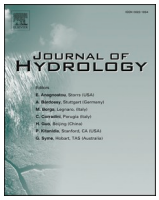
Journal of Hydrology 665 (2026) 134689

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

Machine unlearning: bias correction in neural network downscaled storms

Simon Michael Papalexiou [a,*], Antonios Mamalakis [b,c]

[a] Institute of Global Water Security, Hamburg University of Technology, Hamburg, Germany
[b] Department of Environmental Sciences, University of Virginia, Charlottesville, VA, USA
[c] School of Data Science, University of Virginia, Charlottesville, VA, USA

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief

Keywords:
Precipitation downscaling
Machine learning
Spatiotemporal dependencies
Bias Correction
Wasserstein GAN (WGAN)
Statistical properties

ABSTRACT

Accurate precipitation at fine spatial resolutions is essential for hydrologic modeling and risk assessment, yet most precipitation data are available at coarse scales. Dynamical downscaling can improve spatial resolution but is computationally expensive while statistical downscaling struggles to reproduce high-resolution characteristics. Machine learning has been shown to offer an operational alternative for transforming coarse data into fine-scale fields, especially honoring spatiotemporal precipitation dependencies. Here, we show that combining machine learning with post-processing bias correction approaches—a form of "machine unlearning"— yields improved performance. We evaluate four machine-learning models—Linear Network (LNet), Fully Connected Network (FCNet), Convolutional Neural Network (UNet), and Wasserstein Generative Adversarial Network (WGAN)—for downscaling precipitation using synthetic benchmark storms with known marginal and spatiotemporal properties. Using synthetic fields provides full control over storm characteristics and enables rigorous evaluation. Raw outputs from all models struggle to reproduce wet/dry boundaries, statistics, and extremes, and only WGAN captures the complex spatiotemporal structure of fine-scale storms. We then apply linear and nonlinear bias corrections to enforce zeros, match the mean of positive values, and align full marginal distributions, including tails. This demonstrates that post-processing is crucial for reliable neural network outputs in operational settings. The results highlight WGAN's potential for operational downscaling while emphasizing the need for systematic post-processing and careful validation before real-world application.

1. Introduction

"*Machines take me by surprise with great frequency*." ~ Alan Turing

Spatial downscaling of precipitation aims to generate high-resolution datasets from coarse-scale information. For example, climate models like those in CMIP6 (Coupled Model Intercomparison Project, Phase 6; Eyring et al., 2016) provide precipitation projections at resolutions ranging from 0.1° to over 1°. Such resolutions are inadequate for regional applications requiring finer scales, typically below 4 km (Benestad, 2004; Lucas-Picher et al., 2021; Mamalakis et al., 2017; Rampal et al., 2024). Downscaling is thus critical for applications including evaluating extreme event risks, assessing vulnerabilities in various sectors, formulating adaptation strategies, facilitating governance-level decision-making, and supplying input data for specialized models (Benestad, 2010; Liu et al., 2023).

Spatial downscaling techniques are broadly categorized into statistical and dynamical methods (Rampal et al., 2024). Statistical downscaling uses relationships between large-scale atmospheric variables and observed fine-scale precipitation. This includes regression-based models, weather typing, and analog methods (Maraun et al., 2010; Themeßl et al., 2012; Wilby et al., 1998). A subset is super-resolution downscaling, transforming coarse precipitation fields to high-resolution counterparts without other predictors (Duncan et al., 2022; Leinonen et al., 2021; Mamalakis et al., 2017). Dynamical downscaling employs high-resolution Regional Climate Models (RCMs), like the Weather Research and Forecasting (WRF) model (Powers et al., 2017), nested within global models to simulate fine-scale atmospheric processes. Despite progress, challenges persist in both approaches. Dynamical downscaling is computationally expensive, making it impractical for applications requiring multiple runs, such as uncertainty quantification essential for reliable climate change impact assessment. Statistical downscaling, though less resource-intensive, often struggles to capture the high-resolution dependence structures of precipitation fields and has limited extrapolation capabilities (Maraun et al., 2015; Teutschbein & Seibert, 2013).

