Response to comment on amt-2023-89 by Anonymous Referee #2

This study present a significant advancement on the initial drone-based technique developed by Soderholm (2020), in particular the derivation on melting rates to estimate the original HSD and simulation of hail pad measurements. This is really great work! The authors also improve the underlying methodology and explore the parameter tuning in much more details.

We very much appreciate the positive feedback by Referee #2 and all the raised comments, that help to improve the manuscript. All Referee comments are shown in black, the authors responses in green text color.

It was unfortunate to see the very serious findings of plagiarism from R1. Further, I believe the manuscript needs to be reviewed by a scientific English language editor first before resubmission. The corrections to make to the language were too numerous that I did not document them and instead tried focused on the science content. The text also uses many superfluous words and sentences with repetition that could be removed, significantly reducing the length. I'd encourage the authors to take these actions on board and prepare a resubmission.

Indeed, there are several instance in the manuscript where the text directly referred to an available source without correct citation. We admit that this should be clearly avoided in a scientific publication and want to thank the reviewers for pointing this out. Please refer to the answer to comments by Referee #1 on this issue.

Regarding the raised language and readability issue, the authors of the manuscript tried to improve the whole manuscript text. The mark-up version looks quite messy, but we note that content-wise no critical things changed. We further hope, that you understand, that we cannot go into detail and list all language-based changes, rearrangements and reductions of the text here in this reply.

We also want to note, that at a later stage (if the manuscript gets there), AMT provides the typesetting and English language copy-editing as a service, which should give the final linguistic polish of the manuscript.

Comments/corrections:

Line 24: "asses"

This typo has been corrected.

Line 38-39: Crowd sourced data might not even provide the largest diameter, just the size at some unknown percentile

This is certainly true. Due to rearrangements the updated version about this issue is now located in the second paragraph of the introduction and reads:

"…In addition, hail sensors cannot capture the entire hail size distribution (HSD) due to their small observational area of 0.2 m (Kopp et al., 2023). Similarly, crowdsourced hail reports use predefined

categories (no hail, < 10mm, 10mm, 20mm, 30mm, 50mm and > 70mm) for estimating the hail size, corresponding to an unknown percentile of the actual HSD."

Line 59: The short sentence about chasing needs some more context - who and why?

Due to thorough language changes, text reformulations and reductions, this sentence disappeared. We also do not use anymore the word "chasing" or "chase" in the text. It only remains in the title of Section 2.1.

Line 59: How was the supercell track generated in figure 2a?

The cell track of Figure 2a is based on the TRT (Thunderstorm Radar Tracking) method by Hering et al. (2004). The method is also described in Feldmann et al., 2023. We added the corresponding citations here (lines 52-53):

"The track of the supercell shown in Fig. 2(a) and was generated based on the TRT (Thunderstorm Radar Tracking) algorithm (Feldmann et al., 2023; Hering et al., 2004)."

Line 62: Please add in the values for storm motion to provide some context for comparing against other events.

Thanks for the comment, this is indeed useful information to add. We added now the values for the Bunkers Right storm motion vector and the SFC-6km mean storm relative winds:

"From the hodograph shown in Fig. 1 a storm motion vector of 234° at 13 m s^{-1} (according to Bunkers et al. (2000)) and mean storm relative winds (0–6km) of 71° at 9 m s^{-1} can be derived."

Paragraph ending line 75: I think some work is needed to improve the flow between this paragraph and the next perhaps by moving the limitations of hail impact sensors here as motivation for aerial surveys

We totally agree, that here the readability was not very smooth between these paragraphs. The structure in the introduction also changed a lot. Now the motivation for the new technique introduced by Soderholm et al. 2020, follows the paragraph about the limitations of hail impact sensors and crowdsourced reports. We continue then in line 40 with:

"In order to overcome some of the limitations of automatic hail sensors and crowdsourced reports for estimating the HSD, a new technique, called HailPixel, has been introduced by Soderholm et al. , 2020. They propose …"

Sentence starting line 78: Some duplication in this sentence around "melting", which is mentioned twice

Here we now deleted the term "… to prevent further melting", as it is stated already in earlier within this sentence.

Line 81: Please keep units consistent, either mm or cm

Thanks you for this comment. We now always use mm for the hail dimensions/sizes, also for the mentioned MESHS and crowdsourced values.

