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 #4,

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 major concern 2).
- 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 (shown in the response to the other referee).
- 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 #4

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

Manuscript excerpts

General Assessment

This manuscript presents a novel methodology for retrieving hail size number distributions (HSND) from polarimetric C-band radar data using generalized additive models (GAMs) trained against automatic hailsensor measurements. The work builds upon prior double-moment normalization theory (Ferrone et al., 2024) and extends it to radar-based retrieval, with validation against drone photogrammetry and crowd-sourced reports. The topic is timely and relevant for AMT, and the authors have assembled an impressive multi-source observational dataset spanning multiple years.

However, the manuscript has several significant methodological and interpretational issues that need to be addressed before publication. The validation approach, while ambitious, contains circularities and inconsistencies that undermine confidence in the claimed performance. The uncertainty quantification is insufficient, and several critical assumptions are not adequately justified or tested.

We thank the reviewer for the thorough evaluation of our manuscript and for recognizing the relevance of the topic and the value of the multi-source observational dataset. The comments prompted substantial additional analyses, summarized on the first page of this document. In particular, we expanded the independent evaluation (five drone surveys and crowd-sourced reports from 137 hail-event days), introduced event-specific melting rates, complemented the feature selection with AIC and BIC, and quantified the uncertainty of the retrieval. Detailed replies to each comment are given below.

Major Concerns

1. Spatial Extrapolation of Hailsensor Data (Appendix C)

The spatial extrapolation procedure is the cornerstone of the training dataset but raises serious methodological questions:

The assumption that temporal variations at a fixed sensor represent spatial variability within the storm "segment" is untested and likely problematic. Hailfall is notoriously heterogeneous at small scales; assuming stationarity over 5 minutes and extrapolating up to 200 m is a strong assumption without validation.

Lines 317-318 state that "The added value of the spatial extrapolation was found to be limited, with meaningful contributions extending to a maximum distance of roughly 200 m." If the extrapolation adds so little value, why is it included? This suggests that the method is

effectively training on point measurements against radar volumes that are ~5 orders of magnitude larger—a fundamental scale mismatch.

The extrapolation creates pseudo-replicates in the training data. Multiple extrapolated positions from the same sensor observations are treated as independent training samples, artificially inflating the effective sample size and likely leading to overconfident predictions.

We thank the reviewer for these comments, which helped us clarify the purpose and limits of the spatial extrapolation.

Scale mismatch. We agree that there is a large scale difference between a 0.2 m² hailsensor and a radar resolution volume. This does not prevent a model from being trained on point measurements, as is routinely done when relating radar observations to rain gauges or disdrometers. By using many observations from several hailsensors and events, the GAM represents the average relationship between the radar features and the HSND moments, whereas the local variability that cannot be resolved at the radar scale contributes to the scatter of the observations around this relationship. This scatter is part of the residual uncertainty and is not represented by the fitted relationship itself. This aspect of the scale mismatch, which affects not only the evaluation but also the model training, is now discussed explicitly in Sect. 4.6.

Purpose of the extrapolation. In a first stage of this work, we trained the model directly on the hailsensor observations at the sensor locations, without spatial extrapolation. This configuration yielded lower correlations, both on the training data and in cross-validation, which motivated the development of the extrapolation to reduce the influence of the scale difference between the sensors and the radar. The purpose of the extrapolation is now explained in more detail in Sect. 3.3 (excerpt below): it does not increase the sampling area of the hailsensors, but relates their observations to the radar features at the positions of the storm segment that produced them.

Stationarity assumption. We agree that hailfall is heterogeneous at small scales and that the assumption of a stationary storm structure cannot hold over long distances. For this reason, the maximum extrapolation distance was not prescribed but determined empirically (Appendix D, Fig. A5, already included in the original manuscript). The correlation between the radar features and the extrapolated moments is highest for a maximum extrapolation time of 20 s (about 200 m) and decreases steadily for longer distances, consistent with the small-scale variability of hailfall. The extrapolation was limited accordingly. The sensitivity analysis therefore provides an empirical constraint on the range over which the assumption is supported by the data.

Value of the extrapolation. The sentence quoted by the reviewer was misleading: it did not mean that the extrapolation has little value, but that its validity is limited to short distances. It has been rephrased in Sect. 3.3, where we now also explain the purpose of the procedure.

Pseudo-replicates. We agree that samples from the same sensor, and more generally from the same event, may not be independent. The training dataset comprises 197 samples on 35 event days, with hail recorded by 1 to 13 hailsensors depending on the event day (Appendix C and Table A1). The hailsensor observations of the 2026 season were also checked, but no event satisfied the basic selection criteria of the dataset, such as the minimum of 30 impacts per event (Sect. 2.2.1). This possible dependence does not inflate the evaluation scores, because the event day is treated as the independent unit throughout the analysis: the cross-validation withholds entire event days, so that no sample of a withheld event is used for training (Sect. 3.4.2), and the uncertainty is estimated by resampling entire event days (Sect. 3.4.4). In addition, the shuffled-target test presented in our reply to the next concern (Fig. R1) shows that the selected features cannot be explained by chance correlations.

While revising Appendix C, we also corrected the length of the time interval used to group the hailsensor impacts, which is 30 s, as selected in the sensitivity analysis (Appendix D), and not 20 s as erroneously stated in the original manuscript.

The corresponding passages of the revised manuscript read as follows (Sect. 3.3, Sect. 4.6, and Appendix C, respectively):

"The hailsensor observations were spatially extrapolated along the storm trajectory, using the speed and direction of the associated TRT cell. The extrapolation does not increase the sampling area of the hailsensors and does not solve the sampling limitations associated with their small size: all extrapolated samples remain tied to the hailstones actually recorded by the sensors. Its purpose is to relate these observations, accumulated over the 5-minute radar interval, to the radar features at the positions of the storm segment that produced them. Since the radar features are derived from a 3D composite with 250 m resolution, they can vary over short distances, and the extrapolation provides information on this variability that would otherwise not be available. […] The correlation decreases for extrapolation distances beyond about 200 m (20 s), consistent with the small-scale variability of hailfall; the extrapolation was therefore limited to this distance."
[…]
"The scale difference mentioned in (iii) deserves particular attention, as it already affects the model training: hailsensors sample only 0.2 m², whereas the radar features represent volumes of the order of cubic kilometers. The temporal accumulation over the 5-minute radar interval and the spatial extrapolation along the storm trajectory reduce, but do not remove, this difference. By training on many observations from different sensors and events, the GAM represents the average relationship between radar features and HSND moments, while the local variability that cannot be resolved at the radar scale contributes to the residual scatter of the estimates. A similar mismatch affects the evaluation: radar-based retrievals are representative at the kilometer scale, whereas drone-based measurements and crowd-sourced reports sample much smaller areas."
[…]
"From the resulting transect, in the training of the radar-based model (Sect. 3.4.1), we only used as target HSNDs that count more than 10 hailstones and which the maximum time distance from the actual hailsensor observation to the extrapolated time is 20 s, corresponding to a maximum space distance of approximately 200 m (see Appendix D). In total, the training dataset comprises 197 samples on 35 event days. Depending on the event day, hail was recorded by 1 to 13 hailsensors (Table A1). Since samples from the same event day may not be independent, entire event days were treated as the independent unit in the cross-validation (Sect. 3.4.2) and in the uncertainty estimation (Sect. 3.4.4)."

2. Feature Selection and Model Overfitting

Section 3.4.2 describe the feature selection process, which raises serious red flags:

The features were selected based on cross-validated Spearman correlation with the extrapolated moments (which are derived from the same hailsensor data). This creates a risk of information leakage, especially given the limited number of independent events (35 days).

The model uses 4 features to predict two moments from only 35 independent days. With this limited sample size, the risk of overfitting is high, despite the leave-one-event-out cross-validation. The authors acknowledge that "the limited amount of reference data from hailsensors makes robust cross-validation challenging" (lines 491-492), yet proceed with the selected features.

Table A2 shows only marginal improvements in PCC when adding features (e.g., M4 PCC: 0.64 → 0.65 → 0.66 → 0.68). The incremental gain is modest, and the model complexity increases substantially. The authors should report AIC or BIC to justify the additional features.

