Response letter

Retrieval of the hail size number distribution from polarimetric radar data using the double-moment normalization

Matteo Guidicelli, Alfonso Ferrone, Gionata Ghiggi, Marco Gabella, Urs Germann, and Alexis Berne

Dear Referee #2,

We thank you for the careful and constructive evaluation of our manuscript. Your comments, together with those of the other referee, prompted us to carry out substantial additional analyses and to revise the manuscript extensively. In particular, we have strengthened the independent evaluation of the retrieval, quantified its uncertainty, and provided additional evidence supporting several methodological choices.

The main changes introduced in the revised manuscript are summarized below:

- We substantially expanded the evaluation against crowd-sourced hail reports by including all hail events from 2016 to 2026 (137 event days), including the recent severe hail event of 28 August 2026.
- We extended the drone-based evaluation from two to five independent hail events by including three new drone surveys collected during summer 2026.
- For drone events with repeated surveys, we developed an event-specific method to estimate the melting rate of hailstones as a function of their diameter. These estimates were used in the radar–drone comparison.
- We complemented the feature-selection analysis with the Akaike information criterion (AIC) and Bayesian information criterion (BIC), providing additional evidence that the improvement obtained from each selected predictor justifies the associated increase in model complexity.
- We tested the robustness of the feature selection with a shuffled-target (permutation) test (see our reply to your comment 3).
- We quantified the uncertainty associated with the fitted GAMs using an event-day bootstrap approach and propagated this uncertainty to the retrieved HSND and its derived properties.
- In addition to percentile diameters, we introduced maximum hail diameters defined for specified minimum hailstone number densities. These quantities provide a more directly interpretable description of the upper tail of the retrieved HSND.
- We evaluated model performance as a function of the distance from the nearest radar and across the three hailsensor regions (see our reply to your comment 7).
- We revised the introduction to state the gap addressed by this study earlier and to list its main novelties explicitly.
- We revised the methodological description and discussion to clarify the independence and representativeness of the evaluation datasets, the limitations of the spatial extrapolation, the role of the a posteriori adjustment, and the remaining assumptions and limitations of the retrieval.

Detailed responses to each of your comments are provided below. Your comments are reproduced first, followed by our responses and the corresponding changes made in the manuscript.

We believe that these additions have substantially strengthened the evaluation, transparency, and scientific robustness of the study, and we hope that the revised manuscript is now suitable for publication in *Atmospheric Measurement Techniques*.

On behalf of all co-authors,

Matteo Guidicelli

Comments by Referee #2

Legend: **Referee's comments**; *authors replies*

Manuscript excerpts

This paper is based on the dual-polarization radar of Switzerland. It combines the dual-moment normalization theory and the generalized additive model (GAM) to invert the size distribution of ground hail particles (HSND) using the dual-polarization radar data. It integrates multi-source data such as automatic hail sensors, unmanned aerial vehicle photogrammetry, and crowdsourced hail reports for model training and validation. The research idea is clear, the data is detailed, and the methods are complete. It has clear innovations in the field of radar quantitative hail inversion in complex Alpine terrain. Before publication, there are still some issues that need to be addressed.

We thank the reviewer for the positive assessment of our work and for recognizing its methodological innovation and relevance. We also appreciate the constructive comments provided below, which have helped us substantially improve the manuscript.

The manuscript uses the fixed normalized shape function eliminated by Ferrone et al. (2024), but the applicability of this function within the retrieval framework of this study needs further verification. In different radar bands and under different types of thunderstorms, are the parameters of the normalized function fixed, or does there exist variability? If possible, provide some sensitivity verification experiments.

We thank the reviewer for raising this point. We would first like to clarify that the fixed normalized shape function used in this study was derived by Ferrone et al. (2024) from ground-based hailsensor HSND observations and is therefore independent of radar frequency. Radar data are used in the present framework only to estimate moments 2 and 4, which scale the normalized distribution through inverse double-moment normalization.

The present retrieval was developed and evaluated exclusively using the Swiss C-band radar network. Application to another radar frequency would require dedicated recalibration or retraining of the relationships between the polarimetric radar features and the HSND moments, particularly because electromagnetic scattering properties depend on wavelength. This does not, however, imply a radar-band dependence of the normalized ground-level HSND shape, which is derived independently of radar observations.

The fixed function should be interpreted as an empirical representative shape derived from the available Swiss hailsensor observations, rather than as a universally invariant distribution. Ferrone et al. (2024) showed that double-moment normalization substantially reduced the variability among the available hailsensor events. They also evaluated the method against a drone-derived HSND covering an area several orders of magnitude larger than an individual hailsensor, providing support for its applicability across different observation scales.

Further support is provided by a separate study currently being prepared for submission, in which double-moment normalization was evaluated using drone-derived HSNDs from 11 hail events, including five events outside Switzerland. The results support the relative invariance of the normalized HSND and the applicability of the method across geographical regions and observation scales. Because the complete methodology and results will be presented in that dedicated study, they are not reproduced in the present manuscript.

We have clarified in the revised manuscript that the normalized function is derived independently of radar observations. We have also specified that its relative invariance was demonstrated across events, regions, and observation scales by Ferrone et al. (2024):

> "Leveraging data from the Swiss network of automatic hailsensors, Ferrone et al. (2024) demonstrated the applicability of double-moment normalization to HSNDs and showed that the normalized HSND is relatively invariant among events and regions. The method was also evaluated against a drone-derived HSND covering an area several orders of magnitude larger than an individual hailsensor, supporting its applicability across different observation scales."
> […]
> "The fixed normalized function, fitted by Ferrone et al. (2024) independently of radar observations, is then scaled using these radar-estimated moments to reconstruct HSNDs via (inverse) double-moment normalization."

Are the radar data used for training GAM and the drone/crowd-sourced report used for verification completely independent in terms of time and space? The drone observations may be from the same storm as the radar features and sensor data, which could lead to overly optimistic validation results. If there is spatial correlation among multiple radar data from the same event, it can also result in information leakage.

We thank the reviewer for raising this important point. Within the training dataset, some samples from the same event day may be correlated, for example when nearby hailsensors sample the same part of a storm, whereas other hailsensors on the same day may sample different storms, different parts of a storm, or different times. To prevent information leakage, we used a leave-one-event-out cross-validation: in each iteration, all hailsensor observations of the withheld day, together with the associated radar data, were excluded from model training. Samples from the same event day therefore never occurred in both the training and validation subsets (Sect. 3.4.2).

Regarding the external evaluation datasets, the radar data used for training are restricted to the immediate surroundings of the hailsensors: all selected predictors are aggregated over a vertical column of 1 km radius, so that the radar information associated with a hailsensor extends at most 1 km from its location. The evaluation datasets are largely independent of these training data in time, in space, or both:

– Drone surveys: the three surveys added in the revised manuscript (Winterthur, Hettlingen, and Niederhasli) were conducted in 2026, after the 2018–2025 training period, and are therefore independent in both time and space. The Locarno-Monti survey (28 June 2022) occurred on a training day, but the hailsensors used for training on that day were located more than 12 km away, far outside the 1 km radar columns used for training; the hailsensor closest to the survey recorded only 12 hailstones and was not included in the training dataset. The Entlebuch survey (20 June 2021) is located 1.15 km from the Hasle hailsensor, which recorded the same storm shortly before the survey and was used for training. At the 1 km scale of the radar predictors, the Entlebuch comparison is therefore not fully independent of the training data. However, this concerns a single hailsensor that recorded 31 hailstones, out of 9772 hailstones from 35 event days used for training, and four of the five drone surveys are spatially independent of the training data.

– Crowd-sourced reports: the evaluation now comprises 137 event days from 2016 to 2026, of which 115 do not coincide with any of the 35 hailsensor training days. On the remaining 22 days, the training data cover only the 1 km columns around the few hailsensors that recorded hail, whereas the crowd-sourced comparison extends over the complete hail swaths across Switzerland. These 22 days are now marked in Table A1.