* Corresponding author.
E-mail address: simon.papalexiou@tuhh.de (S.M. Papalexiou).

https://doi.org/10.1016/j.jhydrol.2025.134689
Received 27 January 2025; Received in revised form 22 November 2025; Accepted 26 November 2025
Available online 28 November 2025
0022-1694/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Recently, machine learning (ML) and deep learning (DL) techniques have been explored to address these challenges (Boulaguiem et al., 2022; Chen et al., 2024; González-Abad et al., 2023; Harris et al., 2022; Leinonen et al., 2021; Nishant et al., 2023; Rampal et al., 2022). ML applications include support vector machines, random forests, and neural networks for both dynamical and statistical downscaling. In dynamical contexts, ML models emulate physics-based models, offering computational efficiency once trained (Hobeichi et al., 2023). In statistical downscaling, DL is promising due to its ability to capture nonlinear system behavior. Generative Adversarial Networks (GANs) and conditional GANs (Boulaguiem et al., 2022; Glawion et al., 2023b; Leinonen et al., 2021), in particular, have shown potential in capturing spatial structures of precipitation fields, although challenges like representing extremes remain (Maraun et al., 2015).

Here, we revisit the problem of downscaling precipitation fields, and specifically storms at fine spatiotemporal resolution, using machine and deep learning models. We conduct a comprehensive comparison of various neural network models tasked with downscaling fast-moving storm fields from a coarse 6×6 grid to a fine 60×60 grid—a tenfold downscaling representative of real-world needs (Rampal et al., 2024). We note that as a first step, we here use synthetic storms, so that we have complete control of the experiments (type of storms; velocity fields etc.) and sufficient sample size for training. Specifically, we use stochastically simulated benchmark storms generated by the Complete Stochastic Modeling System (CoSMoS; Papalexiou, 2018) and test the efficacy of downscaling models. For our assessment, we adopt several and thorough performance metrics, specifically relevant for precipitation. In consistency with previous research (see references above), our results show that GANs outperform other models in capturing spatiotemporal dependencies, yet they suffer with the reproduction of other key statistical properties of precipitation, including extremes. We then show that by combining machine learning and post-training bias correction approaches — that can be viewed as a form of "machine unlearning" (Bourtoule et al., 2021; Li et al., 2024)—optimal performance can be attained, a particularly important result for operational applications.

We stress that we use the term machine unlearning in a broader, application-oriented sense to describe the post-processing correction of biases learned by neural networks, rather than the modification of network weights. In the broader ML literature, machine unlearning (MU) typically refers to methods designed to remove or alter previously learned information from trained models (e.g., for privacy, security, or fairness purposes; see e.g., Shaik et al., 2025). Our use of the term highlights that what has been learned by the network may still be biased, requiring systematic "unlearning" through output-level correction.

In the following sections, we start by presenting the algorithm that was used to generate the synthetic storm data and the models that were used for downscaling. In sections 4 and 5, we present the results and the post-training bias correction approaches that we recommend. We discuss our results in section 6, and in section 7, we state our conclusions.

2. Benchmark storms generated using CoSMoS

Here we aim to evaluate the performance of neural networks (NNs) in downscaling storms, or more generally, precipitation fields. To thoroughly investigate this objective, a fully controlled experiment is necessary for accurate assessment. We require fine spatiotemporal resolution storm data, which we will then aggregate to coarser resolutions, using both fine and coarse fields to train the NNs. While radar fields or other observational spatiotemporal data could be utilized, they would not effectively serve the purpose of this study. This is due to potential data quality issues (e.g., Berne & Krajewski, 2013; Ochoa-Rodriguez et al., 2019; Villarini & Krajewski, 2010), and, most importantly, the difficulty in confidently assessing the statistical properties of storms without knowing whether they represent ground truth.