We thank the reviewer for these comments. We address the three points below. In addition, following a suggestion of the other referee, we tested the selection procedure with randomly shuffled targets.

Information leakage. Selecting predictors on the basis of their relationship with the targets is inherent to any supervised model; leakage occurs only if the data used to evaluate the model also influence its fitting or selection. In the stepwise selection, a candidate feature was retained only if it improved the Pearson correlation of the leave-one-event-out cross-validation, i.e., the skill obtained on entire event days withheld from model fitting (Sect. 3.4.2). To further assess whether the selection procedure could produce spurious skill, we performed a shuffled-target test (Fig. R1): 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), compared with 0.61 with the real targets. Moreover, the four selected features fitted to the shuffled targets yielded 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 ($p \leq 0.01$). The selected features therefore reflect a genuine relationship between the radar observations and the hailsensor HSND rather than accidental correlations or leakage. Finally, the drone and crowd-sourced observations, which provide the independent evaluation of the retrieval, were not used in any step of the selection.

Number of features and sample size. The training dataset comprises 197 samples on 35 event days (see our reply to the previous concern). We agree that the event day is the relevant independent unit, and it is the unit withheld in the cross-validation. The flexibility of the GAMs is limited: each predictor is represented by a monotonic cubic spline with five basis functions and a fixed regularization strength (Sect. 3.4.1). This is supported by the small difference between the in-sample and cross-validated correlations (PCC of 0.66 vs. 0.62 for moment 2 and 0.72 vs. 0.68 for moment 4; Table A2). We agree that the limited amount of reference data makes robust discrimination among highly correlated predictors challenging, and we state that the specific selected features may partly reflect this limitation (Sect. 4.1).

Incremental gains and information criteria. As suggested, we computed AIC and BIC for each successive model (Table A2). Both criteria decreased with every added feature and for both moments (ΔAIC between −4.0 and −15.0 and ΔBIC between −1.3 and −8.7), indicating that the improvement in model fit justifies the additional model complexity. We note that the cumulative gain is larger than suggested by moment 4 alone: the cross-validated PCC of moment 2 increases from 0.55 to 0.62 with the four features. For moment 4, the reduction in BIC is small when the second and third features are added (−1.3 and −1.5), indicating that these features mainly improve moment 2, whereas the fourth feature yields the largest reduction for moment 4 (−8.5). This is now stated 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."

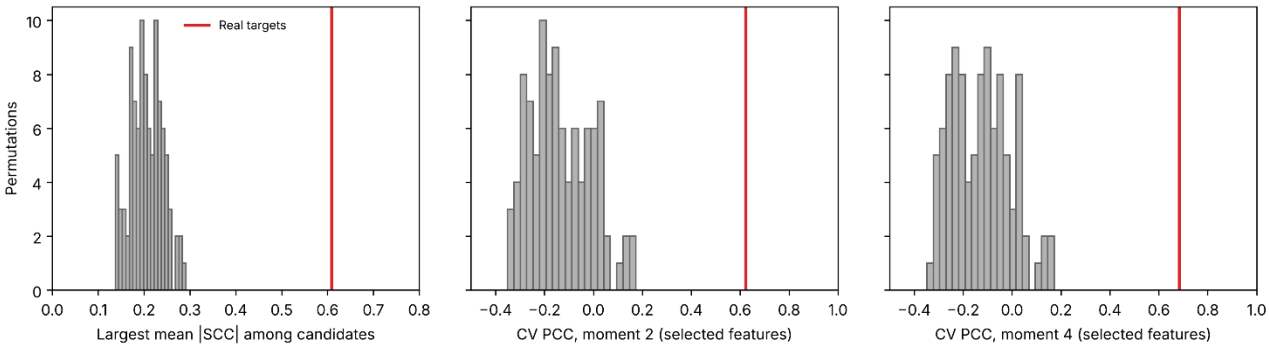

Figure R1: Shuffled-target test (100 permutations; shown to the reviewer only). Grey histograms: distribution over the permutations of (left) the largest mean absolute Spearman correlation between any candidate feature and moments 2 and 4, and of the leave-one-event-out cross-validated Pearson correlation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

3. Posterior Bias Adjustment (Section 3.4.3)

The bias adjustment procedure (quantile mapping + linear regression) is concerning:

This is essentially post-hoc calibration that uses the same training data to correct the model predictions. While applied in cross-validation, the procedure reduces the

The adjusted moments show increased RMSE compared to the unadjusted predictions (Table A2: M2 RMSE from 0.31 → 0.34; M4 from 0.41 → 0.44), yet the authors present the adjusted results as superior. This is counterintuitive—if the adjustment increases RMSE, in what sense is it improving performance?

The authors should clearly distinguish between calibration (improving probabilistic properties) and accuracy (reducing error). Quantile mapping improves distributional alignment at the cost of increased point-wise error, which should be discussed transparently.

We thank the reviewer for these comments, which prompted us to describe the purpose of the adjustment more transparently.

Use of the training data. We agree that the adjustment is a post-hoc calibration. In the cross-validation, however, it was estimated on the training events of each fold only and then applied to the withheld event. The adjusted cross-validation scores reported in Table A2 are therefore not influenced by the withheld data. This is now stated explicitly in Sect. 3.4.3.

Increase in RMSE and calibration versus accuracy. We agree with the reviewer that the adjustment does not improve the point-wise accuracy, and that calibration and accuracy should be clearly distinguished. A regression model minimizing the squared error shrinks its estimates towards the mean: this minimizes the RMSE, but underestimates the spread of the estimates and, in particular, the large values. In cross-validation, the standard deviation of the non-adjusted estimates is 0.25 for moment 2 and 0.38 for moment 4, compared with observed values of 0.39 and 0.56. After the adjustment, it becomes 0.39 and 0.55, while the PCC is preserved and the RMSE increases slightly, from 0.31 to 0.34 and from 0.41 to 0.44 (Table A2). Since the upper tail of the retrieved HSND, which is the most relevant part for hail impacts, is controlled by the estimated moments, we favour a distribution of the estimates consistent with the observations over a minimum RMSE. The effect is visible in the upper tail of the HSND, whose underestimation is reduced by the adjustment (Sect. 4.2, Fig. 6j–m). We do not claim that the adjusted estimates are more accurate in a point-

wise sense, and the revised manuscript now describes the adjustment as an improvement of calibration obtained at the cost of a slightly higher RMSE (Sects. 3.4.3 and 4.2).

The corresponding passages of the revised manuscript read as follows (Sects. 3.4.3 and 4.2, respectively):

> "To further stabilize the correction, we fit a linear regression model between predicted moments and their quantile-mapped counterparts, and use this model to obtain adjusted predictions. In the cross-validation, the adjustment was estimated on the training events of each fold only and then applied to the withheld event, so that the adjusted cross-validation scores are not influenced by the withheld data.
> The purpose of the adjustment is to correct the distribution of the estimated moments, in particular their reduced spread and the resulting underestimation of large values, rather than to reduce the point-wise error. It therefore improves the calibration of the estimates at the cost of a slightly higher RMSE (Sect. 4.2)."
> […]
> "Despite this limitation, the a posteriori adjustment substantially improves the spread of the estimated moments: in cross-validation, the standard deviation of the estimates increases from 0.25 to 0.39 for moment 2 and from 0.38 to 0.55 for moment 4, matching the observed values of 0.39 and 0.56 (Table A2). This improvement is achieved while preserving the PCC of the non-adjusted GAM, but at the cost of a slightly higher RMSE (from 0.31 to 0.34 for moment 2 and from 0.41 to 0.44 for moment 4). This trade-off is expected: a regression model minimizing the squared error shrinks its estimates towards the mean, which minimizes the RMSE but underestimates the spread and, in particular, the large values. Since the upper tail of the retrieved HSND is controlled by the estimated moments, we favour a distribution of the estimates consistent with the observations over a minimum RMSE."

4. Validation Against Drone Data (Section 4.3)

The drone validation contains critical issues:

The melting correction (0.5 mm/min) is arbitrary for the 2022 event, as acknowledged in lines 571-573.