Line 91: The use of "chain" could be improved with "methodology chain"

Instead of “end-to-end chain” we now write in line78: “In Sect. 2 the methodology is presented, …”.

Start of section 2: I find it's often clearer for the reader to be more direct, e.g., "Here we first go into the challenging part" to "First we discuss the"

With respect to a Referee #1 comment, the first 2 sentences at the start of Section 2 were removed. Further, the start of Section 2 now reads different due to rearrangements and language editing of the text.

Lines 119-128: I don't think this paragraph is necessary to support this paper. The methodology of storm chasing is quite specific to the region and the individual.

We agree, that not all information given in this paragraph is needed. Therefore it has been shortened and revised, also with respect to a Referee #1 comment. However, it could be worth sharing of some of the methodological aspects.

Lines 142-147: I feel this is more a reflection on the storm chasing approach that is specific to the authors experiences. It's not necessary to support this paper.

Thanks, and yes we agree, that this paragraph can be removed. It is removed now.

Line 150: Earlier it was started that a 50 MP full frame camera was used. Also MP needs to be expanded.

We checked the specifications and instances in the text. The camera has 45 megapixel and this has been corrected in the abstract. Instead of MP we write now always megapixel.

Line 180: That's an extremely high ISO! Were there any issues with noise or over exposure?

A high ISO value was needed to keep the shutter speed fast and reduce motion blur also in darker scenes during/after the hail storm passed. To keep the flow smooth for the different flights with less camera setting changes, we directly started with such a high ISO. Of course the noise is high but still acceptable for the purpose. For sure many cloudy (whitish) hailstones were overexposed with these settings, but it does not impact the detection of the hail.

Line 183: I'm not sure what the convention is for AMT with notation for number (either comma for full stop). Might be worth checking. Also this needs to be made consistent throughout the text as there's many numbers written within any separator.

During typesetting this will be corrected by the journal to their standards. We had a look through the manuscript so that we are at least consistent in the manuscript for now.

Line 203: ResNet should be expanded in the previous paragraph where it is first introduced

Thanks for the hint, we now expand ResNet (Residual Neural Network) in the previous paragraph: "…uses a Residual Neural Network (ResNet) detection engine described in He et al. (2016)."

Line 221: The sentence starting "The idea behind..." can be merged with the next sentence to make the text less verbose

Some modifications were done to this paragraph. Now this part of the text reads:

"We use 10% randomly selected tiles as reference data (216 tiles). This reference data is further divided into 70% for training (150 tiles) and 15% for the validation (33 tiles) and test data (33 tiles) respectively."

Line 231: "tow"

Corrected to "two".

Line 244: It might be worth explaining to the reader how the term epoch is used for deep learning

For better readability and understanding quite some adjustments have been performed in Section 2.3.2. The epoch time is now better explained within the context:

"Using two images per batch on 1 GPU, a total of 75 batches are needed which represents one epoch time, i.e. to iterate through all available image tiles."

Lines 259-260: I feel this statement about the methodology is already covered in 2.3.1 - "Because we want the test data set to be locked down until we are confident enough about our trained model, we do another division and split a validation set out of the train set. In this scenario we end up with three data sets." Further, "train" should be "training"

The mentioned statement is somehow obsolete and anyway removed/changed now. Where applicable, "train" is substituted with "training" in the right context.

Line 268-269: How extensive was this manual QC to remove non-hail objects? It might have been worthwhile including some tiles with these uncommon non-hail objects in the training.

We agree, that for future model trainings this could be an opportunity to better fine tune the model to distinguish hail from other objects in the image. In our case here it was not very extensive to remove those objects, as there were less than 10. Probably also too few to reliably train on these objects.

Paragraph 270: I think this could be shortened significantly by outlining the parameter space of F1 with a reference.

We thank the referee for his advice to reduce the length of the paragraph following line 207ff.

After major rearrangements and text editing the new paragraph with added references reads now:

“In machine learning, precision and recall (Eq. (1) and Eq. (2)) are commonly used (Powers, 2020). Precision depicts the number of true positive results divided by the total number of positive results. Recall refers to all true positive results divided by the number of all samples that should have been classified (i.e. as visually identified by the experts in the test data set in our case). Precision and recall can be combined in the F1 score in Eq. (3) (Van Rijsbergen, 1979; Goutte and Gaussier, 2005). The F1 score results in values from 0 to 1 where 0 indicates extremely poor performance and 1 refers to a perfect performance of the model.”

line 287: "quadruple variation of the learning rate" could be improved with "four different learning rate values tested"

Yes, your suggested reformulation sounds better and is included now:

“The appearance of four groups in the two plots of Fig. 7 is due to the four different learning rate values tested (Table 1).”

