JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2021MS002765

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 combining two terms enables the neural network to learn the heavy tailed target distribution
- The method improves the relative frequency and categorical skill scores of heavy rainfall

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

P. Hess,
philipp.hess@tum.de

Citation:

Hess, P., & Boers, N. (2022). Deep learning for improving numerical weather prediction of heavy rainfall. *Journal of Advances in Modeling Earth Systems*, *14*, e2021MS002765. https://doi.org/10.1029/2021MS002765

Received 10 AUG 2021
Accepted 12 MAR 2022

Deep Learning for Improving Numerical Weather Prediction of Heavy Rainfall

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

[1]School of Engineering & Design, Technical University of Munich, Munich, Germany, [2]Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany, [3]Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, UK

Abstract The accurate prediction of rainfall, and in particular of the heaviest rainfall events, remains challenging for numerical weather prediction (NWP) models. This may be due to subgrid-scale parameterizations of processes that play a crucial role in the multi-scale dynamics generating rainfall, as well as the strongly intermittent nature and the highly skewed, non-Gaussian distribution of rainfall. Here we show that a U-Net-based deep neural network can learn heavy rainfall events from a NWP ensemble. A frequency-based weighting of the loss function is proposed to enable the learning of heavy rainfall events 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 heavy rainfall events by factors ranging from two to above six, depending on the event magnitude.

Plain Language Summary Modeling 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, that is, included as empirical functions of the resolved variables. This introduces model biases that can lead to an under- or overestimation of heavy rainfall 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 account for heavy rainfall events, and shows substantially improved performance in their prediction.

1. Introduction

Modeling and predicting rainfall, and in particular heavy rainfall events, remains is challenging. The relevant multi-scale dynamics range from small-scale droplet interactions to large-scale weather systems. Further, the high intermittency in space and time, as well the strongly non-Gaussian, right-skewed distribution (Koutsoyiannis, 2004a, 2004b) make accurate predictions difficult.

The thermodynamic Clausius-Clapeyron relation (Allan & Soden, 2008; Donat et al., 2013; Guerreiro et al., 2018), and comprehensive model simulations (Masson-Delmotte, V. et al., 2021) suggest that the frequency and severity of heavy rainfall are expected to increase in a warming atmosphere (Fischer & Knutti, 2016). It should be noted, however, that the spatial patterns of these increases are expected to be heterogeneous and complex (Ali et al., 2018; Traxl et al., 2021). Correspondingly, accurate forecasts of heavy rainfall events will become ever more crucial for disaster prevention and mitigation.

Numerical weather prediction (NWP) models solve the fluid dynamical equations governing the dynamics of the atmosphere. They 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 most of 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, that is, they are written as 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 under- or overestimation of the magnitudes of heavy rainfall events (Wilcox & Donner, 2007).

Post-processing methods are commonly applied to the simulated model output to correct for such biases (Berg et al., 2012; Maraun, 2016; Wilks, 2011; Xu, 1999). Traditional approaches relate the biases to differences in long-term statistics of the simulated and observed variable. Among them, quantile mapping (QM) has become particularly popular for weather and climate model applications (Cannon et al., 2015; Déqué, 2007; Gudmundsson et al., 2012; Tong et al., 2021), as it allows to correct for biases over the entire distribution. While correcting the general long-term statistics, these methods, however, do not directly correct for spatial biases in synchronous events that are both modeled and observed.

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 and climate modeling (Brenowitz & Bretherton, 2019; Rasp et al., 2018; Rasp & Thuerey, 2021; Watt-Meyer et al., 2021; Weyn et al., 2020; Yuval & O'Gorman, 2020) to post-processing NWP output (Grönquist et al., 2021; Rasp & Lerch, 2018).

Here we correct the European Center for Medium-Range Weather Forecasting (ECMWF) (European Centre for Medium-Range Weather Forecasts, 2012) Integrated Forecast System (IFS) for biases in both general statistics and local events, by post-processing its rainfall output with a deep neural network (DNN).

When DNNs are tasked to infer a variable with large intermittency and a heavy-tailed distribution, such as rainfall, the optimization with the widely employed mean squared error (MSE) loss function often leads to a good approximate of the distribution's mean. By simply averaging over a sample batch, the loss is dominated by the most frequent values, while outliers in the tail of the target distribution only have a comparably small contribution. This can lead to blurring of the spatial patterns and a less accurate prediction of the high values in the tail, as the model focuses mainly on accurate predicting the most frequent values near the mean.

For rainfall, this problem has been addressed in different ways, for example, by translating the regression task into a classification problem (Agrawal et al., 2019; Sønderby et al., 2020), by using methods from image quality assessment in computer vision (Tran & Song, 2019), and by employing a weighted loss function (Franch et al., 2020; Shi et al., 2017). 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 the U-Net DNN architecture is able to infer high values in the far right tail of the target distribution from remotely sensed rainfall data. Notably, we use NWP ensemble simulations as input features, which do not exhibit an accurate representation of heavy rainfall events. To capture the heavy rainfall events and the intermittent spatial patterns, we introduce a new loss function, which combines a continuously weighted MSE with a structural similarity measure.