The statement "the HSND obtained via double-moment normalization from hailsensor observations agrees more closely with the radar-based HSND than with the drone-retrieved HSND" (lines 577-578) undermines the value of drone data as independent validation. If the hailsensor (30 m away) matches the radar better than the drone observations, why trust the drone validation?

Figure 7b shows that a 10-minute melting correction gives "great agreement" while the 20-minute correction does not. This is curve-fitting rather than validation, as the melting correction is not independently constrained for this event.

We thank the reviewer for these comments. The drone evaluation has been substantially revised: it now includes five surveys, three of which were conducted in 2026, and the melting correction is based on event-specific melting rates.

Melting correction (first and third points). We agree that the melting rate of 0.5 mm min $^{-1}$ derived from a single event was arbitrary for the 2022 event, and that the comparison of the 10- and 20-minute corrections in the original manuscript was not independently constrained. Both have been removed. Melting rates are now estimated separately for each event with repeated drone surveys of a common area (Entlebuch, Winterthur, Hettlingen, and Niederhasli): hailstones from consecutive surveys were ordered by size and paired by rank, and the decrease in equivalent diameter divided by the time between the surveys provides an estimate of the melting rate, grouped by initial-diameter class (Sect. 3.5.2; Fig. R2, from a companion study that will be submitted shortly). The median rates range from 0.18 to 0.86 mm min $^{-1}$ and differ substantially among events. Only for Locarno-Monti, where a single survey is available, the rate (about 0.26 mm min $^{-1}$) was set to the midpoint of the rates estimated for the two events with the most similar observed

HSNDs (Entlebuch and Winterthur); it was not adjusted to improve the agreement with the radar retrieval. The radar-retrieved HSND is corrected from each radar timestamp to the drone-observation time using the actual elapsed time. Finally, the sensitivity of the comparison to the melting rate is quantified by perturbing the rates by ±0.1 mm min⁻¹ (Figs. 7 and 8). This sensitivity is substantial, in particular at Locarno-Monti, and the melting correction is explicitly discussed as a source of uncertainty of the drone comparison (Sect. 4.3). In addition, at Niederhasli, the time elapsed between the end of the hailfall and the survey is uncertain by about ±5 min, which is now also mentioned in Sect. 4.3.

Hailsensor near the Locarno-Monti survey (second point). The statement quoted by the reviewer has been removed, together with the HSND derived from the hailsensor located near the Locarno-Monti survey. This sensor recorded only 12 hailstones, which is below the minimum number required for the training dataset. We removed it to be coherent with the hailsensor data used throughout the study, since an HSND derived from so few hailstones is not a robust reference. Our intention was not to question the value of the drone observations, but to note that agreement with more than one source of ground observations is encouraging. Drone-based HSNDs are affected by uncertainties in the ground sampling distance and surveyed area, the melting correction, partially hidden hailstones, image resolution and contrast, and the manual annotations used to train the detection model, as discussed in Sect. 4.3. Nevertheless, they provide HSNDs over areas several orders of magnitude larger than a hailsensor and thus constrain at least the order of magnitude of the hail event at a scale closer to that of the radar. With five surveys, including three new events from 2026, the drone comparison now provides a broader evaluation than in the original manuscript (Sect. 4.3).

The corresponding passages of the revised manuscript read as follows (Sects. 3.5.2 and 4.3, respectively):

"Melting rates were estimated from the repeated drone surveys available for Entlebuch, Winterthur, Hettlingen, and Niederhasli. […] 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. […] 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."

[…]

"The drone-derived HSND is affected by uncertainty in the estimated ground sampling distance and surveyed area and in the melting correction (at Niederhasli, for instance, the time elapsed between the end of the hailfall and the survey is uncertain by about ±5 min), as well as by hailstones hidden from view (particularly small stones), image resolution and contrast, and the instance-segmentation retrieval (Portmann et al., 2025). The latter also depends on the manual annotations used as training reference, for which small hailstones can be difficult to identify under low resolution or weak contrast."

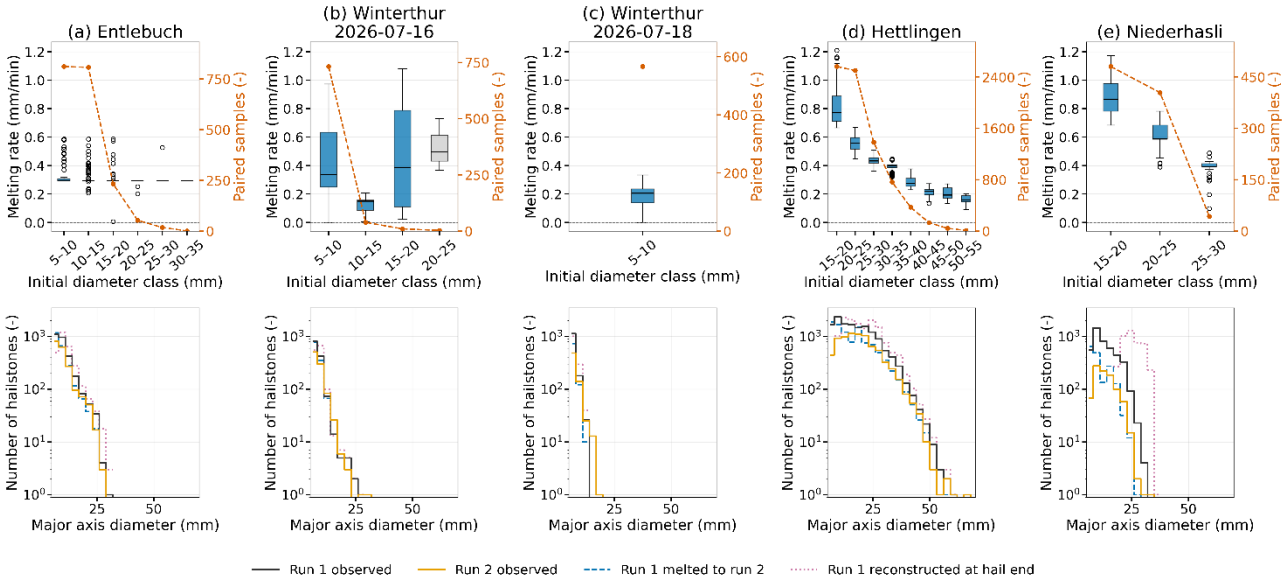

Figure R2: Melting rates estimated from repeated drone surveys (shown to the reviewer only; from a companion study based on drone-retrieved HSNDs, Guidicelli et al., in preparation). Top: melting rates per initial-diameter class (boxplots, left axis) and number of paired hailstones (orange, right axis). 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.

