes large relative to the other elements of the contingency table so F becomes small relative to H. As such the PSS is dominated by hits and misses (the POD) and becomes insensitive to false alarms. The CSI and HSS don't suffer from this issue, which is why they tend to be favoured for rare events. As such, I am not sure that it is fair to say that use of the PSS represents an "equally valid verification procedure" (L347). At the very least, some discussion around this issue with the PSS should be included here.

Thanks for this comment. As you say, the PSS is insensitive to false alarms for rare events. However, also the HSS is more sensitive to a change in misses (c) than to a change in false alarms (b) for rare events. Ebert and Milne (2022) discuss these points extensively and they arrive at three adequacy constraints for skill scores for rare and severe event forecasts (better than chance, direction of fit, and weighing errors). Out of the discussed skill scores, only PSS fulfilled all three constraints (see

their Table 7 below).
We understand that their analysis eventually comes down to the point that for rare and severe events Type II errors (misses, b) are more important than Type I errors (false alarms, c) and only PSS accounts for that. So the behavior you mention for PSS is actually desirable for rare and severe events according to Milne and Ebert. Also, PSS is the only measure with the – according to the authors – correct "worldview", i.e. that the forecast is evaluated against the reality of occurred weather events rather than vice versa (matching the predictions to the world rather than matching the world to the predictions). They discuss this as the omniscient forecaster vs. the omnipotent forecaster.
Of course, it can be discussed if hail counts as rare and severe (obviously it seems less severe than avalanches, the example of Milne and Ebert) and if these argumentations also hold for our case which is not actually a forecast but an evaluation of radar information with ground truth. This is why we regard PSS as equally valid as HSS for our case, because it eventually depends on the perspective of what property of the forecast to value more over another property.
In our opinion, the "perfect" forecast evaluation metric would have to take into account the actual costs of misses and false alarms and evaluate forecasts according to its forecast value (see Richardson, 2003, https://www.ecmwf.int/sites/default/files/elibrary/2003/11922-predictability-and-economic-value.pdf). But this is of course often not known and very case specific.

Table 7. Summary comparison of skill measures in relation to the three adequacy constraints for RSE forecasting.

	Better than chance	Direction of fit	Weighting errors
Accuracy score	No	no preferred direction	no weights
Heidke skill score	Yes	incoherent	no weights
RIOC	Yes	no preferred direction	no weights
Peirce skill score	Yes	correct direction	correct weighting
Clayton skill score	Yes	incorrect direction	incorrect weighting
Dice coefficient (F_1 score)	Not built in	correct and incorrect direction combined	no weights
F_β measure	Not built in	correct and incorrect direction combined	adjustable weighting

In response to your concern, we will avoid the expression 'equally valid' and mention the main argument of Ebert and Milne (2022). This leads to the following changes:

Finally, it is noted that other verification procedures exist than the ones used in this study. Two alternatives and their effect on our results are briefly discussed. First, Ebert and Milne (2022) suggest the use of the Pierce Skill Score (PSS, Peirce, 1884) as alternative to HSS for rare and severe events. One of their arguments in favor of PSS is that it is the only skill measure taking into account that, for rare and severe events, misses tend to be more problematic than false alarms. For more details on this discussion the reader is referred to Ebert and Milne (2022).

- I'd be curious to know how the specific model configurations listed in Table 3 were selected. Was some minimum HSS threshold applied? Or are these just intended as representative examples?

Thanks for this question. These are representative examples. Regarding n_thresh, the choice has to be pragmatic in the sense that the skill would continue to increase the higher n_thresh, but one continuously loses cropland area (see discussion of Figure 6, L 295-310). Regarding the resolution, 8 km has been established as the best choice for field crops and 1km for grapevine. Remains s, the MESHS threshold. There we saw that it changes only marginally in the range from 20 up to about 40 mm. Of course, one could nonetheless identify the threshold for which HSS is maximized. However, I doubt the meaningfulness of this. Rather, the intention is to illustrate the different options a user of such a model has to tweak its skill metrics. One user may want a model with a high POD and does not care about FAR. Another one wants to have more balanced characteristics. And yet another one wants to have a frequency bias close to 1.

