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 2 (RC2)

General comments

Reviewer comment 1

The manuscript presents an interesting approach for identifying non-meteorological echoes in polarimetric weather-radar data. The central idea of training a U-net on synthetically generated mixed radar scenes is original and addresses the important practical difficulty of obtaining sufficiently large pixel-wise labelled datasets. The manuscript is generally clearly written, and the comparison with the operational DWD method indicates that the proposed approach may better preserve weak precipitation and the boundaries of meteorological echoes.

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

Reviewer comment 2

My main concern is the limited size, diversity, and independence of the datasets used for training and testing, as well as the apparent absence of a separate validation dataset. The meteorological winter dataset and the cluttered summer dataset each originate from a single one-hour period. To my understanding, although one radar is withheld from the synthetic training set, data from the remaining radars are collected on the same dates and may contain the same large-scale meteorological or non-meteorological structures. Therefore, the current train–test split mainly evaluates transfer to an unseen radar rather than generalization to independent events and atmospheric conditions. The evaluation on measured data is additionally based on only five manually labelled sweeps, whose exact dates and independence from the data used to develop the synthetic samples should be reported clearly. All figure captions showing radar data should also report the date and timestamp of the displayed event.

Author response: Thank you for this comment. We have fundamentally reworked the train-test split. The data from the withheld radar station is no longer used for testing. Instead, it serves as a validation dataset, used for model selection only. The individual radar images of the test dataset are described in more detail, four of them originate from the 5th and 30th of June, whereas the training, validation and calibration summer sweeps originate from the middle of June.

Reviewer comment 3

I strongly encourage the authors to evaluate the already trained model on at least one fully independent date containing both meteorological and non-meteorological echoes, as this would substantially strengthen the assessment of generalization without requiring changes to the proposed methodology. Such an evaluation would provide an important test of transfer to unseen conditions, although a single additional date would still not be sufficient to demonstrate broad robustness across seasons and event types. I would therefore recommend using an independent and diverse test set that is representative of the range of meteorological and non-meteorological conditions typically encountered in operational data. If additional independently labelled data cannot be obtained, the limitations of the current test design should be discussed explicitly, and the claims regarding generalizability, robustness, and absence of overfitting should be moderated accordingly. Although this limitation is briefly mentioned in lines 738–740, it is a central issue and should receive greater emphasis.

Author response: Thank you for this comment. Two of the three test dates are fully independent: four of the five expert-labeled sweeps were measured on June 5 and June 30, 2022, dates from which no data was used for the generation of the synthetic training data, for training, or for model selection; the fifth sweep, from June 16, 2022, was measured in the hour from which the reflectivity weighting factor was determined. The individual results of the five sweeps are now listed. Labeling further dates was not possible within this revision. The Discussion now states the limitations of the test data explicitly: the statement on overfitting rests on the similar metrics on training, validation, and test data and training was stopped early based on the validation loss. The robustness claim is reduced to the transfer to measured summer sweeps of other days and times of day. The limitation of the training data, that precipitation structures occurring in summer only are not directly represented, is now also pointed out in Section 2.4.

Reviewer comment 4

Given the limited training sample and the complexity of the U-net, I would also encourage the authors to assess whether model performance has saturated, for example through a learning-curve analysis using progressively larger training subsets and a fixed independent and diverse test set.

Author response: The Results section now includes a figure of the validation loss over the epochs of the network training, and this loss is now used for model selection. The selected network is now the network with 64 feature channels (in the first convolutional layer) trained for 90 epochs. A substantially larger test set is however beyond the scope of the current manuscript.

Reviewer comment 5

The manuscript would also benefit from a clearer explanation of how the final network architecture and training hyperparameters were selected. It is currently unclear whether a separate validation set or cross-validation was used for hyperparameter tuning.

Author response: Thank you for this comment. We have added the subsection "Model selection and early stopping" and a table of all hyperparameters. The number of feature channels and training epochs are selected by the loss on a separate validation dataset, the data of the withheld radar station. The remaining hyperparameters were fixed before training and not tuned, which is now stated.

Reviewer comment 6

I would also welcome some analysis of how the U-net relies on the individual radar moments, as this would improve the model interpretability (see specific comment for Sect. 4).