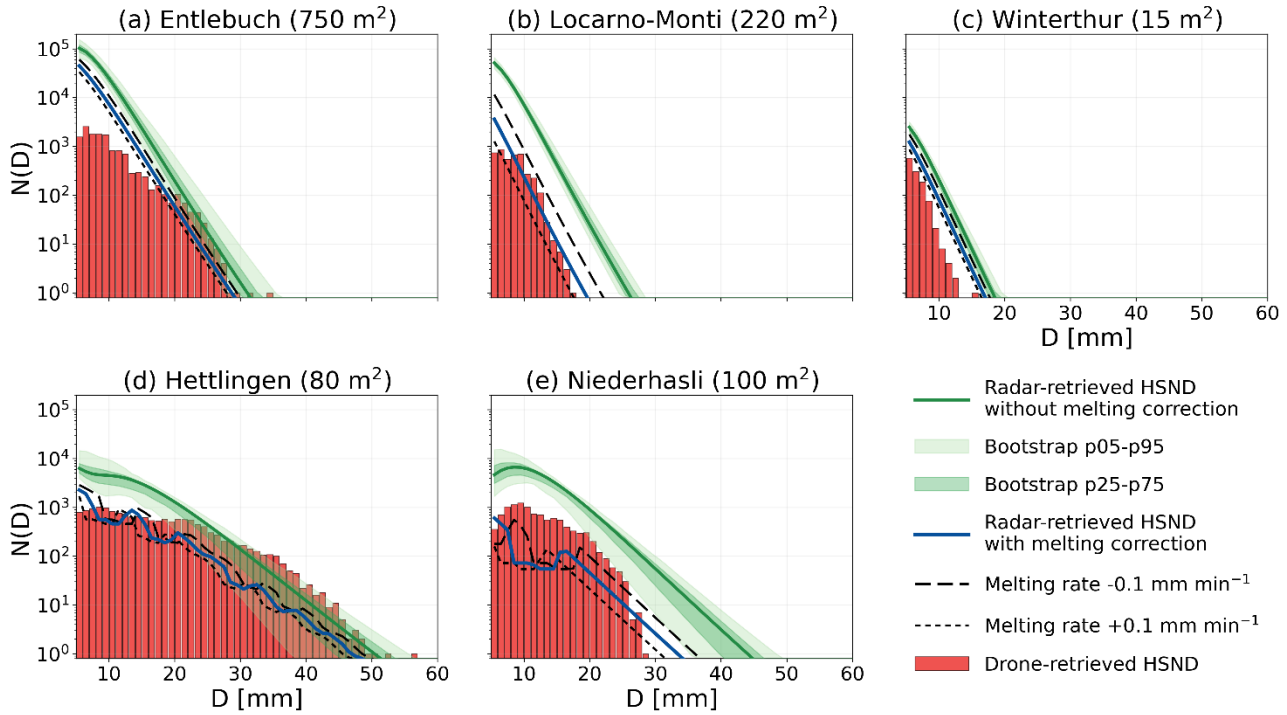

Figure 7 of the revised manuscript: drone-retrieved HSDs compared with the radar-retrieved HSNDs without (green) and with (blue) melting correction, including the sensitivity to a ±0.1 mm min⁻¹ perturbation of the melting rate (black dashed and dotted curves).

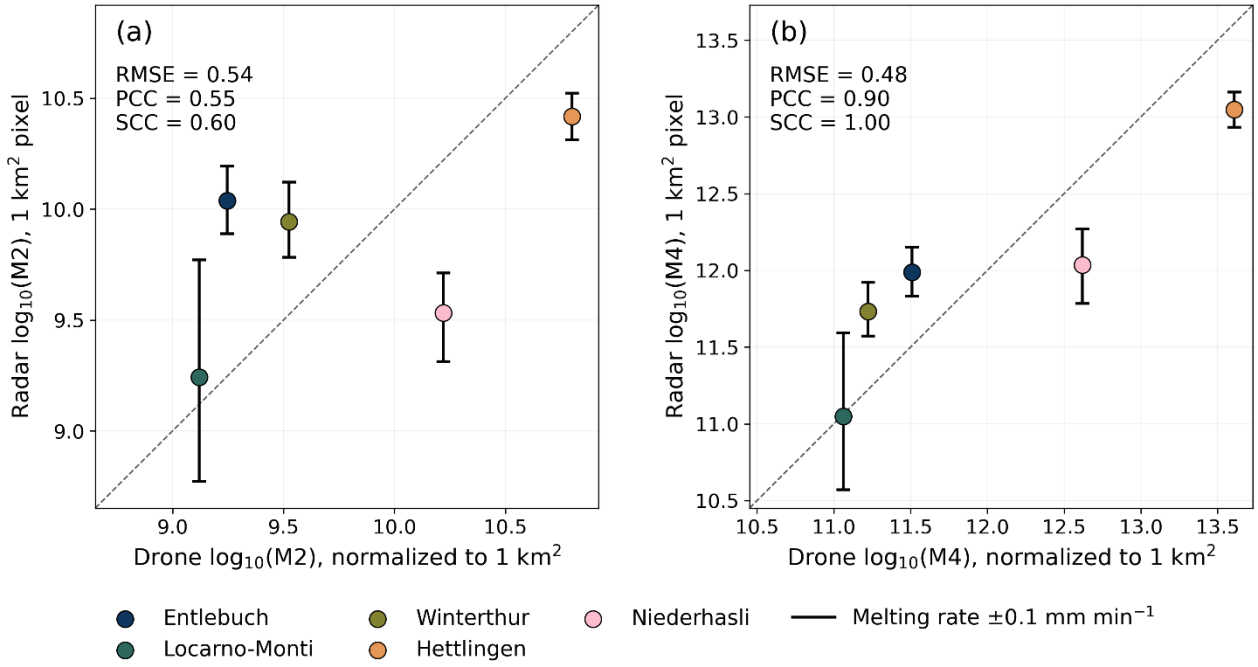

Figure 8 of the revised manuscript: drone-observed versus radar-retrieved moments 2 (a) and 4 (b) for the five drone survey sites; vertical bars show the sensitivity to a ±0.1 mm min $^{-1}$ perturbation of the melting rate.

5. Validation Against Crowd-Sourced Reports (Section 4.4)

The crowd-sourced validation has a data leakage issue:

The spatial aggregation (3-11 km) and the use of maximum values for radar features compared against percentiles of reports creates an apples-to-oranges comparison. MESHS (also a maximum) is a more appropriate comparator. The authors should justify why maximum radar values should be compared against 75th/90th percentiles of reports.

The improved correlation of HSND percentiles over MESHS (Fig. 8) is not surprising: HSND percentiles are tunable parameters (D99.9, D99.99, etc.) that can be adjusted to match the reports, while MESHS is a fixed operational product. The authors should use a fixed percentile (e.g., always D99.9) and avoid cherry-picking the best performer.

We thank the reviewer for these comments. The crowd-sourced evaluation has been substantially extended: it now includes all 137 hail-event days from 2016 to 2026 that passed the selection criteria, in addition to the three events analysed in detail (Sect. 4.4, Figs. 11 and 12, reproduced below).

Maximum radar values versus percentiles of the reports. The comparison aims at relating the largest hail sizes on both sides. On the radar side, the maximum within the neighbourhood of MESHS or of the characteristic diameters describing the tail of the HSND represents the largest hail sizes retrieved in that area. On the report side, crowd-sourced reports are categorical and subjective, and users tend to report the largest hailstones categories. The maximum of the reports is therefore also sensitive to isolated overestimated reports, and the 75th and 90th percentiles were used as a version of the upper part of the reported distribution that is robust to such outliers. However, following the reviewer's comment, the revised manuscript additionally compares all radar products with the maximum of the reports (Fig. 11, dotted lines). The scores are generally poorer for the maximum than for the 75th and 90th percentiles, and they tend to

deteriorate with increasing aggregation radius, consistent with the influence of isolated maximum reports (Sect. 4.4.3). This rationale is now stated in Sect. 3.5.3.

Choice of the characteristic diameter. The purpose of evaluating several characteristic diameters in the original manuscript was to identify which characteristic of the retrieved HSND, if any, best represents the crowd-sourced reports. Following the reviewer's suggestion, the revised manuscript uses a fixed set of characteristic diameters in the main analysis: one percentile (D99.9) and one count-supported maximum diameter, Dmax [1/10 m²], defined as the largest diameter for which the retrieved HSND contains at least one hailstone per 10 m². The latter is more directly interpretable than a high percentile, since it refers to a hailstone density per unit area rather than to a fraction of the total number of hailstones. The sensitivity to the choice of the Dmax threshold, from one hailstone per km² to one hailstone per m², is shown separately (Fig. 12). For all thresholds, the correlations of Dmax with the reports over the 137 event days (PCC of 0.39–0.49) are higher than those of MESHS (0.34–0.39), whereas the RMSE depends on the threshold, as expected since the threshold determines the diameter level of Dmax. The higher correlations obtained with the HSND-derived diameters therefore do not depend on the selection of a particular threshold.

The corresponding passages of the revised manuscript read as follows (Sect. 3.5.3):

"The radar quantities were first summarized over each event day. The event maximum was retained for MESHS and the HSND characteristic diameters, whereas TKE and the number of hailstones larger than 2 cm, Nt,D>2cm, were accumulated over the event. The principal characteristic diameters examined were D99.9 and Dmax [1/10 m²] (Sect. 3.4.6); the other Dmax thresholds were used in a sensitivity analysis (Fig. 12).
The comparison was performed at the centres of the 1-km radar grid cells. Around each grid centre, the reports were spatially aggregated within circular neighbourhoods of 3, 5, 7, 9, and 11 km radius, requiring at least 6, 10, 14, 18, and 22 reports, respectively, and their 75th percentile, 90th percentile, and maximum were calculated. Since crowd-sourced reports are categorical and subjective, and users tend to report the largest hailstone categories, the upper percentiles were used in addition to the maximum to reduce the influence of isolated overestimated reports while still representing the large hailstones. MESHS and the HSND-derived quantities were aggregated over the same neighbourhoods by retaining their maximum value. In this way, the largest hail sizes retrieved within each neighbourhood, described by MESHS or by the characteristic diameters of the HSND tail, are compared with the upper part of the distribution of the reported sizes, of which the 75th and 90th percentiles represent a version that is robust to outliers."

