Point-by-point response to the reviewer comments

Manuscript egusphere-2026-992: *Enhancing weather radar data by removing non-meteorological echoes, using neural networks trained on synthetic weather data*

R. Bölz, T. Kirstein, L. Fuchs, L. Josipović, A. Böhm, U. Blahak, V. Schmidt

We thank both referees for careful reading of the manuscript. Their suggestions and comments helped us to improve the presentation of the material, and we took all of them into consideration when revising the paper.

The most important changes in the revised manuscript are the following:

- The data of the withheld radar station is no longer used for testing. It now serves as validation data, used for early stopping and for model selection described in the new Section 2.6, in which the number of feature channels and the number of training epochs are selected based on the validation loss.
- The network is now tested exclusively on the five expert-labeled measured sweeps. These sweeps are now introduced in more detail, and the performance metrics are reported for each of the five sweeps in a new table.
- A new ablation study covers all seven combinations of the DBZH, ZDR and UDR input channels; the results are reported in a new table and interpreted in the Discussion.
- Composite plots were added to characterize the time periods from which the training data originates.
- The Discussion now states the limitations of the training and test data explicitly.

In the following, we give a point-by-point reply, where each referee comment is quoted and followed by our response.

Response to Referee 1 (RC1)

Reviewer comment RC1 — general assessment

This manuscript presents an interesting idea on how to address the lack of expert-labeled data in classification of weather radar echoes. Using this approach, the authors train a U-net model to classify meteorological and non-meteorological echoes. The issue being addressed is critical to producers and users of radar data, and the authors achieve promising results with their model. However, the small amount of data used to train and validate the model raise questions that should be addressed in the manuscript.

Author response: We thank the referee for careful reading of the manuscript and for their positive feedback.

Reviewer comment 1

One major concern is the small amount of data used in the model training. It appears that the dataset contains data from 3 hours, plus some test data from a time period that is not mentioned. Even if the dataset contains measurements from all 17 radars in the DWD

network, the small temporal windows covered by the dataset raise serious concerns of the resulting model validity. Preferably, the authors should increase the dataset size. If using more data for the model training is not possible, at least the following issues should be addressed with sensitivity tests or discussion in text:

Author response: Thank you for this comment. While we have not increased the amount of data used for this study, in the revised version of the manuscript we have adjusted various aspects on how this data is utilized and described. Most importantly, the data of the withheld radar station is no longer used for model evaluation, and is instead only used as a validation set for model selection. The remaining test data is now described more clearly, with exact dates given and how they relate to the training and validation data. We have clarified that the test data includes four sweeps that are independent of the training data, with the remaining evaluation sweep originating from the period used for the calibration of the synthetic weighting factors. In particular, four of the test sweeps originate from the 5th and 30th of June, whereas the training, validation and calibration summer sweeps originate from the middle of June. In the Discussion we now address that the five test sweeps are unlikely to represent the wide range of rare summer artifacts that the network was not trained on.

Reviewer comment 1.1

How representative are the selected time periods of the conditions they aim to represent? E.g. for cluttered summer measurements it is worth to note that the appearance of insect echoes can vary depending on temperature, diurnal cycle, and annual cycle. Similarly, how representative are the selected winter sweeps?

Author response: We added a more detailed description of the training dataset in the Radar data section and added a composite plot to visualize the different precipitation and insect features. We chose the time period of the winter data to show a variety of meteorological echoes with a minimum of non-meteorological echoes. We chose the time period for the cluttered summer data to show a variety of strong and weak non-meteorological echoes.

Reviewer comment 1.2

How representative are the scaled winter images of summer precipitation?

Author response: The precipitation structures and polarimetric signatures of the scaled winter images remain those of the winter sweeps. However, the network trained exclusively on the synthetic radar images performs similarly well on the synthetic validation data and on the experimentally measured test data (intersection over union of 0.946 and 0.930, respectively), which indicates that the synthetic radar images are representative enough for the network to learn the features of measured mixed radar images. The remaining limitation, i.e., that precipitation structures occurring in summer only are not represented in the training data, is now stated where the generation of the synthetic training data is introduced.

Reviewer comment 1.3

Does the selection of the sweeps require some considerations? If one were to attempt to repeat the study using different measurements, how should one address the dataset selection?

Author response: The time periods of the sweeps were chosen such that all radars showed interesting data during those periods. Loosening this constraint and choosing relevant periods separately for each radar, might improve the variety of the data.

Reviewer comment 1.4

When are the "experimentally measured mixed radar images" measured? How were the images selected?