Author response: Thank you for this suggestion. The revised manuscript now includes an ablation study on the input channels: one network is trained for each of the seven combinations of the DBZH, ZDR, and UDR channels, with otherwise unchanged training, and evaluated on the validation and test datasets. The results are reported in a new table and interpreted in the Discussion: the horizontal reflectivity is the most informative single channel, and the best test performance is obtained for the model trained on DBZH and UDR. When DBZH and UDR are used, adding the ZDR channel does not improve the performance on the test dataset.

Reviewer comment 7

Overall, I find the study promising, but I recommend revisions mainly related to the amount and diversity of the data used for training and evaluation.

Author response: Thank you for this comment. While we have not increased the amount of data used for this study, 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.

Specific comments

Reviewer comment 1

Lines 35–37: The scanned volume is not exactly conical. Please rephrase to indicate that the radar scans "approximately" conical volumes.

Author response: Thank you for this comment, the scanned volume is now introduced as approximately conical.

Reviewer comment 2

Lines 46–64: This paragraph provides a rather extensive textbook-level description of the radar measurement principle, which may not be necessary for the intended readership. I suggest shortening it considerably and retaining only the aspects directly relevant to the problem addressed in the study.

Author response: Thank you for this comment. However, the other reviewer criticized the lack of detail in the description of radar measurements and we have therefore decided not to shorten this paragraph.

Reviewer comment 3

Lines 65–74: This paragraph provides an oversimplified and partly inaccurate description of dual-polarization radar measurements. Dual-polarization radars do not necessarily alternate horizontal and vertical polarization from pulse to pulse, as many operational systems transmit both components simultaneously. In addition, the preferential horizontal orientation applies primarily to oblate raindrops and should not be generalized to all precipitation particles. The description of the derived polarimetric variables as resulting from "differences and their correlations in time" should also be clarified.

Author response: Thank you and we agree with this comment. The description has been changed to discriminate between alternating H and V pulses and simultaneously transmitting and receiving H and V (STAR mode) radars. Actually, the DWD radars operate in STAR mode and this is the whole reason why we use UDR instead of LDR in this study. Both points are now also explicitly clarified in Section 2.1. Also, the description of the aerodynamic behaviour of asymmetric particles has been sharpened. Of course there is not always a perfect horizontal alignment, but the preferentially horizontal orientation (in an average sense) does not only apply to oblate raindrops, but also to snowflakes, pristine ice particles and other asymmetric types, overlaid by more or less tumbling and wobbling (canting angle distribution) depending on the hydrometeor type. We have also added the names and (very) general interpretation of the most commonly used polarimetric radar moments.

Reviewer comment 4

Lines 72 and 96–107: The discussion in lines 96–107 is more closely connected to the actual classification problem than the preceding general description of radar operation. I suggest briefly introducing, possibly already around line 72, why DBZH, ZDR, and UDR provide complementary information for separating meteorological and non-meteorological echoes, while leaving their full definitions to Sect. 2.1.

Author response: We now introduce the most common polarimetric moments and their general interpretation here, but we would like to keep the more in-depth elaboration on their interpretation relating to meteorological/non-meteorological targets in Section 2.1.

Reviewer comment 5

Lines 116–118: The statement that conventional classifiers can incorporate information only in close proximity to the classified bin is not generally true, since spatial and temporal aggregate features may also be included. Please qualify this statement. In my opinion, the clear advantage of the U-net is that feature engineering is not necessary.

Author response: Thank you for this comment, we have adjusted the text accordingly.

Reviewer comment 6

Lines 127–129: Please briefly clarify how the deep-learning approach in Atanbori et al. (2025) dealt with noisy or uncertain labels and why this improved the classification. In particular, what characteristics of the proposed deep learning model made it suitable for learning from noisy training data?

Author response: Thank you for this comment, we have rewritten the sentence to give a more precise overview of the utilized method.

Reviewer comment 7

Lines 142–157: The term “infeasible” may be too categorical; wording such as “extremely laborious” may be more appropriate. The sentence at line 156 describing scaling, rotation, and orientation inversion is too methodological for the Introduction and could be removed.

Author response: Thank you for this comment. We have adjusted the wording accordingly.

Reviewer comment 8

Sect. 2.1 title: The title could be more precise, since the section also describes the radar moments used in the study and not only radar-data acquisition.

Author response: We have changed the title to "Radar data".

