https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

.

Deep Learning Approach Towards Precipitation Nowcasting: Evaluating Regional Extrapolation Capabilities

Tarek Beutler[1], Annette Rudolph[2], Daniel Goehring[1], and Nikki Vercauteren[3]

[1]Department of Computer Science, Freie Universität Berlin, Germany
[2]Department of Geosciences, Freie Universität Berlin, Germany
[3]Department of Geosciences, University of Oslo, Norway

Correspondence: Annette Rudolph (annette.rudolph@met.fu-berlin.de)

Abstract. Precipitation nowcasting refers to the prediction of precipitation intensity in a local region and in a short timeframe up to 6 hours. The evaluation of spatial and temporal information still challenges todays numerical weather prediction models. The increasing possibilities to store and evaluate data combined with the advancements in the developments of artificial intelligence algorithms make it natural to use these methods to improve precipitation nowcasting. In this work a Convolutional Long Short-Term Memory network (ConvLSTM) is applied to Radar data of the German Weather Service. The positive effectiveness of finetuning a network pretrained at a different location and for different precipitation intensity thresholds is demonstrated. Furthermore, in the framework of two case studies the skill scores for the different thresholds are shown for a prediction time up to 100 minutes. The results highlight promising regional extrapolation capabilities for such neural networks for precipitation nowcasting.

1 Introduction

Precipitation nowcasting describes the detailed prediction of rainfall intensity in a local region and a short timeframe (WMO). It is an essential product in many societal contexts, such as when high-impact weather associated with intense rainfall needs to be forecast for traffic and transportation. Established approaches for precipitation nowcasting primarily make use of models based on numerical weather prediction (NWP) and extrapolation of radar echo data (Sun et al., 2014). For probabilistic precipitation nowcasting, ensemble NWP is typically used to derive multiple realistic precipitation forecasts and thereby obtain an estimate of the uncertainty in the forecast, but the precipitation is modelled using an advection equation relying on the radar echo data (Bowler et al., 2006; Pulkkinen et al., 2019). This reliance on the advection model was introduced as a remedy to poor forecasts of precipitation at a few hours lead-time, where non-Gaussianity of the field challenges the data assimilation in NWP (Buehner and Jacques, 2020).

Precipitation nowcasting can be understood as a spatiotemporal forecasting problem, or in other words, a task of making predictions using data that contains both spatial and temporal attributes (Aggarwal, 2015). Such a framework fits with the recent advancements in the deep learning domain, and in fact there has been an increasing usage of machine learning for precipitation nowcasting, see e.g. Agrawal et al. (2019); Ayzel et al. (2020); Ravuri et al. (2021). These deep learning approaches include

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

precipitation nowcasting with U-Net (Ronneberger et al., 2015), which uses only a Convolutional Neural Network (CNN) without any sort of memory. For example U-Net was used by Agrawal et al. (2019) to make predictions with one hour lead time based on Doppler radar data. Similarly, Trebing et al. (2021) use U-Net as a base and introduce Convolutional Block Attention Modules (CBAM) (Woo et al., 2018) into the network. This leads to improvements by focusing the model on important attributes, such as the spatial region of the underlying data. One of the most recent approaches is RainNet (Ayzel et al., 2020) which is loosely based on U-Net and SegNet (Badrinarayanan et al., 2016), the latter using a Fully Convolutional Neural network (FCN) (Long et al., 2014). This approach was tested on German weather data.

In the framework of this study we will consider an alternative method based on a the Convolutional LSTM Network (ConvLSTM) to tackle the challenge of precipitation nowcasting. Its applicability to precipiation nowcasting for the region of Hong Kong has been demonstrated by Shi et al. (2015). The ConvLSTM is a Long Short-Term Memory network (LSTM), first introduced in (Hochreiter and Schmidhuber, 1997), that uses convolutional layers to improve the processing of spatiotemporal correlations. This ConvLSTM predicts precipitation more accurately than state-of-the-art methods based on radar echo extrapolation. Shi et al. (2017) introduced the Trajectory Gated Recurrent Unit (TrajGRU) as an approach to actively learn the location-variant structure of the network. This development enables to capture certain location-variant motion patterns better than the location-invariant ConvLSTM. In a recent development, Luo et al. (2021), introduced the pseudoflow spatial-temporal LSTM (PFST-LSTM), which solves problems inherent to the Convolutional LSTM and GRU structures and outperforms older approaches, such as the TrajGRU, on the CIKM AnalytiCup 2017 and MovingMNIST++ data sets. Nevertheless, the present work will focus on the method of Shi et al. (2017), who also used a standardized benchmark data set and introduced improved error functions to measure the skill of the network.

While a standardized benchmark data set is good for testing machine learning models and comparing them against each other, ideally one wants to use these recent advancements to predict precipitation in applied societal contexts. Users often face problems such as incomplete and smaller sample size of data, as well as regional differences in weather and precipitation. In this work, the TrajGRU network is trained on a German weather data set and then compared to a network trained on the standardized data set from Hong Kong. This latter pretrained network is then finetuned on the German weather data to evaluate the possibilities of regional extrapolation of pretrained networks. Different approaches on finetuning and transfer learning have been published. They usually depend strongly on the problem class. For image classification problems, (Girshick et al., 2014) showed how feature layers can be reused to save training time. In the medical domain, a variety of papers exists on transfer learning for image segmentation using FCNs, as, e.g., in (Karimi et al., 2020) for brain or liver images. In this work, finetuning will be used for the regional extrapolation of the network, following the hypothesis that such regional finetuning can outperform and reduce the costs of training a region-specific network for precipitation nowcasting.

To test the hypothesis, the paper is structured as follows. In Sec. 2, after a short introduction to the German RADOLAN data set, the notwork model, its training as well as the technical settings to evaluate the effect of finetuning on the RADOLAN data is summarized. For the evaluations the Critical Success Index (CSI) and the Heidke Skill Score (HSS) are used. An overview on the skill scores, the error functions and the classifications of the precipitation intensities is given in Sec. 3. The results are presented in Sec. 4, where first the scores of the model trained from scratch and the finetuned model are compared for the

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

five precipitation thresholds. The results show that finetuning not only reaches the same performance of the network trained from scratch, and this significantly faster, but it is also able to improve it, especially for heavy rainfall events. In the second part of Sec. 4, two case studies of precipitaion events are evaluated showing promising scores over a prediction time up to 100 minutes. Finally, the conclusion is given in Sec. 5.

