Supplement of Nat. Hazards Earth Syst. Sci., 25, 41–47, 2025
https://doi.org/10.5194/nhess-25-41-2025-supplement

Natural Hazards and Earth System Sciences
Open Access

Supplement of

Brief communication: Training of AI-based nowcasting models for rainfall early warning should take into account user requirements

Georgy Ayzel and Maik Heistermann

Correspondence to: Maik Heistermann (maik.heistermann@uni-potsdam.de)

Supplement

S1 Further details on model specification

S1.1 RainNet2024

In our study, we employ a common approach that involves utilizing existing deep learning model architectures and transferring them for hypothesis testing in different fields. For this purpose, we use the segmentation models library (Iakubovskii, 2019), which provides a variety of model architectures (e.g., U-Net, LinkNet, PSPNet, FPN), encoder models (various families of neural networks designed for efficient feature extraction, such as those based on VGG, ResNet, and EfficientNet), as well as loss functions and metrics. Despite the proliferation of different model architectures, recent studies in weather forecasting (Andrychowicz et al., 2023; Bodnar et al., 2024) continue to employ U-Net as a primary building block. Therefore, to maintain consistency with RainNet2020 and align with state-of-the-art research, we continue to use the U-Net architecture as the core of RainNet2024. However, unlike RainNet2020 (Ayzel et al., 2020), which utilized a default stack of convolutional layers for feature extraction in the encoder block (referred to as the "backbone" in the segmentation models library), we opted for more sophisticated encoders based on residual neural networks. In our preliminary but extensive benchmarking of different model backbones (results are not shown here), we found that EfficientNetB4 exhibits similar or even superior performance compared to other models based on residual connections (e.g., ResNet34), while introducing significantly fewer model parameters. This aligns with the original findings on EfficientNet performance (Tan and Le, 2019), which have been validated on various datasets for image classification. The choice of the particular version, EfficientNetB4, was justified by the trade-off between model complexity and efficiency. Further exploration of model variants with a higher number of parameters, such as B5 and B6, did not result in significantly improved metrics. Interested readers can find all the details related to the EfficientNet family of models in the original publication Tan and Le (2019).

In RainNet2020, we employed logarithmic transformation of precipitation intensities as a preprocessing step to smooth the data distribution. To mitigate the impact of particularly high values on the loss function, we used the log-cosh function instead of mean squared error. However, during the development of RainNet2024, we discovered that a more straightforward setup with linear scaling as preprocessing and mean squared error as the loss function yielded the best results compared to more complex approaches. We believe this is primarily due to the EfficientNet-based encoder's enhanced efficiency in feature extraction compared to the fully convolutional setup used in RainNet2020.

S1.2 RainNet2024-S

We trained the RainNet2024-S models using the Jaccard loss function, which is a relaxed and differentiable modification of the Critical Success Index (CSI). RainNet2024-S predicts, for each grid cell (pixel), the estimated confidence (ranging from 0 to 1) of exceeding the specified accumulation threshold. For the final binarization (segmentation) of areas where accumulated precipitation is below or above the threshold, it is necessary to determine a threshold for the confidence value. Although the default value is 0.5, this can be optimized numerically to maximize efficiency metrics on a validation dataset. However, as demonstrated by Leinonen et al. (2022), using CSI as a loss function can lead to an uncalibrated classification model in terms of the comparability between predicted confidence and probability estimates (observation frequency). Consequently, the predicted confidence distribution tends to saturate near 0 and 1, making the choice of threshold less critical compared to training with, for example, binary cross-entropy loss. Therefore, we implement the default threshold value of 0.5 to obtain the segmentation mask.

S2 Additional skill scores: POD and FAR

In order to give a more comprehensive view on model performance and inherent trade-offs, we provide additional verification metrics in terms of the probability of detection (POD, also known as hit rate) and false alarm rate (FAR) for the different models and thresholds in Tab. S1 and Tab. S2, respectively. Based on a standard contingency table, POD quantifies the ratio between hits and the sum of hits and misses while FAR corresponds to the ratio between false alarms and the sum of false alarms and correct negatives. As reported in the main manuscript (section 2.4), "non-events" (i.e. the sum of false alarms and correct negatives) outweigh "events" (i.e. the sum of hits and misses) by far: depending on the precipitation threshold, the relative frequency of "event" grid cells in the test data amounts to 4.27 % for the threshold of 5 mm in one hour and decreases further with increasing precipitation thresholds (10 mm: 1.26 %, 15 mm: 0.44 %, 20 mm: 0.19 %, 25 mm: 0.09 %, 30 mm: 0.04 %, 40 mm: 0.01 %). It should be emphasized that POD and FAR should not be interpreted in isolation since either of them can increase/decrease at the cost of the other one. This is why we use the CSI metric as the main verification metric as it considers hits, false alarms, and misses.

