Reviewer #1

Review of
"Hail events in Germany, rare or frequent natural hazards?"
Feb 2025
Tabea Wilke, Katharina Lengfeld, and Markus Schultze

Thank you for the opportunity to read this paper regarding a new hail climatology of Germany. It's promising to see a more detailed evaluation of potential report biases, and refitting of the MESH relationship. I will have to recommend a major review though, as I believe some work is needed in the results to better support the use of VII and evaluate it's skill. I hope you can take on these comments and I'm looking forward to seeing the revisions.

major comments:

line 262: Would it be possible to provide some quantitative evidence of the MESH and VII comparison to back up statements like "MESH overestimates the sizes"? This could involve using a sample of multiple events not included in the fitting and directly comparing reported hail size and radar estimates using the same neighbourhood rules. I'm a little worried that an graphical evaluation based on one event might not provide robust evidence to support this claim. Furthermore, if MESH75 was overestimating hail size, why not try a lower percentile threshold and see what is optimal? Would the mean fitted MESH work better? A statistical evaluation would also let you look at biases as a function of reported hail size... Perhaps VII works well for smaller sizes, but works poorly, compared to MESH, for larger sizes. To reject MESH and focus on VII based on the analysis of one event seems a little rushed, especially given the work you've done to refit MESH.
We added a validation of 8 days of 2024 that were not included in the fitting. We compared different approaches of estimating hail sizes with different metrics. We can see that VII performs best for most of these cases. We added the table with all the values to the appendix and added the following text to the case study:
“To show a more complete picture for different MESH formulas and VII compared to the observations, we have validated all of those algorithms with hail days in 2024 that were not used for the MESH fitting (see A1). From that we can say that VII performs best for hail sizes up to 3 cm. The MESH formula that uses the mean instead of the 75 % percentile is also very good. For larger hail sizes there were only a few observations, but for them, the 75 % formulas of MESH (as well the one of Murillo et al. as our own fit) outperform VII. This result is not unexpected, because MESH is fitted so that 75%, 95% respectively, of the values is lower so that we can be very sure that the resulting hail size is the maximum expected. For most of the observations (75 %) this value is too large, so in total MESH does not perform as well as VII especially, for smaller hail sizes. For the larger observation values MESH performs better because of the assumption for fitting the formula. Both VII and MESH, therefore, take different aspects of hail size into account. We have chosen VII for a larger analysis over the period 2018 - 2023, as it fits better for most of the values.”

Section 3.1 VII: Would it be possible to investigate the performance of this hail size estimate from VII using the reports dataset, and maybe even refitting it, making it more consistent with the analysis performed on MESH?
Thank you for your suggestion. Indeed, exploring this idea could lead to a interesting new study, which is somewhat beyond the scope of our current paper. It would make the long paper even longer, and the question is whether the same methodology as for MESH is suitable for VII, or whether it is better to use a completely new approach for refitting. We appreciate your input and will consider it for future research. We have added it to the outlook. “A similar approach of refitting could be applied to the VII data to increase the performance even further.”

General comment: Would it be possible to integrate your findings about how the "mean of many hail reports is usually very good" with the radar based climatology fitting/evaluation? Perhaps rather than

training/evaluating using one SHI value for every hail observation, it could be worthwhile using the local neighbourhood of hail observations (as a mean etc) to correlate with the SHI value. This could certainly reduce the noise in observations, as you've indicated, and provide a direct benefit to your radar analysis from this study of report biases.

This would be a valuable addition. Investing further in this area could prove advantageous, particularly in terms of pooling observations. Nevertheless, we need to have in mind that the fundamental objective of the MESH approach is to determine the maximum of the expected hail size, we cannot be sure that one of the observations represents the maximum. It is imperative to consider this aspect in the context of warning. We added this aspect to the outlook. “For the refitting the observations could be pooled to use only the mean of neighboring observations and to take only observations into account with neighboring observations available.“

Section 4.7: Can you please relate this back to monthly trends shown in figure 8 and 9, and explore the differences? I think it's important to put this insurance information in context with the observations.

Yes, sure, we added the following to section 4.7:

“We can compare the number of damage reports, which is equally true for the loss expenses, to the number of observations in total (Fig. 8) as well as to the distribution of hail sizes (Fig. 9).

Both, number of damage reports and number of hail observations (Fig. 8), show a clear annual cycle with a peak in the summer months, moderate numbers of reports and observations in spring and autumn and almost no hail events in winter. The maximum of damage reports, however, is in June and July and, therefore, shifted to later in the year compared to the observations with a maximum in May and June. This might be due to the hail size: Fig. 9 indicates that small hail dominates in winter, spring and autumn, while medium sized hail dominates in the summer months probably leading to enhanced damage. Large and giant hail with the highest damage potential almost exclusively occur from May to August. In case of leaving out the extreme hail event in July 2013 in the damage reports, their maximum is in June as well as the maximum for medium to giant hail observations.“

General comment: Did you apply the correction developed by Brook et al. 2024 to the C band data here? I'm still not clear why your C band MESH is overestimating hail size to such an extent given you fitted a new relationship, and the performance seems poor even for smaller sizes. It's also not clear why VII can perform so well with no refitting!