2 Setup

2.1 Data

The RADOLAN data set (radar online adjustment) provided by the German Weather Service (Deutscher Wetterdienst, DWD) is used (Bartels et al., 2004; Winterrath et al., 2012). This data is a reflectivity composite of 17 radar stations in Germany, which provides high definition data in both temporal and spatial resolution. The composite images are taken in a 5 minute interval, are 480 x 480 pixel in size and have a spatial resolution of 1 km per pixel. The years 2017 to 2021 are considered for this study. In particular, the time period 06.04.2017–14.07.2017 is considered as test data set, 15.07.2017 – 07.10.2020 as training data and the data in the time frame 08.10.2020 – 30.03.2021 is taken for the validation. In total this data set contains 251,358 frames over 1099 days. Recently, Kreklow et al. (2020) confirm that the RADOLAN data set has been improved during the last years. Now, typical radar artefacts, orographic and winter precipitation as well as range-dependent attenuation are corrected.

2.2 Network model and its training

All experiments were run using the TrajGRU. This model is based on the Convolutional Gated Recurrent Unit (ConvGRU). The ConvGRU, similiar to the ConvLSTM, is a modified version of the Gated Recurrent Unit (GRU). All inputs and cells are extended by two dimensions that represent the spatial information. Additionally, all matrix multiplications in the cell equations are replaced by convolutional operators.

The TrajGRU aims to solve several issues with the older ConvLSTM model (Shi et al., 2015). The biggest problem is the location invariance of the used convolutions. To understand this, let $\mathcal{N}_{i,j}^{h}$ be the ordered neighboorhod set at location (i, j). These sets are defined by the convolution hyperparameters, for example kernel size, dilation and padding (Yu and Koltun, 2016). Because these hyperparameters are the same by default for each location (i, j), the convolution is location invariant. However, many motion patterns are location variant and thus hard to grasp by the convolutions.

The TrajGRU layer takes the current and previous state to generate a local neighborhood set for each location $\sum_{l=1}^{L}(p_{l,i,j}, q_{l,i,j})$ in each timestamp, where L is the chosen maximum number of local links, thus making it location variant. Optical flows - vector

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

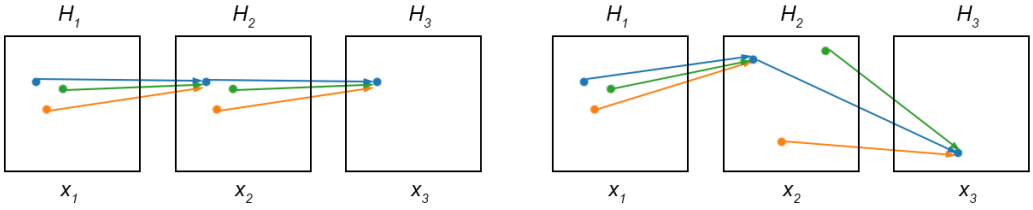

Figure 1. Comparison of the different connection structures. Adapted from Shi et al. (2017).

fields which represent a motion pattern (Gibson, 1950) - are used to store the indices of the locations:

$$U_t, V_t = \gamma(x_t, H_{t-1}) \tag{1}$$

$$z_t = \sigma(W_{xz} * x_t + \sum_{l=1}^{L} W_{hz}^l * \text{warp}(H_{t-1}, U_{tl}, V_{tl})) \tag{2}$$

$$r_t = \sigma(W_{xr} * x_t + \sum_{l=1}^{L} W_{hr}^l * \text{warp}(H_{t-1}, U_{tl}, V_{tl})) \tag{3}$$

$$H'_t = f(W_{xh} * x_t + r_t \circ (\sum_{l=1}^{L} W_{hh}^l * \text{warp}(H_{t-1}, U_{tl}, V_{tl}))) \tag{4}$$

$$H_t = (1 - z_t) \circ H'_t + z_t \circ H_{t-1} \tag{5}$$

Here W denotes the weights, x the input, H the output state, t the current timestep, $\circ$ the Hadamard product. z denotes the GRU update gate, r the reset gate, H the saved and H' the new information. γ is a small neural network, which generates the optical flows stored in U_t and V_t. The warp function selects the locations that are pointed out by U_t and V_t from H_{t-1} through a bilinear sampling kernel (Ilg et al., 2016). This makes it possible to learn the connection structure through the subnetwork γ. Figure 1 shows a comparison of the ConvLSTM and TrajGRU connection structures.

The complete TrajGRU model can be described as an autoencoder, a neural network which first encodes an input sequence to a fixed size, to decode it afterwards and output it as a prediction (Bank et al., 2020). For both encoder and decoder, multiple TrajGRU layers are stacked on top of each other.

In the following, the term finetuning is used to denote the action of taking a model which was already trained on data to some degree, and then training it again on different data while keeping the weights from the previous training. This procedure was shown to improve generalization and reduce training time of networks (Yosinski et al., 2014). The goal is to extrapolate regional differences from one data set to another, without training a completely new network for the considered location.

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

Table 1. Rainfall distribution in the data sets. For the evaluation of the HKO-7 data see Shi et al. (2017).

Rainfall Rate (mm/h)	HKO-7 (%)	RADOLAN (%)	Description
$0 \leq r < 0.5$	90.25	96.54	Very light
$0.5 \leq r < 2$	4.38	2.38	Light
$2 \leq r < 5$	2.46	0.77	Light to moderate
$5 \leq r < 10$	1.35	0.22	Moderate
$10 \leq r < 30$	1.14	0.08	Moderate to heavy
$30 \leq r$	0.42	0.02	Heavy

2.3 Experimental setup

To evaluate the effect of finetuning on the German RADOLAN data, a TrajGRU trained with random weight initialization is compared to one that uses pretrained weights. The latter model was pretrained on the original data set HKO-7, which contains radar echo data collected from 2009 to 2015 by the Hong Kong Observatory, as explained in Shi et al. (2017), while the former was trained using only German RADOLAN data. From now on, we will use the term "trained from scratch" to refer to the randomly initialized model and the term "finetuned model" to refer to the model that uses pretrained weights.

Both models follow the autoencoder structure described by Shi et al. (2017), with three TrajGRU layers in both encoder and forecaster that use 192, 192 and 64 filters respectively. We train both models for 100,000 iterations using the Adam optimizer, an extension of the stochastic gradient descent method (Kingma and Ba, 2014), with a learning rate of 10^{-4}, that decreases by $\gamma = 0.1$ on iteration 30,000 and 60,000. Both models use an input of 5 frames to predict the next 20. All 20 output frames are weighted equally in the evaluation.

Yosinski et al. (2014) have shown that many deep neural networks learn more general features in their first layers, while learning more specific ones in their last layers. For the autoencoder structure used by our models, this would mean that the outer layers learn general features that we hypothesize to be applicable across the world, while the inner layers learn more specific regional features. Because of this we freeze the weights of the outermost TrajGRU layer of both encoder and forecaster for the finetuned model and only train the two innermost layers on the German RADOLAN data afterwards. Other finetuning configurations were tested, such as freezing more layers or none at all, but displayed worse performance.