Specifically, storms at fine spatiotemporal scales have statistical properties that not only challenge simulation but also complicate their assessment and quantification (see e.g., Bacchi & Kottegoda, 1995; Houze, 2018; Moszkowicz, 2000; Niemi et al., 2014). At sub-hourly temporal scales and spatial resolutions, for example, below 10 km, storms exhibit: (1) Advection—storms move across regions with constant or varying velocities; (2) Anisotropy—there is often an axial preference, where storm properties vary depending on the direction of investigation in space; (3) Spatiotemporal intermittency—periods of no rainfall alternate between partially and fully wet regions, forming irregular boundaries between dry and wet areas; (4) Strong spatiotemporal dependence—there is a high correlation with both previous instances and distant locations within the storm; and (5) Skewed and potentially heavy-tailed marginal distributions—positive intensity values follow positively skewed probability distributions (typically J-shaped) with subexponential tails, allowing for more frequent and intense extremes.

Several models in the literature address space–time modeling of precipitation (e.g., Bardossy & Plate, 1992; Leblois & Creutin, 2013; Papalexiou, Serinaldi, & Porcu, 2021; Paschalis et al., 2013; Pegram & Clothier, 2001; Peleg et al., 2017). Here, to generate synthetic storms, we use the Complete Stochastic Modeling System (CoSMoS), which defines storm properties through its marginal distribution, spatiotemporal correlation, advection velocity, and anisotropy. CoSMoS (Papalexiou, 2018) is a flexible framework for simulating hydroclimatic processes like precipitation, temperature, and discharge, tailored to match observed or specified distributions and correlation structures. Its univariate version generates time series, while CoSMoS-2 s (Papalexiou, 2022) enhances precipitation modeling through copula-based dependencies and dedicated wet-dry spells simulation. The multisite version (Papalexiou et al., 2023) models time series across locations, capturing local and large-scale variability. Additionally, CoSMoS can generate static spatiotemporal fields (Papalexiou & Serinaldi, 2020) and simulates spatially varying velocity fields and anisotropy for realistic storm dynamics (Papalexiou, Serinaldi, & Porcu, 2021).

We generate synthetic spatiotemporal storms characterized by a highly skewed distribution, strong spatiotemporal correlation, a constant velocity field, and affine anisotropy. Specifically, the storms have a theoretical probability of zero, $p_0 = 0.7$, while the intensity of positive values follows the $\mathscr{GE}4$ distribution introduced in Papalexiou (2022) with distribution function:

$$F_{\mathscr{GE}4}(x;\beta,\gamma_1,\gamma_2) = 1 - \left(\left(\exp\left(\frac{x}{\beta}\right)^{\gamma_2} - 1 \right)^{\gamma_1/\gamma_2} + 1 \right)^{-\gamma_2/\gamma_1} \quad (1)$$

where $\beta > 0$ is a scale parameter and $\gamma_1 > 0$ and $\gamma_2 > 0$ are shape parameters. Although Equation (1) is not simple, the $\mathscr{GE}4$ distribution intuitively extends the exponential family (the Exponential distribution is a special case), introducing two shape parameters that control the left and right tails and allow flexible representation of both light and heavy rainfall. Here we use $(\beta,\gamma_1,\gamma_2) = (3, 0.8, 1.2)$ resulting in a highly skewed J-shaped probability density function (Fig. 1a), consistent with previous studies on precipitation (e.g., Cavanaugh et al., 2015; Marra et al., 2023; Moustakis et al., 2021; Nerantzaki & Papalexiou, 2019; Papalexiou et al., 2018; Serinaldi & Kilsby, 2014).