No, we did not use the correction developed by Brook et al. 2024. We believe that a correction is not necessary with the new relationship as it should represent the direct correlation of SHI to observation. The simple formula for estimating hail size from VII is based on years of experience of the weather forecasters and is particularly well suited to our data. This may be the reason why its performance is so good. VII does not try to estimate the maximum of hail size, so it fits to most of the observations better, but it can be off in detecting the maximum hail size.

minor comments:

line 15: what year were these figures normalised for? That's important if we're comparing them.

These are the original values, and they have not undergone normalization. Therefore, as you correctly stated, they are not directly comparable. Nevertheless, the numerical values can offer insight into the extent of the damage.

line 26-27: The rich-poor (disparity in wealth) bias can be normalised for though if we know the total sum insured. Is this possible with your datasets?

Indeed, while this is a feasible undertaking, the nature of the data set, which is focused on the postcode area, makes it too coarse to see the wealth disparity for most postcodes. Furthermore, given the potential for hailstorms to impact only part of a large postcode area, it is challenging to adequately account for this in the analysis.

line 32: It's not clear why 90% damages was selected here for discussion. Can you provide more context?
It was meant that roofing systems have the largest fraction on the total loss and make up to 90 % of the total loss. This is clarified in the text.

line 48: Brook et al. 2024 found that C band overestimated hail with respect to S band estimates, not hail reports (Which this could imply). Please clarify that this w.r.t. S band.
We changed
"Originally developed for S-band radars in the US, Brook et. al showed that MESH used with C-band radars tends to overestimate hail sizes"
to
"MESH was originally developed for S-band radars in the US, Brook et. al showed that MESH used with C-band radars tends to overestimate hail sizes compared to S-band radars."

line 52: "was expected to improve" reads like it failed to improve estimation of hail size. I would disagree with this, as the HSDA retrieval has provided a robust way to size hail using polarimetric information (but it's only for S band) and is used in operations.
We changed "was expected to improve" to "gives the opportunity to improve".

lines 57-60: Can you please link how ZDR/kdp signatures of liquid water in updrafts relate to hail size? I'm not aware of any hail size estimates that have been developed from polarimetric signatures like this. Some references would be great.
This was not well worded, so we changed "For the development of large hailstones" to "For the development of the potential of large hailstones" and "thus hail size" to „potential of large hail generation"

line 62-63: For the Junghänel et a. 2016 study, can you please comment on their findings and how it relates to hotspots?
We added a bit more context to this study: Junghänel et al. combined reported data with reflectivity values of radar data, either having a report and a reflectivity value higher than 50 dBZ or having a reflectivity value higher than 55 dBZ only." The findings and hotspots are similar to the other hail studies in Germany, so they were mentioned only once for all of the studies:
"All the hail studies for Germany have the hotspot of hail in southern Baden-Wuerttemberg in common. Further on, they share the north-south increase of hail days."

line 66: Can you please add any background or literature on VII? There hasn't been any discussion of it in the literature review, but it's only of the main retrievals used by this paper! Perhaps you can comment on how it differs from MESH, and why it has not been used in many climatology studies?
It is important to note that VII is not widely recognized for its capabilities in hail size estimation, which serves as a motivation for the present study. Operationally, it has been utilized for years in the context of detecting potential hail occurrences. Consequently, there is a lack of literature addressing its application in the context of hail.

line 135: Why was a minimum threshold of 7.5 mm used? Do you have any past studies which selected this? Or was it something you found suitable through experimentation?
A significant quantity of graupel has been observed in Germany during the initial months of the year. To exclude these from the data set, a threshold of 7.5 millimeters was empirically determined.

line 161: Can you please introduce the DWD app hail size categories here?
As previously mentioned in Section 2.1 of the Data section, "Human observed data," the necessary information has already been introduced. Therefore, we added a reference to Section 2.1 here.

lines 161-162. This sentence, starting with "Therefore, ..." was not clear to me. Please look at reworking.
We added more context to the sentence and changed it from „Therefore, we have selected the reference values as our observation.“ to „Therefore, we have selected the reference values from the category e.g. a report from the category ‚hail of 2 cm‘ was assumed as an observation of a hailstone with a diameter of 2 cm although it might be slightly smaller or larger.“

Figure 3: I'm curious to why the full hail reports dataset (from the 3 sources) was not applied to develop your own calibration of MESH. This may have helped increase the number of larger hailstones in the sample.
The incorporation of the WarnWetter App data and the ESWD data was the only viable option, as the alternative dataset is given only on a daily basis, thereby making it impossible to find the correct SHI value for it. Given that the majority of the ESWD data for the years 2022 and 2023 originates from the WarnWetter app, it was determined that the calibration process would be conducted using solely the WarnWetter app.