3 Metrics for skill evaluation

Because of the highly imbalanced rainfall distribution in both the German RADOLAN and HKO-7 data as shown in Table 1, a weighted loss function is considered, which is the sum of the Balanced Mean Squared Error (B-MSE) and Balanced Mean Absolute Error (B-MAE) as proposed by Shi et al. (2017), which assign a specific weight $w(r)$ to each pixel according to its

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

rainfall rate r. The error functions are calculated as follows:

$$\text{B-MSE} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{480}\sum_{j=1}^{480} w_{n,i,j}(x_{n,i,j} - \hat{x}_{n,i,j})^2 \tag{6}$$

$$\text{B-MAE} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{480}\sum_{j=1}^{480} w_{n,i,j}|x_{n,i,j} - \hat{x}_{n,i,j}|, \tag{7}$$

where N is the total amount of frames and $w_{n,i,j}$ is the calculated weight of pixel x at position (i,j) in frame n given by:

$$w(r) = \begin{cases} 1, & r < 2 \\ 2, & 2 \leq r < 5 \\ 5, & 5 \leq r < 10 \\ 10, & 10 \leq r < 30 \\ 30, & r \geq 30\,. \end{cases} \tag{8}$$

Additionally we calculate two skill scores to evaluate the actual precipitation nowcasting performance: The Critical Success Index (CSI) and the Heidke Skill Score (HSS) (Hogan et al., 2010). Both measure how accurate a prediction is compared to the ground truth. The CSI measures what portion of rain events was predicted correctly, while the HSS measures the accuracy of made predictions compared to a random one. The CSI and HSS are calculated as follows:

$$\text{CSI} = \frac{TP}{TP + FN + FP} \tag{9}$$

$$\text{HSS} = \frac{TP \cdot TN - FN \cdot FP}{(TP+FN)(FN+TN) + (TP+FP)(FP+TN)}\,. \tag{10}$$

For both scores 1 would be a perfect prediction, while 0 means no improvement.

The CSI and the HSS are calculated for different thresholds of the rain rate R (0.5 mm/h, 2 mm/h, 5 mm/h, 10 mm/h and 30 mm/h). For this purpose, each pixel of the prediction and ground truth image gets converted to a binary 0 or 1, based on the current threshold. After this the true positives (TP) (prediction = 1, ground truth = 1), false negatives (FN) (prediction = 0, ground truth = 0), false positives (FP) (prediction = 1, ground truth = 0) and true negatives (TN) (prediction = 0, ground truth = 0) are calculated.

4 Results

4.1 Statistical comparison of the model trained from scratch and the finetuned model

The highest scores of the model trained from scratch and the finetuned model are summarized in Table 2. The finetuned model consistently reaches better results across all scores and measurements. Table 4 shows the results of the Welch t-test (Welch, 1947). It can be recognized that for several of the different skill scores thresholds the difference between the finetuned model and the model trained from scratch are statistically significant ($p < 0.05$). Moreover, Table 3 shows that the finetuned model is

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

Table 2. RADOLAN training results: Comparison of the scores for the model from scratch and the finetuned model. The best results are marked in **bold** font.

Model	CSI					HSS					B-MAE	B-MSE
	$r \geq 0.5$	$r \geq 2$	$r \geq 5$	$r \geq 10$	$r \geq 30$	$r \geq 0.5$	$r \geq 2$	$r \geq 5$	$r \geq 10$	$r \geq 30$		
Scratch	0.5839	0.4866	0.3709	0.288	0.13182	0.7179	0.6385	0.5278	0.4365	0.2319	7079	1990
Finetuned	**0.5853**	**0.4879**	**0.3718**	**0.2883**	**0.151**	**0.7189**	**0.6401**	**0.5286**	**0.4367**	**0.251**	**7038**	**1968**

Table 3. RADOLAN training results: Number of iterations until the models reached the best score of the model trained from scratch. The best scores are marked in **bold** font.

Model	CSI					HSS					B-MAE	B-MSE
	$r \geq 0.5$	$r \geq 2$	$r \geq 5$	$r \geq 10$	$r \geq 30$	$r \geq 0.5$	$r \geq 2$	$r \geq 5$	$r \geq 10$	$r \geq 30$		
Scratch	88000	88000	88000	88000	94000	88000	88000	88000	88000	94000	85000	88000
Finetuned	**32000**	**32000**	**50000**	**50000**	**13000**	**32000**	**32000**	**50000**	**50000**	**13000**	**50000**	**31000**

a lot faster in achieving the best score than the model trained from scratch has. The full training results are visualized in Fig. 2 and in Fig. 3. They show the Mean Critical Success Index and the Mean Heidke Skill Score with their corresponding standard deviation over 100,000 training iterations and for the five different rain rate thresholds 0.5 mm/h, 2.0 mm/h, 5.0 mm/h, 10 mm/h and 30 mm/h. In general, the CSI as well as the HSS clearly indicate better results of the finetuned model for at least the first 50,000 iterations. In the range of 50,000 to 100,000 iterations, the finetuned model shows slightly higher scores. We remark the different scaling of the ordinates in Fig. 2 and Fig. 3. In order to compare the skills of the finetuned model and the model from scratch over the last iterations, the Critical Success Index and the Heidke Skill Score over last 35,000 iterations are additionally represented as box plots, see Fig. 4 and Fig. 5. They confirm the slightly better scores for the finetuned model across all thresholds.

While the differences between the scores of both models are relatively small for most of the rain rate thresholds, there is a significant difference in scores for heavy rain events with $r \geq 30$. The finetuned model shows a 14.5% increase on the CSI with $r \geq 30$ ($\Delta = 0.0192$), and a 8.2% increase on the HSS with $r \geq 30$ ($\Delta = 0.0191$). This can be explained by the unbalanced distribution of the rainfall events between the data sets. The HKO-7 data set has a significantly higher amount of heavy rainfall events compared to RADOLAN (cf. Table 1). Yosinski et al. (2014) show that transfering features between networks can improve generalization on data, even after finetuning the network. This could be an explanation, why the finetuned model has such a big increase in performance for heavy rainfall events over the model trained from scratch.