This modeling approach uses parent Gaussian random fields (GRFs) or fields with different types of dependence, such as t-copula fields. These parent fields are transformed to exhibit desired characteristics, like advection and anisotropy. These transformations alter the original spatiotemporal correlation structure (STCS) of the parent GRF, often making the resulting STCS more complex and difficult to visualize (for more on STCSs see e.g., Gneiting, 2002; Hristopulos, 2020; Porcu et al., 2020). In general, the generated storms are described by their marginal distribution, the STCS of the GRF, the advection velocity field, and anisotropy (for further details, see Papalexiou, Serinaldi, & Porcu, 2021). Here we use the Ali-Mikhail-Haq-Weibull (AMHW) STCS, defined as:

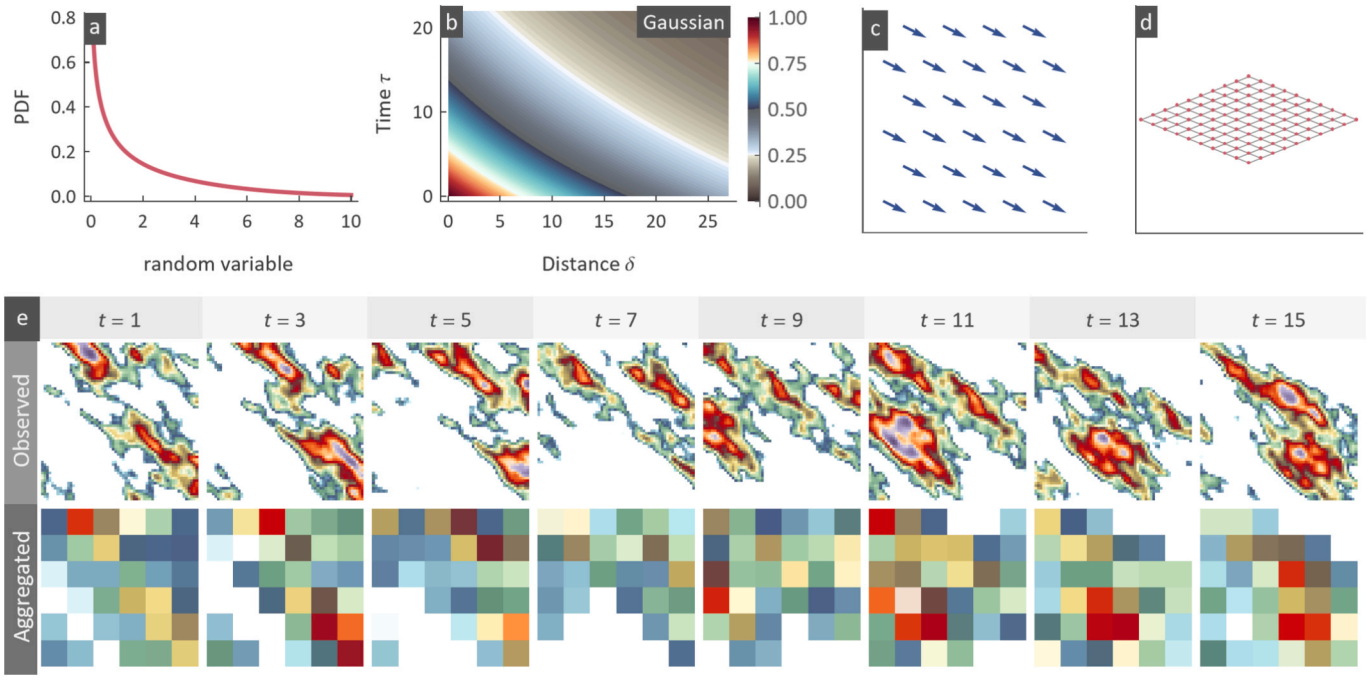


Fig. 1. (a) Probability density function (PDF) of the $\mathscr{GE}4$ distribution for precipitation amounts. (b) Ali-Mikhail-Haq-Weibull spatiotemporal correlation structure for the parent Gaussian random fields. (c) Velocity field for storm advection. (d) Transformed grid showing the affine anisotropy transformation (stretch, compress, rotate). (e) Generated storms used to benchmark NN downscaling performance, with corresponding aggregated storms at a coarser resolution for NN training. See also Animation S1.