Lines 310: "Right tail" should be "Upper tail". I'd also be more specific than saying "smaller devices"

We are now more specific and mention the hail sensor data instead of small devices and use the “upper tail” expression instead of “right tail” throughout the manuscript.

Lines 312: "logarithmic view" is not needed in the main text, this belongs in the figure caption.

Correct not really needed here and thus removed.

Line 313: check use of 'maximal' I think maximum is more suitable.

Thanks for this hint. We checked the whole manuscript for “maximal” and replaced it with “maximum”.

line 343: I'm unclear what the author is trying to assert with "We note that the HSD is considered at the scale of a single hail cell."

We agree with the referee that this sentence is confusing. We have rewritten the introduction of this Section 3.2:

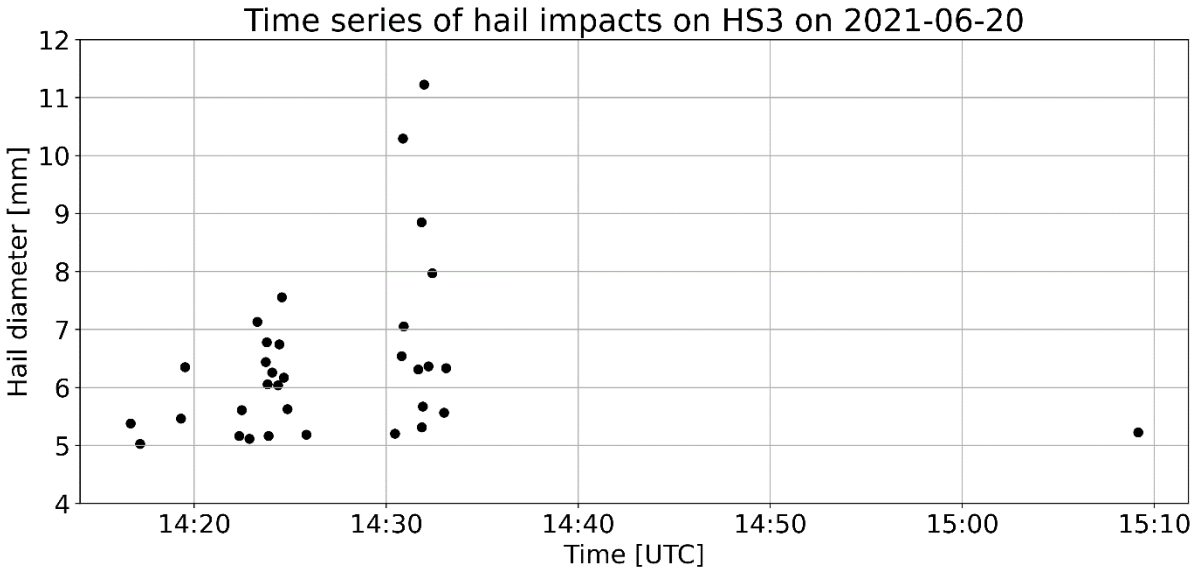
“In this section, we determine the probability of impacts of a given hail diameter on randomly placed hail sensors. 10000 virtual hail sensors with a size of 0.2m2 were distributed across an area of 600m2 within the orthophoto (blue circles in Fig. 5(d)). For each virtual sensor, the HSD was derived and the individual Kernel density estimates (KDE, gray lines) are shown in Fig. 12(a). The KDE was obtained from 7817 virtual sensor areas. The remaining 2183 sensors did not have enough virtual impacts to estimate the KDE. The distribution from the entire 600 m^2 area is shown in black, and the respective quantiles (Q25, Q50 and Q75) from all the virtual sensors in blue, red and green.”

line 354-355: repeated from the start of the paragraph. This is some really nice work too!

Thank you very much! This part was already adapted with respect to a Referee #1 comment.

line 363-364: Can you use the time series information from these disdrometers to separate the two hail events for HS3?

After looking at the time series of impacts for HS3 below (no disdrometer data), the single impact near 15:10 was discarded as it happened 40 minutes after the event and could result from a non-hail object.