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

Table 4. Welch t-test (Welch, 1947) results between the fintuned and trained from scratch models, using the last 35000 iterations. Value sets indicating statistical significance are marked via **bold** font.

t-test	CSI					HSS					B-MAE	B-MSE
	$r \geq 0.5$	$r \geq 2$	$r \geq 5$	$r \geq 10$	$r \geq 30$	$r \geq 0.5$	$r \geq 2$	$r \geq 5$	$r \geq 10$	$r \geq 30$		
t statistic	-1.2402	**-2.4049**	**-3.1831**	-1.4366	**-6.0669**	-1.2289	**-2.5295**	**-3.0331**	-1.2648	**-6.2450**	0.0617	0.8884
p-value	0.2190	**0.0189**	**2.18e-3**	0.1553	**6.15e-8**	0.2232	**1.37e-2**	**3.40e-3**	0.2102	**2.91e-8**	0.951	0.3774

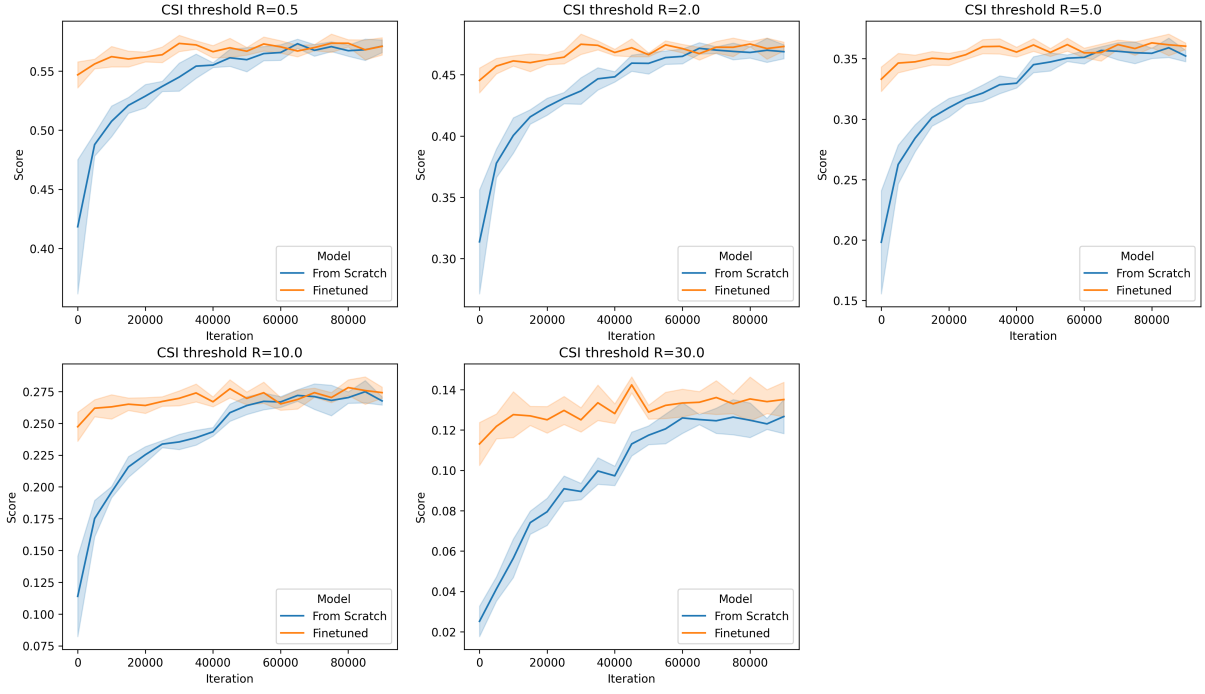

Figure 2. Mean Critical Success Index with standard deviation over all iterations for five different rain rate thresholds.

4.2 Case studies

We compare the model output for the real truth in the frame of two case studies. Using the the RADOLAN data set, we consider frontal systems at 4 May 2017, 14:40:00 UTC and at 12 May 2017, 07:40:00 UTC. These are two exemplary dates of clusters of mainly moderate precipitation crossing Germany, where the data is not part of the training data set of the model. The results of the two case studies are shown in Fig. 6 (a) and (b). The first row of each example is the input data depicted in color scale, the second row depicts the raw input data, the third row shows the ground truth and the fourth row is the model prediction. Each frame, or column, represents a 5 minute timestep. Overall, the location of the precipitation cells are predicted correctly,

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

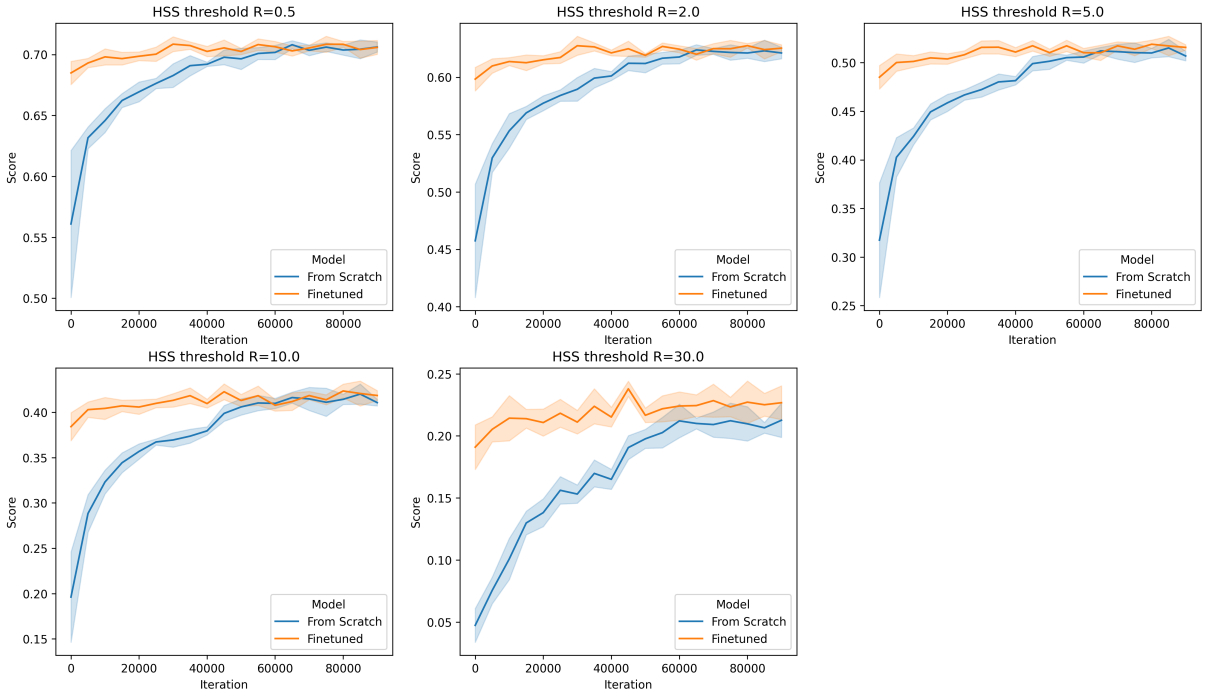

Figure 3. Mean Heidke Skill Score with standard deviation over all iterations for five different rain rate thresholds.

as well as the regions of intense precipitation. While the the precipitation around the outer boundaries of precipitating cells is slightly underestimated by the TrajGRU, the rainfall in regions of intense precipitation is overestimated. This could be caused by the blurred edges of the predictions that are inherent to the convolutional RNN approach (Shi et al., 2015).

The differences between the prediction and ground truth of each sample are shown in Fig. 7 (a) and (b). Regarding both case studies, a small increase of the absolute values of differences with time can be recognized. The differences of the boarders of the precipitating cells are slightly negative, i.e. the precipitation intensity is underestimated. Furthermore, concerning the regions of intense precipitation in Fig. 7 (a), the results of Fig. 6 are confirmed, where intense precipitation seems to be slightly overestimated.