$$\rho_{\mathrm{AMHW}}(\delta,\tau)=\frac{\exp(-(\delta/b_{\mathrm{S}})^{c_{\mathrm{S}}})\exp(-(\tau/b_{\mathrm{T}})^{c_{\mathrm{T}}})}{1-\theta(1-\exp(-(\delta/b_{\mathrm{S}})^{c_{\mathrm{S}}}))(1-\exp(-(\tau/b_{\mathrm{T}})^{c_{\mathrm{T}}}))} \tag{2}$$

where b_{S} and b_{T} are scale parameters, while c_{S}, c_{T}, and θ are shape parameters. The parameters used here for the STCS (Fig. 1b) of the parent GRFs are the $(b_{\mathrm{S}}, c_{\mathrm{S}}, b_{\mathrm{T}}, c_{\mathrm{T}}, \theta) = (25, 1, 20, 1, -1)$. The storms in the model move with an advection velocity $\boldsymbol{v}_{x,y} = (v_x, v_y) = (6, -3)$ causing the field to move rightward and downward at an angle of -26.6° (Fig. 1c) relative to the positive x-axis, covering a distance of about 6.7 units per time step. To introduce anisotropy, we apply an affine transformation (e.g., Chilès & Delfiner, 2009):

$$\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix}=\begin{pmatrix}\kappa_x & 0\\ 0 & \kappa_y\end{pmatrix}\begin{pmatrix}\cos\omega & -\sin\omega\\ \sin\omega & \cos\omega\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix} \tag{3}$$

where $\kappa_x = 2.5$ and $\kappa_y = 1$ are scaling factors for horizontal and vertical directions, respectively, and $\omega = -\pi/4$ applies a counterclockwise rotation. This transformation induces anisotropy, as visualized by the distortion of a regular grid (Fig. 1d). The storm cells, elongated along the -45° line (Fig. 1e), reflect this anisotropic effect, which is also apparent in the field's movement (see Animation S1 in Supplementary Material). In simple terms, this transformation stretches the field horizontally, compresses it vertically, and then rotates it, producing storm structures that are elongated rather than circular—similar to how wind shear might distort a storm system.

In total, we generated 55,000 storm fields on a 60×60 grid, which were aggregated to a 6×6 grid (Fig. 1e). The 50,000 fine- and coarse-scale field pairs were used to train four NNs to downscale coarse fields to finer resolutions. The remaining 5,000 pairs were used to validate the models' performance. Note that, although the storm fields in this study were generated by the CoSMoS model, we refer to these fields hereafter as "observed" or "benchmark" in all figures and text. This avoids ambiguity with the NN-downscaled fields, which are also simulated, and reflects a real-world scenario where one would rely on observations. A schematic overview of the full workflow, including storm generation, neural network training, and bias correction, is provided in Fig. A1.

3. Neural networks used to downscale storms

3.1. Linear networks (LNet)

The simplest network we consider is the Linear Network (LNet), which is used here as a baseline model. The LNet consists of multiple linear regression models that are independent of each other (i.e., they do not share any parameters; Fig. 2a) and are trained at the same time by minimizing the loss function; see Section 3.5. Specifically, if $\hat{\boldsymbol{x}}(t)$ is the vectorized, downscaled high-resolution precipitation field at time t, then the LNet can be summarized as follows:

$$\hat{\boldsymbol{x}}_{3600\times1}(t) = \boldsymbol{b}_{3600\times1} + \boldsymbol{w}_{3600\times36}\,\boldsymbol{y}_{36\times1}(t) \tag{4}$$

with $\boldsymbol{y}(t)$ being the vectorized, coarse precipitation field at time t, and $\boldsymbol{b}$ and $\boldsymbol{w}$ being the parameters of the network to be estimated during training.