We will add a column for the frequency bias and model setups that have a frequency bias close to 1 (with the clear intention that these can be used for climatological analyses of crop damaging hail) and clarify the intention of the table better in the text

Table 3. Most suitable model setups for field crops (aggregate crop class including wheat, maize, rapeseed, and barley) and grapevine.

crop type	parameters			skill metrics			
	resolution (km)	s (mm)	n_{thresh}	POD	FAR	HSS	B
field crops	8	20	100	0.80	0.48	0.53*	1.54
field crops	8	20	20	0.81	0.54	0.49	1.79
field crops	8	20	1	0.82	0.64	0.41	2.27
field crops	8	30	100	0.67	0.42	0.54*	1.13
field crops	8	30	20	0.68	0.49	0.51	1.33
field crops	8	30	1	0.68	0.60	0.44	1.69
field crops	8	34	100	0.59	0.40	0.52	0.99**
field crops	8	34	20	0.61	0.47	0.49	1.15
field crops	8	40	100	0.48	0.35	0.47	0.74
field crops	8	40	20	0.50	0.43	0.47	0.88
field crops	8	40	1	0.50	0.54	0.42	1.09
grapevine	1	20	10	0.70	0.61	0.48*	1.78
grapevine	1	20	1	0.75	0.79	0.30	3.54
grapevine	1	30	10	0.54	0.56	0.47	1.23
grapevine	1	30	1	0.57	0.76	0.32	2.41
grapevine	1	34	10	0.44	0.55	0.42	0.99 **

* highest skill for this crop type. ** frequency bias closest to 1 for this crop type. Note that n_{thresh}=100 is only a sensible choice for the aggregate crop class field crops due to its high cropland density. For individual crop types, lower values like n_{thresh}=20 are to be preferred.

Finally, the key skill metrics of *selected representative* model setups for the best resolution (8 km for field crops, 1km for grapevine) *are shown* in Table 3. In general, MESHS thresholds of 20 and 30 mm outperform MESHS thresholds of 40 mm in particular for higher n_thresh. *In general, a larger n_thresh will yield results closer to the "true" skill of MESHS, i.e., the skill it would have given a gapless hail detection network on the ground, but comes at the cost of reduced number of data points for verification*. The best performing setups for field crops are achieved at 8 km resolution and reach a POD of about 0.8 combined with a FAR of about 0.5 (for s = 20 mm) or a POD around 0.7 combined with an FAR about 0.4 (for s = 30 mm). For grapevine, the best performance is achieved at 1 km and reaches either POD of around 0.7 and a FAR of 0.6 (s = 20 mm) or POD and FAR of around 0.55 (s = 30 mm). *We note again that the suitable threshold depends on the purpose for which the model is used. For climatological purposes, it is important that the frequency bias is close to 1. While thresholds of 20 - 30mm strongly overpredict damage occurrence (B>1), a threshold of 40 m underpredicts it (B<1). The*

MESHS threshold with B closest to 1 is 34 mm for both field crops at 8km as well as grapevine at 1km and is hence recommended to derive accurate climatological frequencies of crop hail damage occurrence.

- In the conclusions (L391-393), you could also note that your results compare reasonably well with verification of the Severe Hail Index (SHI) and Maximum Expected Size of Hail (MESH) metrics (Witt et al. 1998) in several previous studies (Cintineo et al., 2012; Skripniková and Řezáčová, 2014; Kunz and Kugel, 2015; Warren et al., 2020).

  Thanks for this suggestion and the references. We will add that to the conclusions

  These results are comparable to previous verification efforts of MESHS (Nisi et al., 2016), the original Waldvogel et al. (1979) criterion (Puskeiler et al., 2016), as well as MESH (Cintineo et al., 2012; Skripniková and Řezáčová, 2014; Kunz and Kugel, 2015; Warren et al., 2020), although methodological verification approaches substantially differ from ours.

In the following, we also reply to some comments that the reviewer annotated in the manuscript PDF. With the remaining comments in the manuscript we largely agree and will follow the reviewers suggestions.