Compared with the original manuscript, which included two drone surveys and a limited set of crowd-sourced events, the evaluation now relies on five drone surveys, three of which are fully independent of the training period, and on all hail events reported between 2016 and 2026. This substantially strengthens the independence and robustness of the evaluation. The updated map of the observational data (Fig. 1, reproduced below) and the extended Table A1 show the temporal and spatial distribution of the training and

evaluation datasets. Over the 137 event days, the HSND-based characteristic diameters yield substantially lower RMSE and generally higher correlations with the crowd-sourced reports than MESHS, consistently with the three events analysed individually (Figs. 11 and 12 of the revised manuscript, also reproduced below).

The corresponding passages of the revised manuscript read as follows (Sects. 3.4.2, 2.2.2, and 2.2.3, respectively):

"A leave-one-event-out cross-validation was performed to select the features that yield the best performance on unseen hail events. For each iteration, all hailsensor observations from the withheld calendar day and all associated spatially extrapolated radar–hailsensor samples were excluded together from model training. Consequently, possibly correlated samples belonging to the same event day could not occur simultaneously in the training and validation subsets."
[…]
"The drone observations were not used for model training or feature selection. The three drone surveys conducted in 2026 are independent of the 2018–2025 training period. The Locarno-Monti survey occurred on a training day, but the hailsensors used for training on that day were located more than 12 km away; the nearby Locarno-Monti hailsensor recorded only 12 hailstones and did not satisfy the minimum-count criterion for inclusion in model training. The Entlebuch survey is located 1.15 km from the Hasle hailsensor, which was used for training and recorded the same storm shortly before the survey; this comparison is therefore not fully independent of the training data at the 1 km scale of the radar predictors."
[…]
"All available hail-report days from 2016 to 2026 were considered. After applying the report quality controls and the event-selection criteria described in Sect. 3.5.3, 137 event days were retained for the quantitative evaluation. Of these, 115 days do not coincide with any of the 35 hailsensor training days and are therefore independent of the training data in both time and space. On the remaining 22 days (hailsensor days marked with an asterisk in Table A1), the radar data used for training cover only the 1 km columns around the few hailsensors that recorded hail, whereas the crowd-sourced comparison extends over the complete hail swaths across Switzerland."

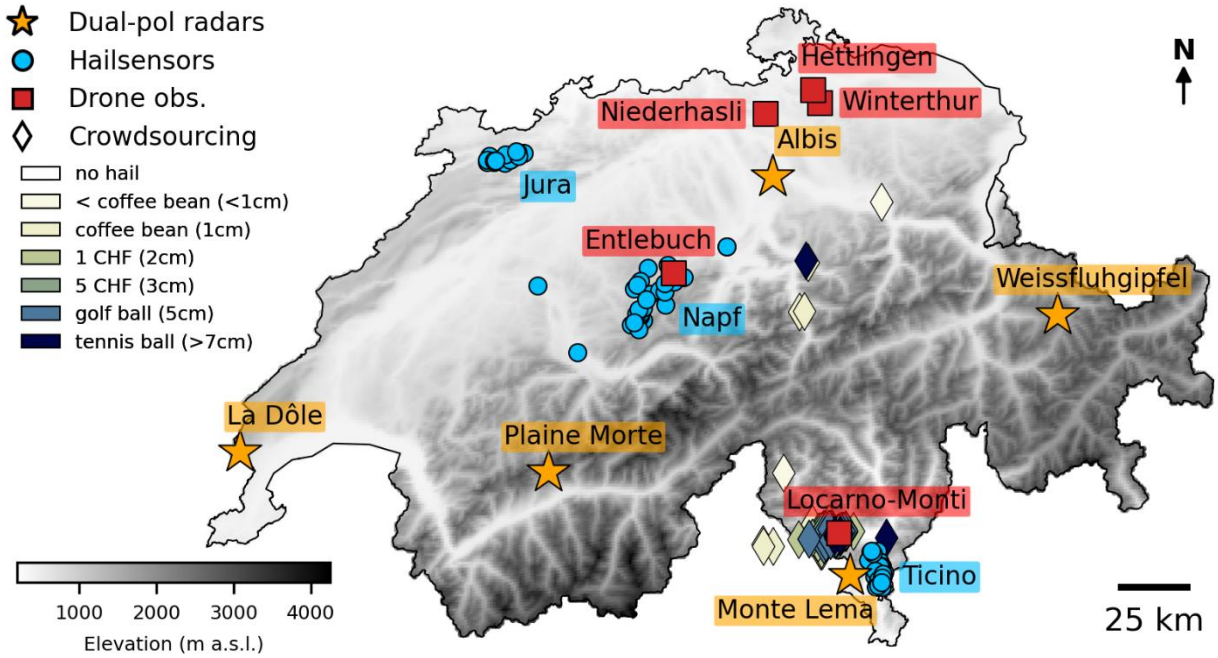

Figure 1 of the revised manuscript: geographic distribution of the observational data, including the five drone survey sites and the three hailsensor regions.

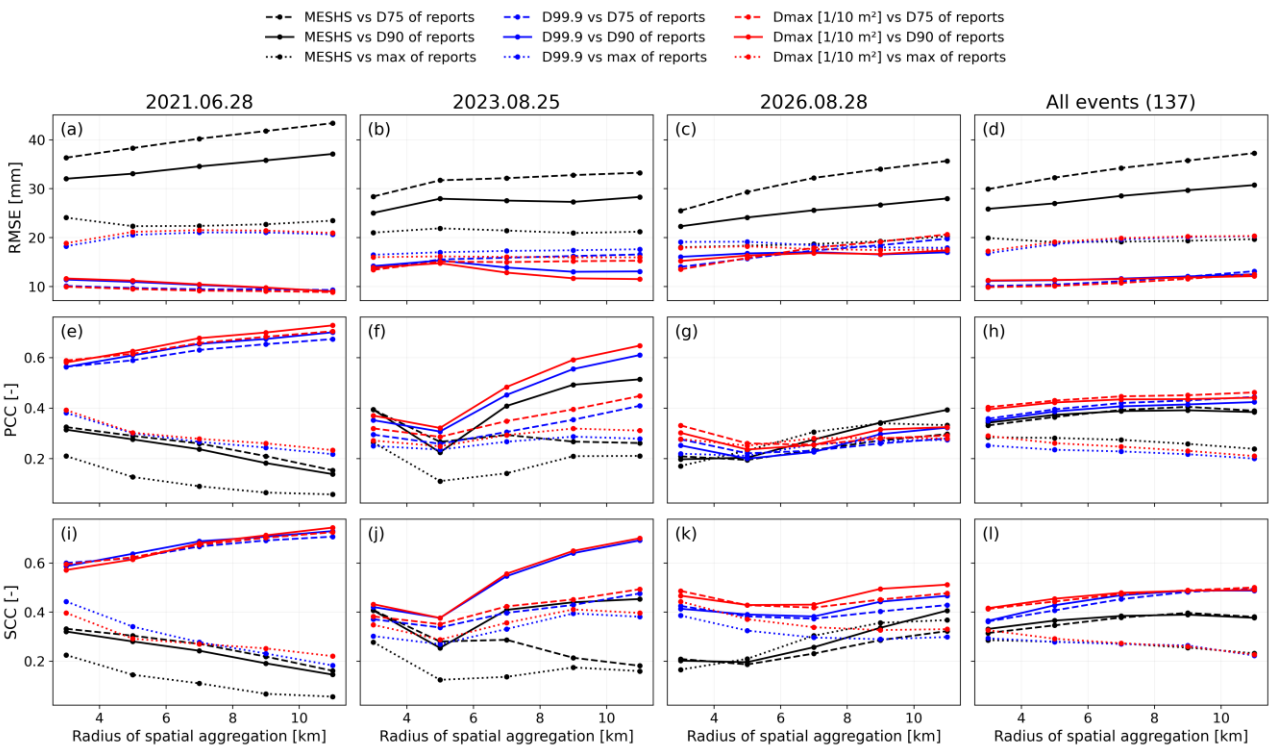

Figure 11 of the revised manuscript: RMSE, PCC, and SCC as functions of the spatial aggregation radius for MESHS, D99.9, and Dmax [1/10 m²] compared with the 75th and 90th percentiles and the maximum of the crowd-sourced reports, for the three individual events and for all 137 event days.