All others impacts were considered to be part of the same event as they were separated by less than 5 minutes each, strongly indicating that they belonged to the same cell (Ref: How observations from automatic hail sensors in Switzerland shed light on local hailfall duration and compare with hailpad measurements, Kopp et al. 2023).

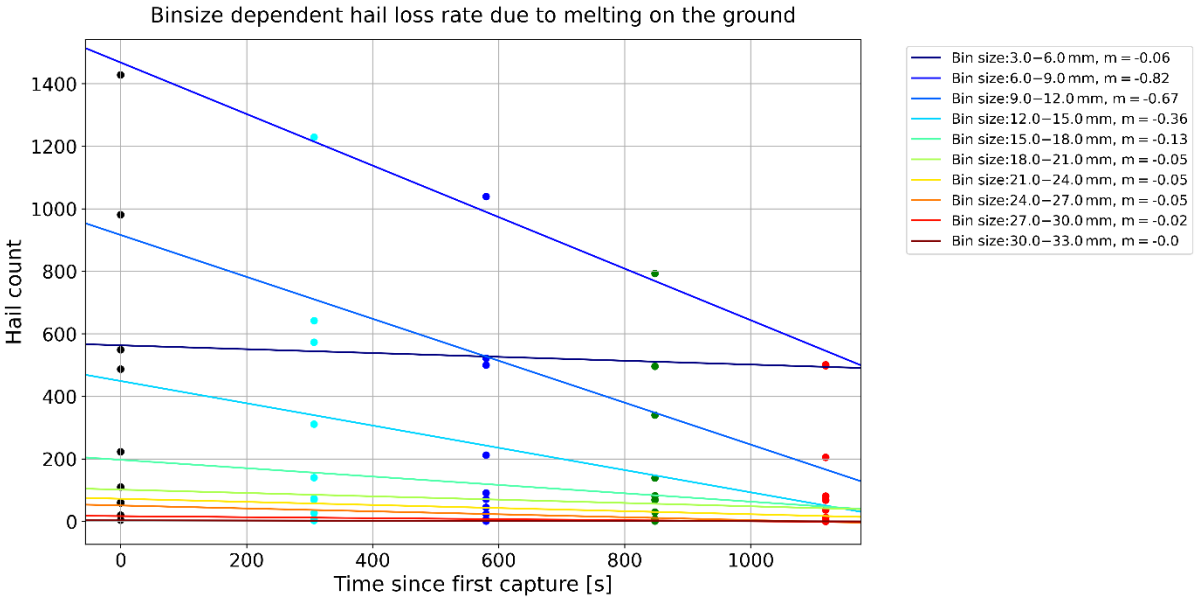
lines 376: please avoid repeating information from the caption "Those hailstones shrink from initially 33 mm to 21 mm, respectively 25.5 mm, during the course of 1119 s."

We keep the information in the caption and removed the sentence in line 276. The whole text are here is also strongly rearranged and reformulated.

lines 385: I feel a more effective plot for the analysis of melting rates would be to use initial size bins, fit a linear fit to each size bin and plot the slope. This would directly show the melting rate for different sizes.

Your explanation for a more effective plot of the melting rate is not totally clear to us and we encourage you to explain it a bit more detailed. At least we tried to go into this direction somehow and therefore want to present a plot here in our answer, but we are not sure if this is what you really meant.

In this plot we show hail counts versus time since first capture for a series of bin sizes. The colored scatter points mark the 5 consecutive flights while the colored lines go with the different bins. Of course one could intuitively think that when the slope of these lines is higher, the higher is also the melting rate within this bin. But this is not necessarily true, because stones change bins and can even over jump a bin between 2 consecutive flights. It is quite obvious in the first bin (3-6mm), which gets steadily filled up by stones from the higher bins and we just do not know exactly from which bin these stones come. Most of them likely from the bin above.



line 388: This analysis of 48 hailstones doesn't seem necessary as you can't confirm a robust result and isn't completely described (where is the hail from, what sizes, etc).

We agree, that this analysis is not profound enough yet to be included in this paper and we thus remove the sentences in lines 387-389. The listed results in the conclusion were adapted accordingly:

" - The evolution of the HSD caused by melting could be monitored during 18.65 min by analysing data from multiple drone flights. A melting rate in the order of 0.5 mm min^{-1} could be estimated."

