Deep Learning for Improving Numerical Weather Prediction of Rainfall Extremes

Philipp Hess[1] and Niklas Boers[2]

[1]Free University of Berlin
[2]Potsdam Institute for Climate Impact Research

November 22, 2022

Abstract

The accurate prediction of rainfall, and in particular rainfall extremes, remains challenging for numerical weather prediction models. This can be attributed to subgrid-scale parameterizations of processes that play a crucial role in the multi-scale dynamics, as well as the strongly intermittent nature and the highly skewed, non-Gaussian distribution of rainfall. Here we show that a specific type of deep neural networks can learn rainfall extremes from a numerical weather prediction ensemble. A frequency-based weighting of the loss function is proposed to enable the learning of extreme values in the distributions' tails. We apply our framework in a post-processing step to correct for errors in the model-predicted rainfall. Our method yields a much more accurate representation of relative rainfall frequencies and improves the forecast skill of extremes by factors ranging from two to above six, depending on the event magnitude.

Deep Learning for Improving Numerical Weather Prediction of Rainfall Extremes

Philipp Hess[1,2], Niklas Boers[1,2,3]

[1]Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, Berlin, 14195, Germany
[2]Potsdam Institute for Climate Impact Research (PIK), Telegraphenberg A31, Potsdam, 14473, Germany
[3]Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, UK

Key Points:

- Correcting biases in the rainfall forecast of a numerical weather prediction ensemble with a deep neural network.
- Training with a weighted loss function enables the neural network to learn the heavy tailed target distribution.
- The method improves the relative frequency and categorical skill scores of rainfall extremes.

Corresponding author: Philipp Hess, `hess@pik-potsdam.de`

Abstract

The accurate prediction of rainfall, and in particular rainfall extremes, remains challenging for numerical weather prediction models. This can be attributed to subgrid-scale parameterizations of processes that play a crucial role in the multi-scale dynamics, as well as the strongly intermittent nature and the highly skewed, non-Gaussian distribution of rainfall. Here we show that a specific type of deep neural networks can learn rainfall extremes from a numerical weather prediction ensemble. A frequency-based weighting of the loss function is proposed to enable the learning of extreme values in the distributions' tails. We apply our framework in a post-processing step to correct for errors in the model-predicted rainfall. Our method yields a much more accurate representation of relative rainfall frequencies and improves the forecast skill of extremes by factors ranging from two to above six, depending on the event magnitude.

Plain Language Summary

Modelling rainfall is challenging because of its large variability in space and time, and its highly skewed distribution. Numerical weather prediction (NWP) models have to be simulated on discretized grids with finite resolution. Although important especially for the generation of rainfall, small-scale processes can therefore not be resolved explicitly and must be paremeterized, i.e. included as empirical functions of the resolved variables. This introduces model biases that can lead to an underestimation of extreme events. Here we apply a deep neural network (DNN) to correct biases in the rainfall forecast of a NWP ensemble. The DNN is optimized with a loss function that includes weights to penalize rare extremes, and shows substantially improved performance in the prediction of extreme rainfall.

1 Introduction

Modelling and predicting rainfall, and in particular its extremes, is challenging because of the relevant multi-scale dynamics ranging from small-scale droplet interactions to large-scale weather systems, the high intermittency in space and time, as well the strongly non-Gaussian, right-skewed distribution (Koutsoyiannis, 2004b, 2004a). With larger spatial averages approximately following a positive trend expected from the thermodynamic Clausius-Clapeyron relation (Allan & Soden, 2008; Donat et al., 2013; Guerreiro et al., 2018), the frequency and severity of extreme rainfall are projected to increase in a warming atmosphere (Fischer & Knutti, 2016), making their accurate prediction even more challenging but also more important.

Numerical weather prediction (NWP) models, solving the fluid dynamical equations governing the dynamics of the atmosphere, are essential for weather forecasting, including the prediction of heavy rainfall events. Despite the large improvements made over the past decades (Bauer et al., 2015), considerable sources of error remain in the models, in particular for rainfall (Boyle & Klein, 2010). Global NWP models, with a resolution of about 20 km, cannot explicitly resolve many of the relevant small-scale processes. These processes need to be included as sub-grid parameterizations, i.e., they are written as empirical functions of the explicitly resolved (grid-scale) variables. These parameterizations of important processes involved in the generation of rainfall introduces biases and errors that can lead to an underestimation of extremes (Kang et al., 2015).

Recent work has shown promising results by including data-driven machine learning methods including neural networks (LeCun et al., 2015), into the traditional NWP workflow. Well-suited applications of neural networks range from data-assimilation (Bocquet et al., 2020), purely data-driven and hybrid weather prediction (Weyn et al., 2020; Rasp & Thuerey, 2021; Brenowitz & Bretherton, 2018; Watt-Meyer et al., 2021) to post-processing NWP output (Rasp & Lerch, 2018; Grönquist et al., 2021).

Here we apply a deep neural network (DNN) to correct the ECMWF (European Centre for Medium-Range Weather Forecasts, 2012) Integrated Forecast System (IFS) for biases by post-processing its rainfall output. When DNNs are tasked to infer a variable with large intermittency and a heavy-tailed distribution, such as rainfall, the optimization with an averaging loss function such as the widely employed mean squared error (MSE) can be expected to lead to a good approximate of the distribution's mean, but an underestimation of the extreme values in the tail. For rainfall, this problem has been addressed in different ways, e.g by translating the regression task into a classification problem (Agrawal et al., 2019; Sønderby et al., 2020), by using methods from computer vision (Tran & Song, 2019), and by employing a weighted loss function (Shi et al., 2017; Franch et al., 2020). The latter being composed of a weighted MSE and mean absolute error (MAE), with a set of five discrete weights determined by binned rainfall intensities. We show that a state-of-the-art DNN architecture is able to infer extreme values in the far right tail of the target distribution from remotely sensed rainfall data using a loss that combines a continuously weighted MSE with a structural similarity measure. Notably, we use NWP ensemble simulations as input features, which do not exhibit an accurate representation of the extremes.