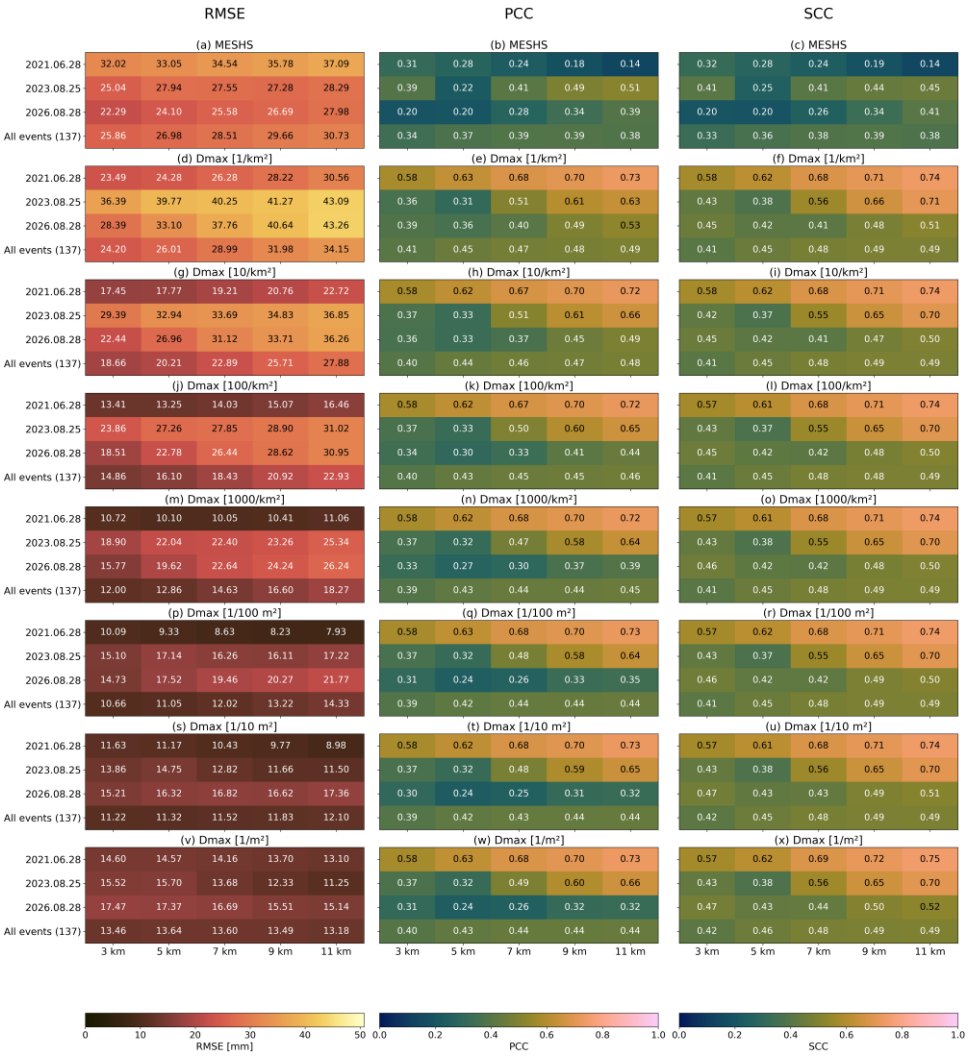

Figure 12 of the revised manuscript: RMSE, PCC, and SCC for MESHS and count-supported maximum diameters compared with the 90th percentile of the crowd-sourced reports.

Feature selection was carried out stepwise based on cross-validation using Spearman correlation, but only 4 features were ultimately selected. Among hundreds of candidate features, this stepwise selection method may amplify accidental correlations, especially in cases with limited sample size. If possible, it is recommended to randomly shuffle the target variables and repeat the feature selection process to see if the selected features remain stable.

We thank the reviewer for this suggestion. In our approach, the risk that the stepwise selection favours accidental correlations is reduced because a feature is retained only if it improves the leave-one-event-out cross-validated score, i.e. the performance on event days that were not used for fitting. We nevertheless further tested the robustness of the selection in two ways.

First, following also the suggestion of Referee #4, the cross-validation-based selection is now complemented by the Akaike (AIC) and Bayesian (BIC) information criteria, computed for each successive model (Sect.

3.4.2 and Table A2). Both criteria decreased with every added feature and for both moments, indicating that the improvement in model fit justifies the additional model complexity. The reduction is small for moment 4 when the second and third features are added, which is now stated in Sect. 4.1.

Second, we performed the test suggested by the reviewer (Fig. R1). When the best predictors are selected among hundreds of candidates, some correlation with the targets can arise by chance alone. To quantify how large such chance correlations can be, the moment-2 and moment-4 targets were jointly and randomly permuted across all training samples 100 times, while the radar features, event labels, and cross-validation folds were kept unchanged. With shuffled targets, the largest Spearman correlation between any candidate feature and the targets reached 0.20 (median over permutations; 5th–95th percentile 0.15–0.26, maximum 0.29), i.e., the level of correlation that can be expected by chance given the number of candidates and samples. With the real targets, it reached 0.61, more than twice the largest value obtained in any permutation. Moreover, the four selected features fitted to the shuffled targets yielded leave-one-event-out cross-validated Pearson correlations of −0.15 (5th–95th percentile −0.30 to 0.07) for moment 2 and −0.13 (−0.30 to 0.07) for moment 4, compared with 0.62 and 0.68 with the real targets, confirming that the cross-validation does not produce skill in the absence of a physical relationship. None of the 100 permutations approached the values obtained with the real targets ($p \leq 0.01$).

These results show that the selected features reflect a genuine relationship between the radar observations and the hailsensor HSND rather than accidental correlations arising from the large number of candidate features. Nevertheless, because many candidate features are strongly correlated with the selected ones and differ mainly in the aggregation radius or reflectivity threshold, the specific variants selected may partly depend on the available data, as acknowledged in Sect. 4.1.

The corresponding passages of the revised manuscript read as follows (Sects. 3.4.2 and 4.1, respectively):

"To complement the cross-validation-based selection, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) were also examined for each successive model to assess whether the improvement obtained by adding a predictor justified the corresponding increase in model complexity. Lower AIC and BIC values indicate a more favorable balance between model fit and complexity."
[…]
"Both AIC and BIC decreased after each successive feature was added, indicating that the improvement in model fit outweighed the corresponding increase in model complexity. The reduction was, however, small for moment 4 when the second and third features were added (ΔBIC of −1.3 and −1.5), indicating that these two features mainly improve the estimation of moment 2, whereas the fourth feature yielded the largest reduction for moment 4. […] Many of the other extracted features are strongly correlated with these predictors and differ primarily in the spatial aggregation scale or the applied reflectivity threshold. Consequently, the selection of these specific features may partly reflect the limited amount of available hailsensor reference data, which makes robust discrimination among highly correlated predictors challenging."

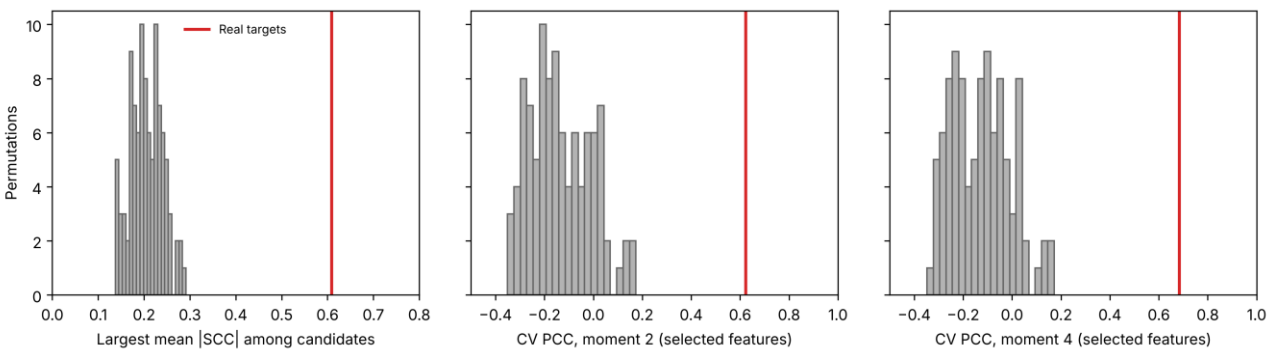

Figure R1: Shuffled-target test (100 permutations; shown to the reviewer only). Grey histograms: distribution over the permutations of (left) the largest mean absolute Spearman correlation between any candidate feature and moments 2 and 4, and of the leave-one-event-out cross-validated Pearson correlation of the four selected features for (middle) moment 2 and (right) moment 4. Red lines: values obtained with the real targets.