Section 3.2: I'm curious how this experiment would go considering only severe hail sizes (e.g., above 20 mm). But there might not be enough information from the hail disdrometers for a comparison.

We do not use disdrometer data, and assume you mean the hail sensor data. As the HSDs in Fig. 2(d) (new Fig. 4) show, there are no hail diameters recorded on the sensors above 20 mm and as you say we thus cannot compare such results.

Section 4 paragraph 1: I don't think this opening paragraph is needed as this information is discussed again later. Further, it doesn't flow well into the second paragraph.

line 395-400: I would clarify that dry growth produces high densities of microscope air bubbles. Wet growth definitely soaks, but it also accretes on top of existing outer later too.

For now, we decided to keep the paragraph but we reformulated it, also to clarify better the raised points. Now it reads:

"A major challenge for drone-based photogrammetry of hail is related to the appearance of the hail within an orthophoto. The hail stones need to show distinct differences from the background. This is not always the case as hail is formed by a combination of dry and wet growth processes, which can lead to varying densities and appearances in the ice. Dry growth produces high densities of microscopic air bubbles that scatter light, while wet growth causes liquid to soak into gaps and accretes on top of existing outer ice to form clearer ice."

line 401: "pure" isn't needed here. Also "In a first step" should be "In a first attempt". Please also update "Second step" in line 408.

We simplified these sentences (line 345 and 352) to:

"First, a simple computer vision approach (without neural networks) was tested to extract the segmentation hail masks. … Second, a deep-learning model (Mask R-CNN) was tested. "

Line 419: I'm unclear how "Also a cropped hailstone binary mask can still lead to the correct major axis length." I would argue that it would lead to a negative bias.

Our wording here was a bit misleading. Of course you are right, in any way this potential cropping reduces either the major or intermediate axis and generates a potential (if the model classifies it as hail) second hailstone. This sentence is removed and we reformulated (starting line 358) it to:

"However in our case large hail was sparse and, as the image tiles cover large areas (500×500 pixels), it is safe to assume that the number of truncated hailstones is very low. Other sources of errors such as false positive detections or missed hailstones likely play a more important role."

Line 422: I wonder if soaking during melting is the main driver of changes in brightness

Regarding one of your comments below, this paragraph was removed.

Line 428: Ryzhkov et al. 2013 uses simulations of melting hail to estimate changes in polarimetric radar information (which is later used to develop a retrieval). So this isn't a radar study of melting hail.

Sorry for the mistake here. We correct and write now:

“The effect of melting hail in the air was studied by Kumjian and Ryzhkov (2008) using polarimetric radar measurements and numerical model investigations were performed by Fraile et al. (2003).”

Paragraph starting on 421: This paragraph feels incomplete. I'd suggest removing it if the authors can't link this into the results.

For now, we follow the suggestion and removed this paragraph.

Line 473: "different ages" should be "the duration".

Due to extensive language changes and text reduction the expression “different ages” disappeared.

Table 4: The information about the different comparison points should be in the text, not the caption.

For sure it is better to move this information into the text part. In the revised manuscript you find it in line 418ff:

“The comparison of drone-based photogrammetry with automatic hail sensors allowed to highlight the advantages and limitations of both approaches in measuring hail (see a summary in 4). We here want to highlight that the clustering problem refers to many hailstones that aggregate on the ground next to each other. This predominantly occurs during hail events with dominating small hail and intense precipitation. The resulting hail clusters pose a problem to the algorithm to differentiate between individual hailstones. An equivalent problem within the automatic hail sensor data is related to the dead time after each hail impact. The dead time is necessary to avoid any interference with subsequent impacts and to perform the retrieval of the data (Kopp et al., 2023). Furthermore, by combining data from both approaches strongly improves the reconstruction of the complete HSD and could further extend our understanding of hailstorms.”

To be compliant with AMT, the red and green color in Table 4 was changed to black. The advantages are now in bold font and the disadvantages in normal font style.

Also the red color in Table 1 was changed to black and bold font style.

Figure 1: Please reduce the number of wind barbs so it's readable! Can you please also annotate the hodograph with the levels or indicate what they are in the figure caption.