2. Materials and Methods

2.1. Integrated Forecast System

Atmospheric variables simulated as reforecasts by a ten-member ensemble of the IFS of the model cycle CY41R2 from the ECMWF (European Centre for Medium-Range Weather Forecasts, 2012) are taken as inputs of the DNN. The data is provided by the ECMWF at three-hourly time steps and 0.5° horizontal resolution. It is initialized twice daily at 06 and 18 UTC with a 12 hr lead time and small perturbations in the initial conditions. In this work, the ensemble mean of the variables is used, since taking the individual ensemble members as inputs would not be computationally feasible at present.

2.2. Training Data

The input features of the DNN are the three-hourly accumulated rainfall and vertical velocities of the IFS ensemble mean at the respective lead time. The forecast consists of three-hourly steps up to 12 hr lead time. The ensemble mean is taken from eleven pressure levels: 200, 250, 300, 400, 500, 600, 700, 800, 900, 950, and 1,000 hPa. The vertical velocity is dynamically linked to rainfall through convective processes and large-scale updrafts of warm, moist air (Müller et al., 2020; O'Gorman & Schneider, 2009; Pfahl et al., 2017). 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. Following (Beck et al., 2019; Rasp et al., 2020) the spatial resolution is regridded to 0.5° using bilinear interpolation to match the IFS grid. The TRMM data is considered to have high

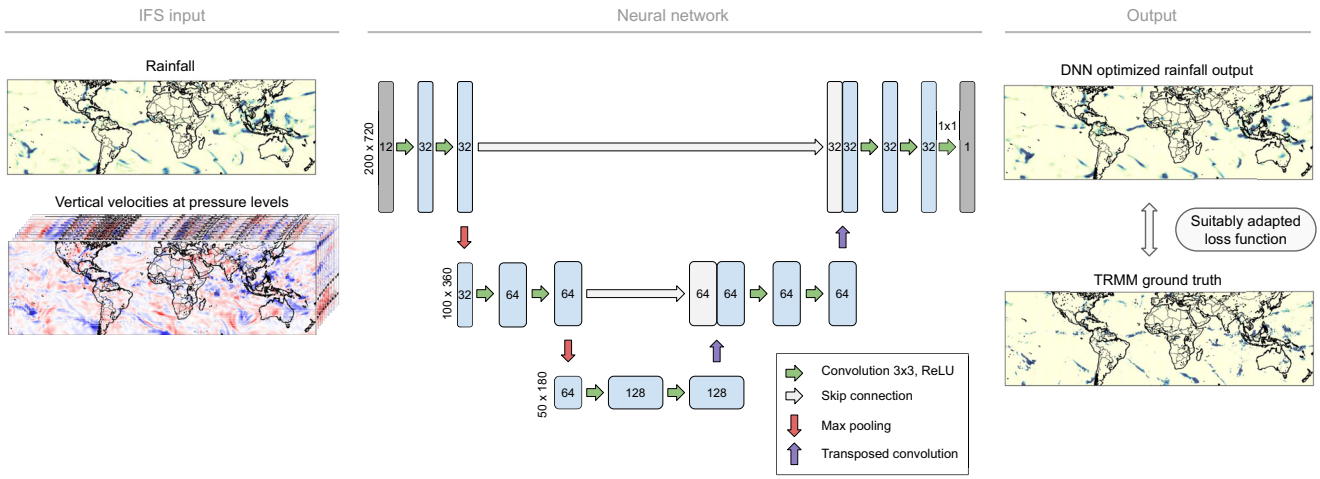

Figure 1. Sketch of the U-Net-based deep neural network (DNN) architecture. IFS output for rainfall and vertical velocities is passed to the DNN, which produces rainfall output optimized to approximate corresponding spatial fields from a satellite-based, quasi-global high-resolution rainfall data set. The number of channels in the DNN is indicated inside each layer. The horizontal dimensions per pooling level are given on the left. The arrows show the operations applied after each layer. Green arrows indicate convolutional operations followed by a ReLU activation function. The skip-connections are shown as gray arrows, transferring the hidden state across the bottleneck. Orange and purple arrows indicate max pooling and transposed convolutions respectively. For a more detailed explanation of a similar sketch we refer the original U-Net publication (Ronneberger et al., 2015).

accuracy especially for heavy rainfall events (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 of 8,096 samples (1998–2008), a validation set containing 2208 samples (2009–2011) to optimize the hyperparameters of the DNN model, and a test set with an equal number of samples 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 S6 in Supporting Information S1, and the period after 2014 was therefore excluded.