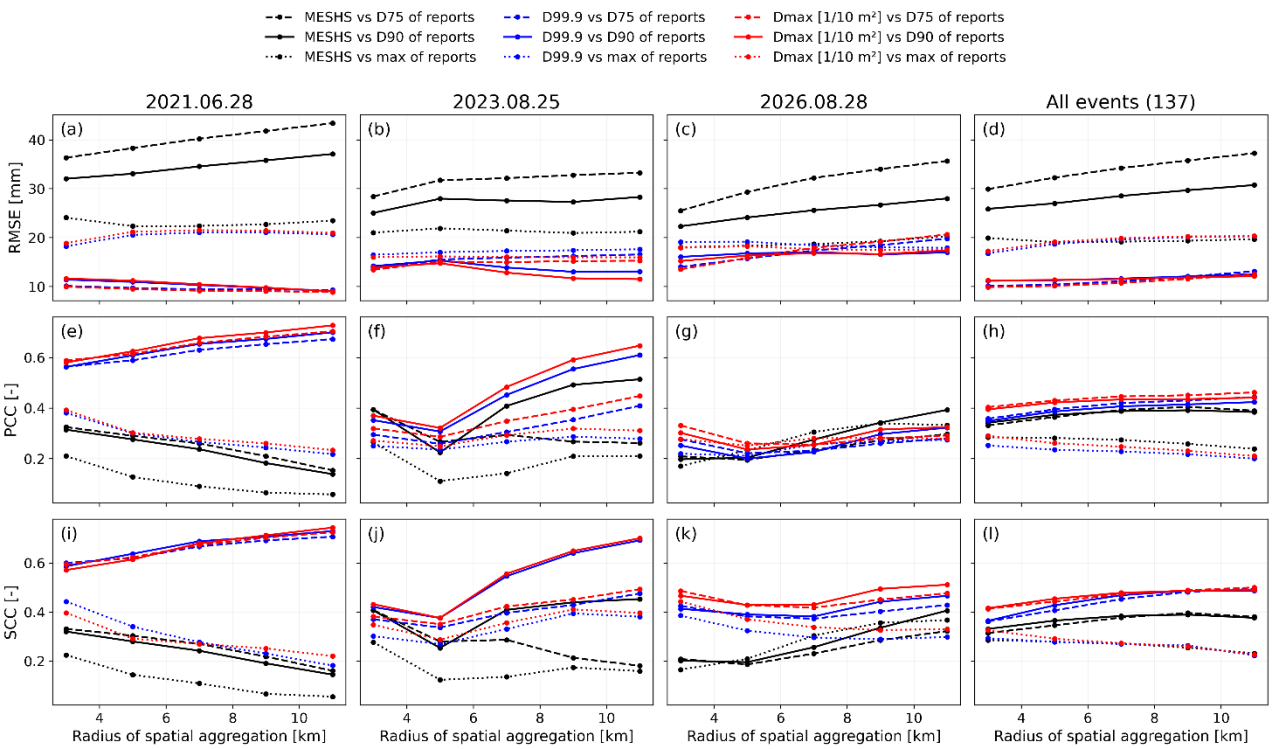

Figure 11 of the revised manuscript: RMSE, PCC, and SCC as functions of the spatial aggregation radius for MESHS, D99.9, and Dmax [1/10 m²] compared with the 75th and 90th percentiles and the maximum of the crowd-sourced reports, for the three individual events and for all 137 event days.

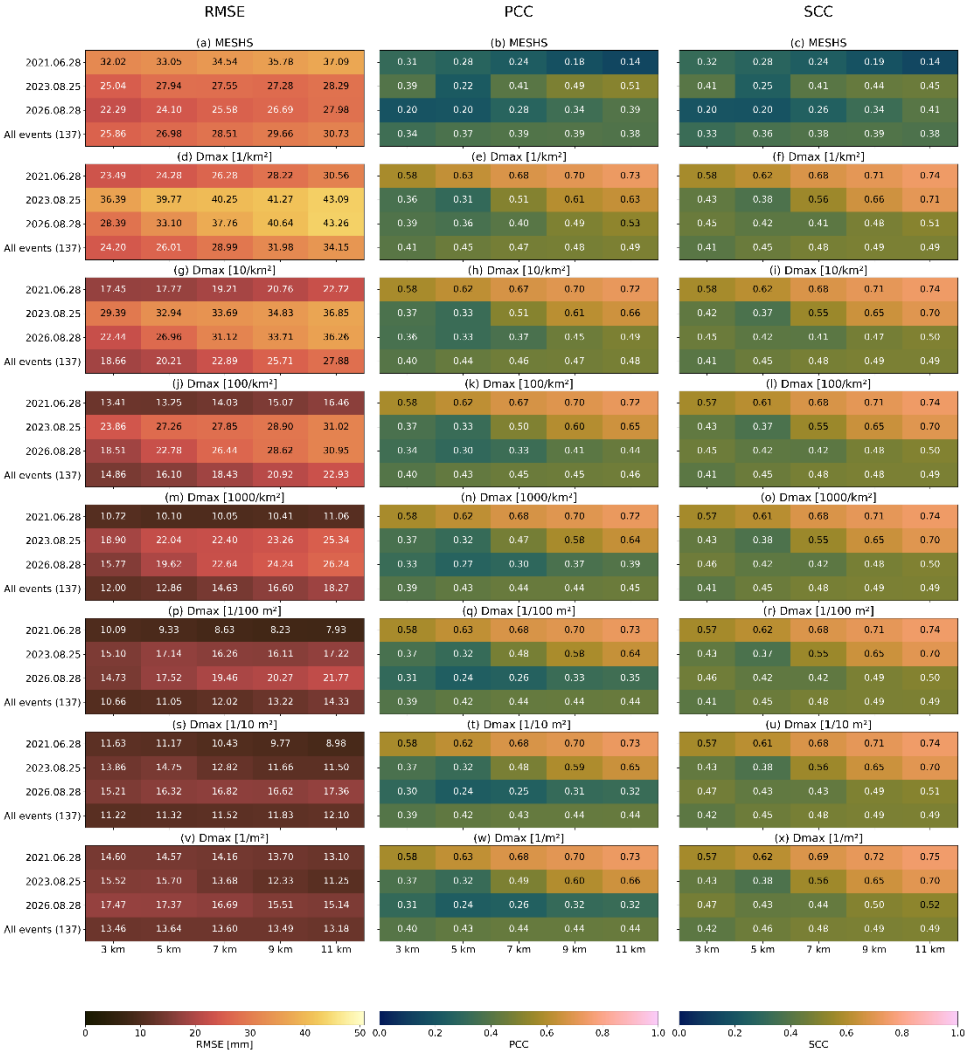

Figure 12 of the revised manuscript: RMSE, PCC, and SCC for MESHS and count-supported maximum diameters with minimum hailstone densities from one per km² to one per m², compared with the 90th percentile of the crowd-sourced reports.

6. Non-Rayleigh Scattering Treatment (Section 4.1.1)

The discussion of non-Rayleigh scattering is inadequate:

Lines 495-496 acknowledge that C-band resonance effects occur for hailstones >2 cm, which directly affects the ZV features used in the GAM. The authors state that moment 2 is "generally less sensitive" and moment 4 "is more directly impacted," but provide no quantitative assessment of how resonance affects their specific features.

Given that the hailsensor data include events with hailstones up to 4.2 cm (Fig. 3l), the GAM is being trained on data that may have nonlinear and non-monotonic relationships between radar features and ground truth due to resonance. The monotonic splines used in the GAM may be inappropriate for these conditions.