Fig. 8 (a)–(d) shows the skill scores with respect to the prediction time in minutes. The scores of the rain rate of the thresholds 0.5 mm/h, 2 mm/h, 5 mm/h, 10 mm/h are depicted. Regarding the CSI of the 0.5 mm/h threshold in Fig. 8 (a) and (c), a score of 0.729 for the 15 minute prediction time, a score of 0.5662 for the 60 minute prediction time and for the 90 minute prediction time still a score of 0.5268 is reached. Particularly for the lower thresholds, a slow decrease of the scores with prediction time can be observed. Stronger decreases at high thresholds can be explained by the small number of data points of high precipitation intensities.

Ravuri et al. (2021) develop an observations-driven approach for probabilistic nowcasting using deep generative models (DGMs) and compared it to the probabilisitc method PySTEPs (Pulkkinen et al., 2019) regarding UK data. Ravuri et al. (2021)

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

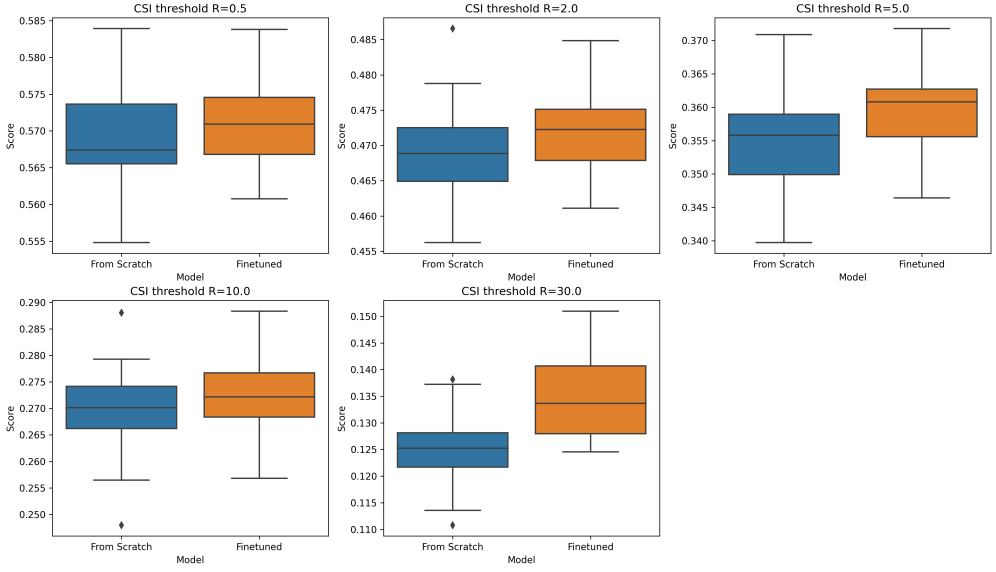

Figure 4. Critical Success Index boxplots over the last 35,000 iterations for five different rain rate thresholds.

consider the precipitation intensity thresholds 1 mm/h, 4 mm/h, 8 mm/h and a prediction interval up to 90 min. Overall, comparing the CSI over the prediction time of Ravuri et al. (2021) with the CSI shown in Fig. 8 (a) the finetuned TrajGRU provides slightly higher scores across the whole time period.

Focusing on the CSI of the 2 mm/h threshold in Fig. 8 (a) and comparing it with the CSI for the 1 mm/h threshold evaluated by Ravuri et al. (2021), the scores are comparable for a prediction time of 4 minutes. After 80 minutes prediction time the CSI of the 2 mm/h threshold obtained with the finetuned TrajGRU shows values greater than 0.4, while the CSI of PySTEPS in Ravuri et al. (2021) are smaller than 0.4. Such small differences in the results can be recognized across all thresholds. Especially for longer prediction times the scores of the finetuned TrajGRU are slightly higher. Overall, as Ravuri et al. (2021) we find that the scores decrease with increasing precipitation thresholds. This can be explained by the decreasing sample sizes at increasing precipitation thresholds. E.g. one cause of the low scores of the 10 mm/h threshold could be the small sample size of only 132 pixels above that threshold across all 20 predicted frames. Both works show skillfull predictions compared to previous works, but as Ravuri et al. (2021) state, heavy rainfall at long lead times are hard to predict. But still, as discussed in the previous paragraph, the positive effect of finetuning is clearly visible for higher precipitation intensities.

Furthermore, Ayzel et al. (2020) considered radar data of Germany, evaluating the models RainNet and Rainymotion for the intensity thresholds 0.125 mm/h, 5 mm/h and 15 mm/h. The authors also observe a decrease of the score with increasing thresholds. The Rainnet model shows clearly higher scores than the model of persistence. Regarding the 5 mm/h threshold of the RainNet model analysed in Ayzel et al. (2020), after a couple of minutes the CSI is about 0.6, but it decreases fast

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

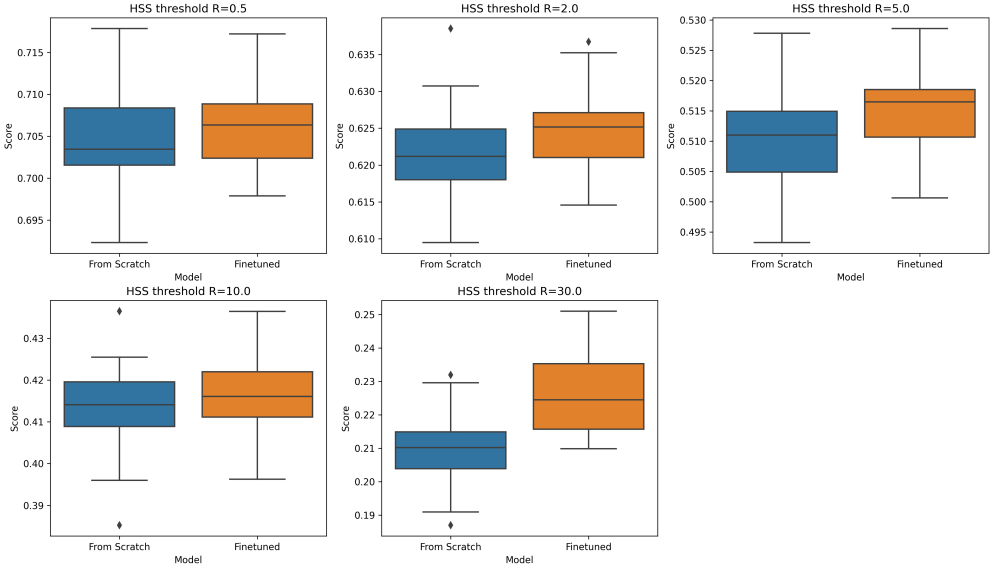

Figure 5. Heidke Skill Score boxplots over the last 35,000 iterations for five different rain rate thresholds.

and reaches values only slightly greater than 0.2 after a 60 minutes prediction time. In comparison, the finetuned TrajGru in Fig. 8 (a) starts with a CSI slightly below 0.6, but the CSI decreases slower. After 60 minutes the CSI still reaches values only slightly below 0.4 and after a prediction time of 100 minutes the CSI is still above 0.3. We remark that in the second example, shown in 8 (c), the skill scores for the higher thresholds, are smaller, which can be explained with a small data basis at low thresholds. Fig. 4 confirms statistically that the finetuned model improves the scores for extreme precipitation threshold. However, a statistical analysis would be wishful to confirm an improving CSI over the prediction time for higher thresholds as shown by the case study in Fig. 8 (a). It can be recognized that the HSS, shown in Fig. 8 (b) and (d), provides higher scores than the CSI depicted in Fig. 8 (a) and (c).