2.3. Definition of Heavy Rainfall Events

We define heavy rainfall 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 cell in terms of percentiles. The percentiles are computed from the entire TRMM time series from 1998 to 2014 of 3-hourly time steps with rainfall amounts above 0.1 [mm/3h]. This allows to determined the event thresholds in the most accurate way by leveraging all the available data, which is important for the heavy rainfall events considered in this study.

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. The U-Net includes 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. 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 and spatial dimensions identical to the global rainfall grid. It 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 found to be optimal for all the models in this study. This effectively reduces the model parameters size compared to the original U-Net. Adding more layers did not lead to improvements, as similarly found in (Grönquist et al., 2021; Weyn et al., 2020). The ADAM optimizer (Kingma & Ba, 2017) was employed for training the networks. We use a batch size of 64, an initial learning rate of 10^{-4}, and early stopping with a patience of 20 epochs without improvement of the loss function on the validation data set to prevent overfitting. The learning

rate is reduced during training using a scheduler. It decreases by a factor of 0.1 after a period of 10 epochs without improvement on the validation loss.

2.5. Loss Function

To improve the training regarding high values and 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) (named WMSE-MS-SSIM in the following), introducing an additional hyperparameter λ. 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 all the network hyperparameters on the validation set using random search with uniform distributions for each loss function. The intervals of the parameter distribution were adapted during the optimization procedure. For the loss in Equation 1, we find through manual evaluation $\alpha = 0.007$, $\beta = 0.048$ and $\lambda = 0.158$ to be optimal with respect to continuous metrics such as root mean square error (RMSE) and mean error (ME) as well as categorical skill scores such as F1 and CSI. 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 emphasise less frequent and high values when training a DNN to estimate radar composite reflectivity from satellite imagery. While the weighted MSE accounts for the skewed rainfall frequency distribution, the MS-SSIM evaluates the mean, standard deviation and covariance in the predicted rainfall output and ground truth. This is done through an iterative comparison of luminance, contrast and structure on different scales by downsampling and low-pass filtering the image signals (see supporting information). It is highly sensitive to blurring in images, as opposed to the MSE loss term. This can be seen for example, in Figure 2 in (Wang et al., 2004), showing a comparison of the sensitivity of MSE and MS-SSIM for different image distortions. Intuitively, one might hope that including the MS-SSIM will improve the spatial patterns of the DNN output, which is important for an accurate reproduction of heavy rainfall events. In our case, we indeed 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, with the role to improve the structural similarity. Further introducing bounds on the weights was crucial for a robust optimization of the network.

2.6. Baseline

We compare our method to two different baselines. A linear ridge regression (Hoerl & Kennard, 1970) with the IFS ensemble mean rainfall of a single grid-cell as input is used as the first baseline model. The regularization constant of 10^{-3} was found to be optimal using the same validation method as for the DNNs. Including the vertical velocity fields did not improve the performance of this baseline model. In addition, we use QM (Déqué, 2007) as a second baseline. The period from 1998 to 2011 is used to estimate the cumulative distribution functions (CDFs) of the simulated F_{hist} and observed F_{obs} data with 750 discretized quantiles, which are found to be optimal. The CDFs are then used to match the corresponding quantiles via

$$\tilde{p}_{sim} = F_{obs}^{-1}\left(F_{hist}(p_{sim})\right). \tag{3}$$

Here, $\tilde{p}_{sim}$ and p_{sim} are the quantile-mapping corrected and simulated rainfall values, respectively.