Quantitative assessment of resonance effects. We agree that resonance affects the radar variables for hailstones larger than about 2 cm at C-band. A quantitative assessment of its effect on our specific features would require scattering simulations for realistic hailstone shapes and densities, propagated through the 3D composite and the feature extraction, which is beyond the scope of this study. Our approach is empirical: the GAMs are trained directly on the relationship between the observed radar features and the HSND moments measured at the ground, so that the average effect of resonance on the features is implicitly contained in this relationship. Moreover, the selected features are integrated over the vertical column or over volumes, and the moments integrate over the whole HSND. Both integrations reduce the sensitivity to the size-dependent oscillations that characterize the scattering of individual hailstones. The remaining effect of resonance is discussed qualitatively in Sect. 4.1.1, where we note that the maximum of Z_V correlates more strongly with moment 2 than with moment 4, consistent with a reduced sensitivity of reflectivity to large hailstones. This limitation is now stated explicitly in Sect. 4.1.1.

Monotonic splines. The monotonic constraint was chosen for robustness. Unconstrained splines produced non-monotonic relationships driven by the few samples available at the ends of the observed range of the moments, i.e., they overfitted the training data. Physically, although resonance causes reflectivity to oscillate with size for individual hailstones, for a distribution of hailstones the integration over sizes smooths these oscillations, and reflectivity continues to increase with the size of the distribution, albeit more slowly than in the Rayleigh regime (Ryzhkov et al., 2013). Similarly, the volume with pHV < 0.97 and Z_H > 50 dBZ is expected to increase with the number and size of hailstones. Importantly, the monotonic constraint does not prevent a saturation of the relationship at large values: it only excludes a decrease, which is not supported by the available data. This justification is now added to Sect. 3.4.1.

The corresponding passages of the revised manuscript read as follows (Sects. 3.4.1 and 4.1.1, respectively):

> "For both moments, the GAM uses four predictor features, each represented by a third-order monotonic cubic spline with five basis functions and a regularization strength (λ) of 0.6 (see Sect. 3.4.2). The final estimate is obtained as the sum of the smooth contributions from all predictors (Eq. 3). The monotonic constraint was chosen for robustness: unconstrained splines produced non-monotonic relationships driven by the few samples available at the ends of the observed range of the moments. The constraint excludes a decrease of the estimated moments with increasing feature values, but still allows a saturation of the relationship, as may be expected for reflectivity-based features at C-band for large hailstones affected by resonance (Ryzhkov et al., 2013)."
> […]
> "Moreover, since the GAMs are trained directly on the relationship between the observed radar features and the HSND moments measured at the ground, the average effect of resonance on the selected features is implicitly contained in the fitted relationships. A quantitative assessment of resonance effects on these features would require scattering simulations for realistic hailstone shapes and densities, propagated through the 3D composite and the feature extraction, which is beyond the scope of this study."

Minor Issues

1. Sample Size and Representativeness

Only 35 days of hailsensor data are used, spanning 2018-2025 (7 years). This is a small sample for a statistical model with 4 features and smooth splines. The authors should report the effective sample size considering the spatial extrapolation (which likely reduces independence).

The training dataset is biased toward intense events (≥30 hailstones per event), which the authors acknowledge. This may explain the systematic overestimation of small hailstones in

the drone comparison. However, the authors do not propose a correction or weighting scheme for this selection bias.

We thank the reviewer for these comments.

Effective sample size. The training dataset comprises 197 samples on 35 event days, as now stated in Appendix C. As discussed in our reply to major concern 1, samples from the same event are not independent, and the event day, rather than the sample, is treated as the independent unit in the cross-validation and in the uncertainty estimation. The effective number of independent units is therefore 35 event days.

Selection bias toward intense events. We agree that the requirement of at least 30 hailstones per event and more than 10 hailstones per 5-minute interval may bias the training dataset toward intense events. This requirement is, however, necessary: given the small sampling area of the hailsensors (0.2 m²), HSND moments estimated from fewer hailstones are dominated by sampling noise and do not provide a reliable reference. This selection may therefore introduce a bias: because only the sensors that recorded many hailstones are retained, the training targets may be biased toward high hailstone numbers. The present evaluation does not allow this effect to be quantified, since the drone-derived HSNDs may themselves underestimate the number of small hailstones, which can be hidden from view or not resolved in the images (Sect. 4.6). In the future, a larger number of drone surveys of both weak and intense events could be used to recalibrate the retrieval; at present, however, the number of surveys is too small for a robust recalibration. This limitation is now stated explicitly in Sect. 4.6.

The corresponding passages of the revised manuscript read as follows (Appendix C and Sect. 4.6, respectively):

"[…] In total, the training dataset comprises 197 samples on 35 event days. Depending on the event day, hail was recorded by 1 to 13 hailsensors (Table A1). […]"
[…]
"The training dataset may be biased toward intense events, since hailsensor observations with few hailstones do not provide reliable estimates of the HSND moments and were excluded (at least 30 hailstones per event were required, Sect. 2.2.1, and more than 10 per 5-min interval, Appendix C). The training targets may therefore be biased toward high hailstone numbers. The present evaluation does not allow this effect to be quantified, since the drone-derived HSNDs may themselves underestimate the number of small hailstones, which can be hidden from view or not resolved in the images. Additional hailsensor and drone observations, in particular of intense events with large hailstones, which are underrepresented in the training dataset (Sect. 4.2), would allow the model to be trained and evaluated over a wider range of hail sizes."

2. Missing Uncertainty Quantification

No uncertainty estimates are provided for the radar-retrieved HSND. The authors mention that "explicit quantification of model uncertainty" is "not addressed" and should be "explored" (lines 635-638). This is a significant limitation for an operational product and should be addressed before publication.

The authors should at minimum provide prediction intervals for the GAM outputs, perhaps using bootstrapping or quantile regression.

We thank the reviewer for this comment and agree that uncertainty estimates are an important complement to the retrieved HSND, in particular in view of a potential future operational use of the retrieval. Following the reviewer's suggestion, we quantified the uncertainty of the GAM outputs with a bootstrap approach and propagated it to the retrieved HSND and to all derived quantities (new Sects. 3.4.4 and 4.5, Figs. 13 and 14, reproduced below).

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, which is summarized by its 5th, 25th, 50th, 75th, and 95th percentiles. Figure 13 shows the resulting intervals for the GAM estimates of the hailsensor observations, and Fig. 14 their propagation to Dmax [1/10 m²], the total kinetic energy, and the number of hailstones larger than 2 cm for the event of 28 June 2021. The bootstrap intervals are also shown for the radar-retrieved HSNDs in the drone comparison (Fig. 7, see our reply to major concern 4).

These intervals quantify the uncertainty of the fitted relationship between the radar predictors and the moments arising from the limited number of training events. The scatter of individual observations around this relationship is quantified separately by the cross-validated RMSE (0.34 and 0.44 in log10 units for moments 2 and 4). The sources of uncertainty not represented by the bootstrap intervals are discussed explicitly in Sect. 4.6. The statement on missing uncertainty quantification has been removed from the manuscript.

