## Review for NHESS-2020-138

### General Comments

This paper presents a 10-year radar-based climatology of hail frequency in a portion of western Europe encompassing France, Germany, Belgium and Luxembourg. The authors combine 2D reflectivity composites from the French and German weather services into a single mosaic with a resolution of 1 $km^2$ and then applying storm cell tracking with a 55 dBZ reflectivity threshold to identify likely hail events. The spatial distribution of hail is analysed, with particular focus on the relation to surface topography (coastlines and mountain ranges), along with diurnal and seasonal variations in different parts of the study domain. The authors also examine the characteristics of the identified cell tracks, including their length, width, duration, and orientation.

The length and spatial extent of this analysis alone make it a novel contribution to the hail climatology literature, which often focuses on smaller regions. The paper is largely well written and the figures are mostly of a high quality. However, I see a number of issues that need to be addressed before this work can be accepted for publication. Chief among these is the unacceptable amount of speculation in the results, particularly when it comes to discussion around the role of surface topography in hailstorm formation. I would also like to see more details regarding the construction of the national radar composites and discussion on the importance of radar calibration errors. Detailed comments are provided below.

### Specific Comments

*Major Comment 1:*

Currently, your results section contains too much speculation, particularly when it comes to the role of surface topography in the formation of hail storms. Examples include L197-207, L223-226, and L249-253. It is fine to note the clear correspondence between high hail frequencies and orographic features, but not to speculate at length on the underlying mechanisms in the absence of detailed observations or numerical simulations (either presented here or in other published studies). A bit of speculation is OK, but this should probably be reserved for the conclusions/discussion, where it can be used to motivate future investigation into physical mechanisms. Alternatively, if you do want to at least start this investigation here, you could use sounding or reanalysis data to examine the flow characteristics (wind speed and direction, Froude number, etc) on hail days in your various subdomains (c.f. section 6 of Kunz and Puskeiler 2010). This would obviously involve a bit of extra work (and additional data), but would make this study more than just “another hail climatology.”

*Major Comment 2:*

I’d like to see a bit more detail regarding how the French and German radar composites are produced. For example, do they use radar from the lower (or lowest unblocked) tilt only or do they compute a column-maximum reflectivity across all tilts? How are reflectivities combined in regions of overlapping coverage? Is the nearest radar used, the one with the lowest unblocked beam, or is a more complex quality index applied? Is any account taken for variations in beam diameter with range or differences in beam width between different radars? Such information is important for the reader to understand limitations in the composite product.

*Major Comment 3:*
Radar miscalibration is a major issue in many operational networks and can lead to significant inhomogeneities across large study domains. It is also a very tricky problem to overcome, although methods do exist (see, for example, Louf et al. 2019). However, for the purpose of this study I think you just need to mention it as a potential source of error in your results. Specifically, differences in radar calibration across the study domain may lead to an overestimation of the relative hail frequency in some regions (where radars are calibrated too high) and underestimation of the relative hail frequency in others (where radars are calibrated too low).

L3: My first question on seeing the study period is why does it end six years ago in 2014? Was one (or both) of the national composites not available for later dates? This information should be provided in section 2.

L26-36: You should also mention radar-based hail climatologies from other parts of the world, such as the USA (Cinineo et al. 2012) and Australia (Soderholm et al. 2017, Warren et al. 2020).

L39: I would argue that the issue isn't that these satellite- and model-based methods are "not as straightforward as those based on radar reflectivity". Rather it is that the link between the observed quantities and hail occurrence at the surface is less direct than it is with radar-based measurements.

L50: Your study provides information about the frequency of hail but not its intensity; as such this statement should be modified.

L53-54: Section 4 also presents results on seasonal and diurnal variations in hail frequency and the characteristics of the hail cells. Maybe mention this here.

L64: Why 2015 when your study covers 2005-2014? If the number or type of radars changed during your study period this should be mentioned.

L69-70: What map projection (coordinate system) does the French mosaic use?

L72-73: This sentence needs rephrasing. What sort of quality checks are performed?

L76: Does this mean that there is a gap in the data from mid-June to late-July 2009? Are there any other gaps during the study period? These will need to be accounted for if you estimate annual hail frequencies, as recommended below.

L77: Why did you bother processing the coarse resolution data from 1999 to 2004 when your study period only starts in 2005?

L80: Again, why discuss the state of the network in a year that falls outside your study period? How many radars were operational during 2005 to 2014 and did this number change?