Reviewer comment 9

Lines 214–215: Please use the standard term “copolar cross-correlation coefficient”.

Author response: Thank you, this is indeed more precise and we changed it.

Reviewer comment 10

Fig. 1. A short description of the main differences among the winter, mixed-summer, and cluttered-summer examples would be helpful when introducing Fig 1 for the first time.

Author response: Thank you for this suggestion. We added composite plots along with some description in the text to illustrate the nature of the meteorological and non-meteorological echoes of the training data.

Reviewer comment 11

Lines 226–239: The temporal coverage of the dataset is quite limited. It would be useful to acknowledge this when discussing the representativeness of the training samples and the generalization of the proposed method. Please discuss more explicitly how representative the selected winter sweeps are of the diversity of meteorological situations. Since the meteorological component of the synthetic training data originates from a single winter period, the model is likely not exposed to sufficiently diverse precipitation structures and polarimetric signatures (e.g. summer convection). If possible, please provide examples addressing performance under meteorological conditions that are poorly represented in the training data.

Author response: Thank you for this comment, in the Discussion we now state that the five test sweeps are unlikely to represent the wide range of rare summer artifacts that the network was not trained on. Despite this, the overall network performance on the manually labeled summer sweeps (two from the 5th of June, one from the 16th of June, and two from the 30th of June) is quite promising.

Reviewer comment 12

Sect. 2.3: Please clarify how the final network configuration and architecture were selected. Were alternative architectures or different U-net configurations tested, and was the chosen setup based on a preliminary sensitivity analysis or evaluated against validation data?

Author response: Thank you for this comment. We believe that the scope of the manuscript does not allow for a comprehensive comparison of a wide variety of network architectures. From our perspective, the U-net is an appropriate first choice, though there are likely architectures that could outperform the chosen architecture. However, the manuscript now includes a study on hyperparameter choice, in line with the available data. In particular, the hyperparameter that was varied is the number of feature channels in the first convolutional layer, with possible values of 8,16,32 and 64. The other hyperparameters are fixed design decisions, largely adapted from the original U-net paper. The new subsection "Model selection and early stopping" describes the selection procedure, and the validation-loss curves for these different values are shown in the Results section.

Reviewer comment 13

Lines 387–392: Please clarify whether the choice of $\lambda_w \in [5, 8]$ was supported by a quantitative comparison in addition to visual inspection of the representative cases.

Author response: The choice was supported by comparing the DBZH values of manually sampled pixels within similar meteorological structures of the synthetic and the respective mixed summer radar image, for each of the four representative cases. The manuscript now states this clearly. We believe a purely quantitative evaluation is difficult, as comparing the statistics of different meteorological events could be misleading.

Reviewer comment 14

Fig. 4: Since -40 dBZ is used as the placeholder value, the color scale should extend down to -40 dBZ. Alternatively, use an "under" indicator to make clear that all values below -20 dBZ share the same color; otherwise, placeholder values and clipped low-reflectivity values cannot be distinguished. The same should be applied to Figs. 7; 8.

Author response: We have added "under" and "over" indicators to the colorbars of the respective figures.

Reviewer comment 15

Lines 409–417: When labelling the ground truth data, please briefly clarify that pixels containing both meteorological and non-meteorological contributions are intentionally labelled as meteorological because the meteorological signal is assumed to dominate.

Author response: This is now clarified where the ground truth image is defined. Pixels that are also included in the cluttered summer radar image, and whose synthetic reflectivity values therefore contain both contributions, are labeled as meteorological. In this way, every pixel containing meteorological echoes is labeled as such. Typically, the non-meteorological contribution to such pixels is also small compared to the meteorological one; for example, insect echoes of 10 dBZ mixed with rain echoes of 20 dBZ yield a total of 20.4 dBZ, due to the addition in linear scale. However, at the edge of weather events, the DBZH can be similar to that of non-meteorological echoes, therefore we did not mention this as the reason in the manuscript.

Reviewer comment 16

Lines 440-442 - test design: For a more robust and independent assessment, the model should ideally be evaluated on data from a completely independent date, with no radar data from that date or event used during training or model selection. The current split by radar station primarily assesses generalization to an unseen radar rather than to unseen meteorological and clutter conditions. If this cannot be done, the resulting limitation for generalization should be discussed explicitly. Also, please specify which radar has been used for network evaluation.