3.2: MESH: Would it be possible to report on the number of samples in each category when fitting SHI with reports, which is important to better understand the potential skewness of the fit.
Yes sure, we added the following to the text: “In total we had 2403 samples for 0.5 cm, 4232 samples for 1 cm, 1979 samples for 2 cm, 750 samples for 3 cm and 192 samples for 5cm.”

Figure 3: Please describe the features in the plot: vertical bars, dots and shaded area.
We changed the figure caption from „Comparison of all power laws to a violin plot including its median of the data from the German WarnWetter app for the year 2023.“ to „Comparison of all power laws of SHI (colored lines) to observation values (blue dots) in a violin plot visualizing the amount of observations (shaded area) including its median SHI (black vertical bar) of the data from the German WarnWetter app for the years 2022 and 2023.“

line 189: "larger diameter" could be improved with "maximum dimension/axis", as we're not measuring spheres.
You are right, we changed it to maximum dimension.

Section 4.2: would it be possible to justify why results where analysed by age groups? What hypothesis are you testing here?
We added the following “We were interested in determining whether there is a discrepancy among age groups due to differing degrees of experience with the digital world or its attendant technologies.”

Figure 4: Can you please list the sizes (maximum dimension) and type (oblate vs spherical) of hailstones shown here. This would be help the reader to make better use of this image.
Sure, we added it to the figure caption: from left to right: round 0.5 cm, oval 7 cm, round 2 cm, oval 5 cm, round 5 cm, round 7 cm

Figure 5: Again, please details all the features of the plot for the reader.
We changed the caption to “The distribution with its median of answers (black vertical bar) for the question ‘How large is a 2€ coin?’ In the different age groups. The correct answer is shown as red dashed line. The mean and standard deviation for each group are presented in the figure on the right. The answers are shown in three different visualizations: The upper one: a kernel density estimation for all answers in the age group showing the different distributions of answers; the middle one: a box plot showing the quartiles and outliers; the last one: the answers as single dots showing all single values.“

line 225: Can you comments on how this bias changes with hail size? It seems larger hailstones have a smaller relative bias?

We added “The bias remains constant irrespective of hail size; however, the standard deviation undergoes a reduction for larger hail sizes.” to the text.

Figure 8 : Please comment on the total number of hail reports used in this analysis for context.
We added it to the caption of the figure: “Hail observations in its (a) diurnal cycle based on data from ESWD (3769) and WarnWetter app (21 231) and (b) annual
cycle based on data from station observations (25 719), ESWD (3769) and WarnWetter app (21 231).“

line 235: Is it possible to normalise this dataset to reduce the effects of the much larger DWD starting in 2021? I'm concerned these stats shown in figure 8 and 9 might be impacted by the strong annual variability of hail for the years when the much larger DWD app datasets started.
Regrettably, normalizing the data is not feasible due to the varying temporal changes exhibited by the disparate observation datasets. A similar phenomenon like the starting DWD dataset is observed in the ESWD dataset in earlier years. Excluding the years 2021-2023 (corresponding to the WarnWetter app data or the dark blue stacks) results in a similar annual cycle, albeit with a shift in the peak from June to May. The diurnal cycle would be considerably more poorly defined due to the lack of data, as only 3769 observations of the ESWD would remain. To address this limitation, the data has been normalized against all occurrences, as the y-axis represents the relative occurrence.

line 240: What is the "h3 grid of level 6"?
We changed the „h3 grid of level 6“ to „hexagonal grid with an average hexagonal size of 36.13 km^2 (h3 grid of level 6)“

Figure 10: Did you investigate using a larger grid size? There are many grids points with no data which make it difficult to see any hotpots, especially with the colourmap used (maybe it's just me!). Using a larger grid might help bring out the underlying trends.
Indeed, an investigation was conducted, but we decided for the finer grid. We changed the colormap a bit and hope it is now easier to see the hotspots/absence of hotspots. – The objective of this figure is to demonstrate the absence of a trend. As illustrated in the plot a), there is a conspicuous presence of hotspots. Conversely, the hotspots are no longer discernible in plot c), as it has been normalized with respect to population. Consequently, the observed hotspots can be attributed to the urban reporting bias.

lines 257-259: Please reword, the meaning of this sentence is not clear to me.
We changed “Assuming the presence of hail for both algorithms, only if their value is greater than 7.5 mm, we might not account for reports of the category “smaller than 1 cm”.” to
“In instances where the value of MESH/VII exceeds 7.5 mm, the presence of hail is assumed. Consequently, it is highly improbable that we can provide coverage for hail reports categorized as "smaller than 1 cm."”

line 282: While it is mostly likely that supercells provided the hail larger than 5cm, it would be good to see evidence, such as mesocyclone tracks.

