Review comments for “Severe hail detection with C-band dual-polarisation radars using convolutional neural networks” by Forcadell et al.

**Indications of authors:** answers to each comment are in blue. Text in quotes shows portions of the modified text. Underlined text shows the modifications.

The authors would like to thank the reviewer for his/her thorough analysis of the paper that will definitely increase its quality. We hope that readers will find the answers useful.

This work constructed a convolutional neural network (CNN) based hail occurrence model trained on dual-polarisation radar data over France, and the training “truth” combines three ground datasets including a crow-sourcing one with careful screening and quality control. In addition to the radar measured variables, some traditional hail prediction proxies are also included as the input features. The target is to predict either severe hail (flag = 1) or rain/small hail (flag = 0). The machine learning (ML) model performance is then comprehensively compared to previously used hail detection proxies for performance statistics, and the feature importance for the model is also thoroughly evaluated. The CNN model outperforms all 6 traditional proxies in all evaluated metrics.

This work is carefully designed and thoroughly conducted. The quality-control of the training “truth” dataset involves a great amount of work, which is highly appreciated (I don’t find the open science statement, but do think it would be very valuable if the training dataset can be published somewhere). The ML model architecture selection and fine-tuning are carefully executed. The results are concrete.

However, I do feel the design of the work has limited contributions to advance science, mainly because of the concern that the input features include the 6 proxies, and one of them dominates the determination processes according to the feature importance rank (and that’s partially because some other highly correlated proxies are removed before feature ranking, or otherwise, they’ll all rank high). So scientifically speaking, the new ML model is a “smart improved version 2.0” of the previous proxies. Since processing input features reads like not easy (e.g., interpolation using two adjacent radars to reconstruct the 3D fields, and then interpolate to 2D images, etc.), I doubt the applicability of the new ML model to operational use, given the traditional proxy data seem to be much easier to be calculated and the performance the best two traditional indices are only slightly worse (Fig. 13 and Table 8). In your revised version, I’d strongly recommend the authors adding one paragraph in the discussion or summary about their thoughts of the scientific merit and applicability of their work to the future.

Authors’ comment:

It is important to note that we did not compare the “raw” hail proxies (MESH, POSH, POH) and our CNN approach, but a heavily “optimized” version of them. In Table 7, the performance of the five best models with feature thresholds (i.e. beta_x) that lead to the best performance are shown. On top of this feature threshold, an area threshold was implemented for the hail proxies to further increase their performance and add fairness to the comparison with the CNN. The study shows that, indeed, when these two thresholds are well tuned, the performance of existing hail proxies can be drastically increased to (nearly) reach the performance of the CNN approach. Newly added scores show that the CNN beats the existing hail proxies by a certain margin (see Table 7).

On the operational use of the CNN, the necessary time to produce the input features is easily counterbalanced by the fact that the inference is performed on detected storm cells only. Hence, the inference is only performed around cell centroids rather than the whole radar domain, drastically decreasing the processing time of volumetric radar data.

A discussion has been added in the Conclusion:

"[...] is recommended to examine cells that have produced reflectivities of at least 45 dBZ. The cell-identification algorithm and the production of input features for the CNN may require a greater investment of computational time and resources than existing hail proxies. The necessary 3D interpolation can be particularly costly. However, this additional computational time can be offset in real-time by the cell-identification algorithm. The input features can be generated [...]"

Another minor concern is the length. It's a bit too lengthy right now, easily causing readers missing highlights of your work. I'd suggest moving some definition of the common ML terminologies (ROC, AUC, confusion matrix) to the appendix, as well as detailed procedures of the QC of your training "truth". This comment being shared with another reviewer, the following parts have been moved to Appendix to improve the readability of the article: the description of the second cell identification algorithm and the storm-mode assessment.

Minor caveats:

1. Reconstructed 2D images from one closest radar and a second-closest radar are both used. But it was  never discussed (or I might have missed) what are the differences for the prediction? Is it more practically to just use the closed image? Or does the result really sensitive to the distance between radar location and event location? Only the "Polarimetry" features in Table 3 are produced independently for each case using each of the two nearest radars. The other features are extracted from the 3D grid, which is created using both radars. The addition of "Polarimetry" features of the second-nearest radar was a way to augment the dataset and to make a prediction in operational use even if data missed from the first radar. It was also a way for the CNN to be less sensitive to the distance to the radar as situations where the radar is far away (by more than 130km) are used for the training, mainly coming from the second-nearest radar. Differences in the prediction were not particularly explored depending on the radar used for the "Polarimetry" features. It is expected that the predictions from the second radar using the CNN might be less robust than using the nearest radar in most of the cases, particularly if the storm is already well sampled by the first radar, or if artefacts prevent a correct sampling of the storm by the second radar (PBB, attenuation). Fig. 15 shows an inference example with the "Polarimetry" features produced with the first radar only, because it is indeed more practical. Ideally, two samples for each centroid should be created and the prediction averaged among both: one with the "Polarimetry" features from the first radar, and another with the "Polarimetry" features from the second radar.

2. Starting from comparing ML results to various previous hail detection proxy variables (Section 4.2), the ResNet is dropped for discussion. Why? Is it because training ResNet takes significantly longer time than training a ConvNet? All CNNs tested in this study took relatively the same time to train, mainly because the dataset can be considered “small”. If the dataset was 10 times bigger, the question of the resources needed to train the ResNet would have arisen. The ResNet is dropped from the discussion during the tuning phase (see section 4.1). It is dropped because it reaches the same performance on the validation dataset compared to other models, but at the cost of being much more complex. More complex models on small datasets have a tendency to overfit as they are exposed to learning noise in the data rather than important features. It is shown in the evolution of the validation loss during consecutive epochs (Fig. 09), where huge oscillations are rapidly visible for the ResNet.
3. How was “SHI” defined? It was never clear to me. If it has an explicit analytical format involving radar measured quantities, then it’s a “smart use” of radar measurements, and you can directly use “SHI” as your training input feature.The definition of the SHI has been added to the text in the form of Equations in section 2.5:

$$\mathrm{SHI} = 0.1 \int_{H_0}^{H_t} W_T(H_T)\dot{E}dH,$$

with

$$\dot{E} = 5 \times 10^{-6} \times 10^{0.084Z} W(z),$$

$$W_T(H) = \begin{cases} 0 & \text{for } H \leq H_0 \\ \dfrac{H - H_0}{H_{-20} - H_0} & \text{for } H_0 < H \leq H_{-20}, \\ 1 & \text{for } H \geq H_{-20} \end{cases}$$

$$W(Z) = \begin{cases} 0 & \text{for } Z_H \leq Z_L \\ \dfrac{Z_H - Z_L}{Z_U - Z_L} & \text{for } Z_L < Z_H < Z_U, \\ 1 & \text{for } Z_H \geq Z_U \end{cases}$$

The SHI could have been added as an input feature as well, while removing its highly correlated counterparts. It is anyway indirectly implemented in MESH proxies and would be perfectly Spearman correlated to MESH. The importance of SHI in the input can be seen as the importance of MESH, and vice versa.