Author response: The dates, times, and radar stations of the five sweeps are now stated where the test dataset is defined in the Results section and in the new per-sweep table: 09:00 UTC on June 5, 2022 (Essen and Neuhaus), 09:00 UTC on June 16, 2022 (Dresden), and 16:00 UTC on June 30, 2022 (Neuheilenbach and Türkheim). We also stated that the five test sweeps were selected to cover a variety of scenarios: stratiform and convective precipitation, convection close to the radar and at longer ranges, morning and afternoon data, and different radars. They were labeled by a radar meteorologist of DWD.

Reviewer comment 1.5

As far as I can tell, Figure 5d shows differential attenuation that is unlikely to be present in any winter measurement and thus would not appear in your training material. How does the model perform for this case or other similar artefacts in summer measurements that are not present in winter?

Author response: Yes, you are correct. The differential attenuation seen in Figure 5d would not be present in the winter dataset. It also does not occur in any of the five test sweeps. The Discussion now states this as a limitation. The scene is the ZDR image of the Memmingen sweep at 09:55 UTC on June 16, 2022, one of the mixed summer sweeps from which the weighting factor was determined.

Reviewer comment 2.1

The process for selecting the scaling factor seem rather arbitrary. When comparing the images, did you compare mean values, max values etc?

Author response: We have clarified the description in the manuscript. We did not compare mean or maximum values of the whole images. Instead, we manually sampled pixels within similar meteorological structures of the synthetic and the mixed summer radar image. We believe that a purely quantitative evaluation is difficult, as comparing the statistics of different meteorological events could be misleading.

Reviewer comment 2.2

How representative do you expect the scaling determined in the described way to be over longer time periods? Is it impacted by calibration differences etc?

Author response: The scaling is determined to be representative for sweeps from June. Our training includes a random variation of the factor for each training sample, resulting in differences of up to 3 dB. This should offset radar calibration, which the DWD aims to set at 1 dB around the true value.

Reviewer comment 2.3

Ideally, the scaling factor should be related to some physical factor or explanation, e.g. differences between Z-R and Z-S relations rather than subjective selection.

Author response: Assuming that the difference between summer and winter observations arises primarily because measurements are taken below the melting layer (in water) in the summer case and above the melting layer (in ice) in the winter case, the discrepancy in reflectivity values is caused by the differing dielectric constants of water and ice. For liquid raindrops, the dielectric factor $|K|^2$ is approximately 0.93, whereas it is roughly 0.18 for ice. This results in a difference of approximately a factor of 5, which is very close to our scaling factors of 5-8. However, the actual difference also depends on the weather in the two periods. We therefore determined the weighting factor from the data itself.

Reviewer comment 2.4

Assigning of UDR when creating the synthetic images: did you test other approaches to assign the value, e.g. weighted sum of the original images? Now, as far as can I follow, a radar gate in the synthetic image has contributions of both the winter and summer images in DBZH and ZDR but not in UDR? Do the resulting UDR values in the synthetic images follow a similar distribution (on their own and joint distributions with DBZH/ZDR) as experimentally measured values?

Author response: Your reading is correct, and we have now stated this more clearly. For pixels included in both radar images, the synthetic DBZH and ZDR values combine both contributions, whereas the synthetic UDR value is taken from the winter radar image only. We did not test other assignment schemes such as a weighted sum, since there is no physically obvious way of combining the two depolarization ratios, as the manuscript now states. However, the manuscript now includes an ablation study on the input channels: adding the UDR channel to DBZH improves the intersection over union on the five expert-labeled test sweeps. In the single channel case, the UDR channel does appear least performant when evaluated on the validation or test dataset.

Reviewer comment 3.1

There is no validation dataset. Typically, ML model training should include a training dataset used to train the model, validation dataset for hyperparameter selection and monitoring training convergence (i.e., selecting the best model outcome among all possibilities), and test dataset for independent validation of the selected model. Since there is no validation dataset, how is the training convergence monitored?

Author response: Thank you for this comment. The data of the withheld radar station, which was previously referred to as the synthetic test data, now serves as validation dataset. This validation dataset is used for early stopping and to select the number of feature channels and training epochs, see the new subsection "Model selection and early stopping". The five expert-labeled measured sweeps form the test dataset and are used neither for training nor for model selection. All networks were retrained with this scheme, and all reported metrics were recomputed.

Reviewer comment 3.2

Given that the training and test datasets are temporally overlapping, how was information leakage between the datasets reduced? The test dataset being from a different radar site does not automatically remove information leakage, as the precipitation areas move and same precipitation could easily be present in the measurements of multiple radars; we can also expect that precipitation within one hour is correlated within different areas inside Germany. Additionally, the measurement range is said to be 150km; do the images from

multiple radars overlap?