L82-83: Figure 2 suggests that all of Germany is covered by the radar composite, with no gaps. The locations that you mention (the far north near the Danish border and southeastern Bavaria) are only covered by a single radar, but they will still surely feature in the composite.

L88-91: This type of data compression is quite common. Since the resolution of the data is quite high (0.5 dB) I don't think this needs to be discussed. It would only be worth mentioning if there were only a few reflectivity levels (as in Puskeiler et al. 2016).

L98: I'm not sure what you mean by the "standard coordinate system". What map projection is used? I'm guessing it differs from the ones used in the French and German composites. Was any account taken of this difference? Given that each domain covers around 1000 km, there could be some distortions introduced in this procedure.

L113-114: You say that "only reflectivity in the range of 35 to 70 dBZ was considered in the analyses", but all of your analysis considers a single threshold of 55 dBZ, so does this filtering really matter? Or are you saying that reflectivities below 35 dBZ or above 70 dBZ were set as missing values?

L114-116: This explanation is a little confusing. Looking back at Puskeiler's paper I see that reflectivities have to be >45 dBZ and 5 dBZ or more above the values at the neighbouring grid points to be filtered using Eq. 2. Please rephrase to make this clearer. Also the method doesn't really use a range of 2 km; rather it considers the 8 neighbouring grid points.

L118: What do you define as "a high reflectivity value"?

L119-120: I don't understand what you mean when you say "Reflectivity values near neighboring countries were evaluated and calibrate [sic] with radar stations close to the border." Please elaborate.

L123-125: I've personally never heard of lightning causing spurious radar signatures. If this is a real thing, surely it would represent an argument *against* using lightning data to filter out such signatures?

L125-128: While hailstorms typically do produce lightning, I am not aware of any work that shows that this lightning is always cloud-to-ground, which is the only type that you consider in your analysis. As such it is possible that you may have inadvertently filtered out hailstorms that produced only intracloud lightning. This should be noted as a caveat of the method described here.

L140-141: Is tracking only applied to reflectivity areas of ≥55 dBZ? This is a very high threshold for defining convective cells and is likely to lead to much shorter tracks than one would achieve using a more typical threshold such as 35 dBZ. It will also lead to an unrepresentative estimate of the location of convective initiation (Fig. 9), since developing cells may travel some distance before they achieve reflectivities as high as 55 dBZ. In my view, a better approach would be to identify and track cells using a lower threshold, but then only retain those that reach a reflectivity of at least 55 dBZ. This would also allow you to perform a comparative analysis of hailstorms and non-hailstorms. Perhaps this is outside of the scope of the present study, but it would certainly be a nice avenue for future work. At the very least you should note the caveats of using such a high reflectivity threshold for cell identification and tracking.

L145-148: Looking at Fig. 3a from Puskeiler et al. (2016), the difference in HSS between reflectivity thresholds of 55 and 56 dBZ is very small (both are around 0.6). You should note the corresponding values of POD and FAR: 0.7 and 0.4, respectively. While these values demonstrate reasonable skill, they also indicate that 30 % of observed hail events are missed while 40 % of those predicted are false alarms. This provides some idea of the uncertainty in your climatology.

L149-159: More details are needed concerning the tracking methodology. For example, what are the intensity and size criteria that are used to match cells between scans? How is a significantly different cell area defined for the purpose of identifying merges and splits? It might help to add a figure illustrating the process schematically.

L165: How is the horizontal wind field estimated? Also, I would describe it as a field of motion vectors, since storms do not move with the wind at any particular level.

L171: You say that ESWD reports were located close to the centre of the storm tracks “in most cases”. What percentage of reports were not covered by the tracks? Can you comment on possible reasons (e.g. reflectivity <55 dBZ, erroneous report location/day)?

L180-182: It is good to reiterate that the 55 dBZ reflectivity threshold doesn’t guarantee that hail occurred (and similarly, the absence of such high reflectivities doesn’t guarantee that hail didn’t occur). At this point you could refer back to section 3.3 where the results of Puskeiler et al. (2016) were discussed.

L191-194: If the mistral is cold and dry, is it really relevant to hailstorm formation?

L263: Hail frequencies are also lower over the high terrain of the Alps and Pyrenees, which is consistent with the results of Nisi et al. (2016, 2018).

L275: Does the average include only pixels within the area covered by radar? If not, this is how it should be done, otherwise you will artificially lower the average. You also shouldn’t include points over the ocean, since these have been masked out in the map plots.

L299-300: I wouldn't say this result is particularly surprising. Large hail damage simply requires a few storms passing over densely populated areas, whereas Fig. 5 is considering the average number of hailstorms over a very large area.