When comparing the drone data, the melting correction for the radar inversion of HSND was performed. However, this rate (0.5 mm/min) was only derived from one event in 2021 and was directly applied to another event in 2022. The environmental temperature, humidity, and ground conditions of the two events were different, and this assumption may lead to biased estimation.

We thank the reviewer for this comment. As already noted in the original manuscript, a melting rate estimated for one event may not be representative of another event with different environmental and ground conditions. Following the reviewer's comment, the melting correction has been revised as follows.

First, melting rates are now estimated separately for each event for which repeated drone surveys of a common area were available (Entlebuch, Winterthur, Hettlingen, and Niederhasli). Within the common area, the hailstones retrieved from two consecutive surveys were ordered by size and paired by rank; the decrease in equivalent diameter of each pair, divided by the time between the surveys, provides an estimate of the melting rate. The rates were grouped by initial-diameter class, and the median rate of each class was used to define a diameter-dependent correction (Sect. 3.5.2). The estimates, which are part of a companion study based on drone-retrieved HSNDs that will be submitted shortly, are shown in Fig. R2.

The median rates range from 0.18 to 0.86 mm min $^{-1}$. They decrease with increasing diameter at Hettlingen and Niederhasli, whereas a rate of 0.30 mm min $^{-1}$ was obtained for all diameter classes at Entlebuch. At Winterthur, sufficient paired samples were available only for the 5–10 mm class, and the corresponding rate (0.21 mm min $^{-1}$) was applied to all diameters. For hailstones of 15–20 mm, the rates differ by almost a factor of three between Entlebuch (0.30 mm min $^{-1}$) and Hettlingen or Niederhasli (0.78 and 0.86 mm min $^{-1}$). This variability confirms the reviewer's concern that a melting rate should be transferred from one event to another only with care, and motivated the event-specific correction. The value of 0.5 mm min $^{-1}$ used in the original manuscript has been replaced accordingly.

Only one suitable survey was available at Locarno-Monti. For this event, we used a constant rate of about 0.26 mm min $^{-1}$, the midpoint between the rates estimated at Entlebuch and Winterthur, whose observed HSNDs are the most similar to that at Locarno-Monti among the events with repeated surveys. The melting correction at Locarno-Monti is therefore more uncertain than at the other sites, although the rates were taken from the events most likely to be similar to it.

Second, the sensitivity of the comparison to the melting correction is now quantified by perturbing the melting rates by ±0.1 mm min $^{-1}$ (Sect. 3.5.2). The resulting ranges are shown as dashed and dotted curves in Fig. 7 and as vertical bars in Fig. 8 of the revised manuscript. The sensitivity is substantial, in particular at Locarno-Monti, where no event-specific estimate was available, and the melting correction is discussed as a source of uncertainty of the drone comparison in Sect. 4.3, where we now also mention that, at Niederhasli, the time elapsed between the end of the hailfall and the survey is uncertain by about ±5 min. For moment 4, however, the sensitivity ranges remain small compared with the range of values among events, which spans about two orders of magnitude, so that the strong correlation between drone- and radar-derived moment 4 (PCC = 0.90, SCC = 1.00) is not driven by the choice of the melting rate.

The corresponding passages of the revised manuscript read as follows (Sect. 3.5.2):

"Melting rates were estimated from the repeated drone surveys available for Entlebuch, Winterthur, Hettlingen, and Niederhasli. An HSND was retrieved independently from each survey. Within the common area, hailstones from two consecutive surveys were ordered by size and paired by rank; the decrease in equivalent diameter of each pair, divided by the time between the surveys, provided an estimate of the melting rate. The resulting rates were grouped by initial-diameter class, and the median rate in each class was used to define the diameter-dependent melting correction. The median rates ranged from 0.18 to 0.86 mm min^{-1} and decreased with increasing diameter at Hettlingen and Niederhasli, whereas a rate of 0.30 mm min^{-1} was obtained for all diameter classes at Entlebuch. At Winterthur, sufficient paired samples were available only for the 5–10 mm class, and the corresponding rate of 0.21 mm min^{-1} was applied to all

diameters. Hailstones smaller than the smallest diameter class with paired samples were assigned the rate of that class. At Locarno-Monti, where only one suitable survey was available, we used a constant rate of about 0.26 mm min^{-1}, the midpoint between the rates estimated at Entlebuch and Winterthur, whose observed HSNDs were the most similar to that at Locarno-Monti among the events with repeated surveys. The melting correction at Locarno-Monti is therefore more uncertain than at the other sites, although the rates were taken from the events most likely to be similar to it."

[…]

"The accumulated radar HSND was corrected from each radar timestamp to the drone-observation time using these diameter-dependent melting rates. An additional ±0.1 mm min^{-1} perturbation was applied to quantify sensitivity to the melting-rate estimate."

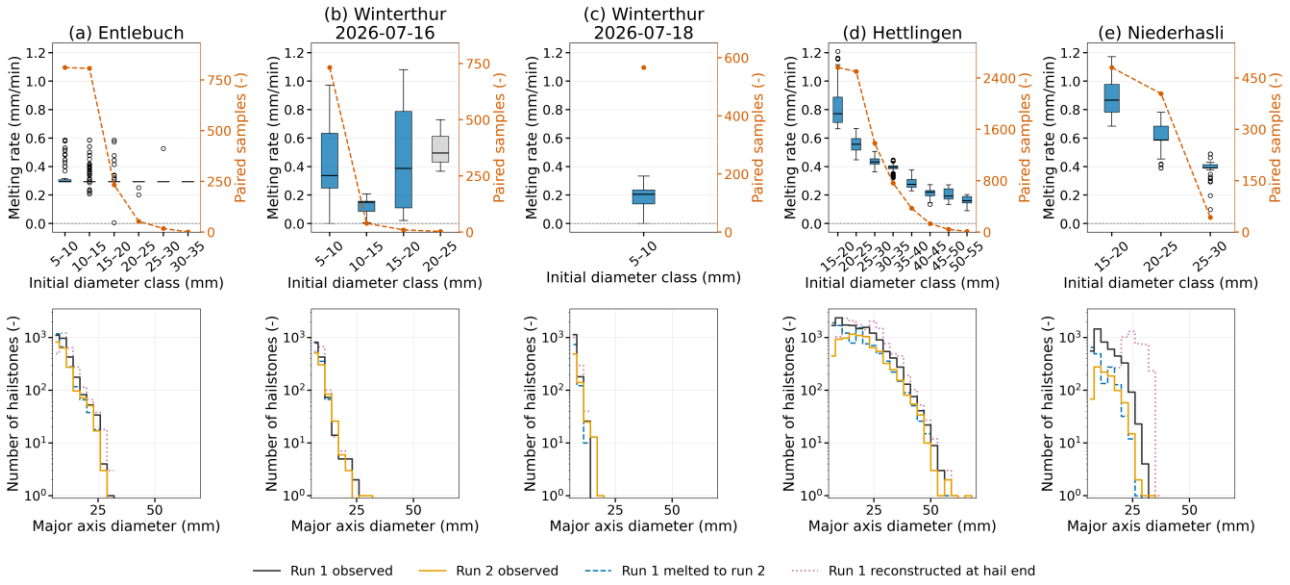

Figure R2: Melting rates estimated from repeated drone surveys (shown to the reviewer only; from a companion study based on drone-retrieved HSNDs, Guidicelli et al., in preparation). Top: melting rates per initial-diameter class (boxplots, left axis) and number of paired hailstones (orange, right axis). Bottom: observed HSDs of the first (black) and second (orange) survey, the first survey melted to the time of the second (blue dashed), and the first survey reconstructed at the end of the hailfall (pink dotted). Panel (b) refers to a survey at Winterthur on 16 July 2026 that is not used in the present study.