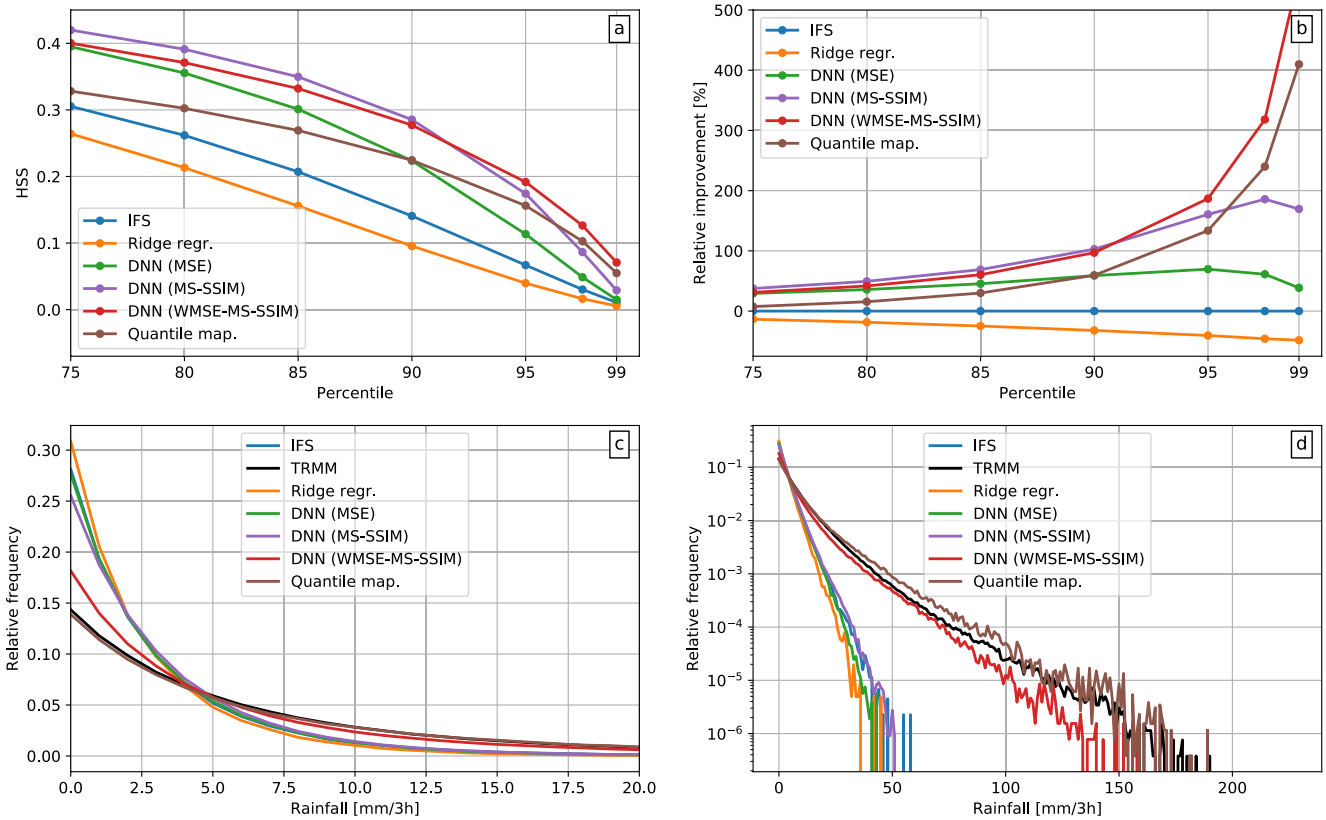

Figure 2. Relative rainfall frequencies and categorical heavy rainfall event forecast scores for the different post-processing models compared to the IFS. (a) The Heidke Skill Score (HSS) for events above increasing percentile thresholds is shown for the IFS (blue), ridge regression (orange), DNN trained with the mean squared error loss (green), the MS-SSIM loss (purple), with the WMSE-MS-SSIM loss proposed here (red) and quantile mapping (brown). A HSS greater than zero implies an improvement over a random forecast, and HSS = 1 would imply a perfect forecast (see supporting information). (b) The relative improvement of the HSS for the different machine learning methods over the IFS mean, is shown in percentages. Histograms of three-hourly rainfall event magnitudes are shown on a linear *y*-axis (c) and a logarithmic *y*-axis (d) for Tropical Rainfall Measurement Mission (black), IFS (blue), ridge regression (orange), DNN trained with the MSE loss (green), the MS-SSIM loss (purple) and the WMSE-MS-SSIM loss (red). The bins were chosen to be evenly spaced with a width of 1 mm/day.

3. Results

3.1. Evaluation of the Continuous Forecast Skill of the Deep Learning Model

The evaluation results reported in the following are computed on the test data set. 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 (Figures 2a and 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 WMSE-MS-SSIM loss function in Equation 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 heavy rainfall events in the tail. The ridge regression shows the largest bias toward low rainfall rates, hence not improving the IFS output at all. Applying QM to the IFS output on the other hand leads to an accurate representation of rainfall frequencies over the entire range of values - also for low values, as expected by construction.

We assess the continuous forecast skill of the different models by computing the RMSE, 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

Table 1
Continuous Validation Statistics Are Given for the Integrated Forecast System Ensemble Mean, Quantile Mapping, 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	-
Qantile map.	-	P	2.071	−42.1	0.149	14.9	0.511	42.3
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	WMSE-MS-SSIM	P, W	1.439	1.2	**0.135**	**22.9**	**0.545**	**51.8**

a threshold of 0.1 [mm/3h] have been excluded before applying the error metrics. Rainfall on such low scales cannot be measured accurately by satellite-based remote sensing (Huffman et al., 2007). Hence, completely dry times are not represented in the error statistics. 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 WMSE-MS-SSIM 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 substantially affecting the performance of the model. The WMSE-MS-SSIM loss function combined with 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%. Besides the metrics discussed above, rainfall maps produced by the IFS, DNN and TRMM are shown in Figure S1 in Supporting Information S1 for a qualitative comparison. While QM is not able to reduce the RMSE of the IFS, it strongly reduces the ME and leads to high similarity values, reflected in the CW-SSIM.