L309: For simplicity, I would make all of the subdomains exactly the same shape and size. It looks like you would only need to modify boxes 11 and 13 for this to be the case. I would also suggest using a consistent 3-letter identifier for all regions, rather than numbers. These could be listed in a key/legend in Fig. 3 or in a separate table. The following would be my suggestions for the identifiers: 1 = NWG (North West Germany), 2 = NEG (North-East Germany), 3 = BEL (Belgium), 4 = WCG (West-Central Germany), 5 = ECG (East-Central Germany), 6 = NWF (North-West France), 7 = IDF (Île-de-France), 8 = LUX (Luxembourg), 9 = BAV (Bavaria), 10 = WCF (West-Central France), 11 = MAS (Massive Central), 12 = SWF (South-West France), 13 = MED (Mediterannean).

L316: It doesn't make sense to simply accumulate the number of hail days over all grid points in each subdomain. For one thing, some of the subdomains contain large areas that are over the sea and/or outside of radar coverage, which will give them a lower number than subdomains that are entirely over land and within radar coverage. It is also very difficult to interpret what these numbers mean. Instead, you should calculate the average number of hail days over all points with data (i.e. excluding those over the ocean and outside of radar coverage). This approach will give a much fairer comparison between the different regions. It's a good idea to use a moving average; however, from Fig. 3 it appears that you consider the preceding 10 days for this average. Instead, I would recommend using a 15-day moving average centred on the day in question (i.e. ±7 days).

L363: Again, you should consider the average number of hail days for each hour, not the total number of days over the domain. Also, why only consider the first time that a reflectivity of 55 dBZ is detected? Surely you should consider all times with 55 dBZ or higher in order to properly capture the diurnal cycle of hail (not just its initiation)?

L397-409: The problem with this analysis is that it assumes that the first detection of reflectivity ≥55 dBZ corresponds to the initiation of convection. In fact, developing convection may travel some distance before it reaches such an intensity, particularly in the presence of strong background flow (which would be expected in high-shear environments that favour severe storms). This is one reason why it would be advantageous to use a lower reflectivity threshold for identifying and tracking convective cells. At the very least, this caveat needs to be mentioned.

L418: You can probably just say lengths "between 10 and 20 km". Using 10.1 km as the lower bound implicitly excludes tracks with a length >10 km but <10.1 km. Alternatively, if you want to be more precise, you could define a variable $L$ to represent track length and then use "$10 < L \leq 20$ km" to represent this particular bin. Either way, the same change should be applied on L420.

L423-426: How did these studies define hail cells? If they used a lower reflectivity threshold, they are likely to get longer hail tracks because they will be including storms at earlier and later stages of their life cycles and are also less likely to break up tracks where the reflectivity temporarily drops below 55 dBZ.

L427-428: While the results for hail track duration may be similar to those for track length, it would still be nice to include the results in Fig. 10. You could also (or alternatively) combine length and duration to compute storm motion estimates for each cell and examine the distribution of this.

L429-432: It would make more sense to compute the track width as the average diameter or the cell (computed over its lifetime) in the direction perpendicular to its movement, since this will actually correspond to the width of the underlying hail swath (under the assumption that hail falls where reflectivity ≥55 dBZ). If you make this change, I would strongly encourage you to include the results in Fig. 10.

L433-436: Again, rather than considering cells at a particular time, why not use the whole swath? For orientation, you could compute it simply as the angle between the first and last points in the trajectory (similar to how you define track length). Alternatively, you could apply a line of best fit to the set of points defining the trajectory. I would recommend a Theil-Sen fit for this purpose, as it is less sensitive to outliers for small sample sizes, compared to a linear least squares fit. Also, you should note that the angle is defined as the direction from which the storm is coming and is measured clockwise from north.

L467-468: The key issue with the lack of 3D radar data is the inability to use more sophisticated proxies such as those based on echo-top height (e.g. POH) or vertical integrals of reflectivity (e.g. MESH), which generally show higher skill in hail prediction (e.g. Skripniková and Řezáčová 2015; Kunz and Kugel 2015; Puskeiler et al. 2016).

Figure 3: I would recommend presenting these results as annual hail frequencies rather than counts of the total number of hail days. This will make it easier to compare your results with climatologies for different periods (including the maps for 2006 and 2010 in Fig. 6) and in other parts of the world (e.g. Cintineo et al. 2012; Warren et al. 2020). Also, you should mask out those grid points that fall outside of radar coverage, as shown in Fig. 2.