Indeed, the comparison of hail and hail sizes to mesocyclone tracks and severity would be another interesting topic. We investigated the number of mesocyclones in Germany for the same time as the VII (see Fig 1). Large hail most likely appears through mesocyclones, but not every mesocyclone may

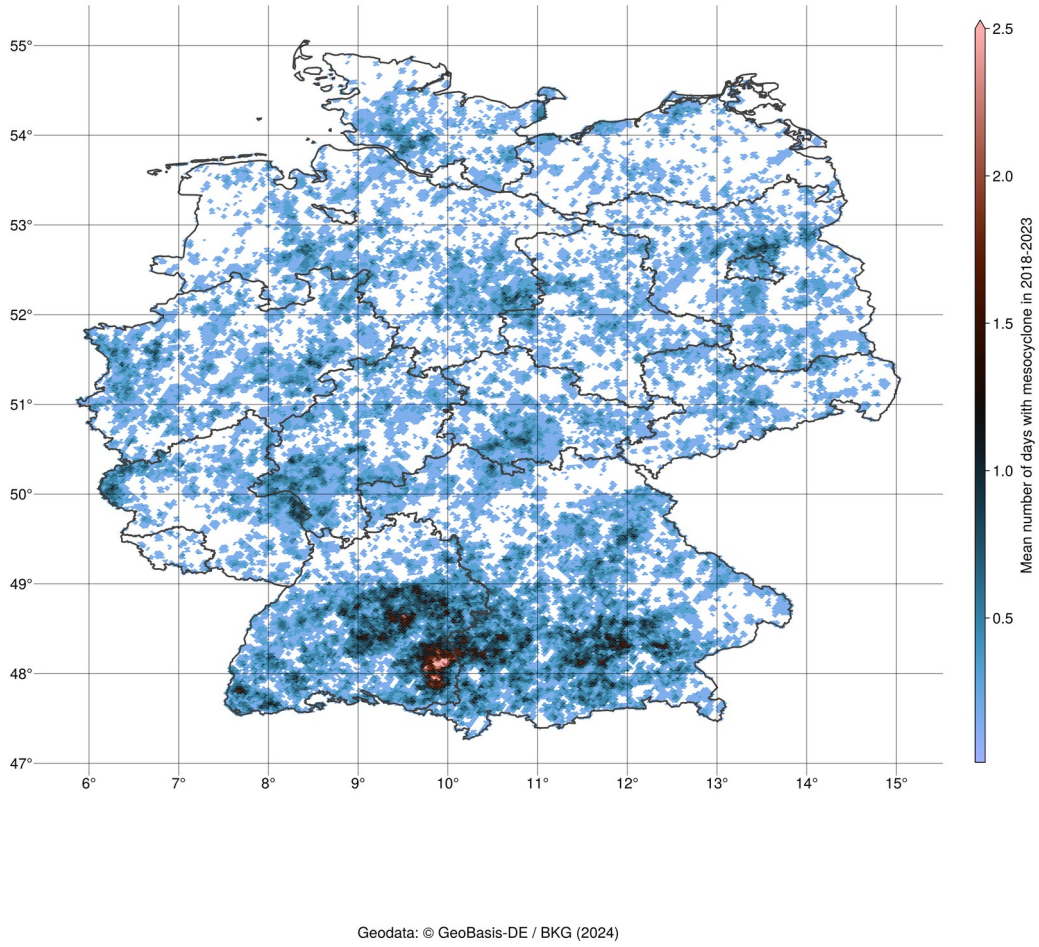

Figure 1: Mean number of mesocyclones from 2018 - 2023 in Germany.

produce hail, therefore, a deeper comparison would be necessary. As this comparison would go beyond the scope of the current study, we changed
“There are only very few hail days where individual supercells produced hailstones larger than 5 cm”
to
“There are only very few hail days with hailstones larger than 5 cm. Over southern Bavaria, individual cell tracks are discernible, suggesting that the most severe hail events are caused by isolated extreme convective cells.”
and did not show this figure.

Figure 11: For the colorbar used in panels (a) and (b), can the colourbar ticks, and ideally a discrete colourmap, match the 4 levels used to plot hail observations. This would greatly improve readability.

This is a great idea, we changed it in the manuscript.

Figure 16 caption: I'm not sure what the authors mean by "ascending over" and "descending from". It might be simpler to put the month on each panel and leave any commentary to the manuscript (to avoid repeating it).

We left the explanation out of the figure caption, but left the a-f in the plot and the month in the caption.

Line 337-338: I would advise not using "perfectly" unless there has been some statistical evaluation.
You are right we changed it to “fit quite well”.