3.2. Evaluation of the Forecast Skill of the Deep Learning Model for Heavy Rainfall Events

To evaluate the forecast skill for heavy rainfall 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 in Supporting Information S1). A detailed definition of the events is given in Section 2.3 and the skill scores are defined in the Supporting Information. Table 2 summarizes the skill scores for events above the 95th percentile. The HSS, defined in the SI (Text S2 in Supporting Information S1), 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 Figures S2 to S5 in Supporting Information S1. 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). Quantile mapping results in HSS, F1 and CSI values higher than for the MSE-trained DNN, but stays below the other two networks. While QM leads to a high probability of detection, it also shows a large FAR score indicating a high number of false positives. The DNN trained using the

Table 2
Event-Based Forecast Skill Scores for Rainfall Events Above the 95th Percentile

Model	Loss	HSS	%	F1	%	CSI	%	POD	%	FAR	%
IFS	-	0.067	-	0.069	-	0.036	-	0.041	-	0.778	-
Quantile map.	-	0.156	133	0.161	135	0.088	144.4	0.163	299	0.840	−8
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	WMSE-MS-SSIM	**0.192**	**187**	**0.195**	**183**	**0.108**	**200**	**0.139**	**239**	0.673	13

Note. The percentage columns give the relative improvement over the IFS mean for each error metric and skill score.

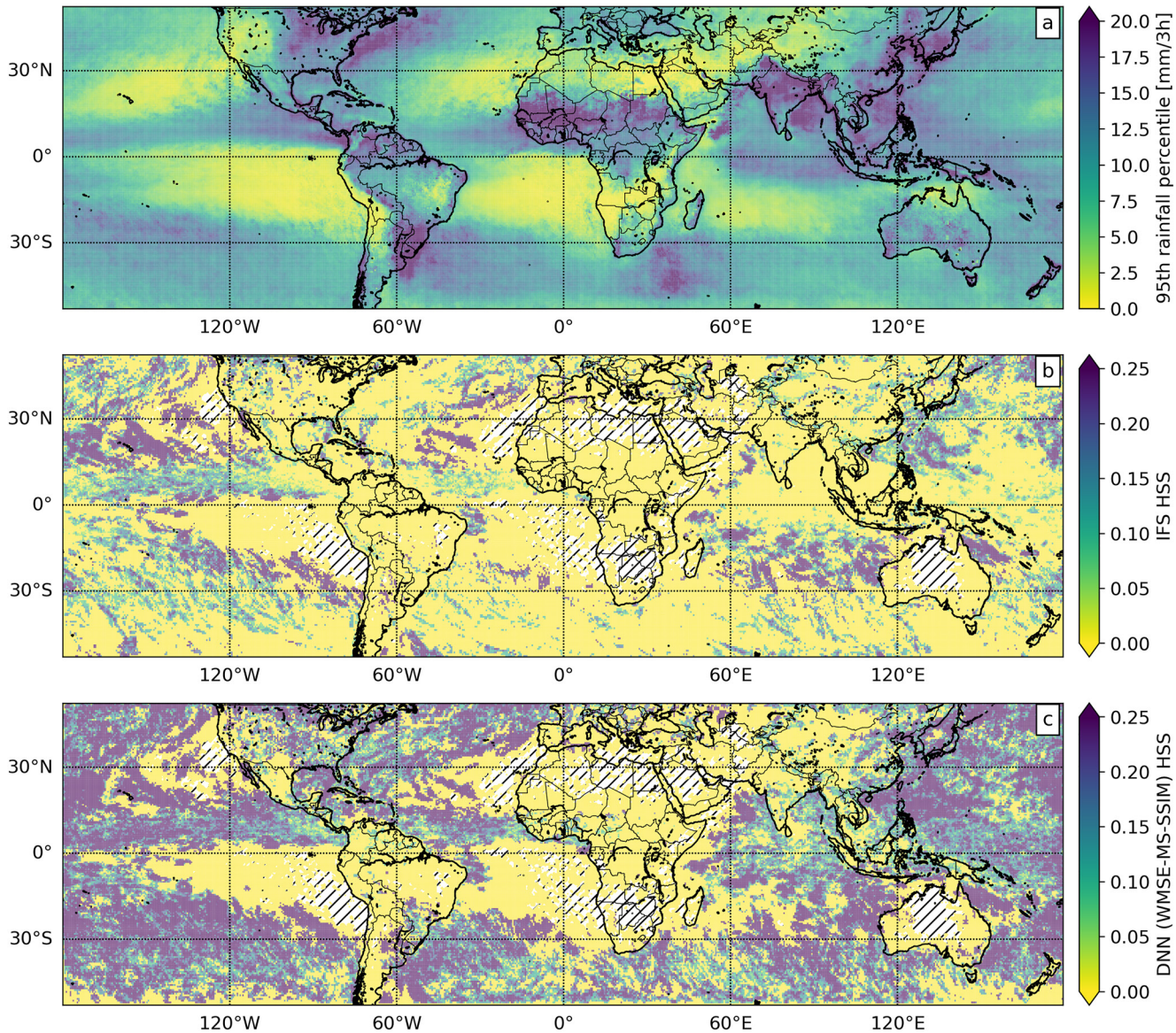

Figure 3. Spatial distribution of the 95th rainfall percentile and Heidke Skill Score (HSS) for events above the 95th percentile. (a) The 95th percentile of the rainfall distribution at each grid cell of the Tropical Rainfall Measurement Mission data set. (b) The spatially resolved HSS for the IFS mean. (c) The spatially resolved HSS for the deep neural network post-processed forecast, trained with the proposed WMSE-MS-SSIM loss. Hatched areas indicate grid-cells where the HSS could not be evaluated. This is due to the low number of wet times in these locations, so that the percentile thresholds could not be determined.