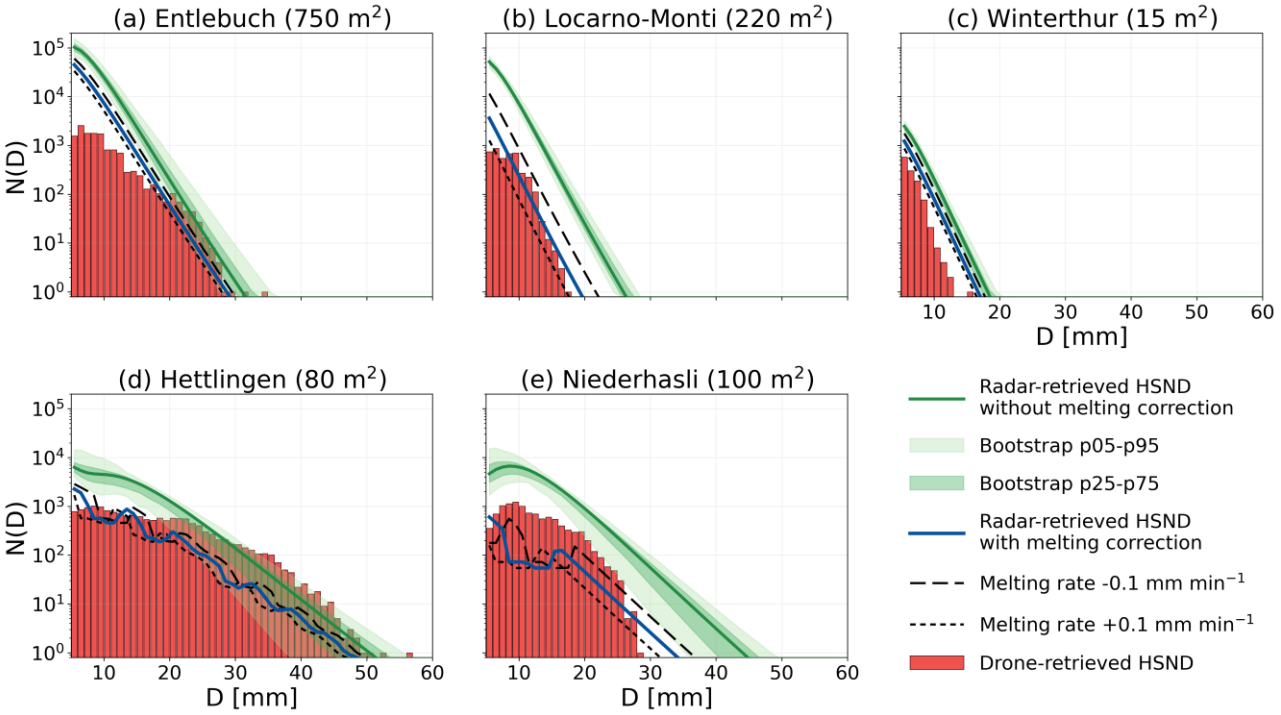

Figure 7 of the revised manuscript: drone-retrieved HSDs compared with the radar-retrieved HSNDs without (green) and with (blue) melting correction, including the sensitivity to a ±0.1 mm min⁻¹ perturbation of the melting rate (black dashed and dotted curves).

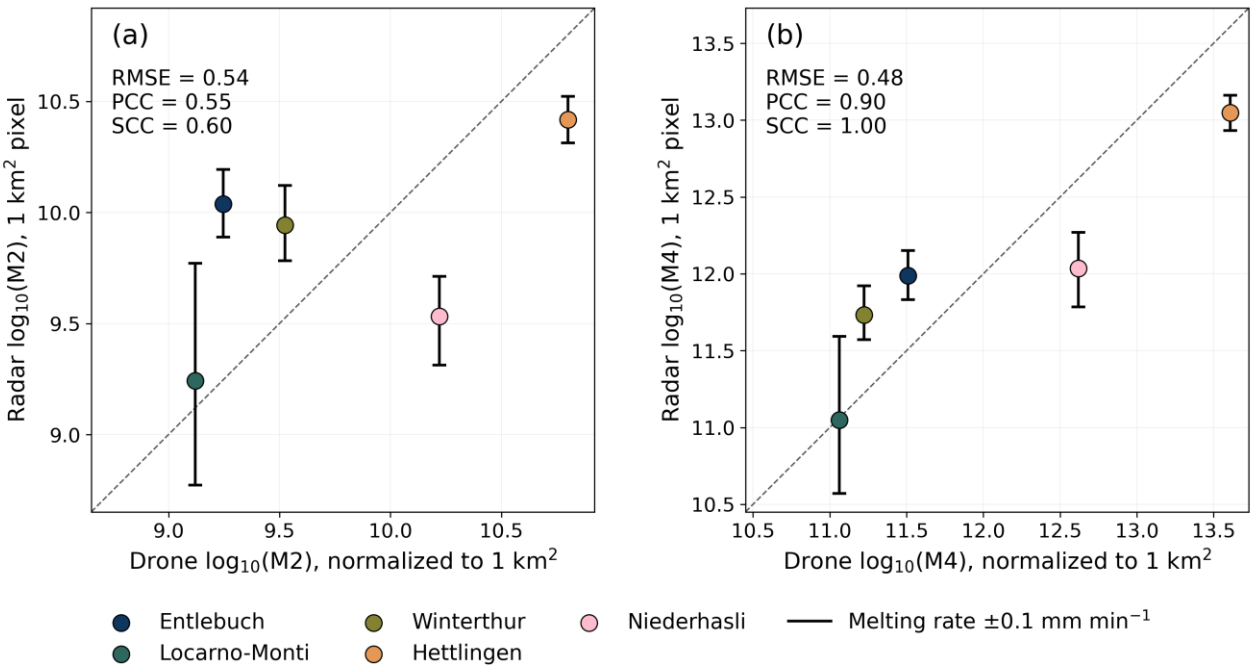

Figure 8 of the revised manuscript: drone-observed versus radar-retrieved moments 2 (a) and 4 (b) for the five drone survey sites; vertical bars show the sensitivity to a ±0.1 mm min⁻¹ perturbation of the melting rate.

The GAM model only provides quantitative estimates without any uncertainty intervals (such as uncertainty range, confidence interval, etc.). It is suggested to provide standard deviation or percentile intervals for model predictions.

We thank the reviewer for this important suggestion. We agree that uncertainty estimates are an important complement to the point estimates, and the revised manuscript now provides percentile intervals for the GAM estimates and for all quantities derived from the retrieved HSND (new Sects. 3.4.4 and 4.5, and Figs. 13 and 14).

The uncertainty is estimated with a non-parametric event-day bootstrap. Complete hailsensor event days are sampled with replacement, which preserves the dependence among observations from the same event. For each of 100 bootstrap realizations, the GAMs for moments 2 and 4 are refitted with the selected predictors and fixed hyperparameters, and the a posteriori adjustment is re-estimated on the resampled data. The resulting moment pairs are propagated through the inverse double-moment normalization, yielding an ensemble of HSNDs and derived quantities (e.g., characteristic diameters, total kinetic energy, and number of hailstones larger than 2 cm). The ensemble is summarized by its 5th, 25th, 50th, 75th, and 95th percentiles.

Figure 13 shows the GAM estimates for the hailsensor observations together with the 50 % and 90 % bootstrap intervals. The intervals vary among observations and tend to be wider for some of the larger observed values. Figure 14 shows how this uncertainty propagates to the HSND-derived quantities for the event of 28 June 2021. The absolute interquartile range increases where the retrieved values are larger, and the relative uncertainty differs markedly among quantities: it is small for Dmax [1/10 m²] but large for the total kinetic energy and the number of hailstones larger than 2 cm. Uncertainty therefore varies both in space and among quantities and cannot be represented by a single event-wide value. The bootstrap intervals are also shown for the radar-retrieved HSNDs in the drone comparison (Fig. 7, green envelopes; see our reply to the previous comment).