2 Materials and Methods

2.1 Integrated forecast system

Atmospheric variables simulated by an ensemble of the Integrated Forecast System (IFS) from the European Center for Medium-Range Weather Forecasting (ECMWF) (European Centre for Medium-Range Weather Forecasts, 2012) are taken as inputs of the DNN. The ensemble consists of ten members with a spatial resolution of 0.5625° (or approximately 63 km) and 137 vertical levels. It is initialized twice daily at 06 and 18 UTC with a 12 hour lead time and small perturbations in the initial conditions. In this work, the ensemble mean of the variables is used, which is provided at three-hourly time steps and 0.5° horizontal resolution.

2.2 Training data

The input features of the DNN are the three-hourly accumulated rainfall and vertical velocities of the IFS ensemble mean. The latter is taken from eleven pressure levels: 200, 250, 300, 400, 500, 600, 700, 800, 900, 950, and 1000 hPa. The vertical velocity is dynamically linked to rainfall through convective processes and large-scale updrafts of warm, moist air (Pfahl et al., 2017; Müller et al., 2020). The satellite-based Tropical Rainfall Measurement Mission (TRMM) 3B42 V7 product (Huffman et al., 2007) is used as a training ground truth at three-hourly temporal resolution and is regridded to 0.5° by bilinear interpolation using the the Climate Data Operator (CDO) software (Schulzweida, 2019), to match the IFS grid. The TRMM data is considered to have high accuracy especially for heavy rainfall extremes (Boers et al., 2015). The geographic region of this study is the entire spatial coverage of the TRMM product, which ranges from 50° S to 50° N and 180° W to 180° W. Further, the June, July and August season is used and split into a training set (1998-2008), a validation set (2009-2011) to optimize the hyperparameters of the DNN model, and a test set for evaluation (2012-2014). Although the TRMM product is continued till present, a change of the satellites in 2014 has introduced significant biases, as shown in Figure S5, and the period after 2014 was therefore excluded.

2.3 Definition of rainfall extremes

We define extreme events as those 3-hourly time steps for which the rainfall sums exceed a pre-defined threshold. This threshold is determined individually for each grid

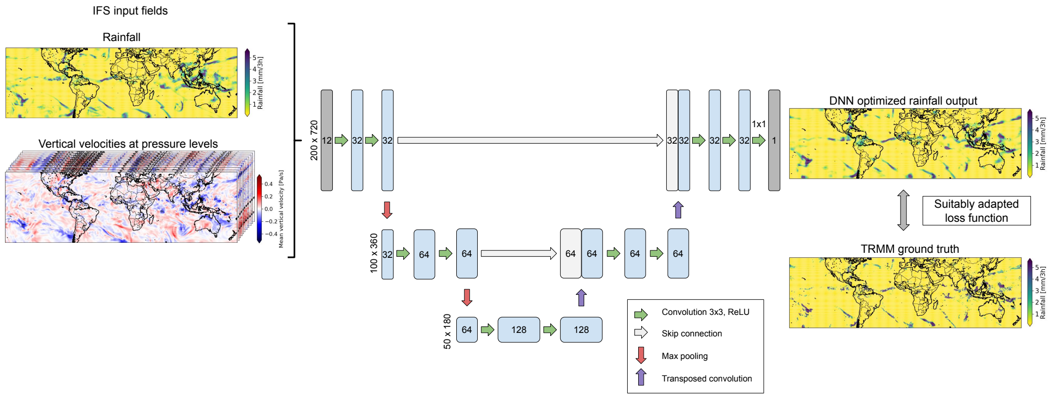

Figure 1. Sketch of the U-Net-based DNN architecture, the number of channels is indicated inside each layer. The horizontal dimensions per pooling level are given on the left.

cell in terms of percentiles, computed on the sets of 3-hourly time steps with rainfall amounts above 0.1 [mm/3h].

2.4 Neural network architecture

The DNN architecture is based on the U-Net (Ronneberger et al., 2015), a convolutional neural network that can capture multi-scale spatial patterns through a combination of pooling operations for large-scale feature extraction and skip-connections to preserve small-scale, high-frequency information. The U-Net architecture has shown good performance in weather prediction and post-processing tasks (Grönquist et al., 2021; Weyn et al., 2020). The model, shown in Figure 1, takes the standardized spatial fields of the atmospheric variables as input, where the number of 12 input channels equals the number of variables times the corresponding number of pressure levels. The output layer has a single channel representing the rainfall rates and applies a rectified linear unit (ReLU) to ensure non-negative output values. The number of weights per layer is reduced by half compared to the original model from (Ronneberger et al., 2015), and only two max pooling operations are applied since a larger model size did not improve the performance. The ADAM optimizer (Kingma & Ba, 2017) was used for training the network together with a batch size of 64, a learning rate of 10^{-4} and early stopping to prevent overfitting.