5 Conclusion

The deep learning algorithm TrajGRU (Trajectory Gated Recurrent Unit) for precipitation nowcasting is applied to radar data of the German Weather Service focusing on the evaluation of the effectiveness of finetuning of the RNN-based network. A pretrained network is finetuned for application on regionally different data and compared to a network trained from scratch on the new region. The TrajGRU is developed by Shi et al. (2017), who improved their own Convolutional Long Short-Term Memory (LSTM) network for radar data (Shi et al., 2015) by the inclusion of error functions. The authors demonstrate its applicability to a standardized data set of the Chinese Weather Service. Here, the scores CSI and HSS are evaluated for different

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

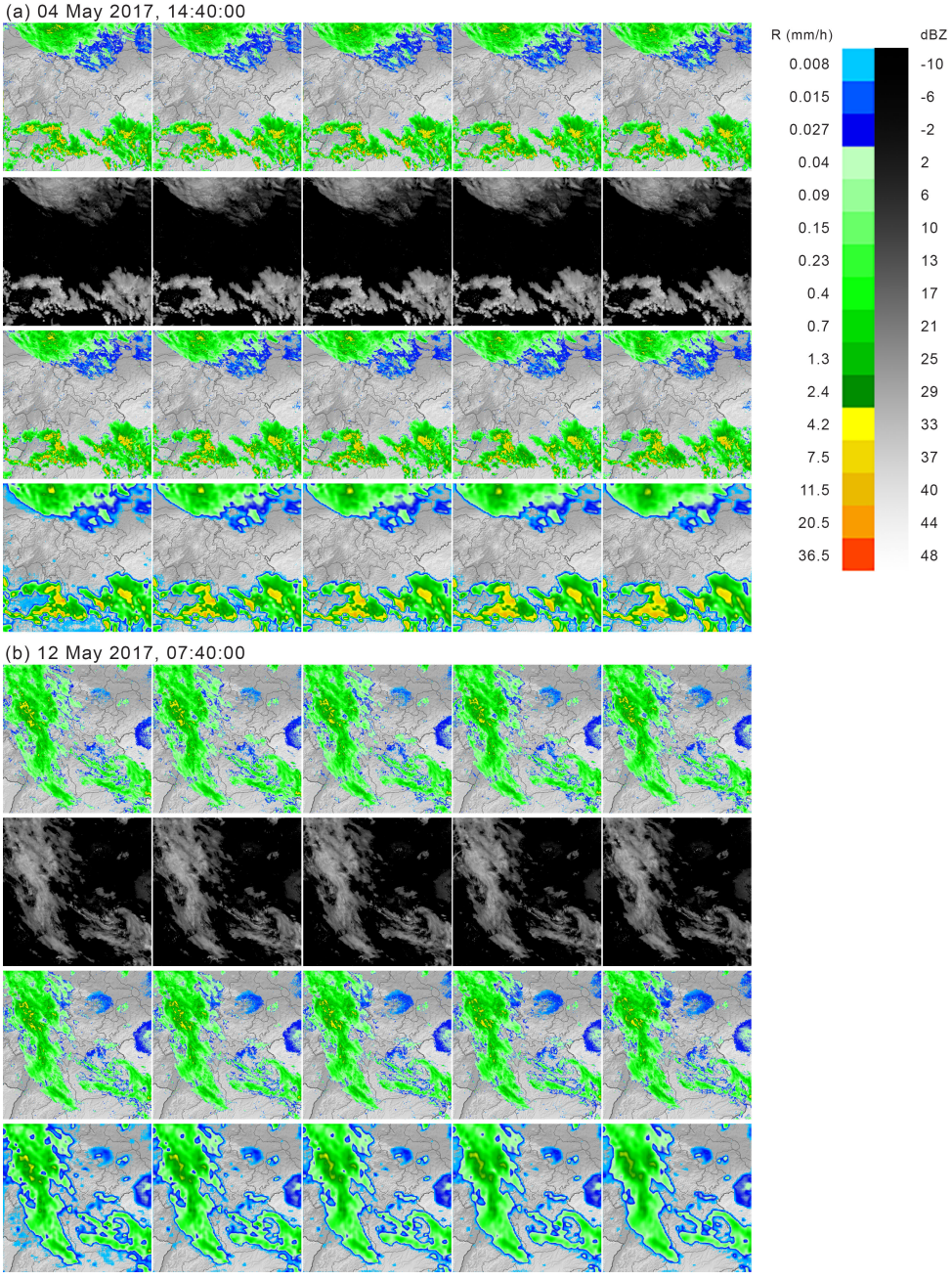

Figure 6. Two samples from the finetuned TrajGRU on RADOLAN data selected outside of the training data set for the 4th (a) and 12th (b) of May 2017. The first row for each panel respectively is the input data in color scale over the borders of Germany. The second row is the raw input data. The third row is the ground truth. The fourth row is the model prediction.