These intervals quantify the uncertainty of the fitted relationship between the radar predictors and the moments arising from the limited number of training events. We now discuss explicitly the sources of uncertainty not represented by these intervals, in particular the residual scatter of individual observations around the fitted relationship, quantified by the cross-validated RMSE (Sect. 4.6).

The corresponding passages of the revised manuscript read as follows (Sects. 3.4.4 and 4.6, respectively):

"The parameter uncertainty of the fitted GAMs, arising from the finite number of hailsensor event days available for training, was estimated using a non-parametric event-day bootstrap. Complete hailsensor event days were sampled with replacement, thereby preserving the dependence among observations from the same event. For each of 100 bootstrap realizations, the GAMs for moments 2 and 4 were refitted using the selected predictors and fixed hyperparameters, and the a posteriori adjustment (Sect. 3.4.3) was re-estimated on the resampled data.
The moment estimates from each realization were propagated through the inverse double-moment normalization to obtain an ensemble of retrieved HSNDs and derived quantities. The ensemble was summarized using its 5th, 25th, 50th, 75th, and 95th percentiles, while the estimate from the GAM fitted to the complete training dataset was retained as the reference output. For the quantities derived from the HSND (Sect. 3.4.6), the percentiles were obtained by computing each quantity from every bootstrap HSND and taking the percentiles over the resulting values, rather than by computing the quantity from a percentile HSND. The sources of uncertainty not represented by these intervals are discussed in Sect. 4.6."
[…]
"These intervals represent the uncertainty of the fitted relationship between the radar predictors and the expected moments. They do not include the residual scatter of individual observations around this relationship (cross-validated RMSE of 0.34 and 0.44 in log10 units for M2 and M4, respectively; Table A2), which reflects both the natural variability of ground-level HSNDs for given radar features and the sampling noise of the 0.2 m² hailsensors. Further sources not represented can be grouped along the retrieval chain: (i) the radar input, including measurement errors, attenuation, non-Rayleigh (resonance) scattering by large hailstones, and beam-geometry effects on the derived features; (ii) the model structure, as the selected predictors, hyperparameters, and normalized shape $\hat{h}(x)$ are fixed in all realizations; and (iii) the reference data and their representativeness, including errors in the hailsensor diameter estimation, the melting of

hailstones during their fall, which is not explicitly accounted for in the training, and the scale difference between the sensor-based targets and the 1 km² product."

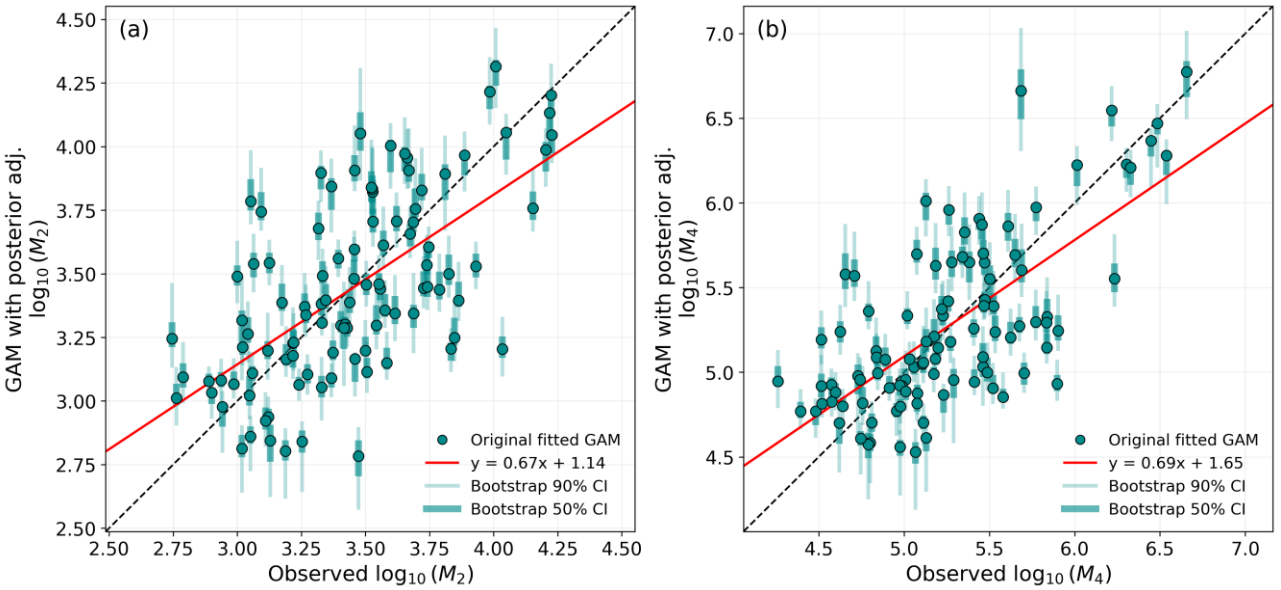

Figure 13 of the revised manuscript: observed and GAM-estimated log10(M2) (a) and log10(M4) (b) for the hailsensor observations, with the 25th–75th (thick) and 5th–95th (thin) percentile ranges of the 100 event-day bootstrap realizations.

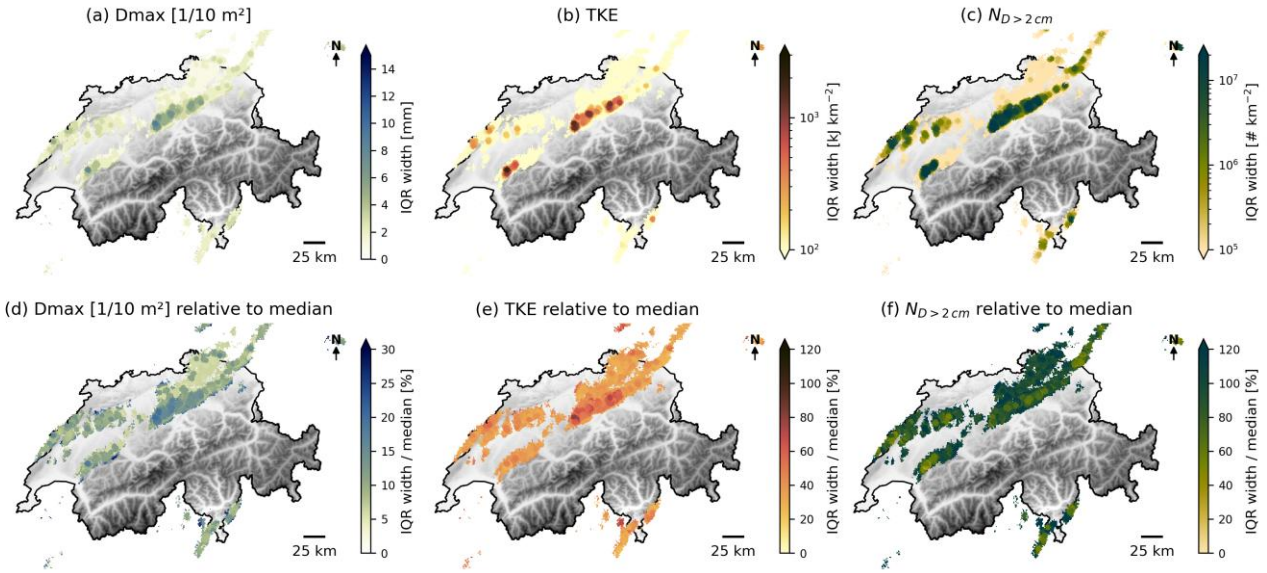

Figure 14 of the revised manuscript: interquartile-range width of the bootstrap realizations for Dmax [1/10 m²] (a), total kinetic energy (b), and number of hailstones larger than 2 cm (c) for the event of 28 June 2021, and the corresponding widths relative to the bootstrap median (d–f). The colour scales of (a) and (d) are narrower than those of the other panels and than the Dmax scale in Figs. 9 and 10.

Five radars are distributed in different areas, with differences in altitude and terrain conditions. When merging the detection data from the five radars, how are the effective

detection range of the radars, the radar blind zones, terrain clutter, and the influence of rugged terrain considered, as well as how to unify the characteristic data from different heights?