2.5 Loss function

To improve the training regarding extreme values and the intermittency, we propose the weighted loss function

$$L_\lambda(y, \hat{y}) = \frac{\lambda}{N} \sum_{i=1}^{N} w(y_i)(y_i - \hat{y}_i)^2 + (1 - \lambda)\text{MS-SSIM}(y, \hat{y}), \tag{1}$$

where N is the number of training examples, w is a weight function and y and $\hat{y}$ are the target and prediction, respectively. The cost function is thus a convex sum of the weighted MSE and the so-called multi-scale structural similarity measure MS-SSIM (Wang et al., 2003), introducing an additional hyperparameter λ. The MS-SSIM quantifies the structural similarity between two images. This is done through an interative comparison of luminance, contrast and structure on different scales by downsampling and low-pass filtering the image signals (see supporting information). The weights w are defined as

$$w(y_i) = \min\left(\alpha e^{\beta y_i}, 1\right), \tag{2}$$

where α and β are hyperparameters. We optimize the hyperparameters on the validation set and set them to $\alpha = 0.007$, $\beta = 0.048$ and $\lambda = 0.158$. Since the relative frequency of 3-hourly rainfall events decreases approximately exponentially with increasing magnitude, the weights aim to account for the statistical imbalance. Ebert-Uphoff et al. (Ebert-Uphoff & Hilburn, 2020) also use an exponentially weighted MSE loss to emphasis rare and high values when training a DNN to estimate radar composite reflectivity from satellite imagery. In our case, we find that only optimizing with the weighted MSE leads to large biases which can be removed through the addition of the MS-SSIM into the loss. Further introducing bounds on the weights was crucial for a robust optimization of the network.

2.6 Baseline

A linear ridge regression (Hoerl & Kennard, 1970) with the IFS ensemble mean rainfall of a single grid-cell as input is used as a baseline model. Including the vertical velocity fields did not improve the performance of this baseline model.

3 Results

3.1 Evaluation of the continuous forecast skill of the deep learning model

We first compare the histograms of the relative frequencies of the 3-hourly rainfall values for the outputs from IFS, the different post-processing models, and the ground truth given by the TRMM remote sensing product (Figure 2a, 2b). The histograms of grid-cell values are computed over the entire part of the globe covered by the TRMM data (50°S to 50°N) and test set period. Training the DNN with an MSE or a MS-SSIM loss leads to a similar rainfall frequency distribution as the IFS ensemble mean and the linear ridge regression baseline, with over-representation of low rainfall frequencies and underestimation of the tail, as compared to the observational TRMM target. Training with the CW loss function in Eq. (1), instead, enables the DNN to infer a distribution that is substantially closer to the target distribution. The frequencies of low rainfall rates are correctly reduced, while at the same time achieving a better statistical representation of the extremes in the tail. The ridge regression shows the largest bias towards low rainfall rates, hence not improving the IFS output at all.

We assess the continuous forecast skill of the different models by computing the root mean square error (RMSE), mean error (ME) and the complex-wavelet structural similarity index (CW-SSIM) (Sampat et al., 2009) (see supporting information). The CW-SSIM allows a structural comparison of two images that is insensitive to small non-structural transformations such as rotation and translation, but sensitive to structural changes such as sharpness. Time steps with rainfall below a threshold of 0.1 [mm/3h] have been excluded before applying the error metrics since rainfall on such low scales cannot be measured accurately by satellite-based remote sensing (Huffman et al., 2007). The results are summarized in Table 1 as averages of the absolute cell-wise metrics. Training the DNN with the MS-SSIM leads to the lowest RMSE, while the CW loss function shows a ME similar to the MS-SSIM, and the highest structural similarity. Processing the IFS output with the ridge regression does not lead to improvements. Omitting rainfall from the input features and thus purely focusing on the vertical wind velocities W is not significantly affecting the performance of the model. The weighted loss function combined with

Table 1. Continuous validation statistics are given for the IFS ensemble mean, ridge regression and the DNNs trained with different loss functions and the input variables rainfall (P) and vertical velocity (W) from the IFS.

Model	Loss	Input	RMSE	%	ME	%	CW-SSIM	%
IFS	-	-	1.457	-	0.175	-	0.359	-
Ridge Regr.	MSE	P	1.473	-1.1	0.209	-19.4	0.359	0
DNN	MSE	W	1.375	5.6	0.165	5.7	0.388	8.1
DNN	MSE	P, W	1.372	5.8	0.166	5.1	0.395	10
DNN	MS-SSIM	P, W	**1.368**	**6.1**	0.136	22.3	0.441	22.8
DNN	CW	P, W	1.439	1.2	**0.135**	**22.9**	**0.545**	**51.8**

the MS-SSIM leads to an improvement of the ME by almost 23% and an improvement of the CW-SSIM metric by more than 50%.

3.2 Evaluation of the forecast skill of the deep learning model for extreme events

To evaluate the forecast skill for extreme events, categorical statistics can be computed from the contingency table containing the true positives and negatives, as well as the false positives and negatives (Table S1). A detailed definition of the events and skill scores is given in the supporting information. Table 2 summarizes the skill scores for events above the 95th percentile. The Heidke Skill Score (HSS), which is equal to zero for a random forecast and equal to one for a perfect forecast, is shown in Figure 2c for thresholds ranging from the 75th to the 99th percentile; corresponding results for the other scores are given in the supplementary Figures S1 to S4. The DNNs improve the scores compared to the IFS mean and ridge regression, in particular for events above the 90th and higher percentiles (Figure 2c). The DNN trained using the MS-SSIM alone as loss shows the highest scores below the 95th threshold. The proposed CW loss leads to significant improvements even above the 95th percentile (improving the IFS forecast by 192% in terms of the HSS) and yields the only skilful forecast for events above the 99th percentile (improving the IFS forecast by more than 500% in terms of the HSS). Note that the FAR score is not as strongly improved as the other skills, indicating slightly more frequent false alarms when optimizing with the CW loss. We attribute this to the highly localized, intermittent nature of rainfall extremes and emphasize that - in view of the results for the other error metrics - the increased number of false positives is more than balanced by the increased number of true positives. The DNN trained with the combined weight (CW) introduced above leads to substantial improvements also for the spatial patterns of extremes, in particular for regions with stronger extreme rainfall events (Figure 3).