- Technically, your study hasn't considered variations in hail intensity but rather variations in the threshold used to identify hail damage. However, I'm not sure how to express this succinctly for the title.
  We agree, that we have considered hail intensity thresholds rather than varying hail intensity. However, similarly, we have considered variations in cropland density threshold, rather than varying cropland density. For brevity, we leave the title as is.

- L253: I don't quite follow this sentence; maybe rephrase.

  We rephrased the sentence to

  Hence, the average FAR at 1 km grid points with 1 field is higher (wheat: 82%, grapevine: 79%) than at gridpoints with 10 fields (wheat: 74%, grapevine: 60%). In conclusion, if cropland density strongly reduces with coarser resolution, FAR will increase accordingly.

- Fig 1: It would be helpful to plot the other verification metrics (POD, FAR, CSI, and FB) as well, to help explain the trends in HSS.

  We agree that this would be helpful but it would also make the Figure less readable. Given the contribution of other metrics is discussed extensively in Fig. 2 and 3 we keep this figure as is.

Reviewer 2 (Tomeu Rigo)

The manuscript introduces the relationship between a radar-based hail size estimating product and its affectation on different types of crops in Switzerland. It is well-written and easy to follow. The main issue of the research is the lack of a physical or scientific explanation of the results, presenting a simple statistical analysis. This point reduces the work potential, which can be notably improved if the Authors consider some elements associated with the MESHS (Maximum Expected Severe Hail Size).

Thanks for the comment. We tried our best to respond to your comments and to add more physical / scientific explanations, keeping in mind that a main goal of the study was to bring together the various data sources to build a working and scientifically sound hail damage footprint model. A main part of this challenge was to explore which configurations are most suitable to bring these data together, which is why the statistical analysis / verification is the core element of the study.

In the following lines, I present my suggestions/questions/doubts regarding the manuscript:

- Is CLIMADA the acronym of something?

  Yes, it stands for CLIMate ADAptation. We will add this explanation.

  To make it accessible to stakeholders, the model is implemented in the open-source natural catastrophe modelling platform CLIMADA (CLIMate ADAptation, Aznar-Siguan and Bresch, 2019)

- L22, 23: The authors should avoid technical reports if scientific references are available (SwissRe, 2021, 2022). Here are some of the multiple possibilities:

Rana, V. S., Sharma, S., Rana, N., Sharma, U., Patiyal, V., Banita, & Prasad, H. (2022). Management of hailstorms under a changing climate in agriculture: a review. Environmental Chemistry Letters, 20(6), 3971-3991.

Gobbo, S., Ghiraldini, A., Dramis, A., Dal Ferro, N., & Morari, F. (2021). Estimation of hail damage using crop models and remote sensing. Remote Sensing, 13(14), 2655.

Bell, J. R., Gebremichael, E., Molthan, A. L., Schultz, L. A., Meyer, F. J., Hain, C. R., ... & Payne, K. C. (2020). Complementing optical remote sensing with synthetic aperture radar observations of hail damage swaths to agricultural crops in the central United States. Journal of Applied Meteorology and Climatology, 59(4), 665-685.

Thanks for these valuable references which will help to root our introduction further in the existing scientific literature. Regarding the reports by SwissRe: We agree that such reports should not be used as basis for scientific studies, however, we consider them valuable because insurance companies are the only ones being able to put a price tag on hail damages on the larger scale. We therefore would

like to keep the SwissRe (2021) reference but remove the SwissRe (2022) reference and use the suggested additional studies you mentioned.

It will read as follows

Hail storms frequently cause severe damage to agriculture and infrastructure in various places across the globe (Bell et al. 2020, Allen et al., 2020, Gobbo et al. 2021, Rana et al. 2022). In fact, severe convective storms (which include hailstorms) are among the costliest perils worldwide (SwissRe, 2021).

- Add a reference regarding MESHS (L43)

  Apologies for the confusion. The Betschart and Hering study is also the reference for MESHS which is not clear from the text. We will add it also behind MESHS and also add Germann et al. 2022.

- What is the physical basis of these products? (L44)

  Thanks for this question. We assume you are referring to the physical meaningfulness of using the height difference between a reflectivity level and the melting level height? We will add an explanatory sentence regarding this point

  Both are based on a height difference between the melting level and the height of a certain reflectivity level, a criterion introduced by Waldvogel et al. (1979). These products are physically meaningful as they relate to the vertical extent of the thunderstorm updraft above the melting level, representing the hail growth zone. The greater the vertical extent is associated with a higher the likelihood of hail as well as an increase in the potential size of hailstones.