MS-SSIM alone as loss shows the highest scores below the 95th threshold. The proposed WMSE-MS-SSIM loss leads to significant improvements even above the 95th percentile (improving the IFS forecast by 192% in terms of the HSS) and yields the most skillfull 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 WMSE-MS-SSIM loss. We attribute this to the highly localized, intermittent nature of heavy rainfall events 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 WMSE-MS-SSIM loss introduced above leads to substantial improvements also for the spatial patterns of heavy rainfall events. In particular for regions with stronger heavy rainfall events (Figure 3) the skill improvement increases. This is also visible in Figure 4 showing longitudinal averages of the 95th rainfall percentile and the HSS scores of the IFS and the DNNs trained with MSE and the WMSE-MS-SSIM loss function. There remain regions, however, where the HSS is not substantially improved. These are

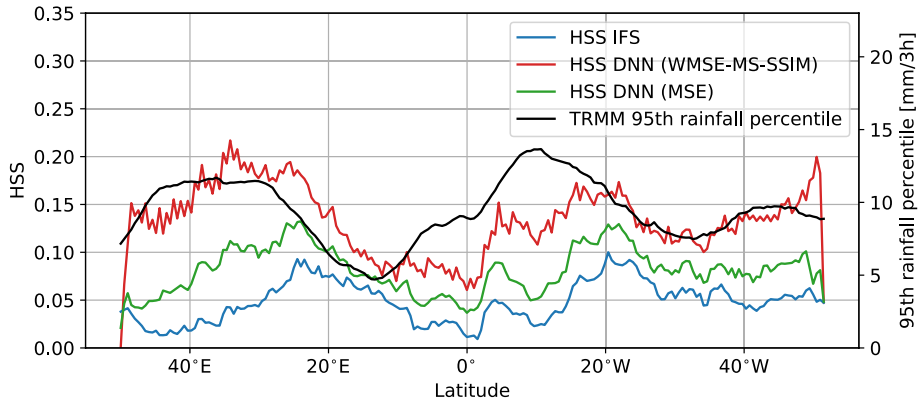

Figure 4. Zonal mean of the Heidke Skill Score for events above the 95th percentile for the Integrated Forecast System mean (blue), the deep neural network (DNNs) trained with the mean squared error (MSE) (green) and WMSE-MS-SSIM loss (red). The zonal mean of the Tropical Rainfall Measurement Mission (TRMM) 95th rainfall percentiles are shown in black. *Note.* That the averaged HSS of the DNN (WMSE-MS-SSIM) is approximately proportional to the 95th rainfall percentiles of TRMM.

mainly given by areas for which the IFS itself already has particularly low forecast skill (Figure S7 in Supporting Information S1), including a large fraction of the land masses between 30°S and 30°N. The rainfall frequencies are still improved in this region (Figure S8 in Supporting Information S1), although less strongly than over the entire global domain.

4. Discussion

We introduced a DNN to model heavy rainfall 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 (WMSE-MS-SSIM). The proposed WMSE-MS-SSIM loss function is found to substantially improve the training with respect to high values 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-Net. Moreover, our results suggest that the U-Net architecture is indeed capable of capturing the multi-scale spatial structure of rainfall accurately.

The WMSE-MS-SSIM loss substantially improves relative rainfall frequencies in the DNN output, the ME and CW-SSIM of overall rainfall fields, as well as categorical skill scores for heavy rainfall events above the 90th and higher percentile, with strongly increasing relative rate of improvement for higher thresholds. As seen in Figure 3 and Figure 4 the skill improvement follows largely regions with higher rainfall percentiles. A possible explanation could be that the WMSE-MS-SSIM loss is particularly successful at locations with high rainfall values. This is supported be the seemingly lower correlation of the MSE-trained DNN's HSS with the rainfall percentiles as shown in Figure 4. In regions, where the IFS predictions are not much better than a random forecast (Heidke skill score close to 0), our DNN-based post-processing that uses these IFS prediction as input is not be able to improve these IFS forecasts significantly. A direction for future research could thus be the improvement of our method in tropical regions where the skill is lower than in higher latitudes.