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. […]"

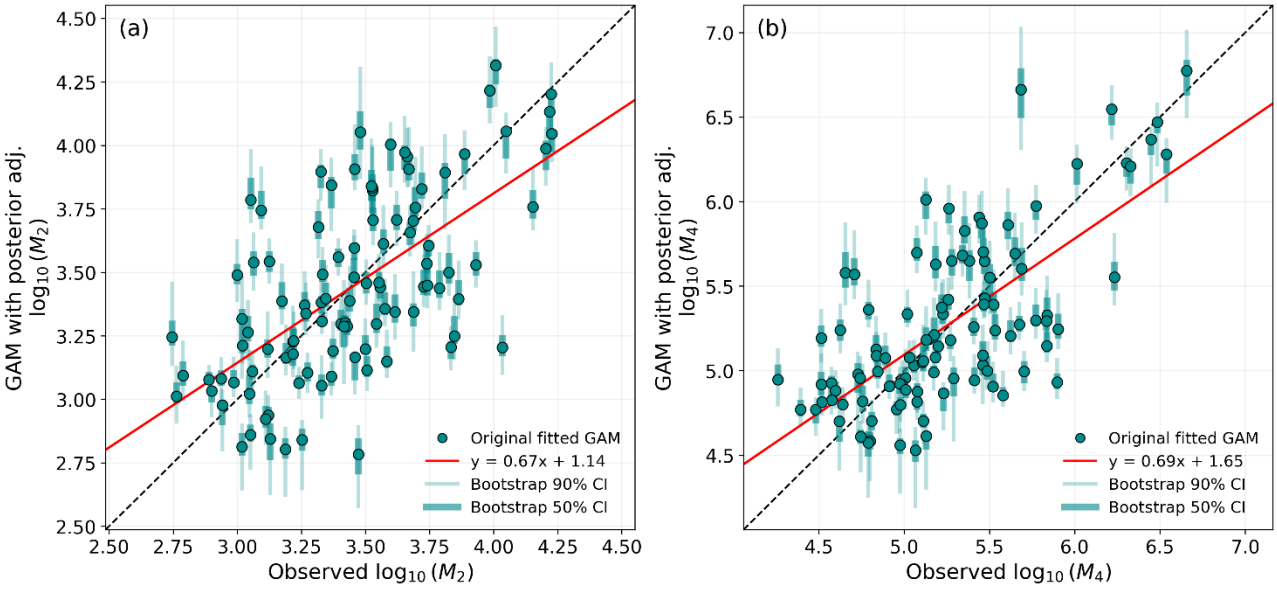

Figure 13 of the revised manuscript: observed and GAM-estimated log10(M2) (a) and log10(M4) (b) for the hailsensor observations, with the 25th–75th (thick) and 5th–95th (thin) percentile ranges of the 100 event-day bootstrap realizations.

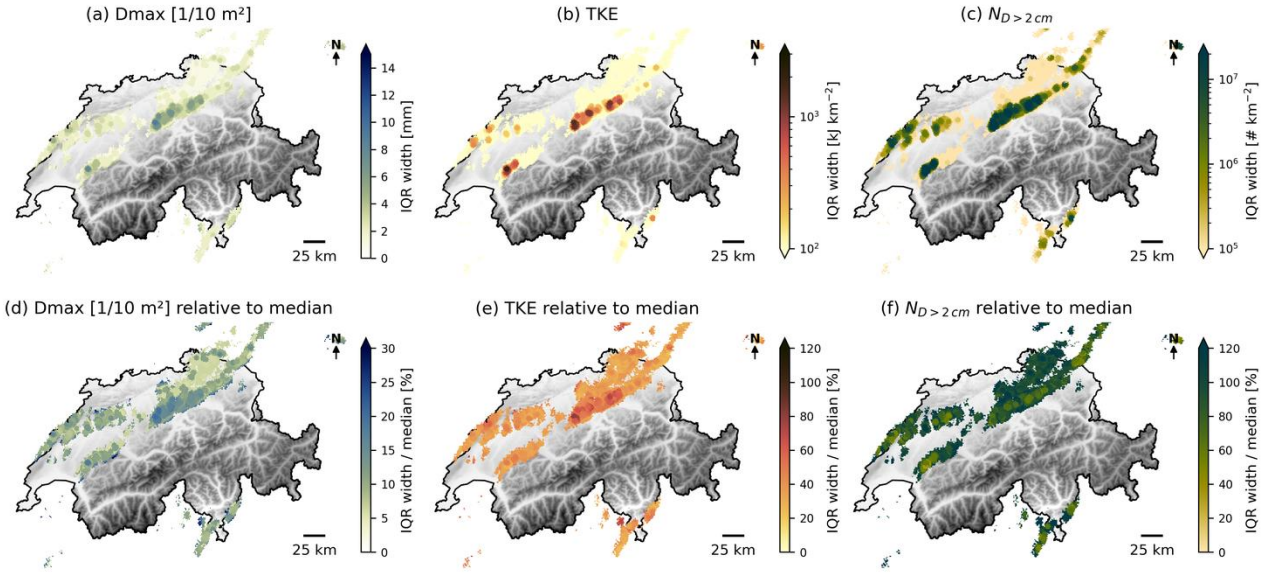

Figure 14 of the revised manuscript: interquartile-range width of the bootstrap realizations for Dmax [1/10 m²] (a), total kinetic energy (b), and number of hailstones larger than 2 cm (c) for the event of 28 June 2021, and the corresponding widths relative to the bootstrap median (d–f). The colour scales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

3. Physical Interpretability of Selected Features

The selected features (particularly the ρHV < 0.97 and ZH > 50 dBZ volume) are physically plausible (mixed-phase regions with high reflectivity). However, the sum of ZV and maximum ZV features are highly correlated (as noted in Fig. 4), and the physical distinction between

them is not clearly explained. The authors should discuss whether these features could be combined or reduced without loss of skill.

The thresholds (ρHV < 0.97, ZH > 50 dBZ, ZV > 30 dBZ) were likely optimized empirically. The authors should explain how these thresholds were chosen and whether they are specific to the Swiss radar calibration or transferable.

We thank the reviewer for these comments.

Maximum and sum of Z_V. We agree that the two features are correlated. They nevertheless describe different aspects of the vertical column: the sum integrates the depth and intensity of the region where Z_V exceeds 30 dBZ, whereas the maximum represents its peak value. The maximum was retained by the stepwise selection because it improved the cross-validated PCC beyond that obtained with the sum, and both AIC and BIC decreased when it was added (ΔAIC of −7.1 and −4.0, ΔBIC of −4.4 and −1.5 for moments 2 and 4, respectively; Table A2). The reduction is small for moment 4, indicating that the maximum mainly improves the estimation of moment 2. This is now discussed in Sects. 4.1 and 4.1.2.

Choice of thresholds. The thresholds were not prescribed: the candidate features were computed for several thresholds of each variable, e.g., ρHV below 0.97, 0.95, or 0.93 and Z_H above 50, 45, 40, or 35 dBZ, as well as within sub-regions defined by Z_H thresholds of 20, 30, 40, and 50 dBZ (Sect. 3.2). The thresholds of the selected features resulted from the automatic selection among these candidates. Since they refer to polarimetric variables with a physical meaning, they are expected to be applicable to other well-calibrated C-band radars. However, the radar features also depend on the scanning strategy and beam geometry, which affect in particular the volume-based features, so that applying the model to another radar network would require adapting the features to its scanning strategy, for example by emulating that strategy with the Swiss radar data and testing the sensitivity of the retrieval against the Swiss ground observations, or a re-training if needed. This is now stated in Sect. 4.1.2.

The corresponding passages of the revised manuscript read as follows (Sect. 4.1.2):

"Although the maximum and the sum of Z_V in the vertical column are correlated, they describe different aspects of the column: the sum integrates the depth and intensity of the region where Z_V exceeds 30 dBZ, whereas the maximum represents its peak value. The maximum was retained by the stepwise selection because it improved the cross-validated PCC beyond that obtained with the sum, and both AIC and BIC decreased when it was added (Table A2).
[…]
The thresholds of the selected features (ρHV < 0.97, Z_H > 50 dBZ, and Z_V > 30 dBZ) were not prescribed, but resulted from the automatic selection among features computed for several candidate thresholds (Sect. 3.2). Since they refer to polarimetric variables with a physical meaning, they are expected to be applicable to other well-calibrated C-band radars. However, the radar features also depend on the scanning strategy and on the beam geometry, which affect in particular the volume-based features, so that the application of the model to another radar network would require adapting the features to its scanning strategy, for example by emulating that strategy with the Swiss radar data and testing the sensitivity of the retrieval against the Swiss ground observations, or a re-training if needed."

4. MESHS Comparison

The comparison with MESHS (Section 4.4) is valuable, but MESHS is a single-polarization product. Comparing a dual-polarization model against a single-polarization baseline is not a fair test; the comparison should be against state-of-the-art dual-polarization hail products.

The authors should also compare against POH (Probability of Hail) to demonstrate improvement in detection, not just size estimation.

We thank the reviewer for these comments.