3.2. Fully connected networks (FCNet)

The second model we consider is a fully connected artificial neural network (FCNet; also known as dense network or multi-layer perceptron). It consists of input and output layers and several hidden layers (see Fig. 2b). The input and output layers are vectorized versions of the coarse and high-resolution precipitation fields, i.e., $\boldsymbol{y}(t)$ and $\hat{\boldsymbol{x}}(t)$, respectively. Each of the hidden layers consists of neurons where the outputs from all the neurons of the previous layer are used as inputs and passed through a nonlinear function, known as activation function. Specifically, if z_q^m is the output of the neuron q in the previous layer m, then the output of the neuron r in the next layer $m+1$ is:

$$z_r^{m+1} = H\left(b_r + \sum_q w_{qr} z_q^m\right) \tag{5}$$

where w_{qr} is the weight connecting the neurons $q \to r$ and b_r is known as the bias term. These weights and bias terms for the entire network

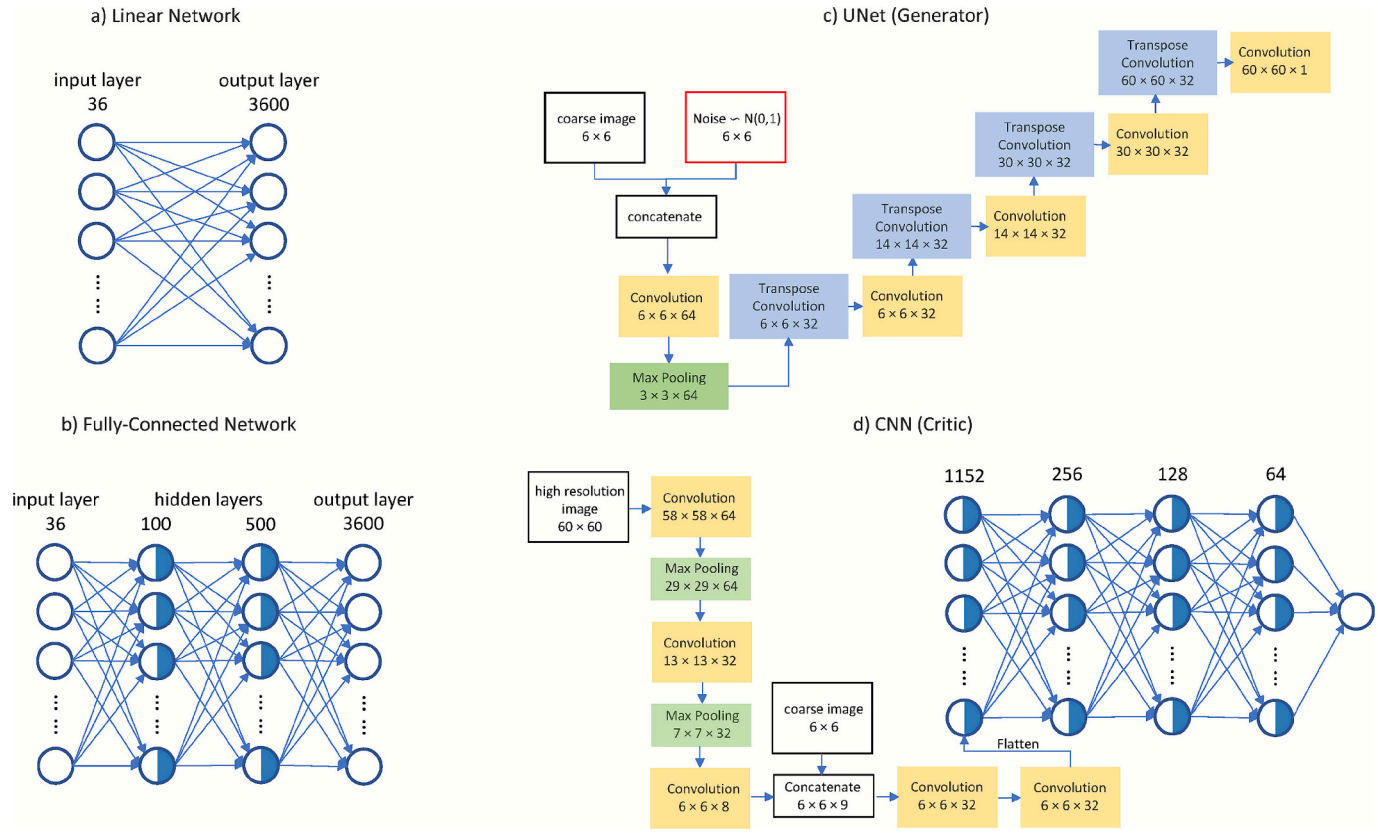


Fig. 2. Model architectures used in the study: (a) Linear Network (LNet), (b) Fully Connected Network (FCNet), (c) UNet – Generator (the generator component of the Wasserstein Generative Adversarial Network, WGAN), and (d) Convolutional Neural Network – Critic (CNN – Critic, the critic used in the adversarial training).