4 Discussion

We introduced a DNN to model rainfall extremes from short-range numerical weather ensemble forecasts. To address the strong statistical imbalance of the training data, a loss function is introduced that combines a weighted MSE with a structural similarity measure. The proposed combined loss function (CW) is found to substantially improve the training with respect to extremes compared to using the MSE and MS-SSIM individually, which are two commonly used loss functions. For comparison, we show that post-processing the IFS mean with a ridge regression model does not lead to any improvements. This motivates the importance of a non-linear DNN architecture such as the U-

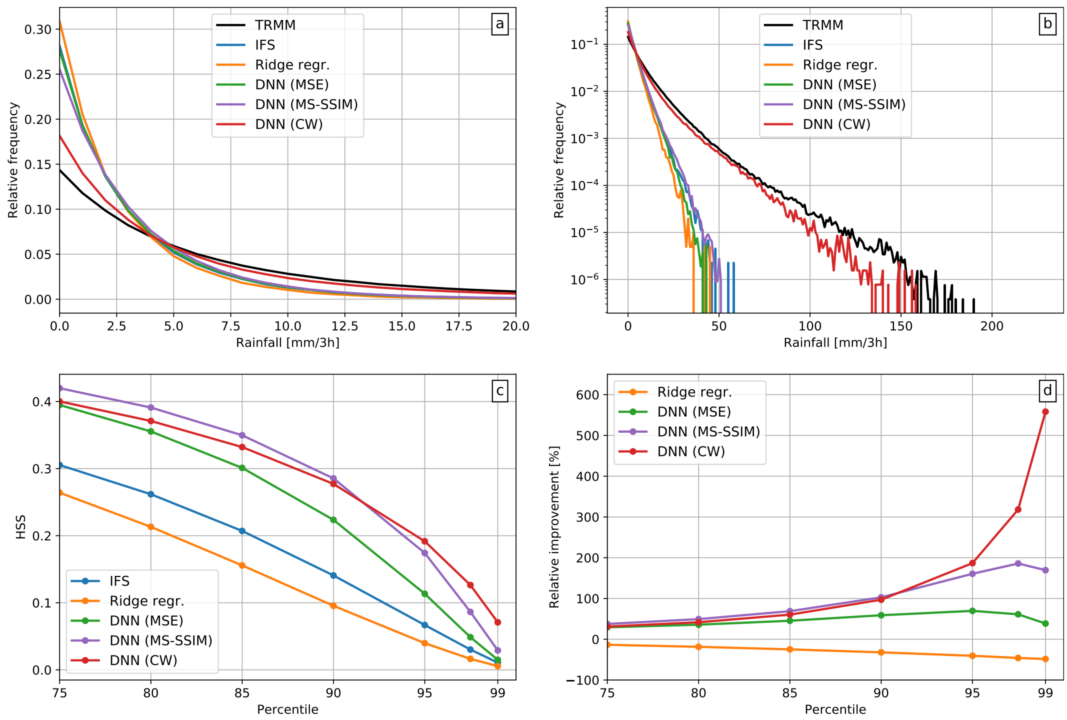

Figure 2. Relative rainfall frequencies and categorical extreme rainfall forecast scores for the different post-processing models compared to the IFS. Histograms of three-hourly rainfall event magnitudes are shown on a linear y-axis (a) and a logarithmic y-axis (b) for TRMM (black), IFS (blue), ridge regression (orange), DNN trained with the MSE loss (green), the MS-SSIM loss (purple) and the CW loss (red). (c) The Heidke Skill Score (HSS) for events above increasing percentile thresholds is shown for the IFS (blue), ridge regression (orange), DNN trained with the MSE loss (green), the MS-SSIM loss (purple), and with the CW loss proposed here (red). A HSS greater than zero implies an improvement over a random forecast, and HSS = 1 would imply a perfect forecast (see supporting information). (d) The relative improvement of the different machine learning methods over the IFS mean, in percentages.

Table 2. Event-based forecast skill scores for rainfall events above the 95th percentile. The percentage columns give the relative improvement over the IFS mean for each error metric and skill score.

Model	Loss	HSS	%	F1	%	CSI	%	POD	%	FAR	%
IFS	-	0.067	-	0.069	-	0.036	-	0.041	-	0.778	-
Ridge Regr.	MSE	0.040	-40	0.041	-41	0.021	-42	0.022	-46	0.775	0
DNN	MSE	0.113	69	0.115	67	0.061	69	0.066	61	**0.567**	**27**
DNN	MS-SSIM	0.174	160	0.177	157	0.097	169	0.115	180	0.622	20
DNN	CW	**0.192**	**187**	**0.195**	**183**	**0.108**	**200**	**0.139**	**239**	0.673	13