As noted by several authors, a single metric that captures all characteristics of a forecast does not exist, which renders the evaluation of purely data-driven weather forecasts particularly challenging (Ebert-Uphoff & Hilburn, 2020; Rasp & Thuerey, 2021; Ravuri et al., 2021). In particular, physical consistency, that is often assumed in established metrics, is not always guaranteed in neural network-based predictions. In this study, the DNN performance was manually evaluated using several metrics, both continuous and categorical. We believe that the development of more suitable and comprehensive evaluation metrics, or combinations thereof, will be an important direction of future research. It would enable a fully automatic hyperparameter tuning with respect to the various forecast qualities, which is, however, outside the scope of this study.

Taking the mean of the IFS ensemble is expected to damp the high rainfall values in the forecast. Hence, the results of the IFS shown here do not represent the skill of single ensemble members to forecast heavy rainfall events, which do not show this damping. The comparison of the bias correction methods presented in this study to the IFS ensemble mean therefore aim to show the respective relative improvements. Nevertheless, our results demonstrate the ability of the proposed DNN architecture to learn high rainfall values that are not produced by the existing precipitation parameterization of the IFS model, and to substantially improve their prediction.

The satellite-based TRMM rainfall data is used in this study as a ground truth. However, since different observational rainfall datasets usually agree only on much larger spatial and time scales than considered in this study, this should not be taken literally. The high resolution chosen for this study is important to capture the intermittent variability of heavy rainfall events. Since our method can be retrained in a flexible manner, it is possible to re-calibrate it to other observational data set as well. This allows for a continuous update of the DNN once more accurate observational datasets become available.

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 high values from the vertical velocity alone. This is also in agreement with previous works (Müller et al., 2020; O'Gorman & Schneider, 2009) on the link of the vertical velocity to heavy rainfall events.

An improved structural similarity in terms of the CW-SSIM is achieved when using the WMSE-MS-SSIM loss, compared to using the MS-SSIM alone as loss function. Adding the weighted MSE to the MS-SSIM loss might not be expected to increase the overall structural similarity of the DNN output. We speculate that the increased structural similarity we found might be related to the DNNs ability to improve the standard deviation that is measured in the CW-SSIM and to perform a transformation of the rainfall distribution similar to the QM method. Both our DNN and QM show accurate frequency distributions as well as relatively high structural similarity compared to the other models. Nevertheless, when trained using the WMSE-MS-SSIM loss, the forecast skill of our DNN-based post processing outperforms all other methods including QM, for almost all continuous and event-based validation metrics (see Tables 1 and 2).

A qualitative comparison of the DNN output with the TRMM target (Figure S1 in Supporting Information S1) shows that there remain small-scale features that are not captured by our method. This lack of high-frequency details in the output can be attributed to the deterministic nature of our neural network model. Here, generative models that learn stochastic functions and are therefore able to produce realistic small-scale features might offer a direction toward further improvements. However, producing stochastic small-scale features does not necessarily lead to a better forecast skill of the model, in particular for high rainfall events (Ravuri et al., 2021). We therefore believe that the results presented here could also be relevant for such stochastic approaches.

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.

With ongoing global warming, the characteristics of heavy rainfall events are expected to change. To account for this non-stationarity, the training of the proposed method can in principle be continued over time when new training data becomes available. Further, making use of the entire IFS ensemble will allow to incorporate uncertainties into the framework, which are essential for operational forecasting of heavy rainfall events.

Data Availability Statement

Data pre-processing was done using the Climate Data Operator (CDO) software (Schulzweida, 2019) for regridding as well as the Xarray 0.15.1 package. The Pytroch 1.7.0 (Paszke et al., 2019) source code for training and data processing will be available at (https://zenodo.org/badge/latestdoi/457716105) on publication. The QM method was implemented using the Xclim 0.25.0 Python package (Logan et al., 2021). 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).

Acknowledgments
The authors would like to thank the three anonymous reviewers for their helpful comments and suggestions. 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. Open access funding enabled and organized by Projekt DEAL.

References

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

Ali, H., Fowler, H. J., & Mishra, V. (2018). Global observational evidence of strong linkage between dew point temperature and precipitation extremes. *Geophysical Research Letters*, *45*(22), 12–320. https://doi.org/10.1029/2018gl080557

Allan, R. P., & Soden, B. J. (2008). Atmospheric warming and the amplification of precipitation extremes. *Science*, *321*(5895), 1481–1484. https://doi.org/10.1126/science.1160787

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., et al. (2019). Mswep v2 global 3-hourly 0.1 precipitation: Methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500. https://doi.org/10.1175/bams-d-17-0138.1

Berg, P., Feldmann, H., & Panitz, H.-J. (2012). Bias correction of high resolution regional climate model data. *Journal of Hydrology*, *448*, 80–92. https://doi.org/10.1016/j.jhydrol.2012.04.026

Bocquet, M., Farchi, A., & Malartic, Q. (2020). Online learning of both state and dynamics using ensemble kalman filters. *Foundations of Data Science*, *3*(3), 305. https://doi.org/10.3934/fods.2020015