constitute the model parameters and are learned during training. The function H is a nonlinear function that is called activation function. There are many different choices that one could use for H, and although for specific applications some types might be more suitable than others, generally the choice is arbitrary, or it can be considered as a hyper-parameter. It is by using this nonlinear function H that fully connected networks can capture nonlinear system behavior. This is especially the case for deep networks with many hidden layers. After we experimented with a few different architectures, we chose to use an FCNet with two layers with 100 and 500 neurons, respectively. We chose to use the ReLU function as an activation function for the hidden layers and the "linear" function ($H_{\text{linear}}(u) = u$) for the output layer. The ReLU function is a popular choice, and its formula is given by:

$$H_{\text{ReLU}}(u) = \begin{cases} 0, & u \leqslant 0 \\ u, & u > 0 \end{cases} \tag{6}$$

Note that the LNet is effectively an FCNet with no hidden layers.

3.3. Convolutional network (UNet)

A Convolutional Neural Network (CNN) is a type of DL model designed to process data with a grid-like structure, such as images or spatiotemporal fields, making it especially suited to pattern recognition and downscaling tasks (LeCun et al., 2015). A CNN is composed of layers that apply convolutional operations to extract hierarchical features from input data. The initial layers may capture simple patterns, like edges and textures, while deeper layers recognize complex features and structures, such as shapes or objects. Each convolutional layer utilizes a set of filters (also known as "kernels") that slide over the input data, performing element-wise multiplications, summing the results and passing them through an activation function to produce feature maps. These element-wise multiplications are essentially similar to Equation (5), but w_{qr} and b_r are fixed for each filter. Typically, in each convolutional layer, many filters are used, to allow for different types of edges or shapes to be captured. Similarly to the previous models, the parameters w_{qr} and b_r of each filter of each layer are being estimated during training to maximize predictive performance.

Max pooling layers follow the convolutional layers to downsample the feature maps, by taking the maximum value from a 2×2 filter that slides over the entire feature map, reducing spatial dimensions and computational requirements while retaining critical information.

This hierarchical approach enables CNNs to capture spatial dependencies and complex, non-linear relationships within the data, making them highly effective for the downscaling task considered in this study. By adjusting the depth and number of filters, CNNs can be tailored to capture the intricate patterns within diverse data types. For the downscaling task, we aim to increase the resolution of the CNN input and we use a sequence of convolutional, pooling and reverse convolutional layers, as shown in Fig. 2c. The dimensions of the output feature map of each layer is given in Fig. 2c, and for each convolution or reverse convolution, we are using ReLU activation functions. Because of the shape of the architecture depicted in Fig. 2c, such a network is usually called UNet, due to its resemblance to the English letter "U".

3.4. Wasserstein generative adversarial networks (WGAN)

We finally consider a Generative Adversarial Network (GAN). A GAN is a type of deep learning model composed of two neural networks: a generator and a discriminator. The generator's job is to create new data instances that mimic real data, while the discriminator attempts to differentiate between real and generated data. The two networks engage in a "game" where the generator continuously improves its outputs to fool the discriminator, while the discriminator improves its ability to detect fake data. Over many iterations, this adversarial process helps the

generator produce highly realistic outputs, making GANs popular for applications like generative AI, image synthesis, style transfer, and data augmentation.