As you suggested, we reduced the number of wind barbs in the Skew-T plot. We changed the color for the layers in the hodograph (0, 1, 3, 5, 10 km) and they are defined now in the figure caption. Further the label on the hodograph axis was added and we changed from m/s to kt units. All unit styles are now conform to AMT (no "/" anymore). We removed the interpretation of the sounding from the caption and added the MetPy software reference. Please note there is a slight change in the absolute values for e.g. CAPE, CIN, SRH likely due to the use of a newer version of MetPy.

Figure 2: Please check all the text and annotations in this figure can be read at 100% zoom when rendered. The font size is also not consistent across the subplots. The white box and magenta cross in (b) are not visible at 100% zoom and I also can't find HS1 in subplot (c). Finally, I'd suggest not repeating the same information in both the caption and the main text; for example "corresponding to the MeteoSwiss app categories: smaller than coffee bean, coffee bean, 1 CHF coin, 5 CHF coin and tennis ball), are given." is repeated in both.

As the figure was made out of 4 separate plots and rearranged afterwards it was difficult to get the font sizes consistent. We suggest to split the Figure in 3 separate Figures and keeping only (a) and (b) together with consistent font sizes.

In Figure 2(a) we removed the magenta circle from the storm track as it was hard to see and adjust the caption. The given time information of the cell should be enough to follow the story. Regarding the caption information for Figure 2(b) we removed: "~~corresponding to the MeteoSwiss app categories: smaller than coffee bean, coffee bean, 1 CHF coin, 5 CHF coin and tennis ball~~)", as this information is already given in the text part.

HS1 sensor is the closest to the soccer field towards the south and drawn as cyan circle. We enlarged the sizes of all markers a bit (by 20%) and changed from magenta to black cross for the soccer field in old Figure 2(c), which should be better visible. In Figure 2(b) we switched from white to black rectangle for the map zoom area of new Figure 3 (old Fig. 2(c)).

The new Figure 4 shows the HSDs from the 3 hail sensor that recorded hail. Please note, that for HS3 the distribution changed a little bit because we now cut the last impact (likely from a non-hail object). The recording time is now only 16 minutes, rather than 52 minutes, which is also more consistent to the other recording times for the studied hail cell.

Figure 3: Please just describe colors as their proper names (e.g., dark red, light grey and green). Also, there overlap in (d) is significant and I can't really get much from it. Can you just show the center locations perhaps?

Color names were changed from "whitish" to "light grey", "greenish" to "green" and "reddish" to "dark red".

With the added alpha transparency in subplot (d) one at least sea where there is a lot of overlap present as the blue gets darker. From our point of view it makes more sense to show the actual size of the virtual sensor areas with respect to the soccer field area than just the center points of the sensors. However, we also show here the plot version with the smaller center points of the

locations, for further discussion. We are up to add/exchange it at a later stage, if you still think it is worthwhile.

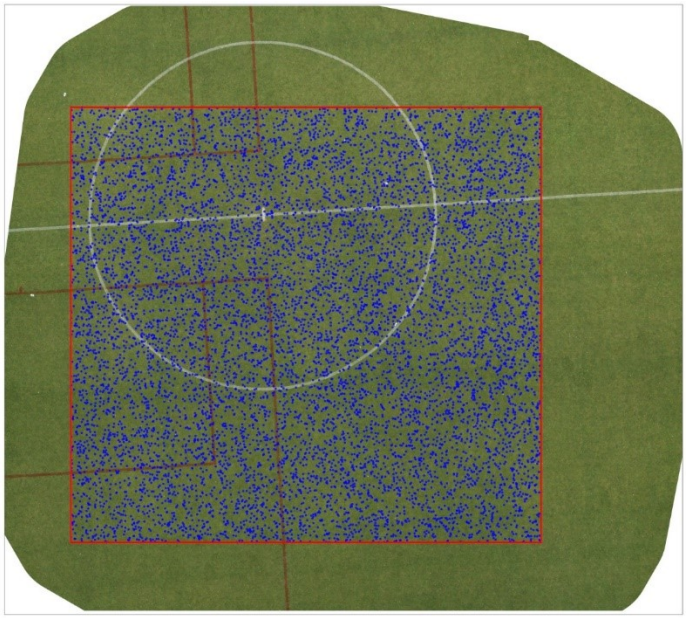


Figure 4: Is this really a spaghetti plot? I would describe this as line plots.

We changed the nomenclature from “spaghetti” to “line”.