Author response: Thank you for this comment. We no longer use the withheld radar station for testing. Its data now serves as validation dataset, used for early stopping and model selection, see the new subsection "Model selection and early stopping". The network is evaluated on the five expert-labeled measured sweeps, whose dates, times, and radar stations are stated at the definition of the test dataset and in the per-sweep performance table. The withheld validation station is Essen, which the manuscript now names in Sect. 2.5.

Reviewer comment 17

Sect. 2.5 - model and hyperparameter selection: Please clarify how the final architecture was selected and how the training hyperparameters were tuned. Was a separate validation set, cross-validation procedure employed? In particular, please justify choices such as the sampling probabilities of 0.75, 0.125, the batch size, the number of training iterations, and other fixed values. Care should be taken not to use the final test datasets for hyperparameter selection.

Author response: Thank you for this comment. In the new version of the manuscript, the number of feature channels and training length are selected by the loss on a separate validation dataset, see the new subsection "Model selection and early stopping". The sampling probabilities, the batch size, and the learning rate are default values that we fixed before training and did not tune; the manuscript now states this and lists them in a table. Since the networks accuracy on winter images (train and validation) is substantially higher, future work would likely assume lower ratios for them in the training data as they contain less signal during training. The five expert-labeled measured sweeps, four of which originate from a date

two weeks different than the dates of the training data, are used as test data and are not used for training or for model selection. The remaining expert labeled sweep is from the same date as the sweeps used for the calibration of the synthetic scaling factors, which is why the test scores are now listed individually for each sweep as well.

Reviewer comment 18

Sect. 2.5 - learning curve: Given the limited amount of training data and the high complexity of the U-net architecture, it would be informative to include a learning-curve analysis in which the model is trained using progressively larger subsets of the training dataset and evaluated on the same independent test set. If performance is still improving as the training-set size increases, this would indicate that the model is data-limited and that additional training data could further improve its skill. Such an analysis would, however, require a test set that is sufficiently representative of the range of meteorological and non-meteorological conditions encountered in operational radar data.

Author response: Thank you for this comment. We have added a learning curve for the validation losses of the different network hyperparameter choices. However, due to the large amount of training runs and added complexity, we have left the training dataset fixed.

Reviewer comment 19

Loss function: Please check the binary cross-entropy formula. Should the second term be $(1 - x) \log(1 - y)$?

Author response: Thank you for spotting this error. We have adjusted the formula accordingly. The training code used the standard binary cross-entropy loss of PyTorch, BCEWithLogitsLoss, so the error was limited to the manuscript and does not affect the reported results.

Reviewer comment 20

Sections 3.1 and 3.2: These sections describe the evaluation metrics and the reference DWD classification method rather than presenting results. I strongly recommend moving both subsections to Materials and Methods and starting the Results section with the network evaluation.

Author response: Thank you for this comment, we have moved the evaluation methodology to the methods section. The SotA method is described in its own section.

Reviewer comment 21

Lines 637-644 - dates of labelled sweeps: Please report the exact dates and times of the manually labelled sweeps and clarify whether they are independent in time and date from the data used for training and for developing the synthetic-data procedure. This description should also be included in the Method section rather than the results.

Author response: The dates, times, and radar stations of the five sweeps are now stated at the definition of the test dataset and in a per-sweep table. Four sweeps were measured on June 5 and June 30, 2022, dates from which no data was used for the generation of the synthetic training data, for training, or for model selection. The sweep from June 16, 2022

was measured in the hour from which the reflectivity weighting factor was determined, which the manuscript now discloses.

Reviewer comment 22

Sect. 3.3: The evaluation on measured data is based on only five manually labelled sweeps. Please provide more information on the selected cases. In Fig. 9, please report the variability of performance across individual sweeps in addition to the overall scores. This would help assess how robust the reported improvement over the DWD method is.

Author response: Thank you for this comment. The metrics of the network and the SotA method are now listed for each of the five test sweeps in a per-sweep table, with date, time, and radar station. The intersection over union of the network ranges from 0.855 to 0.974 and is higher for each of the five sweeps than that of the SotA method, which ranges from 0.776 to 0.855.

Reviewer comment 23

Line 660: Unless statistical significance was formally assessed, please replace "significantly worse" with wording such as "substantially worse".