Figure 4: For the zoomed in view of the Massif Central region, it looks like you've just cut out and blown up a section of the fairly coarse-resolution image on the right. As such it's very difficult to make out the details in both the orography and the hail frequency contours. I don't think you need the map of the full study domain as this is already shown in Fig. 1. Instead, I would make this a multi-panel figure, showing zoomed-in views (at an appropriately high resolution) for several of the hail hotspots visible in Fig. 3 and discussed in section 4.1.

Figure 5: Please present these data as the actual number of hail days for each year rather than the difference from the mean (which can easily be inferred).

Figure 7: As noted above, rather than the total number of hail days in each subdomain, you should plot the mean number of hail days (excluding points that are over the sea or outside radar coverage). The same change should be applied to Fig. 8.

Figure 9: It would be nice to show these plots for other hours, rather than just 02 and 18 LT. For example, you could group the data into 3h blocks (00-03, 03-06, 06-09, 09-12, 12-15, 15-18, 18-21, and 21-00 LT) and present the results as an 8 panel figure.

**Technical Corrections**

I am not sure what the standard is for this journal, but in English (both American and British), a period is used as the decimal separator and a comma (or sometimes a thin space) is used to break up numbers of ten thousand or higher. For example, twelve-thousand three-hundred and forty-five point six would be written as 12,345.6.

L10-11: Change “spatially most extended” to simply “longest”.

L12: Change “implied” to “produced” or “were associated with”.

L12-13: Change “2 Billions Euros” to “€2 billion”.

L20: This is not the correct use of “respectively” - it should only be used when describing two or more items that refer back to a previous statement. For example, “in northern Germany and southern France, hail occurs most frequently in August and June, respectively”. The same comment applies to L96 and L269.

L23: Change “major part” to “majority”.

L31: Puskeiler et al. (2016) consider the years 2005-2011, not 2004-2014.

L34: Change “criterions” to “criteria”. Also “echo top” should be two words.

L46: Change “allows” to “allow”.

L47: “sea” shouldn’t be capitalized.

L213: Change “weaker” to “lower”.

L229: Get rid of “recently”.

L317: I think you meant to put “(Figure 7)” at the start of this line.

L386: The definition of overshooting tops should be given when they are introduced on L37.

L394: This doesn’t need to be a new paragraph.

L421: Get rid of “including squall lines”. Also, it should be “MCSs”.

L446: “...from <u>French and German</u> national radar composites...”

L448: Duplication of “Mason”.

L557: Get rid of “to” before “orography”.

L466-467: Change “allows obtaining hail proxies” to “provides a proxy for hail occurrence”.

Figure 1: Please use different line thicknesses, styles or colors to distinguish between country and state/distinct borders. The same applies to Fig. 3, 4, and 6.

Figure 6: The colour bar is incorrectly labelled. The number of hail days in a single year will always be an integer, so you don’t need the range or the decimal place (i.e. the labels should just be 1, 2, 3, …, 12).

Figure 10: The x axes of these plots are incorrectly labelled. Each bin corresponds to a range of values, so the tick labels should be located under the ticks to illustrate this. So, for example, for panel (a) the ticks should be labelled 0, 10, 20, …, 310.

**References**

Cintineo, J.L., Smith, T.M., Lakshmanan, V., Brooks, H.E. and Ortega, K.L. (2012) An objective high-resolution hail climatology of the contiguous United States. *Weather and Forecasting*, **27**, 1235–1248.

Louf, V., Protat, A., Warren, R.A., Collis, S.M., Wolff, D.B, Raunyiar, S., Jakob, C. and Petersen, W.A. (2019) An Integrated Approach to Weather Radar Calibration and Monitoring Using Ground Clutter and Satellite Comparisons. *Journal of Atmospheric and Oceanic Technology*, **36**, 17–39.

Skripniková, K. and Řezáčová, D. (2014) Radar-based hail detection. *Atmospheric Research*, **144**, 175–185.

Soderholm, J., McGowan, H., Richter, H., Walsh, K., Weckwerth, T. and Coleman, M. (2017) An 18-year climatology of hailstorm trends and related drivers across southeast Queensland, Australia. *Quarterly Journal of the Royal Meteorological Society*, **143**, 1123–1135.

Warren, R.A., Ramsay, H.A., Siems, S.T., Peter, J.R., Protat, A. and Pillalamarri, A. (2020) Radar-based climatology of damaging hailstorms in Brisbane and Sydney, Australia. *Quarterly Journal of the Royal Meteorological Society*, **146**, 505–530.