Review of egusphere-2024-1336

This is a review of "Severe hail detection with C-band dual-polarisation radars using convolutional neural networks" by Forcadell et al.

The study analyses the performance of convolutional neural networks (CNN) in detecting severe (> 2cm) from non-severe (< 2cm) hail cases. The hail cases come from different sources of observations (ESWD hail reports, ANELFA hailpad networks, MeteoFrance mobile app). Three CNN architectures are compared and trained using 19 different radar-derived features on images of four different input sizes. Features comprise existing radar-based hail proxies and polarimetric radar variables subject to ongoing research. The proxy maximum estimated size of hail (MESH) is found to be the most beneficial to the CNN as an input feature. The study also shows that existing hail proxies can be adjusted using a threshold value and a threshold area to achieve similar performance to that of the CNN for severe hail detection. Ten fitted CNN models in inference mode are used as an ensemble on hail event, allowing to compute a probability of severe hail with a related uncertainty.

This a very complete study containing several interesting results. A very limited number of studies have used CNN in the context of hail detection and the use of both existing hail proxies and polarimetric variables as input features to the CNN is innovative. The research questions are not clearly stated in the introduction. Multiple data sources are considered, and several treatments are applied to the data to identify severe and nonsevere hail cases. This makes section 2 (data and method) very dense and sometimes unclear. The choice of the CNN architecture, of the input size resolution and of the relevant features is sound and well documented. The comparison of the CNN with the "optimized" existing hail proxies not only helps assess the CNN performance in a transparent way but is also relevant for the operational use of the hail proxies themselves. The results are clearly presented and thoroughly discussed, as well as well-illustrated and summarized by good-quality figures.

The paper is highly relevant to AMT and I strongly recommend its publication, although some important revisions outlined below are required.

General Comments

The introduction lacks a precise description of what the authors want to investigate. Specific research questions should be listed in the introduction, and the corresponding results should be discussed with respect to those questions. This would help the authors make their point.

Section 2 should be thoroughly revised and restructured. For example, I did not fully understand the approach for constructing the rain/small hail datasets. Moreover, it contains descriptions of elements (the 2nd cell tracking algorithm used for the case study, the storm modes classification) that are not key to the paper and blur the understanding of the essential points of the methodology. Here is a list of suggestions:

- In the introduction: explain why the two datasets are needed (we have this information only at L186).
- Rename "rain or small hail" to "non-severe" for clarity (or any other shorter name)

- Use “case” instead of “report” because both reports and hailpads are used to identify severe and non-severe situations.
- Section 2.1 should be merged with 2.6 to have all information on polarimetric radar data together.
- Section 2.2: The first cell identification algorithm should be described in the section discussing the identification of non-severe hail cases. The 2nd cell identification algorithm is used only at the end of the results section for a single case study. Its description could be moved to an appendix.
- Section 2.3 is called severe-hail reports but contains a description of the three datasets used to identify both severe and non-severe cases (ESWD, ANELFA, MeteoFrance app) and is a mix between describing those datasets and how severe hail is identified.
- I suggest separating the description of the 3 datasets into a dedicated section, followed by two sections describing the identification of severe and non-severe hail cases, where the specific filtering of each dataset for this study is described. The section dedicated to non-severe hail cases would include a description of the cell tracking algorithm.
- Section 2.4 should be rewritten using a step-by-step approach.
- The categorization in storm modes made in section 2.5 is not used in the results nor further discussed. It should be moved to an appendix to improve the flow (sections 2 and 3 are already dense).

Specific comments

L21: “targets” – the authors could be more specific by explicitly naming the targets of interest: hydrometeors such as raindrops or hailstones.

L27: “a given content” - Do you mean an equivalent scanned volume?

L46: every detection or forecast technique will have false alarms. What level of false alarms are found for those techniques? Are they relatively high? Be more specific.