With regard to POD (the higher the better), the RainNet2024-S models outperform all competitors for thresholds $\leq$ 20 mm/h. For larger thresholds, persistence is the superior model in terms of POD. RainNet2024 rates second best up to a threshold of 10 mm only. With regard to FAR (the lower the better), RainNet2024-S is superior only for a threshold of 5 mm/h. For higher thresholds, RainNet2024 takes over the lead while RainNet2024-S rates second best. Altogether, the results for POD and FAR suggest that the RainNet2024-S models learned to preserve areas of high intensity, but have a tendency of misplacing them (due to the low frequency of event grid cells) while RainNet2024 tends to underpredict high rainfall accumulations ($\geq$ 15 mm/a) anyway. As expected, Persistence scores in terms of POD for high thresholds, but always at the cost of very poor FAR values.

Table S1. Hit rate (also known as probability of detection, POD) for the different investigated precipitation accumulation thresholds and models (RainNet2020 is excluded due to its obviously low performance).

Threshold (mm/h)	Persistence	PySteps	RainNet2024	RainNet2024-S
5	0.357	0.421	0.487	0.551
10	0.249	0.255	0.261	0.372
15	0.208	0.166	0.131	0.271
20	0.182	0.108	0.057	0.282
25	0.163	0.072	0.023	0.144
30	0.152	0.051	0.011	0.128
40	0.130	0.027	0.004	0.096

Table S2. False alarm rate (FAR) for the different investigated precipitation accumulation thresholds and models (RainNet2020 is excluded due to its obviously low performance).

Threshold (mm/h)	Persistence	PySteps	RainNet2024	RainNet2024-S
5	0.639	0.535	0.446	0.398
10	0.815	0.744	0.535	0.548
15	0.897	0.855	0.601	0.660
20	0.939	0.917	0.649	0.809
25	0.962	0.950	0.694	0.775
30	0.974	0.968	0.732	0.899
40	0.985	0.987	0.858	0.951

S3 Confidence intervals for CSI metric

For Fig. 2 of the main manuscript, we confirmed that the shown mean CSI for all combinations of models and thresholds were significantly different (except for Persistence and PySteps at a threshold of 20 mm/h). The evaluation of significance was based on the 90 % confidence intervals of the mean CSI values (as shown in Tab. S3). These confidence intervals were obtained by means of bootstrapping (or resampling) for each combination of model and threshold, based on the following procedure:

- we computed the CSI_i on each test dataset i;
- from these CSI_i, we randomly sampled 1000 values (with replacement), and computed the mean $\overline{CSI}_j$ from this sample;
- the previous step was repeated 100 times;
- finally, we obtained the 5th and the 95th percentiles (P_5, P_{95}) as the boundaries of the confidence interval from all 100 realisations of $\overline{CSI}_j$.

Table S3. Mean CSI on test dataset and corresponding 90 % confidence interval (P_5 and P_{95} correspond to 5th and 95th percentile).

Model	Threshold (mm/h)	Mean	P_5	P_{95}
Persistence	5	0.215	0.214	0.217
	10	0.113	0.112	0.114
	15	0.069	0.068	0.070
	20	0.044	0.043	0.045
	25	0.030	0.029	0.030
	30	0.021	0.021	0.022
	40	0.012	0.012	0.013
PySteps	5	0.285	0.283	0.287
	10	0.142	0.140	0.143
	15	0.077	0.076	0.078
	20	0.042	0.041	0.043
	25	0.023	0.023	0.024
	30	0.014	0.013	0.015
	40	0.005	0.005	0.005
RainNet2020	5	0.205	0.202	0.207
	10	0.028	0.026	0.028
	15	0.000	0.000	0.000
	20	0.000	0.000	0.000
	25	0.000	0.000	0.000
	30	0.000	0.000	0.000
	40	0.000	0.000	0.000
RainNet2024	5	0.340	0.338	0.342
	10	0.182	0.180	0.184
	15	0.094	0.092	0.096
	20	0.042	0.041	0.043
	25	0.016	0.016	0.017
	30	0.006	0.006	0.007
	40	0.001	0.001	0.001
RainNet2024-S	5	0.402	0.400	0.404
	10	0.246	0.243	0.248
	15	0.160	0.158	0.163
	20	0.113	0.111	0.115
	25	0.076	0.074	0.078
	30	0.043	0.041	0.044
	40	0.021	0.020	0.021

References

Andrychowicz, M., Espeholt, L., Li, D., Merchant, S., Merose, A., Zyda, F., Agrawal, S., and Kalchbrenner, N.: Deep learning for day forecasts from sparse observations, arXiv preprint arXiv:2306.06079, 2023.

Ayzel, G., Scheffer, T., and Heistermann, M.: RainNet v1.0: a convolutional neural network for radar-based precipitation nowcasting, Geoscientific Model Development, 13, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020, 2020.

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., et al.: Aurora: A foundation model of the atmosphere, arXiv preprint arXiv:2405.13063, 2024.

Iakubovskii, P.: Segmentation Models, https://github.com/qubvel/segmentation_models, 2019.

Leinonen, J., Hamann, U., and Germann, U.: Seamless lightning nowcasting with recurrent-convolutional deep learning, Artificial Intelligence for the Earth Systems, 1, e220 043, 2022.

Tan, M. and Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, pp. 6105–6114, PMLR, 2019.