Boers, N., Bookhagen, B., Marengo, J., Marwan, N., vonStorch, 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. https://doi.org/10.1175/jcli-d-14-00340.1

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*, D23113. https://doi.org/10.1029/2010jd014262

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. https://doi.org/10.1029/2019ms001711

Cannon, A. J., Sobie, S. R., & Murdock, T. Q. (2015). Bias correction of gcm precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? *Journal of Climate*, *28*(17), 6938–6959. https://doi.org/10.1175/jcli-d-14-00754.1

Déqué, M. (2007). Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, *57*(1–2), 16–26. https://doi.org/10.1016/j.gloplacha.2006.11.030

Donat, M., Alexander, L., Yang, H., Durre, I., Vose, R., Dunn, R., et al. (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. https://doi.org/10.1002/jgrd.50150

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. https://doi.org/10.1175/bams-d-20-0097.1

European Centre for Medium-Range Weather Forecasts. (2012). *The ECMWF ensemble prediction system*. Retrieved from 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. https://doi.org/10.1038/nclimate3110

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. https://doi.org/10.3390/atmos11030267

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. https://doi.org/10.1098/rsta.2020.0092

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., & Engen-Skaugen, T. (2012). Downscaling rcm precipitation to the station scale using statistical transformations–a comparison of methods. *Hydrology and Earth System Sciences*, *16*(9), 3383–3390. https://doi.org/10.5194/hess-16-3383-2012

Guerreiro, S. B., Fowler, H. J., Barbero, R., Westra, S., Lenderink, G., Blenkinsop, S., et al. (2018). Detection of continental-scale intensification of hourly rainfall extremes. *Nature Climate Change*, *8*(9), 803–807. https://doi.org/10.1038/s41558-018-0245-3

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*, 1999–2049.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., et al. (2007). The trmm multisatellite precipitation analysis (tmpa): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, *8*(1), 38–55. https://doi.org/10.1175/jhm560.1

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

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

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

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

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

Logan, T., Bourgault, P., Smith, T. J., Huard, D., Biner, S., Labonté, M.-P., et al. (2021). Ouranosinc/xclim: V0.31.0. Zenodo. [Data set]. https://doi.org/10.5281/zenodo.5649661

Maraun, D. (2016). Bias correcting climate change simulations-a critical review. *Current Climate Change Reports*, *2*(4), 211–220. https://doi.org/10.1007/s40641-016-0050-x

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., et al. (2021). *IPCC, 2021: Climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. In Press.

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. https://doi.org/10.1080/16000870.2020.1811535

O'Gorman, P. A., & Schneider, T. (2009). The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *Proceedings of the National Academy of Sciences*, *106*(35), 14773–14777. https://doi.org/10.1073/pnas.0907610106

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (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* (Vol. 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. https://doi.org/10.1038/nclimate3287

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002203. https://doi.org/10.1029/2020ms002203

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/mwr-d-18-0187.1

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

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. https://doi.org/10.1029/2020ms002405

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skillful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. https://doi.org/10.1038/s41586-021-03854-z

Ronneberger, O., Fischer, P., & Brox, T. (2015). In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.U-net: Convolutional networks for biomedical image segmentation.

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. https://doi.org/10.1109/tip.2009.2025923

Schulzweida, U. (2019). Cdo user guide. (Version 1.9. 6). *Max Planck Institute for Meteorology*, *53*, 20146. https://doi.org/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. *Advances in Neural Information Processing Systems*, *30*.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., et al. (2020). *Metnet: A neural weather model for precipitation forecasting*. arXiv preprint arXiv:2003.12140.

Tong, Y., Gao, X., Han, Z., Xu, Y., Xu, Y., & Giorgi, F. (2021). Bias correction of temperature and precipitation over China for rcm simulations using the qm and qdm methods. *Climate Dynamics*, *57*(5), 1425–1443. https://doi.org/10.1007/s00382-020-05447-4

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. https://doi.org/10.3390/atmos10050244

Traxl, D., Boers, N., Rheinwalt, A., & Bookhagen, B. (2021). The role of cyclonic activity in tropical temperature-rainfall scaling. *Nature Communications*, *12*(1), 1–9. https://doi.org/10.1038/s41467-021-27111-z

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/tip.2003.819861

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

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J. J., et al. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555. https://doi.org/10.1029/2021GL092555

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. https://doi.org/10.1029/2020ms002109

Wilcox, E. M., & Donner, L. J. (2007). The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model. *Journal of Climate*, *20*(1), 53–69. https://doi.org/10.1175/jcli3987.1

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences* (Vol. 100). Academic press.

Xu, C.-y. (1999). From gcms to river flow: A review of downscaling methods and hydrologic modelling approaches. *Progress in Physical Geography*, *23*(2), 229–249. https://doi.org/10.1177/030913339902300204

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-020-17142-3