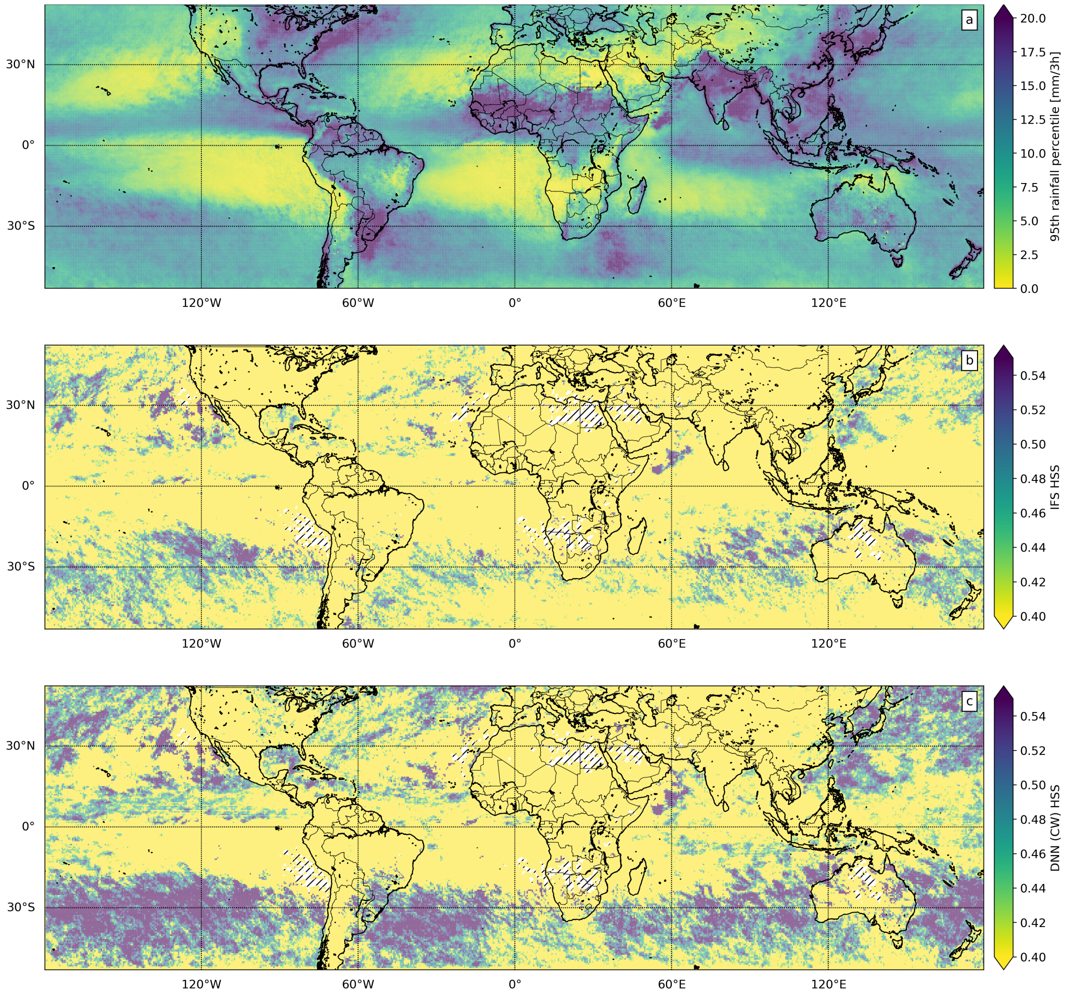

Figure 3. Spatial distribution of the 95th rainfall percentile and HSS for events above the 95th percentile. (a) The 95th percentile of the rainfall distribution at each grid cell of the TRMM dataset. (b) The spatially resolved HSS for the IFS mean. (c) The spatially resolved HSS for the DNN post-processed forecast, trained with the proposed CW loss. Hatched areas indicate grid-cells where the HSS could not be evaluated because no extreme events occurred in these locations.

Net. Moreover, our results suggest that the U-Net architecture is indeed capable of capturing the multi-scale spatial structure of rainfall accurately.

The CW loss substantially improves relative rainfall frequencies in the DNN output, the mean error and structural similarity of overall rainfall fields, as well as categorical skill scores for extreme events above the 90th and higher percentile, with strongly increasing rate of improvement for higher thresholds.

Taking the mean of the IFS ensemble is expected to damp the extremes in the forecast. Hence, the results of the IFS shown here do not represent the skill of single ensemble members to forecast extremes. Nevertheless, our results demonstrate the ability of the proposed DNN architecture to learn extremes that are not resolved in the input features, and to substantially improve their prediction.

Interestingly, the error statistics did not change significantly when rainfall was excluded and only the vertical wind speed were considered as input features. This indicates that the DNN can learn a good representation of rainfall and especially its extremes from the vertical velocity alone.

Similarly surprising is the improved structural similarity when using the CW loss, compared to using the MS-SSIM alone as loss function. Although the considered forecast has a high temporal resolution of three hours, the forecast lead time of up to twelve hours is still comparably short. With applications to disaster prevention in mind, an extension of the study to longer forecast lead times will be an important direction for future research. Further, making use of the entire IFS ensemble will allow to incorporate uncertainties into the framework that are essential for operational forecasting of extreme events.

Acknowledgments

The authors acknowledge funding by the Volkswagen Foundation, as well as the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for supporting this project by providing resources on the high performance computer system at the Potsdam Institute for Climate Impact Research. The Pytroch 1.7.0 (Paszke et al., 2019) source code for training and data processing is available at [github-link when accepted]. The IFS training data is available for download at the Copernicus Climate Change Service (C3S) (Hersbach et al., 2020) (`https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels`). The TRMM (TMPA) data can be obtained at the Goddard Earth Sciences Data and Information Services Center (GES DISC) (Leptoukh, 2005) (`https://disc.gsfc.nasa.gov/datasets/TRMM\_3B42\_7/summary`).

References

Agrawal, S., Barrington, L., Bromberg, C., Burge, J., Gazen, C., & Hickey, J. (2019). *Machine learning for precipitation nowcasting from radar images.*

Allan, R. P., & Soden, B. J. (2008). Atmospheric warming and the amplification of precipitation extremes. *Science*, *321*(5895), 1481–1484.

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55.