We thank the reviewer for this question. As described in Sect. 3.1 and Appendix B, the five radars are not combined at the level of derived products or radar features: their polar measurements are merged into a single 3D Cartesian composite with 250 m resolution before any feature is extracted. All features are then computed from this unified composite. The aspects raised by the reviewer are addressed as follows:

– Effective detection range and blind zones: the polar measurements of each radar (20 elevation angles) are first mapped onto a common regular 3D grid. A voxel is filled only if it lies within the sampling volume of a radar gate, i.e., within half the range resolution and within the beamwidth in azimuth and elevation, and voxels beyond the range allowed by the beam geometry and grid resolution are excluded. Voxels that are not sampled by any radar therefore remain empty, and no values are extrapolated into blind zones. Only at high elevation angles is the half-beamwidth enlarged, for gridding purposes, to avoid gaps close to the radars (Fig. A1 and Appendix B).

– Overlapping coverage: where a voxel is observed by more than one radar, the reflectivity and all other polarimetric variables of the radar with the highest horizontal reflectivity are retained, as this radar is assumed to be the least affected by attenuation. Because partial beam shielding also reduces the measured reflectivity, this criterion tends to favour the radar with the least obstructed view of the voxel. An example of the radar selected for each voxel is shown in Fig. A2.

– Ground clutter and terrain: the radar data used are the operational quality-controlled measurements of MeteoSwiss, which were further corrected for ground clutter following Joss and Lee (1995) (Sect. 2.1).

– Data from different heights: since the measurements of all radars are mapped onto the same 3D grid, all features are computed on the same vertical columns, independently of which radar observed each voxel. The selected features (e.g., the maximum and the sum of Z_V in the column, and the volume with pHV < 0.97 and Z_H > 50 dBZ) are therefore defined consistently across the whole domain.

Nevertheless, the merging cannot fully compensate for the effects of radar sampling geometry in complex terrain. Beam broadening at far range, the high altitude of the lowest beams, and beam blocking may lead to an underestimation of volume-based features, as discussed in Sect. 4.6. The dependence of the retrieval on the distance from the radar is evaluated in our reply to the next comment.

The corresponding passages of the revised manuscript read as follows (Sects. 3.1 and 4.6, respectively):

"For each radar, a 3D composite is first created by interpolating its measurements onto a regular 3D grid of 250 m resolution, chosen as a compromise between computational efficiency and high spatial detail that could potentially explain local ground-level variability in observed HSNDs. The individual radar composites are then merged into a single unified 3D product, retaining for each voxel the reflectivity and all other polarimetric variables from the radar with the highest horizontal reflectivity, which is assumed to be least affected by attenuation. Details on the procedure are reported in Appendix B, where Fig. A2 illustrates the selected radar in the final 3D composite for the event of 25 August 2023 between 19:20 and 19:25 UTC, over the hailsensor located in Locarno-Monti (Ticino)."
[…]
"Some of the radar-based features used in this study are inherently influenced by radar sampling geometry. At far ranges, the widening of the radar beam reduces vertical resolution, which can prevent storm tops from being properly captured. Conversely, the minimum elevation angle may not adequately sample the lower storm regions, particularly given the high altitude of Swiss radars and beam blocking by complex terrain. These effects can lead to underestimation of volume-based features. In addition, the increasing radar resolution volume with distance can affect polarimetric variables, which may differ from values observed closer to the radar."

The hailsensors, drone, and crowdsourced reports show distinct cluster characteristics. They mainly focuses on three centers in this study. Among them, their observations in Locarno Monti is closest to the Monte Lema radar, while the other two clusters are farther away from the radar. It is suggested to conduct some sensitivity tests to discuss the dependence of the model's prediction results on the radar's detection distance.

We thank the reviewer for this suggestion. We performed the suggested sensitivity analysis using the leave-one-event-out cross-validation errors of the hailsensor observations, grouped by hailsensor region and as a function of the distance to the closest radar (Figs. R3 and R4, shown to the reviewer only). The hailsensors are located at about 10–15 km (Ticino), 45–65 km (Napf), and 85–97 km (Jura) from the closest radar.

The median errors are close to zero in all three regions for moment 4, whereas moment 2 shows a slight underestimation in the Jura region (median error of about −0.2 in log10 units; Fig. R3a, c). The absolute errors in log10 units increase weakly with distance, with linear fits increasing from about 0.21 to 0.33 for moment 2 and from about 0.29 to 0.39 for moment 4 between 10 and 95 km (Fig. R3b, d). The absolute relative errors show no clear trend with distance and are dominated by a few large outliers occurring in all three regions (Fig. R4). In all cases, the spread of the errors within each region is much larger than the differences between regions.

The weak increase of the errors with distance and the slight underestimation of moment 2 in the Jura region are consistent with the expected effects of beam broadening and of the higher altitude of the lowest beams at far range, which can lead to an underestimation of volume-based features (Sect. 4.6). However, distance and region are confounded in this analysis: each region lies at a distinct distance from the radars, is sampled on partly different event days, and the Jura region contains only a small number of observations. The effect of radar distance therefore cannot be separated from regional differences in storm characteristics with the available data. Overall, the analysis does not indicate a strong dependence of the retrieval performance on the distance from the radar within the range covered by the hailsensors. Given the absence of a clear range dependence and the length of the manuscript, we decided not to include these figures in the revised manuscript.

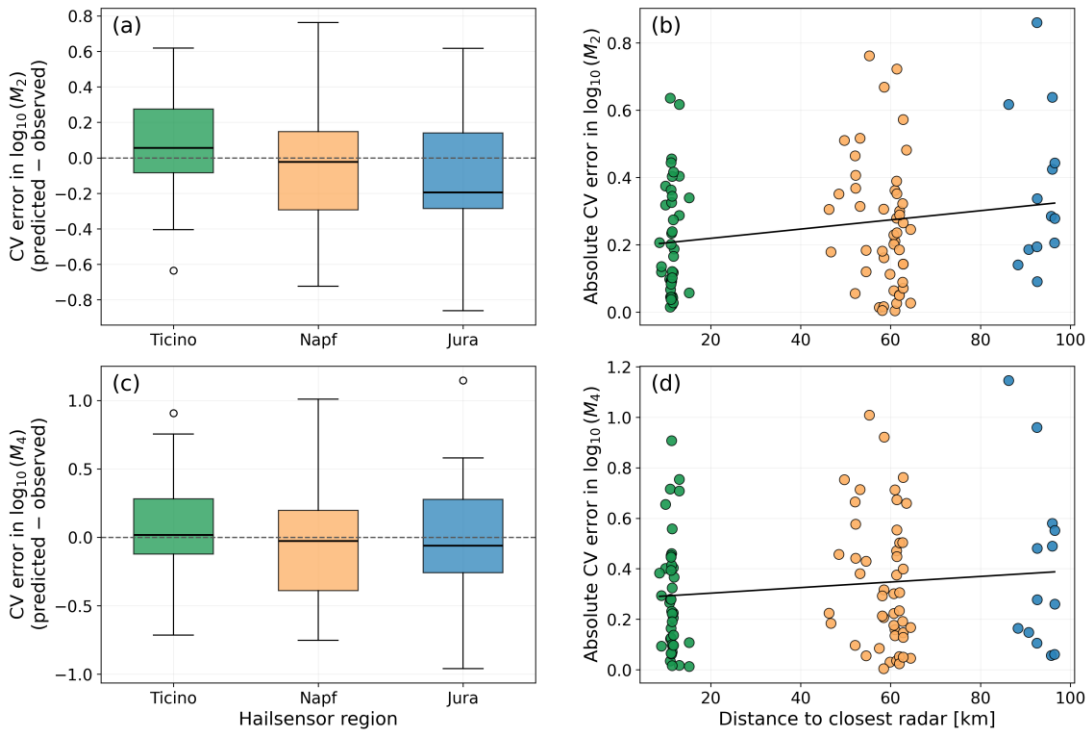

Figure R3: Leave-one-event-out cross-validation errors in log10 units for moment 2 (a, b) and moment 4 (c, d), grouped by hailsensor region (a, c) and as a function of the distance to the closest radar (b, d; black lines: linear fits).