Figure 8: HSD should be expanded in the title. and # replaced with "hail count"

In Figure 8, HSD is written as “Hail size distribution” and “#Hail” is replaced by “Hail count”.

We also changed the title in Figure 7 accordingly to “Comparison of hail size distributions with the test data set” and avoid the acronym HSD.

Figure 9: Font sizes is not consistent for percentiles and I think the colors change? I could extend these lines such that the text sits on top of the highest bars (with no overlap).

The alpha value of the color changed with the percentiles. Now we revised the figure in the way, that we have consistent font sizes for the text which belongs to the percentiles. The text is simplified and made consistent to Figure 10 by the use of e.g. “Q5” instead of “5th percentile”. The line color does not change anymore with the quantiles and the height of the lines are now filling the whole plot. Because there can be 2 quantiles on one bar, we should not put the text on top of the bars. The line color for the quantiles in subplot (b) was changed from orange to blue and is now consistent to (a).

In the caption we write: “The vertical blue dashed lines indicate the position of the particular quantiles with respect to the major axis (Q5, Q25, Q50, Q75, Q95) and projected aspect ratio (Q5, Q25, Q50, Q60).”

Figure 10: (a) Q25 and Q75 lines should ideally be different colors. Issue with percentile font sizes again. X-axis labels on (c) are also a bit too close with that font size. Caption: "from virtually and random placed hail sensors" reads better as "from simulated hail sensors at random locations"

Yes, it is a good idea to change colors for Q25 and Q75 in (a). Now they are shown by blue and green dashed lines. The font size for the axis labels has been reduced and now the labels on subplot (c) are better separated. Like for Fig. 9 we now use the quantile nomenclature "Q" and not "percentile" anymore and adjusted the font sizes and the Q-lines (red) fill up the whole y-range of the axis. Please note, the total impacts on HS3 are now 32 (33 before) - see comment on line 363-364.

In the caption we use the terminology of quantiles and we substitute "…from virtually and random placed hail sensors…" with "…from simulated hail sensors at random locations…".

Figure 11: Can you please add the time since first capture above the columns of images? This will be useful to info the reader about the duration since first capture. The final sentence of the caption could be improved with "During the 1119 s these hailstones shrink about 12 mm (upper row) and 7.5 mm (lower row) in their major axis length."

To improve the Figure 11, we included the mm units for the major/minor sizes and we also added the time variable since first capture t_c [s] in the title of each hailstone subplot.

As suggested, the final sentence has been changed to "During the 1119 s these hailstones shrink about 12 mm (upper row) and 7.5 mm (lower row) in their major axis length.".

Figure 12: I don't think the log view adds much value to this analysis because the sample size in the upper tail is so small. The main message is carried well by plot (a). I'd suggest replacing the use of "map" with "flight" or "survey". I'd also suggest changing "secure" to "capture". Finally, how many hailstones are in this sample during flight 1?

Thanks for your advice to restrict the presentation to plot (a). The new title of this plot is now "'KDE evolution due to melting". Further we use the word "flight" instead of "map" and do not use "secure", but "capture".

The number of hail samples is summarized in Table 2 for all five surveys. During flight one we have 3925 hailstone samples for the melting analysis. This information is now included also in the caption of new Fig. 14 (old Fig. 12).

General comment:

I suspect that hail is most likely to fall such that the major and intermediate axes are visible from drone imagery. The minimum axis is most likely orientated to the vertical as the centre of mass is lowest to the ground at this most, and therefore has a (likely) high stability. This should be considered when discussing aspect ratio as a function of the major and minor axis.

Thanks for this very important comment, which is similar to a Referee #1 comment on the aspect ratios.

In Section 3.1 and the image caption of Fig. 9 (new Fig. 11), we now speak of projected aspect ratios:

"The projected hail aspect ratios indicate that the majority of hailstones have equal axis lengths (Fig. 11(b)) and 75% of the hailstones have projected aspect ratios higher than 0.75."

Further, we added a paragraph about the aspect ratios within the discussions (see answer to Referee #1 comment).

References

Kopp, J., Manzato, A., Hering, A., Germann, U., and Martius, O.: How observations from automatic hail sensors in Switzerland shed light on local hailfall duration and compare with hailpad measurements, Atmos. Meas. Tech., 16, 3487–3503, https://doi.org/10.5194/amt-16-3487-2023, 2023.