Author response: We have replaced most occurrences of "significant" with "substantial" throughout the manuscript.

Reviewer comment 24

Sect. 4: The manuscript currently provides limited insight into how the U-net uses the individual radar moments and essentially attributes the better performance than the DWD method to the larger spatial neighborhood used by the U-net. Some information and discussion about the feature importance would be interesting. For example, a permutation-based feature-importance analysis on an independent test set, obtained by perturbing each input channel separately and measuring the corresponding performance decrease, would improve the model interpretability and help quantify the model's reliance on DBZH, ZDR, and UDR, without requiring retraining.

Author response: Thank you for this comment, we chose an ablation study over the suggested permutation-based analysis. A permutation-based analysis measures only how much the already trained network relies on a channel, so a channel whose information is also carried by the other channels could appear important even though it carries no additional information. Most importantly, permuting one channel produces physically inconsistent inputs that the network never encountered during training, which can distort the measured performance decrease. We therefore trained one network for each of the seven combinations of the DBZH, ZDR, and UDR channels, with otherwise unchanged training; the results are reported in a new table and interpreted in the Discussion.

Reviewer comment 25

Lines 719–727: The explanation linking the reduced performance near the radar to incorrectly labelled non-meteorological echoes in the winter data is plausible but appears speculative. Please clarify whether this was verified quantitatively or present it explicitly as a hypothesis.

Author response: We agree that this explanation is a hypothesis. It is grounded in the preprocessing of the winter radar images, which visibly retain non-meteorological echoes close to the radar (Figure 3). As a result, the ground truth is incorrectly labeled, and we would thus expect network performance in those areas to be degraded. We did not verify the explanation quantitatively, and this is now clearly stated in the Discussion.

Reviewer comment 26

Lines 728–740: The claims regarding generalizability, robustness across different times of day, and the absence of overfitting should be moderated. Although one radar is excluded from training, data from the other radars originate from the same dates and may observe the same large-scale meteorological or non-meteorological structures. The test set is therefore not fully independent in terms of events or atmospheric conditions. Together with the additional evaluation on only five measured sweeps, the current results demonstrate transfer to an unseen radar more than generalization to unseen events. A large, fully independent test dataset would make the conclusions much more robust if the reported performance were confirmed.

Author response: Thank you for this comment. The data of the withheld radar station now serves as validation dataset only, and no claim about generalization is based on it. The statements on generalizability in the Discussion now rest on the five expert-labeled test sweeps only. The robustness claim is reduced to the transfer to measured summer sweeps of other days and daytimes, with explicit mention that evaluating the network on rare summer artifacts requires more test data. However, due to the difficulty of labeling experimentally measured mixed sweeps, a large test dataset is out of scope for the current study. For future studies, one could make more use of independent but not mixed summer sweeps and partially labeled mixed sweeps.

Reviewer comment 27

Lines 770–776: The suggestions to include consecutive time steps and multiple elevation angles are repeated twice. Please merge these sentences to avoid redundancy.

Author response: Thank you for this comment, the conclusion now states this outlook once.

Technical corrections

Reviewer comment technical corrections

All figure captions showing radar data should report the date and timestamp of the displayed event. — Please revise references to figures and sections according to the journal style, using "Fig." and "Sect." in running text. — Line 74: add the missing full stop "." at the end of the sentence. — Line 119: Please check whether "annual structure" should read "annular structure". — Line 144: consider removing the hyphen in "multiple-radar moments". — Lines 160–165: The final sentence could be slightly smoother, for example: "Section 4 discusses the results, and Section 5 presents the conclusions." — Line 164: define the abbreviation of Deutscher Wetterdienst (DWD) at first occurrence. — Lines 167–168: Please rephrase the reference to Sect. 2.1. — Line 237: Fig. 1 is already introduced at line 224. Please remove one of the two references to avoid repetition. — Line 271: consider writing "... connected components consisting of several pixels (Fig. 1)." — Line 380: insert a space after the full

stop in “images.To”. — Line 752: write “. . . motion (Fig. 10).” — Line 761: Please use the abbreviation DWD after it has been introduced in the manuscript.

Author response: Thank you for careful reading of the manuscript. We have changed the manuscript accordingly. To keep the section self-contained, we did not include the abbreviation DWD in the conclusions.