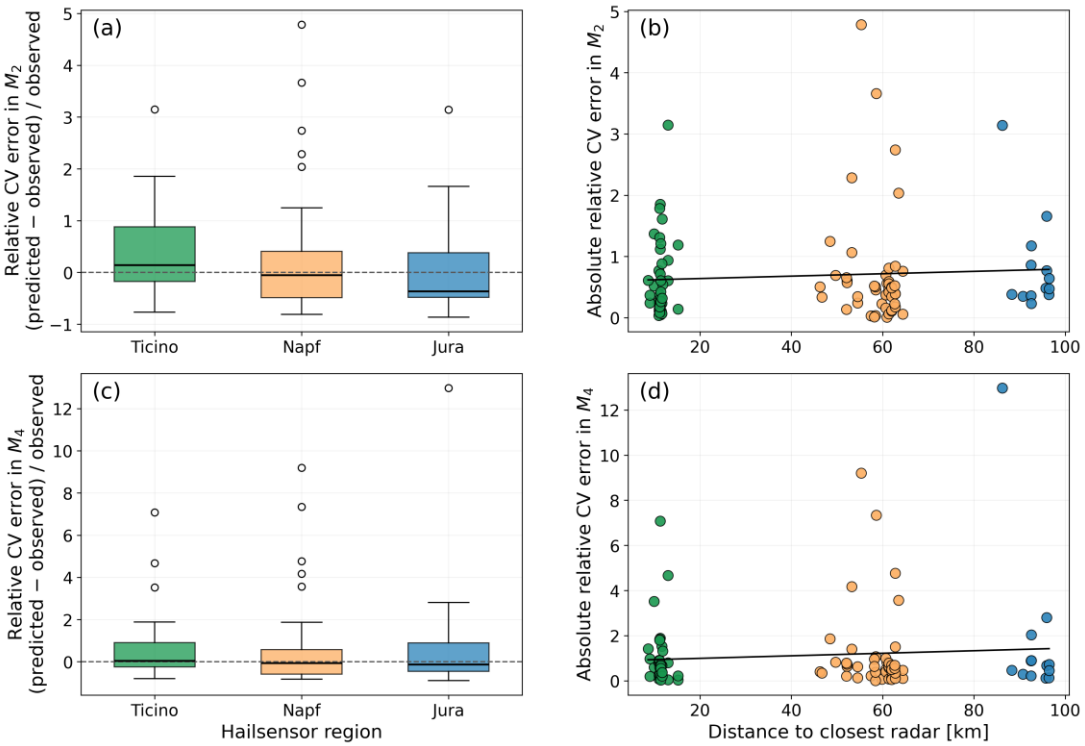

Figure R4: As Fig. R3, but for the relative cross-validation errors, (predicted − observed)/observed, of moments 2 and 4.

Extreme hailstones are more destructive in nature, although the samples of extreme hailstones are extremely limited. It is still recommended to verify the performance of the model in Large-diameter hailstones events. Additionally, hailstones and strong convective precipitation share certain similarities in radar characteristics, and it is necessary to verify whether the model can still predict the presence of hailstones in extreme precipitation events.

We thank the reviewer for this comment. We agree that large and extreme hail is the most relevant for damage and, at the same time, the most difficult to evaluate given the limited number of observations. The revised manuscript evaluates the performance for large hail with all three ground-based datasets:

– Hailsensors: events with hailstones larger than 2 cm are underrepresented in the training dataset, and the model performance decreases for increasing diameters (Sect. 4.2). The a posteriori adjustment reduces the underestimation of the upper tail of the HSND, although the cross-validation results still underestimate large hail. We state this limitation explicitly.

– Drone surveys: the newly included survey at Hettlingen (28 August 2026) observed hailstones with equivalent diameters up to 56.5 mm (Table A1), the largest in our evaluation datasets. The melting-corrected radar-retrieved HSND reproduces the observed distribution up to about 50 mm (Fig. 7d), and Hettlingen has the highest drone-observed and radar-retrieved moment 4 among the five surveys (Fig. 8b).

– Crowd-sourced reports: the event of 28 June 2021, for which severe damage was documented at the ground, is analysed in detail (Sect. 4.4.1, Figs. 9 and 10). While MESHS assigns high maximum diameters to both main hail tracks, the HSND-derived characteristic diameters, total kinetic energy, and number of hailstones larger than 2 cm reproduce more distinctly the contrast in severity between the two tracks indicated by the crowd-sourced reports. In addition, the evaluation over all 137 event days includes the reports of the largest size class (D > 37 mm).

Nevertheless, hail larger than about 6 cm is not represented in our evaluation datasets (the largest hailstone observed by the drones, at Hettlingen, measured 56.5 mm), and the performance of the model for such sizes cannot be verified with the available data. As stated in Sect. 4.6, additional drone and hailsensor observations of extreme events are required for a more thorough assessment.

Regarding the distinction between hail and heavy rain, we note that the proposed model is not a hail-detection algorithm: it estimates the HSND under the assumption that hail is present, while hail detection relies on the operational POH product. This point is now discussed in a new paragraph of Sect. 4.6 (reproduced below). Unlike the HSND retrieval, which includes two ρHV-based predictors, MESHS depends only on the height of the 50 dBZ echo top above the 0 °C isotherm and has no polarimetric information to discriminate between hail and rain. This may explain why, for the event of 28 June 2021, MESHS assigns high maximum diameters to both main tracks, whereas the HSND-derived quantities reproduce the weaker severity of the second track indicated by the crowd-sourced reports and the reported building damage (Schmid et al., 2024, Fig. 10i). In the detailed view of the main corridor, the 50 dBZ echo-top height of Z_V is high along the entire track, whereas the ρHV-based volume is largest where the largest hail was reported (Sect. 4.4.1, Figs. 9 and 10).

The corresponding passages of the revised manuscript read as follows (Sects. 2.1.2, 3.5.3, and 4.6, respectively):

> "In this study, POH is used as a reference for hail detection, while MESHS serves as a benchmark for comparison with the characteristic diameters of our radar-retrieved HSND (Sect. 3.4.6)."
> […]
> "The radar data were restricted to grid cells within the TRT thunderstorm-cell areas where POH was at least 90 % and both MESHS and the HSND retrieval were valid; days without such grid cells were discarded."
> […]
> "The proposed retrieval is not a hail-detection algorithm: it estimates the HSND under the assumption that hail is present, while hail detection relies on the operational POH product (Sect. 2.1.2). In intense rain without hail, the two ρHV-based predictors, which are sensitive to the decrease in ρHV associated with melting hail or hail growing in the wet regime (Sect. 4.1.2), are expected to be smaller than in hail. Nevertheless, the retrieval would likely produce non-zero estimates where the rain is intense, because the maximum and the sum of Z_V in the column increase with reflectivity. For this reason, the HSND-derived quantities were considered only where POH was at least 90 % (Sect. 3.5.3), and the retrieval should be combined with a hail-detection criterion such as POH in applications."

The introduction section provides a detailed review of existing radar methods and ground observation techniques. However, the presentation of the innovative aspects of this study and the introduction of the specific technical approach are somewhat slow. Please streamline the introduction and emphasize the core innovations.

We thank the reviewer for this comment. We have kept the review of existing radar methods and ground observation techniques, which we consider useful for readers from both the radar and the ground-observation communities. To introduce the objective of the study earlier, the first paragraph of the introduction now states the gap addressed by this work, namely that current operational radar products do not provide the full distribution of hail sizes and numbers. In addition, the main novelties of the work are now explicitly listed in the paragraph presenting this study.

The corresponding passages of the revised manuscript read as follows:

“Weather radar provides the spatial coverage needed to observe hail at these scales; however, current operational radar products provide only the probability of hail or an estimate of the maximum hail size, rather than the full distribution of hail sizes and numbers.”
[…]
“The main novelties of this work are: (i) the retrieval of the full HSND and associated features (e.g., kinetic energy), together with an estimate of their uncertainty, at high spatiotemporal resolution (1 km, 5 minutes) across Switzerland, overcoming the spatial limitations of ground-based measurements; and (ii) a large and diverse reference dataset combining an automatic hailsensor network for model training with drone surveys and crowd-sourced reports from 137 hail-event days for an independent evaluation.”