Author response: We agree with this comment. The measurement ranges of neighboring radars overlap, and the same precipitation can appear in the sweeps of several radars within the same hour, so the split by radar station does not yield an independent test set. We therefore no longer use the held-out radar for testing. The synthetic datasets generated from its sweeps are now the validation datasets, used for monitoring the convergence of the training process and for model selection; no claim about generalization is based on them.

Reviewer comment 3.3

It is unclear how are the two winter and summer measurements used to a create synthetic image selected? Randomly, matching time from start of the measurement period, something else? How does this selection impact the dataset and model skill? Do you only combine images from a single radar site?

Author response: Thank you for this comment. The synthetic training dataset contains the combinations of all winter and all cluttered summer radar images of the training stations. The pairs are not matched by radar station or measurement time, and the manuscript states this more clearly now. During training, the samples are drawn uniformly from this dataset, so every combination occurs with the same probability. This selection multiplies the number of combinations by the number of radar stations used for training, and we have not tested more limited selection schemes.

Reviewer comment 4.1

How is the model training convergence monitored and how do you decide if the training has converged?

Author response: The new subsection "Model selection and early stopping" describes the stopping criterion. The training is stopped once the validation loss has not decreased over three consecutive intervals of five epochs, and after 100 epochs at the latest. The validation losses over the epochs are shown in a new figure, and determine the final model selection.

Reviewer comment 4.2

Please list all relevant hyperparameters used in the training (e.g. learning rate), and refer to any relevant libraries used in the model implementation and training.

Author response: The relevant hyperparameters are now listed in a table after describing model training. The deep learning library is now also named in the manuscript.

Reviewer comment 5.1

The colormap of ZDR measurements should be limited to show only the interval of interest, which the authors state to be around 0 dB to 20dB (not starting from -20dB)

Author response: Values below 0 dB occur in the data regularly, especially for meteorological echoes. Therefore we have not adjusted the shown interval.

Reviewer comment 5.2

It would be better to show the excluded radar gates with some color that is visible to aid the reader in interpreting the figures

Author response: Thank you for this comment. We have adjusted the colormap of the differential reflectivity to not include white. As a result, the pixels of white color can now be distinguished from all included pixels.

Reviewer comment 5.3

I'm not sure if there is a need to repeat the range ring labels in every image; the images would be less cluttered if those were removed especially in the smaller images

Author response: We agree that it is not strictly necessary to show the ring labels on every plot. However, we feel this makes clear that the shown images always represent the entire sweep.

Reviewer comment 6

I would appreciate more specificity on the description of radar measurements in introduction. There is also some repetition in the descriptions in the introduction and Section 2.1 that could be reduced. Specific comments:

Author response: Thank you for this comment. To address it, we had to compromise with the other reviewer, who had criticized the too extensive description of radar measurements. While we now introduce the most important polarimetric moments in the introduction and remove repetitions in Section 2.1, we have decided to not add more general details of radar measurements.

Reviewer comment 6.1

Lines 51-54: If talking about radar moments, it would be better to name them, e.g. "compute so-called radar moments, such as radar reflectivity representing the strength of the signal" etc.

Author response: The introduction now names the most common polarimetric radar moments together with their general interpretation in relation to hydrometeor type, asymmetry and spatial orientation distribution.

Reviewer comment 6.2

Lines 65-66: I would interpret this to mean radar systems with waveguide switches; how about radar systems that transmit H and V polarizations simultaneously?

Author response: You are right, thank you! We added a description of the technology with simultaneous transmit and receive (STAR) of H and V polarization and the fact that such radars have some technical advantages (simpler design) but cannot measure the linear depolarization ratio. Actually, the DWD radars and the data used in this paper are using STAR mode, which is now also clearly stated at the beginning of Section 2.1. It is the whole reason why we use UDR instead of LDR, which we now explicitly state in Section 2.1. Again thank you, this helped to clarify things!

Reviewer comment 6.3

Lines 219-221: this seems repetitive

Author response: We agree with this comment and have streamlined the wording.

Reviewer comment 7

The description of the state-of-the-art method in section 3.2 is confusing. It would be better to order the description so that steps are described in the order that they are performed. For example, paragraph starting on L609 should be after the eligible pixels are first mentioned, and the paragraph starting on L626 should follow them

Author response: Thank you for this comment. We have changed the wording in the introductory sentences in l.600-604 to make the order more clear.

Reviewer comment 8

Eq. 1: I assume θ denotes the azimuthal angle? This should be mentioned in the text

Author response: We have clarified the meaning of both variables in Eq. 1, and also changed the variable name for the azimuthal angle since θ is also used for the network parameters.