Bocquet, M., Farchi, A., & Malartic, Q. (2020). Online learning of both state and dynamics using ensemble kalman filters. *Foundations of Data Science*, *0*(0), 0. Retrieved from `http://dx.doi.org/10.3934/fods.2020015` doi: 10.3934/fods.2020015

Boers, N., Bookhagen, B., Marengo, J., Marwan, N., von Storch, J.-S., & Kurths, J. (2015). Extreme rainfall of the south american monsoon system: a dataset comparison using complex networks. *Journal of Climate*, *28*(3), 1031–1056.

Boyle, J., & Klein, S. A. (2010). Impact of horizontal resolution on climate model forecasts of tropical precipitation and diabatic heating for the twp-ice period. *Journal of Geophysical Research: Atmospheres*, *115*(D23).

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298.

Donat, M., Alexander, L., Yang, H., Durre, I., Vose, R., Dunn, R., . . . others (2013). Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The hadex2 dataset. *Journal of Geophysical Research: Atmospheres*, *118*(5), 2098–2118.

Ebert-Uphoff, I., & Hilburn, K. (2020). Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society*, *101*(12), E2149–E2170.

European Centre for Medium-Range Weather Forecasts. (2012). *The ECMWF ensemble prediction system.* https://www.ecmwf.int/sites/default/files/the_ECMWF_Ensemble_prediction_system.pdf.

Fischer, E. M., & Knutti, R. (2016). Observed heavy precipitation increase confirms theory and early models. *Nature Climate Change*, *6*(11), 986–991.

Franch, G., Nerini, D., Pendesini, M., Coviello, L., Jurman, G., & Furlanello, C. (2020). Precipitation nowcasting with orographic enhanced stacked generalization: Improving deep learning predictions on extreme events. *Atmosphere*, *11*(3), 267.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200092.

Guerreiro, S. B., Fowler, H. J., Barbero, R., Westra, S., Lenderink, G., Blenkinsop, S., . . . Li, X.-F. (2018). Detection of continental-scale intensification of hourly rainfall extremes. *Nature Climate Change*, *8*(9), 803–807.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., . . . others (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., . . . Stocker, E. F. (2007). The trmm multisatellite precipitation analysis (tmpa): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of hydrometeorology*, *8*(1), 38–55.

Kang, I.-S., Yang, Y.-M., & Tao, W.-K. (2015). Gcms with implicit and explicit representation of cloud microphysics for simulation of extreme precipitation frequency. *Climate Dynamics*, *45*(1-2), 325–335.

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization.*

Koutsoyiannis, D. (2004a, aug). Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records. *Hydrological Sciences Journal*, *49*(4), 591–610. doi: 10.1623/hysj.49.4.591.54424

Koutsoyiannis, D. (2004b, aug). Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. *Hydrological Sciences Journal*, *49*(4), 575–590. doi: 10.1623/hysj.49.4.575.54430

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Leptoukh, G. (2005). Nasa remote sensing data in earth sciences: Processing, archiving, distribution, applications at the ges disc. In *Proc. of the 31st intl symposium of remote sensing of environment.*

Müller, A., Niedrich, B., & Névir, P. (2020). Three-dimensional potential vorticity structures for extreme precipitation events on the convective scale. *Tellus*

A: Dynamic Meteorology and Oceanography, *72*(1), 1–20.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.
Pfahl, S., O'Gorman, P. A., & Fischer, E. M. (2017). Understanding the regional pattern of projected future changes in extreme precipitation. *Nature Climate Change*, *7*(6), 423–427.
Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900.
Rasp, S., & Thuerey, N. (2021). Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: A new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002405.
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).
Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., & Markey, M. K. (2009). Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, *18*(11), 2385–2401.
Schulzweida, U. (2019, October). *Cdo user guide.* Retrieved from `https://doi.org/10.5281/zenodo.3539275` doi: 10.5281/zenodo.3539275
Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., kin Wong, W., & chun Woo, W. (2017). *Deep learning for precipitation nowcasting: A benchmark and a new model.*
Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., … Kalchbrenner, N. (2020). *Metnet: A neural weather model for precipitation forecasting.*
Tran, Q.-K., & Song, S.-k. (2019). Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks. *Atmosphere*, *10*(5), 244.
Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003* (Vol. 2, pp. 1398–1402).
Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J. J., … Bretherton, C. S. (2021). *Correcting weather and climate models by machine learning nudged historical simulations.*
Weyn, J. A., Durran, D. R., & Caruana, R. (2020). Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002109.

Supporting Information for "Deep Learning for Improving Numerical Weather Prediction of Rainfall Extremes"

Philipp Hess[1,2], Niklas Boers[1,2,3]

[1]Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, Berlin, 14195, Germany

[2]Potsdam Institute for Climate Impact Research (PIK), Telegraphenberg A31, Potsdam, 14473, Germany

[3]Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, UK

Contents of this file

1. Text S1 to S2
2. Figures S1 to S5
3. Table S1

Text S1. The root mean square error (RMSE) and mean error (ME) are defined as,

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{1}$$

$$\text{ME} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i), \tag{2}$$

where N is the number of training examples, y is the TRMM target and $\hat{y}$ is the modelled rainfall output. The multi-scale structural similarity measure (MS-SSIM)(Wang et al., 2003) quantifies the structural similarity between two images, in our case two spatial rain-