- L57: There exists a physical reason for changing the spatial scale linked with the product definition. You need to go deeper into this point.

  Thanks for this remark. We will add a sentence on the physical reasons for changing the spatial resolution

  A strong dependence of the skill on the spatial scale can be expected, as lowering the spatial resolution (or increasing the tolerated distance) increases the likelihood of overlap between radar signals and damages (Holleman et al., 2000; Schmid et al., 2023). The physical reasons for this are the horizontal drift of hail by wind (e.g. Schiesser, 1990) and limitations in the spatial accuracy of radar-based hail observations, which rely on storm-related proxies to infer hail sizes on the ground (Betschart and Hering, 2012).

- L 64: The word "without" appears repeated twice

  Thanks for spotting. We will remove it once.

- Paragraph L 60-68: I found the introduction referring to radar analysis of hail-crop affectation.

Ok. We were not sure if this comment asked for some modification. We interpreted it that it didn't and left this paragraph unchanged.

- L93-96: From these lines, I understand that the authors did not consider the 5 radars for the whole period, isn't it? Please clarify this point.

  No, the events considered in this study occurred all after 2016, so when all 5 radars were already in place.

- L104: But not avoid it. Do you know if the hailstorms occurring during that period are more or less severe than usual?

  Thanks. Yes, of course this can not be avoided. However, for the 12 events that we looked at, there was no splitting of a single storm into two consecutive days. The hailstorms are not more or less severe than similar storms in the observational period (2002-2021) with the clear exception of 28.06.2021, which was an extreme hail event both in terms of spatial extent and local hail intensity.

- L112: However, this is not consistent with the storm size and the hail path. Please, explain better this point.

  Thanks for this remark. If we understand you correctly, you argue that selecting the maximum for the spatial aggregation artificially inflates storm size? We agree that this is the case. However, we found that it was the most meaningful way of aggregating MESHS. The idea is that a MESHS value at several km altitude in a 1x1 km grid cell is not necessarily representative for what happens directly below at the ground. This MESHS value could well be materialized in terms of hail on the ground shifted by several km (due to wind drift). Hence, the maximum MESHS value is taken because that is considered the relevant value for the occurrence of damage on the ground.

- L152: Why 1000 times exactly?

  1000 times was chosen pragmatically because it is large enough to provide a good statistical distribution and small enough to deal with limited computational capacity.

- L158: I assume that you say that two hailstorms occurred over the same place in the same event, is this correct? Clarify.

  Apologies for this inclarity. No, this means that two individual events on different days occurred when some of the crops had already been harvested. They are 08 July 2017 (barley harvested) and 01 August 2017 (wheat, barley, rapeseed harvested), see also Table 1 of the paper.

  We will adjust the text as follows

  Two hail events (08 July 2017 and 01 August 2017) occurred when at least one of the crops had already been completely or to a large extent harvested.

- L168: What "magnitude" is presented in these fields?

  We do not fully understand the question. The gridded damage data represents the number of damaged crop fields (individual farming plots) and the exposure data the total number of crop fields present.

- Section 2.5: Are you considering all the pixels of the full coverage for each event, or only those probably affected by the hailstorm?

  Thanks for this remark. We consider all grid cells with exposure.

  To measure the model skill at different spatial resolutions and for different hail intensity thresholds, we use a 2 x 2 contingency table computed based on the joint distribution of predictions and observations on all grid cells with a non-zero exposures (Table 2, c.f. Wilks, 2019)

- L213: Have you analysed the hailstorm type for each case? Are they similar?

  We didn't attempt to differentiate the selected storms with respect to their characteristics (spatial extent, intensity, etc.) or the environmental conditions in which they were formed (see e.g. Zhou et al., 2021, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2021GL095485), as this was not the main goal of our investigation. This could the topic of a follow-up investigation.

- L227-235: Explain the behaviour of the skill scores depending on when you change the spatial resolution of the radar product.