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

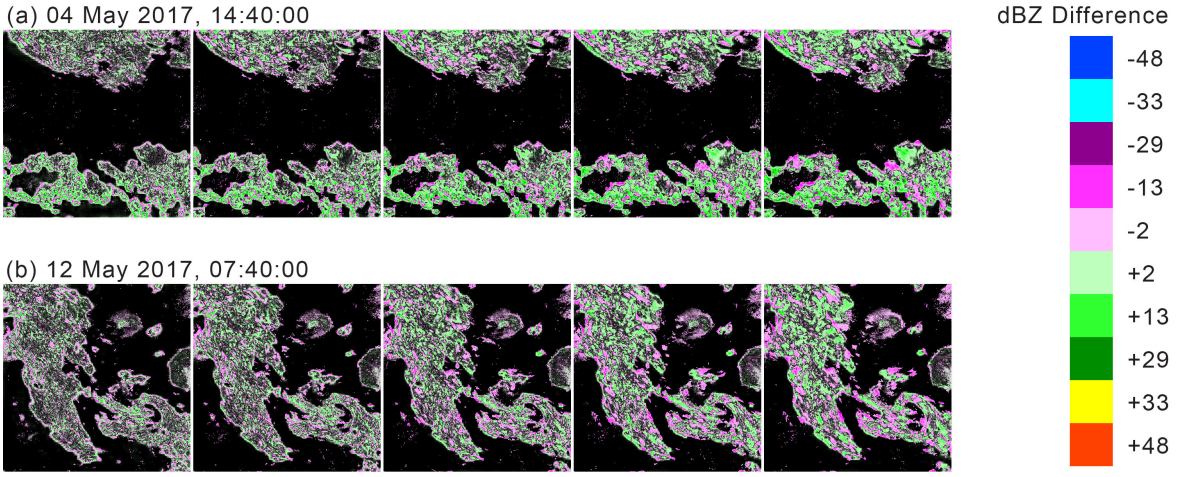

Figure 7. Two samples from the finetuned TrajGRU on RADOLAN data selected outside of the training data set for the 4th (a) and 12th (b) of May 2017. It shows the difference between prediction and ground truth. The color black represents true negatives, so pixels where no rain event occurred and no prediction was made.

thresholds of precipitation intensities (0.5 mm/h, 2 mm/h, 5 mm/h, 10 mm/h and 30 mm/h). For all thresholds, we find that a finetuned network pretrained at a different location reaches similar performance than a fully newly trained network for the region of interest, but is significantly faster. In some cases finetuning can actually increase performance over a model trained from scratch, showing that features transfer very well between data sets, even if taken from entirely different regions. The positive effect of finetuning is especially distinct for extreme precipitation events. Yosinski et al. (2014) show that transfering features between networks can improve generalization on data, even after finetuning the network. This could be an explanation, why the finetuned model has such a big increase in performance for heavy rainfall events over the model trained from scratch.

Furthermore, two case studies underline the effectiveness of the finetuned TrajGRU model. Comparing the here obtained results with recent publications on deep learning algorithms to precipitation nowcasting based on radar data (Ayzel et al., 2020; Ravuri et al., 2021) the finetuned TrajGRU shows slightly higher scores with less decrease with prediction time. While Chen et al. (2020); Ayzel et al. (2020) show that their models are suitable to predict precipitation up to 60 minutes (Chen et al., 2020; Ayzel et al., 2020), we achieve comparable scores for a 100 minute prediction time. A statistical analysis of the prediction time for more than the here presented two case studies would be wishful for future research.

To further optimize the results in future studies, different finetuning methods and finetuning hyperparameters could be taken into account. Moreover, different data sets covering further regions should be investigated. Future work could also include analyzing the ability to generalize and feature transfer between different approaches to precipitation nowcasting with deep learning. Overall, the here presented results of the finetuned TrajGRU highlight promising regional extrapolation capabilities for such neural networks for precipitation nowcasting and suggest further investigations of this method to finally improve precipitation nowcasting.

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

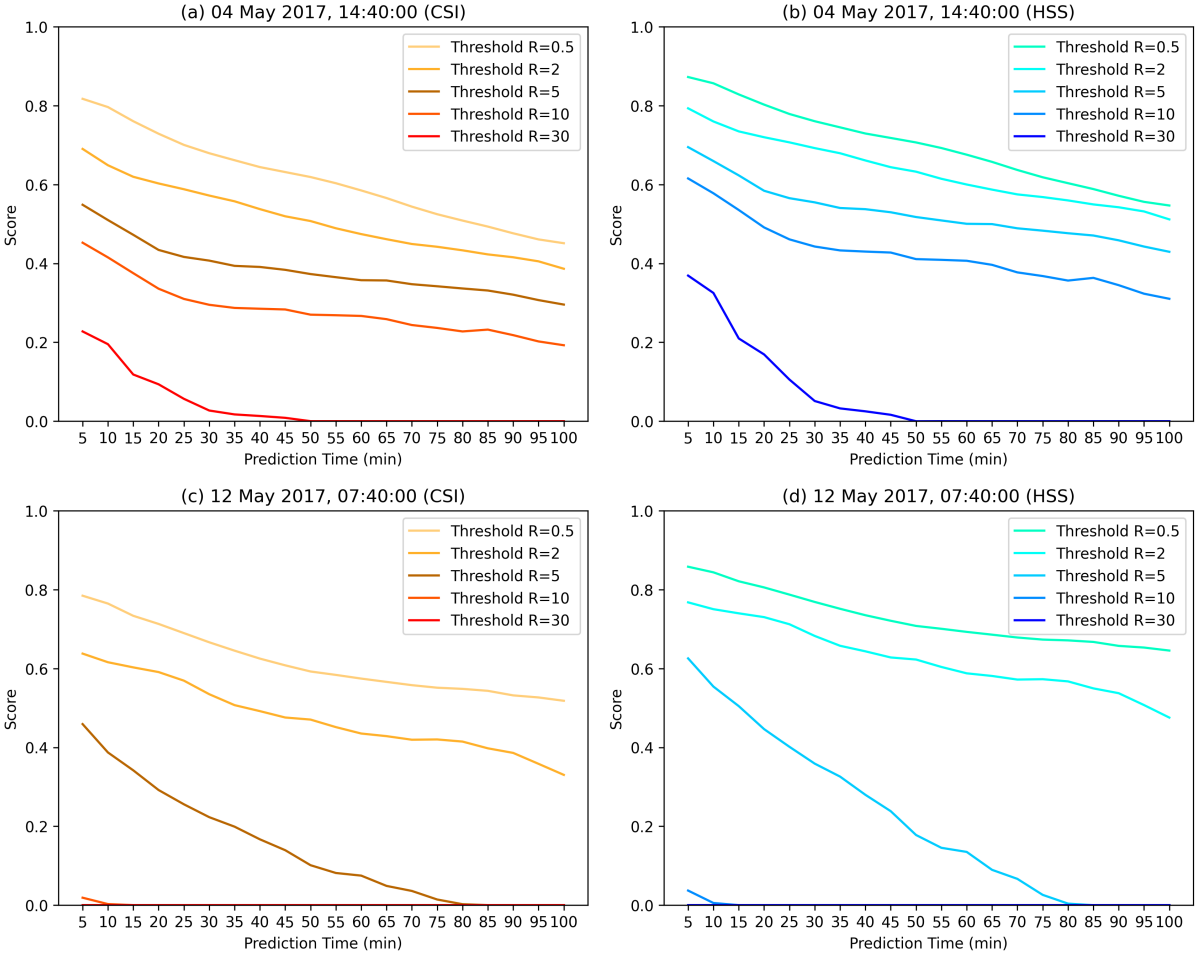

Figure 8. CSI and HSS skill scores over five thresholds for both samples for the 4th (a, b) and 12th (c, d) of May 2017, broken down by prediction time.