Corresponding author: Philipp Hess, hess@pik-potsdam.de

fall maps, as sets of N grid-cells, i.e. $\mathbf{y} = \{y_i | i = 1, 2, ..., N\}$ and $\hat{\mathbf{y}} = \{\hat{y}_i | i = 1, 2, ..., N\}$. The MS-SSIM then iteratively computes three measures, for luminance $l(\mathbf{y}, \hat{\mathbf{y}})$, contrast $c(\mathbf{y}, \hat{\mathbf{y}})$ and structure $s(\mathbf{y}, \hat{\mathbf{y}})$ by successively downsampling and low-pass filtering the image signals. The three measures are defined as

$$l(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\mu_y \mu_{\hat{y}} + C_1}{\mu_y^2 + \mu_{\hat{y}}^2 + C_1}, \tag{3}$$

$$c(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\sigma_y \sigma_{\hat{y}} + C_2}{\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2}, \tag{4}$$

$$s(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sigma_{y\hat{y}} + C_3}{\sigma_y \sigma_{\hat{y}} + C_3}, \tag{5}$$

where μ_y is the mean, σ_y the standard deviation of $\mathbf{y}$ and $\sigma_{y,\hat{y}}$ the covariance of $\mathbf{y}$ and $\hat{\mathbf{y}}$. The small constants C_1, C_2, and C_3 are inlcuded to improve the stability. The MS-SSIM can then be written as,

$$\text{MS-SSIM}(\mathbf{y}, \hat{\mathbf{y}}) = [l_M(\mathbf{y}, \hat{\mathbf{y}})]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(\mathbf{y}, \hat{\mathbf{y}})]^{\beta_j} \cdot [s_j(\mathbf{y}, \hat{\mathbf{y}})]^{\gamma_j}, \tag{6}$$

where M denotes the number downsampling iterations. The exponents α_M, β_j and γ_j can be adjusted to give different weights to the measures, but are set to $\alpha_j = \beta_j = \gamma_j$. The complex wavelet structural similarity (CW-SSIM)(Sampat et al., 2009), extends the idea of structural similarity to the complex wavelet domain. The motivation behind it is that structural changes between two images, such as small rotations or translations will lead to a constant relative phase shift in the coefficients of a complex wavelet transform. Therefore, the CW-SSIM is constructed in such a way that it is insensitive to *relative* phase shifts and magnitude distortions. On the other hand it is sensitive to non-structural transformations in images, such as changes in sharpness, that will lead to phase shifts in the coefficients. The CW-SSIM is defined as

$$\text{CW-SSIM}(\mathbf{c}_y, \mathbf{c}_{\hat{y}}) = \frac{2|\sum_{i=1}^{N} c_{y,i} c_{\hat{y},i}^*| + C}{\sum_{i=1}^{N} |c_{y,i}|^2 + \sum_{i=1}^{N} |c_{\hat{y},i}|^2 + C}, \tag{7}$$

where $\mathbf{c}_y = \{c_{y,i} | i = 1, 2, ..., N\}$ and $\mathbf{c}_{\hat{y}} = \{c_{\hat{y},i} | i = 1, 2, ..., N\}$ are two sets of complex wavelet coefficients obtained at the same spatial location and wavelet subbands of the two images being compared. The asterix denotes the complex conjugate and C is a small constant for stability.

Text S2. We quantify the forecast skill of extreme events with categorical skill scores commonly used in meteorology and machine learning, such as the critical success index (CSI), probability of detection (POD), false alarm ratio (FAR), F1 and Heidke skill score (HSS). These skill scores can be computed from the contingency table (see Table S1). The table classifies event forecast outcomes into true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). Based on these categories, the skill scores can be defined as

$$\begin{aligned}
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\
\text{F1} &= 2\frac{\text{Precision Precision}}{\text{Precision} + \text{Precision}}, \\
\text{HSS} &= \frac{2(\text{TP TN} - \text{FP FN})}{(\text{TP} + \text{FN})(\text{FN} + \text{TN}) + (\text{TP} + \text{FP})(\text{FP} + \text{TN})}, \\
\text{CSI} &= \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}, \\
\text{POD} &= \text{Recall}, \\
\text{FAR} &= \frac{\text{FP}}{\text{FP} + \text{TP}}.
\end{aligned}$$

The recall score computes the proportion of relevant events that were classified correctly and precision gives the fraction of positive classifications that were correct. The F1 score combines precision and recall as a harmonic mean and is commonly used in machine learning to evaluate predictions on strongly imbalanced data. The Heidke Skill Score

(HSS) evaluates the accuracy of event predictions, e.g. rainfall extremes, relative to a random forecast and can also be used for strongly imbalanced classes. The critical success (CSI) relates the accuracy of event predictions to the actually observed events, without accounting for correct negative predictions. The probability of detection (POD) and false alarm ratio (FAR) scores should be assessed together, where the former is defined identically to the recall score. Since POD ignores false alarms, the false alarms ratio (FAR) can be used to evaluate these.

References

Sampat, M. P., Wang, Z., Gupta, S., Bovik, A. C., & Markey, M. K. (2009). Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, *18*(11), 2385–2401.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003* (Vol. 2, pp. 1398–1402).

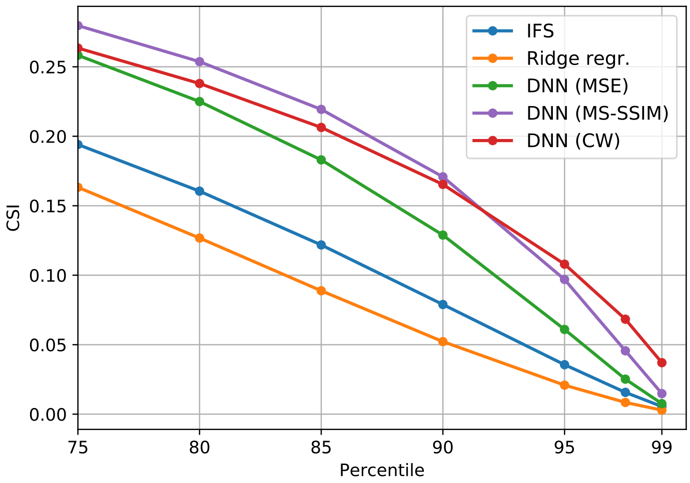

Figure S1. The critical success index (CSI) for rainfall events above the 75th percentile threshold.

