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).

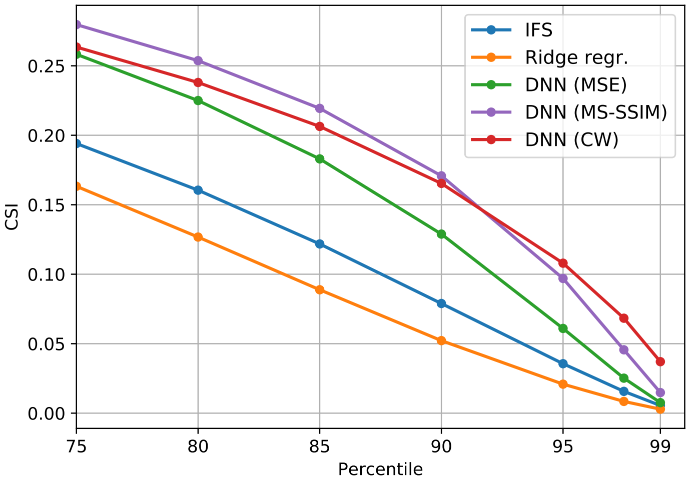

Figure S1. The critical success index (CSI) for rainfall events above the 75th percentile threshold.

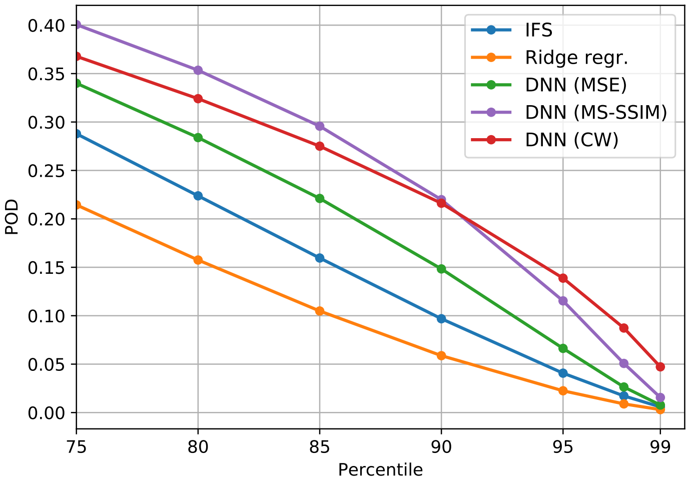

Figure S2. The probability of detection (POD) of rainfall events above the 75th percentile threshold.

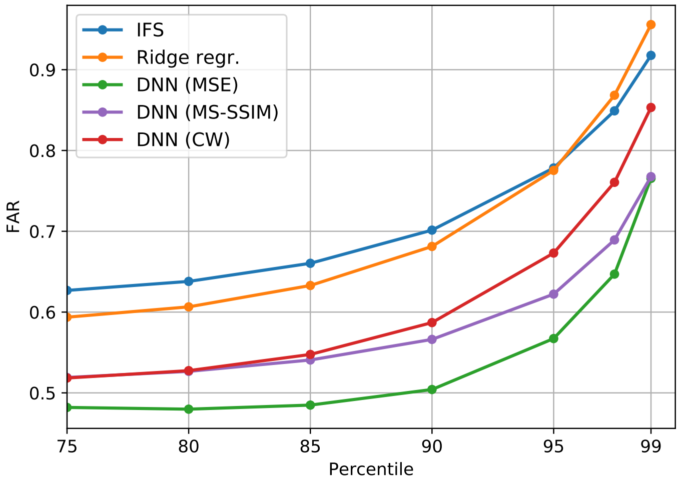

Figure S3. The false alarm ratio (FAR) of rainfall events above the 75th percentile threshold.

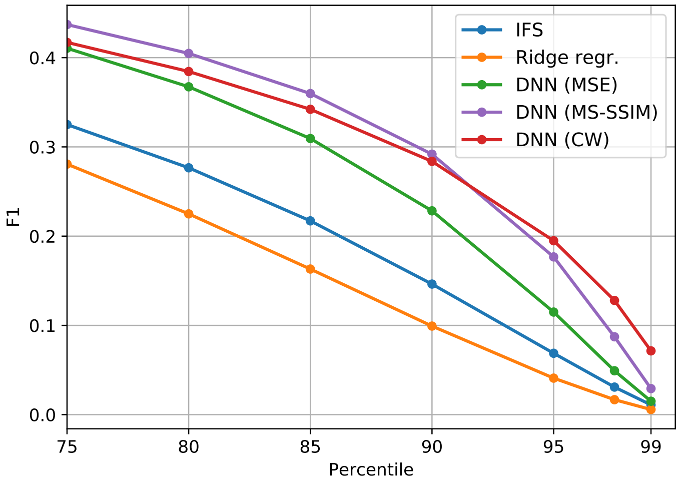

Figure S4. The F1 score for rainfall events above the 75th percentile threshold.

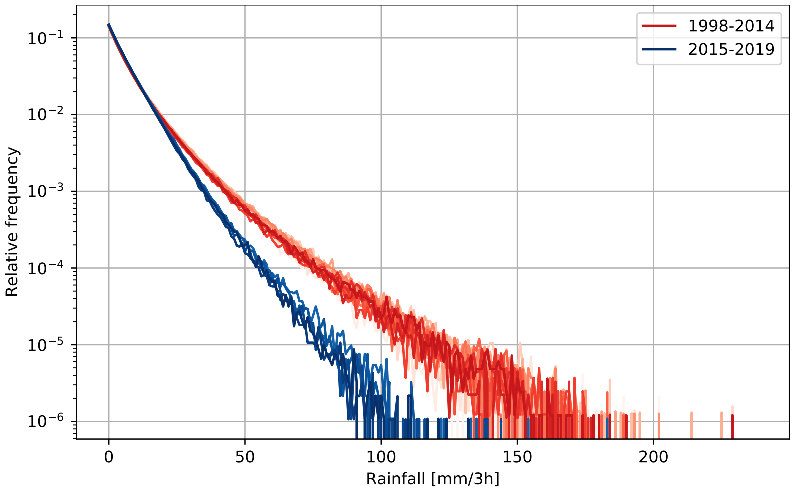

Figure S5. The histograms of grid-cell values show here are computed over the entire part of the globe covered by the TRMM data (50°S to 50°N) and for single years. The histograms of years before 2015 are colored in red and for years thereafter in blue.

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)