  Thanks for noting that there is no clear explanation of this aspect in this section. The explanation of this is currently mainly in the discussion (L320 ff). But we agree that it would be beneficial to add some explanations also at the location you mention. We will edit this paragraph as follows:

  The shifted forecast (15 June 2019, Fig. 2d) greatly improves (HSS=0.42) due to a substantially higher POD (0.93) and a lower FAR (0.67). This is mainly because the coarser resolution compensates for the spatial shift of a few km, which turns misses and false alarms into hits. The overforecast (28 June 2021, Fig. 2e) also improves (HSS=0.38) but mostly because of a lower FAR (0.57) while POD remains unchanged. The coarser resolution effectively eliminates the red 'holes' of false alarms that occur between the blue areas of hits. However, its impact on misses is limited, considering they were already minimal at the 1km resolution. Similarly, the more cohesive damage footprint at 8km, in comparison to 1km, contributes to a reduced FAR (0.38) for the good forecast (12 July 2021, Fig. 2f). However, the overall skill remains largely unchanged due to the increased significance of individual scattered damage reports over the Swiss Plateau, resulting in a lower POD (0.55; without these reports, POD would be 0.73). HSS increases for all 10 considered events if spatial resolution is reduced (Fig. 2g). However, there are substantial differences in the magnitude of the increase between events. The FAR reduces with lower resolution for all events and POD increases for 7 out of 10 events. POD does not increase for events where it is already very high (e.g., the overforecast) or where many misses are located far away from the modeled damage footprint (e.g. the good forecast).

- L245: Add a map (preferably at the beginning of section 2) of the study region including the location of the places appearing in the text. (The Authors should consider that the reader can provide from a distant place and needs to be as familiar as possible with the geography of the region when he/she reads your research)

  Thanks for this suggestion. Apologies for the missing geographical context. We will add a map including the names appearing in the text and refer to the map in the corresponding sections of the text as shown below

  In this study, single-polarization radar data products on a 1x1 km regular grid are used to characterize hail hazard. The Swiss radar network consists of 5 dual-polarization Doppler C-band radars (black dots in Fig. 1) and is in place in this form since 2016 (Germann et al., 2016).

  This means that a grid cell is considered as exposure if it contains at least one field of the considered crop type (shown for field crops and grapevine in Fig. 1).

  Many false alarms appear in the regions with low density of grapevine while hits populate the regions with high grapevine density notably at the shores of Lake Geneva and the Three-Lake-Region in Western Switzerland and Lake Zurich in the North-East.