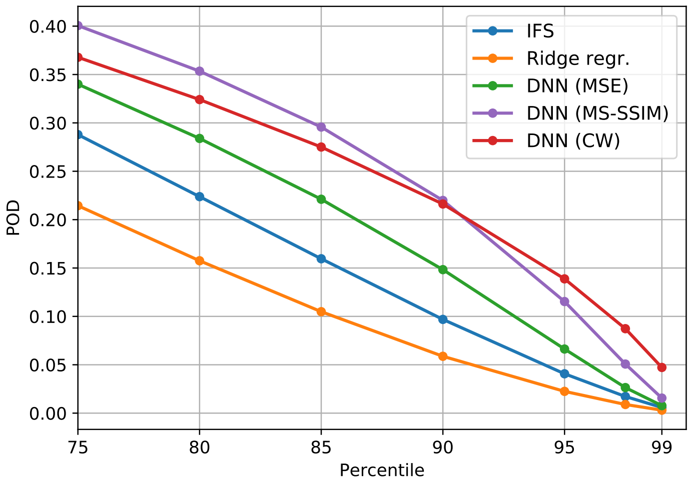

Figure S2. The probability of detection (POD) of rainfall events above the 75th percentile threshold.

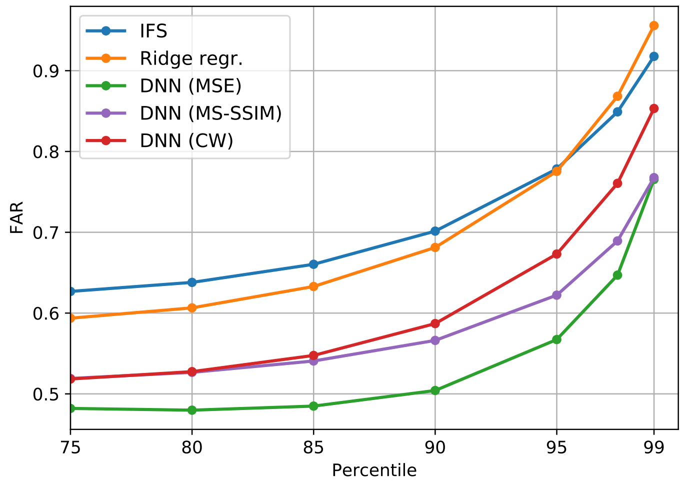

Figure S3. The false alarm ratio (FAR) of rainfall events above the 75th percentile threshold.

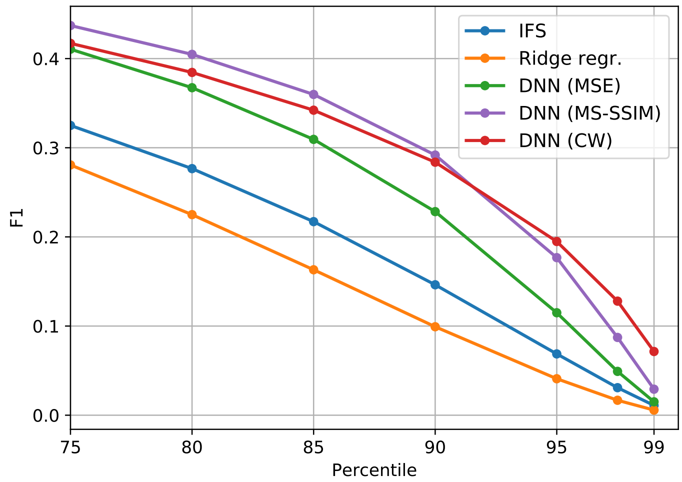

Figure S4. The F1 score for rainfall events above the 75th percentile threshold.

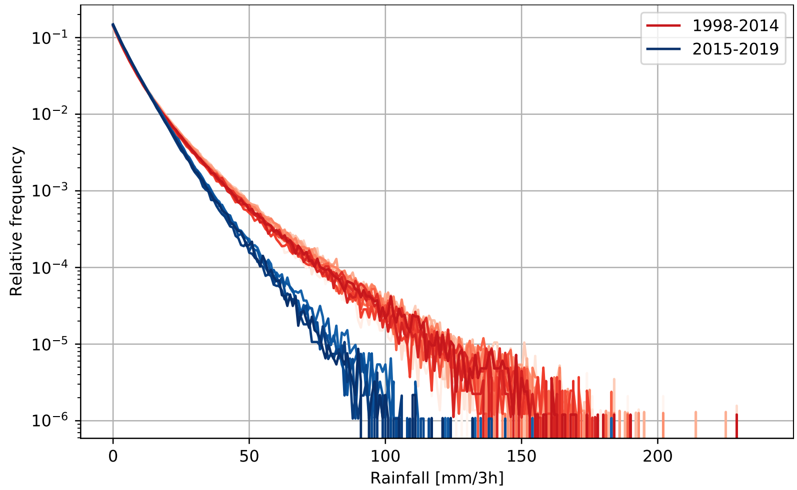

Figure S5. The histograms of grid-cell values show here are computed over the entire part of the globe covered by the TRMM data (50°S to 50°N) and for single years. The histograms of years before 2015 are colored in red and for years thereafter in blue.

Table S1. Contingency table of forecast outcomes for binary events.

	Observed	Not observed
Forecasted	True positive (TP)	False positive (FP)
Not forecasted	False negative (FN)	True negative (TN)