Comparison with dual-polarization products. MESHS is the operational hail-size product of MeteoSwiss, and no operational dual-polarization hail-size product exists in Switzerland. MESHS is widely used in Switzerland for hail monitoring, damage modelling, and climatological analyses, which makes it a natural and well-known reference for a new product intended for the same applications. To our knowledge, no dual-polarization retrieval of the full HSND has been developed elsewhere. Existing dual-polarization hail-size algorithms, such as the hail size discrimination algorithm (Ryzhkov et al., 2013; Ortega et al., 2016), were developed for S-band radars and provide hail-size classes rather than continuous diameters or distributions. They would first need to be adapted and calibrated for the Swiss C-band network before a fair comparison could be made. Even then, a comparison with a categorical maximum size would only address a small part of the information provided by the HSND. We now explain the choice of MESHS as a reference in Sect. 2.1.2.

Comparison with POH. The proposed model is not a hail-detection algorithm: it estimates the HSND under the assumption that hail is present, and hail detection relies on POH. In the comparison with the crowd-sourced reports, the HSND-derived quantities were considered only where POH was at least 90 % (Sect. 3.5.3). A comparison with POH in terms of detection skill is therefore not applicable, since POH is used as an input to define where the retrieval is evaluated.

The corresponding passage of the revised manuscript reads as follows (Sect. 2.1.2):

> "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). MESHS is the operational hail-size product of MeteoSwiss and is widely used in Switzerland, which makes it a natural benchmark. Dual-polarization algorithms, such as the hail size discrimination algorithm (Ryzhkov et al., 2013; Ortega et al., 2016), would be a possible alternative benchmark, but they were developed for S-band radars, provide hail-size classes rather than continuous diameters or distributions, and no comparable algorithm has been implemented in Switzerland."

5. Temporal Resolution and Advection

The 5-minute temporal resolution of the radar composite is coarse relative to the 20-s grouping of hailsensor data (Appendix C). The authors should discuss whether this temporal mismatch affects the feature-target relationships.

The advection correction (Appendix C) uses TRT storm motion, but hailstones have different fall speeds than the storm cell. The 30-s time offset (Appendix D) found empirically is plausible, but the authors should justify this based on physical fall speeds (e.g., 20-40 m/s fall speed × 30 s ≈ 600-1200 m aloft, which seems large relative to the assumed 200 m extrapolation distance).

We thank the reviewer for these comments.

Temporal resolution. This point is closely related to the scale mismatch discussed in our reply to major concern 1. The grouping of the hailsensor impacts (30 s, corrected from the 20 s erroneously stated in the original manuscript) is not the temporal resolution of the comparison: it is only used to position the impacts along the storm trajectory. The grouped impacts are then accumulated over the same 5-minute interval represented by the radar composite (Appendix C), so that both the radar features and the training targets

refer to the same 5-minute period. The remaining variability within this period, which cannot be resolved by the radar, contributes to the residual scatter of the estimates, as now discussed in Sect. 4.6.

Time offset and advection. The 30 s offset is the time lag between the 5-minute window of the radar data and that of the hailsensor data, i.e., the radar window precedes the hailsensor window by 30 s. It compensates for the time needed by the hailstones to fall from the region observed by the radar to the ground, and was determined empirically as the value maximizing the correlation between radar features and hailsensor moments, among offsets between 0 and 60 s (Appendix D, Fig. A4). Multiplied by the fall speed, the offset corresponds to a vertical fall distance, whereas the 200 m limit of the spatial extrapolation is a horizontal distance along the storm motion; the two quantities are therefore not directly comparable. The horizontal displacement of hailstones during their fall is instead accounted for by aggregating the radar features over a radius around each grid point: radii from 250 m to 9 km were tested (Sect. 3.2), and only features aggregated over a 1 km radius were retained by the selection. In addition, the selected features summarize the entire vertical column, so that the radar information on hail aloft is not limited to the height corresponding to the offset. We acknowledge that a constant offset is an approximation, since the actual fall time depends on hailstone size and on the height at which the hailstones are observed.

6. Writing and Clarity

The manuscript is generally well-written but is too long (46 pages with appendices). The extensive appendices (A-E) contain essential methodological details that should be integrated into the main text. The current structure makes it difficult for readers to follow the workflow (Fig. 2 provides an overview, but the details are scattered).

We thank the reviewer for this comment. We share the concern about the length of the manuscript. At the same time, the revision required several additional analyses requested by both referees, such as the extended drone and crowd-sourced evaluations, the information criteria, and the uncertainty quantification, which inevitably increased the length of the main text. Integrating the appendices into the main text would further lengthen the core sections of the manuscript.

We therefore kept the current structure, in which the main text describes each step of the workflow and its key choices, while the appendices contain the technical details needed for reproducibility: the merging of the five radars into a 3D composite (Appendix B), the spatial extrapolation of the hailsensor data (Appendix C), and the sensitivity study used to set its parameters (Appendix D). Each step is summarized in the main text with a reference to the corresponding appendix, and Fig. 2 provides an overview of the complete workflow. We believe that this structure allows readers to follow the workflow without being interrupted by technical details, while keeping these details available to readers interested in reproducing the method.

The abstract should more clearly state the limitations of the approach (bias toward intense events, uncertainty quantification missing, small sample size).

We thank the reviewer for this suggestion. The abstract now states the main limitations of the approach, namely the small number of training event days with large hailstones, and the possible bias of the training data toward intense hail events. In addition, the abstract now mentions that the uncertainty of the retrieval was quantified with an event-day bootstrap, which addresses the third limitation mentioned by the reviewer.

The corresponding passages of the revised abstract read as follows:

"[…] The uncertainty due to the limited number of training event days was quantified with an event-day bootstrap and propagated to the retrieved HSND. […] The main limitations of the approach are the small number of training event days with large hailstones, and a possible bias of the training data toward intense

hail events. Because hail is rare, further observations remain crucial for refining the retrievals and constraining their uncertainty."

Specific Technical Corrections

1. Line 370-371: The phrase "the limited amount of data for large hail diameters prevents the GAM from constructing well-adapted spline functions" contradicts the earlier claim that the model "generalizes well." This should be rephrased to acknowledge the limitation more clearly.

We thank the reviewer for pointing out this inconsistency. We have rephrased both passages: the cross-validation is now described as selecting the features that yield the best performance on unseen hail events, and the limitation for large hailstones is now stated explicitly.

The corresponding passages of the revised manuscript read as follows (Sects. 3.4.2 and 3.4.3, respectively):

"A leave-one-event-out cross-validation was performed to select the features that yield the best performance on unseen hail events. […]"
[…]
"In addition, large hailstones are underrepresented in the training data, which limits the ability of the GAM to represent the upper part of the range of the moments. As a consequence, large moments tend to be underestimated, also in cross-validation."

2. Line 407: The notation "D99.9, D99.99, D99.999, D99.9999" is confusing. For a dataset with 30-100 hailstones per event, the 99.99th percentile corresponds to a probability level that is not statistically meaningful. The authors should use return levels or N-year exceedance estimates instead.

We thank the reviewer for this comment. We would like to clarify that the percentiles are not computed from the 30–100 hailstones recorded by a hailsensor during an event, but from the HSND retrieved by the radar over the 1 km² reference area, which typically comprises several orders of magnitude more hailstones. The 99.9th percentile is therefore well defined. Nevertheless, following this comment and comment 5 of the major concerns, the revised manuscript uses only one percentile, D99.9, together with a count-supported maximum diameter, Dmax [1/10 m²], defined as the largest diameter for which the retrieved HSND contains at least one hailstone per 10 m². The latter is directly related to a hailstone density per unit area and is therefore easier to interpret than a high percentile. The list of higher percentiles has been removed (Sect. 3.4.6). Return levels or N-year exceedance estimates would characterize the frequency of hail events at a given location over many years, rather than the size distribution within a single event, and are therefore not applicable here.

The corresponding passage of the revised manuscript reads as follows (Sect. 3.4.6):

"Characteristic diameters describing the upper tail of the HSND. These comprise (i) percentile diameters, defined as the diameters below which a specified fraction of the total hailstone population is found (e.g., the 99.9th percentile, D99.9), and (ii) count-supported maximum diameters (Dmax), defined as the largest 1 mm diameter-bin centre for which the retrieved number of hailstones over the 1 km² reference area reaches a prescribed minimum count."