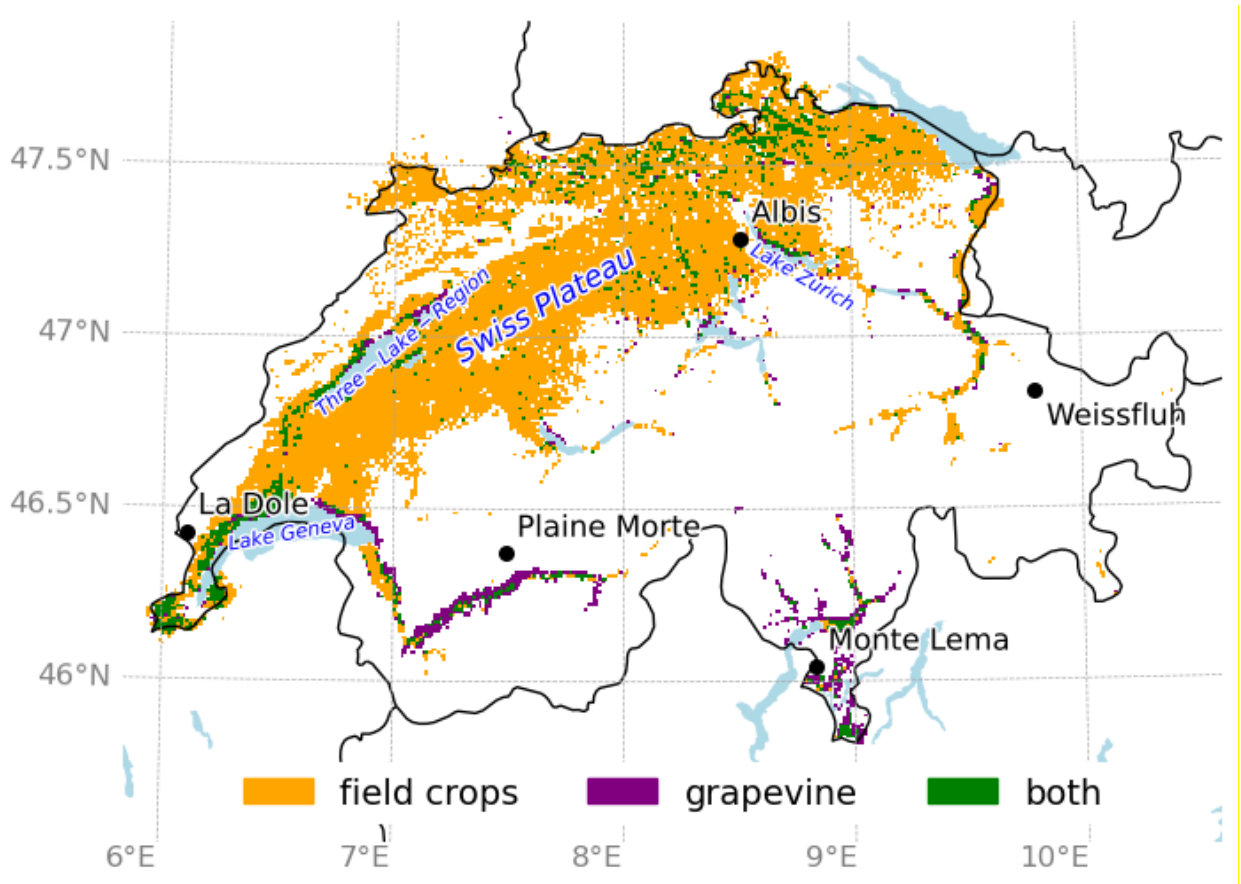

Figure 1: Study region showing exposed field crops (orange), grapevine (purple), and both of the two categories (green), at 1km resolution with at least 1 field per km2, the five radar locations (black dots). Regions / Lakes appearing in the discussion of the results are named in blue.

- L251: I miss more modern references than 1975 ones.

  Thanks. We agree that more modern references are desirable. However, observational studies describing the very local (few 100 m) variability of hail within an individual hail storm are very rare. However, we found another, more recent study (Ortega et al 2009) and will add this reference

  Ortega, K. L., T. M. Smith, K. L. Manross, K. A. Scharfenberg, A. Witt, A. G. Kolodziej, and J. J. Gourley, 2009: THE SEVERE HAZARDS ANALYSIS AND VERIFICATION EXPERIMENT. Bull. Amer. Meteor. Soc., 90, 1519–1530, https://doi.org/10.1175/2009BAMS2815.1.

- Section 3.2: Have you taken into account the fact that if two hailstorms occurred over the same region in a very short time period, the insurance reports would be lower than expected in the second case?

  Thanks for this point. No, we did not explicitly account for this. First, such a case is relatively unlikely (not impossible though!). The insurance reports are usually made through field visits a week or so after the event. So if an event occurs at the same location after the reports have been made, it will probably affect the damage degree reported in the second case. However, since we only look at damage footprints (damage occurrence yes or no) but not measures of damage degree, such a situation would have negligible effects on our results.

- L330: Do you think there are similitudes with studies referring to the density population relationship with hail affectation?

  Thanks for this question. I'm not sure which studies you are referring to but when hail occurrence is evaluated with crowed-sourced hail reports (e.g. Barras et al. 2019) I would suspect the population density plays an important role.

- The caption of figure 3: Simplify it, indicating only the differences with figure 2.

  Thanks, good suggestion. We will simplify it accordingly. The new legend will read as follows (note, Fig. 3 is now Fig. 4 because of the newly introduced Fig. 1.:

  Figure 4: As Fig. 3 but for grapevine and (a,d) 15 June 2019, (b,e) 28 June 2021, and (c,f) 24 July 2021 and g) all 12 events with recorded damages. Grey hatched boxes in panel (g) show events with a modeled damage footprint below 80km2.