Code and data availability. The current version of the source code is available on GitHub: https://github.com/Nakroma/Precipitation-Nowcasting (Beutler, 2022a) under the MIT license. The exact version of the code used to produce the results used in this paper is archived on Zenodo at https://doi.org/10.5281/zenodo.6636422 (Beutler, 2022b), the input data that is available at https://doi.org/10.5281/zenodo.6634851 (Beutler, 2022c) and the weights of the compared models are available at https://doi.org/10.5281/zenodo.6634844 (Beutler, 2022d).

Author contributions. All authors designed the study and contributed to the paper. TB designed/carried out the experiments and wrote large parts of the paper. AR wrote Sec. 4.2, AR, NV and DG supervised the study and co-authored the paper.

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research has been funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 'Scaling Cascades in Complex Systems, Project Number 235221301, Projects A01, 'Coupling a multiscale stochastic precipitation model to large scale atmospheric flow dynamics' and B07 'Selfsimilar structures in turbulent flows and the construction of LES closures'.

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

References

Aggarwal, C. C.: Data Mining: The Textbook, Springer, 2015.

Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., and Hickey, J.: Machine Learning for Precipitation Nowcasting from Radar Images, 2019.

Ayzel, G., Scheffer, T., and Heistermann, M.: RainNet v1. 0: a convolutional neural network for radar-based precipitation nowcasting, Geosci. Model Dev., 13, 2631–2644, 2020.

Badrinarayanan, V., Kendall, A., and Cipolla, R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, 2016.

Bank, D., Koenigstein, N., and Giryes, R.: Autoencoders, 2020.

Bartels, H., Weigl, E., Reich, T., Lang, P., Wagner, A., Kohler, O., and Gerlach, N.: Projekt RADOLAN-Routineverfahren zur Online-Aneichung der Radarniederschlagsdaten mit Hilfe von automatischen Bodenniederschlagsstationen (Ombrometer): zusammenfassender Abschlussbericht für die Projektlaufzeit von 1997 bis 2004, 2004.

Beutler, T.: Precipitation-Nowcasting, https://github.com/Nakroma/Precipitation-Nowcasting, 2022a.

Beutler, T.: Nakroma/Precipitation-Nowcasting: v1.0.0, https://doi.org/10.5281/zenodo.6636422, 2022b.

Beutler, T.: Deep Learning approach towards Precipitation Nowcasting: RADOLAN, https://doi.org/10.5281/zenodo.6634851, 2022c.

Beutler, T.: Deep Learning approach towards Precipitation Nowcasting: Model weights, https://doi.org/10.5281/zenodo.6634844, 2022d.

Bowler, N. E., Pierce, C. E., and Seed, A. W.: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP, Q. J. R. Meteorol. Soc., 132, 2127–2155, https://doi.org/10.1256/qj.04.100, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1256/qj.04.100, 2006.

Buehner, M. and Jacques, D.: Non-Gaussian Deterministic Assimilation of Radar-Derived Precipitation Accumulations, Mon. Weather Rev., 148, 783–808, https://doi.org/10.1175/MWR-D-19-0199.1, publisher: American Meteorological Society Section: Monthly Weather Review, 2020.

Chen, L., Cao, Y., Ma, L., and Zhang, J.: A deep learning-based methodology for precipitation nowcasting with radar, Earth Space Sci., 7, e2019EA000 812, 2020.

Gibson, J. J.: The Perception of the Visual World, Houghton Mifflin, 1950.

Girshick, R., Donahue, J., Darrell, T., and Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 00, pp. 580–587, https://doi.org/10.1109/CVPR.2014.81, 2014.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., 9, 1735–1780, 1997.

Hogan, R. J., Ferro, C. A., Jolliffe, I. T., and Stephenso, D. B.: Equitability Revisited: Why the 'Equitable Threat Score' is not equitable, Weather Forecast., 25, 2010.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T.: FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks, 2016.

Karimi, D., Warfield, S. K., and Gholipour, A.: Critical Assessment of Transfer Learning for Medical Image Segmentation with Fully Convolutional Neural Networks, https://doi.org/10.48550/ARXIV.2006.00356, 2020.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://doi.org/10.48550/ARXIV.1412.6980, 2014.

https://doi.org/10.5194/egusphere-2022-440
Preprint. Discussion started: 21 June 2022

Kreklow, J., Tetzlaff, B., Burkhard, B., and Kuhnt, G.: Radar-Based Precipitation Climatology in Germany—Developments, Uncertainties and Potentials, Atmosphere, 11, 217, 2020.

Long, J., Shelhamer, E., and Darrell, T.: Fully Convolutional Networks for Semantic Segmentation, CoRR, abs/1411.4038, http://arxiv.org/abs/1411.4038, 2014.

Luo, C., Li, X., and Ye, Y.: PFST-LSTM: A SpatioTemporal LSTM Model With Pseudoflow Prediction for Precipitation Nowcasting, EEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 14, 843–857, https://doi.org/10.1109/JSTARS.2020.3040648, 2021.

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., and Foresti, L.: Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1. 0), Geosci. Model Dev., 12, 4185–4219, 2019.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al.: Skilful precipitation nowcasting using deep generative models of radar, Nature, 597, 672–677, 2021.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., kin Wong, W., and chun Woo, W.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, 2015.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., kin Wong, W., and chun Woo, W.: Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model, 2017.

Sun, J., Xue, M., Wilson, J. W., Zawadzki, I., Ballard, S. P., Onvlee-Hooimeyer, J., Joe, P., Barker, D. M., Li, P.-W., Golding, B., Xu, M., and Pinto, J.: Use of NWP for Nowcasting Convective Precipitation: Recent Progress and Challenges, Bull. Am. Meteorol. Soc., 95, 409 – 426, https://doi.org/10.1175/BAMS-D-11-00263.1, 2014.

Trebing, K., Stanczyk, T., and Mehrkanoon, S.: SmaAt-UNet: Precipitation Nowcasting using a Small Attention-UNet Architecture, 2021.

Welch, B. L.: The generalization of 'Student's' problem when several different population variances are involved, Biometrika, 34, 28–35, https://doi.org/10.1093/biomet/34.1-2.28, 1947.

Winterrath, T., Rosenow, W., and Weigl, E.: On the DWD quantitative precipitation analysis and nowcasting system for real-time application in German flood risk management, IAHS-AISH publication, pp. 323–329, 2012.

WMO: Nowcasting, https://old.wmo.int/extranet/pages/prog/amp/pwsp/Nowcasting.htm, last access: 8 February 2022.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S.: CBAM: Convolutional Block Attention Module, 2018.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H.: How transferable are features in deep neural networks?, 2014.

Yu, F. and Koltun, V.: Multi-Scale Context Aggregation by Dilated Convolutions, in: International Conference on Learning Representations (ICLR), 2016.