L86: in terms of hail detection as an image-based problem, the authors should consider mentioning and briefly discussing the two following references :

Gagne, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. Monthly Weather Review, 147(8), 2827–2845. https://doi.org/10.1175/MWR-D-18-0316.1

Gagne, D. J., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., & Xue, M. (2017). Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. Weather and Forecasting, 32(5), 1819–1840. https://doi.org/10.1175/ WAF-D-17-0010.1

L101: According to Fig. 1 Southeast France is covered by S and X-Band radars. Did you remove this region from the study?

L103: So this means that the time resolution of the radar variables is 15 minutes?

L104: What is the maximum distance from the radar (radius) that is considered?

L107: “the data is not corrected for advection…”, but “corrected radar data was collected above hail reports”. Please explain what corrections were made.

L114: What is the national reflectivity composite? Please explain briefly how it is computed.

L143: What reflectivity threshold is used here? Why use reflectivity from the nearest radar to filter ESWD reports and the Morel and Sénési (2002) cell identification algorithm to filter the Meteo France app crowdsourced report? Please explain.

L153: If hailpads reports are not used for severe hail, move this paragraph to the rain or small-hail reports (section 2.4)

L180: Is this difference in frequency a relevant factor for the study? If so, I would discuss it in more detail. If not, I would not mention it.

L184: The Meteo France crowdsourced dataset is fully described in the severe hail reports, but then it is said that it is not used to identify severe hail reports. This is confusing.

L194-L204: Why not write the filtering criteria directly in the text to avoid redundancy?

L194 + L 199: So a maximum reflectivity of at least 45 dBZ?

L199: The use of “locations” is confusing. Why not use reports (or cases)?

L204-210: This is not clear to me. The goal is to build a dataset of confirmed cases of non-severe hail that does not overlap with severe hail cases. On L203, it is said that forbidden areas are defined around all hail reports, so how are the “negative” reports outside the forbidden areas in Fig. 3 obtained? There should not be any report left. How is a hailpad with a maximum hail size > 2cm classified?

L203: 120km x 120km - I guess that you used 120km x 120km to avoid any overlap between the 60km x 60km neighborhoods around the reports. However, in section 3.2 you find that 30 km x 30 km input images contain sufficient information for the CNN. Knowing that, did you try reducing the forbidden areas to 60 km x 60 km squares to potentially increase the number of rain or small hail cases? If you do so, how many cases does this add?

L205: In the previous section, 64051 reports are mentioned. How do you get to 62854?

L209: I would distinguish the uncertainty associated with collaborative reports from the one associated with hailpads. A collaborative report can be a joke or an error and you need to filter them out (using radar reflectivity or another approach) to improve your confidence that hail indeed occurred. However, if a hailpad has multiple dents, then you are virtually certain that hail occurred; the only uncertainty you are left with is related to the time of occurrence, which is estimated by an observer.

L280: 250 m x 250 m is the horizontal resolution and 500 m the vertical?

L281: Do you have grid points that are covered by only one radar? What do you do in this case?

L282: It is not clear to me how the ROI is calculated. Can you show an example?

L295: The polarimetric data is computed for C-Band radar only. Do you only consider C-Band here as well?

Table 3: Hail proxies are named hail algorithms in section 2.7. Please use the same name throughout the paper to avoid confusion.

L328: input features: How are the features computed with respect to time? Do you use the closest radar timestep from the report time or an aggregation (maximum, average) over a time window made of several time steps? Did you test different time windows? How does it influence your results?

L348: “these features”, do you mean all features or only the polarimetric ones?

L350: Is it the same interpolation that is described in section 2.6 or another one? This is not clear.

L367: How do you get from 1169 and 2605 to 2335 and 5188?

L384: Is it possible to include an illustration or a description of the ResNet architecture for comparison with those of the SmallConveNet and ConvNet? Why did you choose to compare those three CNN specifically? Can you shortly explain what are the main differences between the three CNN?

L384: input sizes: The initial image size is 60 km x 60 km. How do you get from this size to the different input sizes? (max pooling, average pooling, a window centered on the report location?)

L423: By class above or equal to small hail, do you mean the small, medium, and large classes?

L440: Please indicate the min/max value of the AUC-ROC and AUC-Pr.Re, ie. what is the best achievable performance according to those metrics (e.g.: 1)?

L445: What about the results for the 50 km x 50 km input size?

L464: “larger images”: which resolution/size?

L549: Explain what each figure and table show (Fig. 13 shows the ROC and Pr.Re curves for the Conv fitted models and the hail proxies, while Table 7 shows the corresponding AUC values. Introduce Fig. 14 and Table 8 when they are discussed.

Figure 10: “Models were also trained with an input size of 50km x 50km, but no amelioration was obtained (not shown).” This should be mentioned at the beginning of the section, not in the figure caption.

L476: Move to the beginning of the section.

L485: random sample: did you limit your selection to pixels where at least one of the feature was not zero?

L526: What do you mean by significance here? Importance?

L526: “finer texture of the field”: I understood that all the features had the same horizontal resolution (250m x 250m). What do you mean by texture?

Table 7: The caption misses a description of the beta_x column.

> A column with the corresponding beta_Ax values would be extremely relevant
>
> Several studies analyzing the skill of hail algorithms use the Critical Success Index (CSI) and the HSS (Heidke Skill Score), which can be easily computed from the contingency table. It would be interesting to have those scores in Table 7 for the best beta_x and beta_Ax pair of each score (+ those of the ConvNet).

L561: What do you mean by local here?

L565: Again, what is the meaning of local in this context?

L565: Why is the beta_x value for MESH_95 = 33 mm in the text and 31 mm in Table 7?

L578: “varies significantly depending on the hail class.” Why don't you show the results for each class instead of the average? It only adds two lines to the table and the reader will have the complete information.

L606: It's not clear what accordingly refers to. You could write “…the threshold value is optimized”.

L611: Do you have the predictions of MESH_95 in this case for comparison? As this is an important feature of the ConvNet, it guess that it should be similar.

Figure 14: For the two POH and POSH, why don't you show the full range of thresholds up to 100%? This is interesting information. Change the y-axis ticks labels to a multiple of 10 for readability.

L638: “utilising the radar information of a unique timestep.”: This should be mentioned in the method section (see my previous remark for L328).

Figure 15: On the bottom right there are two circles without grey lines. What are they?

L646: Did you look at other hail events in detail or only this one? What were the results? This is out of curiosity and doesn’t need to be included in the paper.

Technical corrections

L44: references should be sorted by year of publication.

L59: add “radar variables” after dual-polarization

L62: references should be sorted by year of publication.

L77: replace “traction” by attraction

L77: “In the work of Wang et al. (2018), they developed a CNN applied to...” → Wang et al. (2018) applied a CNN to...

L80: “In the work of Shi et al. (2020), they tracked...” → Shi et al. (2020) tracked...

L84: Ackermann et al. (2024) trained...

L97: “state-of-the-art”... hail detection methods?`

Table 1: I would write explicitly the 3 lowest angles on each row to avoid confusion (even if they are the same).

Figure 1 caption: in the pdf, ESWD reports appear in grey-blue instead of grey, whereas reports from the Météo-France app appear in grey instead of black.

L104: “the **three** upper elevation angles”.

L106: “corrected **for** attenuation”.

L153: hailpad in one word.

L223: **between**

Figure 7: The label of the grey contour is not visible, consider using another color (e.g. red). Use the same scale for both colorbars for direct comparison. The max value for rain or small hail cases is higher than for severe hail.

Table 5: True **N**egative

L457: **decreasing** instead of increasing?

L458: Define ResNet18 on L384.

L522: **sensitive** instead of sensible.

L531: remove rho_HV

L548: State of the art is subjective. Replace with an objective word, e.g.: “Comparison of CNN with **hail proxies/algorithms**”.

L551: **high** instead of strong

Figure 15: the green used severe hail reports appear dark and difficult to see. You could use the same green as in Fig.3 for the reports.