Wasserstein GANs (WGANs) are a refined version of GANs designed to address issues such as training instability and mode collapse (Arjovsky et al., 2017; Gulrajani et al., 2017), where the generator might produce limited or repetitive outputs. WGANs employ the Wasserstein distance (Arjovsky et al., 2017; Gulrajani et al., 2017) instead of the standard binary cross-entropy loss used in regular GANs. This distance metric provides a more stable measure of similarity between the generated data and real data distributions. By using a critic network instead of a traditional discriminator, WGANs enable smoother training, leading to more stable and diverse outputs. These improvements make WGANs especially useful in engineering applications that require high-quality data generation, like synthetic creation for rare events, simulating complex physical systems, and climate downscaling (Glawion et al., 2023b; González-Abad et al., 2023; Leinonen et al., 2021).

The CNN architecture of the critic network is shown in Fig. 2d. Apart from the convolutional and max pooling layers, we also use fully connected layers as depicted in the figure. In this study, we use a conditional WGAN, where the critic considers both the high-resolution image (real or generated) and the corresponding low-resolution image and assesses if these two could be a realistic pair, based on the training data (Glawion et al., 2023; Leinonen et al., 2021). We use the trained UNet of the previous section as the initial version of the generator, to be further trained and refined through the adversarial game. After the adversarial training is finished, the final UNet (generator) is used to downscale coarse precipitation fields. For the remainder of the study, we refer to the results of the generator as "WGAN". Note that during evaluation we fixed the latent vector z, making the generator mapping deterministic. Thus, all reported metrics are computed on a single realization per coarse field. Generating ensembles by resampling z is straightforward but was not pursued here to keep comparisons across models and post-processing steps strictly one-to-one.

3.5. Training process

To train the different networks, we used stochastic gradient descent (LeCun et al., 2015; Rumelhart et al., 1986) and a batch size of 32. The training data of 50,000 instances are split into training (40,000) and validation data (10,000). The validation data is used to determine the optimal values of the hyper-parameters (e.g., learning rate, drop-out probability) and the training data to estimate the parameters of each network. After the optimal hyper-parameters have been determined, each network is trained again from scratch, using 500 epochs and an early stopping algorithm, with the patience parameter being set to 30 epochs. The early stopping algorithm monitors the validation loss function and terminates the training if there has been no decrease in the validation loss for 30 consecutive epochs, to avoid overfitting.

The loss function that we used for the LNet, FCNet and UNet is the mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{N}\sum_{t=1}^{N}\frac{\|\boldsymbol{x}(t) - \widehat{\boldsymbol{x}}(t)\|_2^2}{3600} \tag{7}$$

where N is the batch size (i.e., in our case, $N = 32$), $\boldsymbol{x}(t)$ is the actual (ground truth) vectorized version of the high-resolution observed precipitation field at time t, and $\widehat{\boldsymbol{x}}(t)$ is the prediction by the network.

For the WGAN, the generator and the critic are trained at the same time, aiming to minimize the corresponding losses. The loss functions for the critic and the generator, respectively, are (here, we drop the notation for the time dimension for simplicity):

$$L_C = C(G(\boldsymbol{y},\boldsymbol{z}),\boldsymbol{y}) - C(\boldsymbol{x},\boldsymbol{y}) + \gamma\left(\|\nabla_{\tilde{\boldsymbol{x}}} C(\tilde{\boldsymbol{x}},\boldsymbol{y})\|_2 - 1\right)^2 \tag{8}$$

with $\tilde{\boldsymbol{x}} = (1-\varepsilon)\boldsymbol{x} + \varepsilon\, G(\boldsymbol{y},\boldsymbol{z})$

$$L_G = -C(G(\boldsymbol{y},\boldsymbol{z}),\boldsymbol{y}) \tag{9}$$

where C is the critic network, $\boldsymbol{x},\boldsymbol{y}$ are the high resolution and low resolution observed precipitation fields respectively, G is the generator network and $\boldsymbol{z}$ is the noise field. For the gradient penalty in the critic's loss (Gulrajani et al., 2017), we use $\gamma = 10$, and samples $\tilde{\boldsymbol{x}}$ are random weighted-averages between observed and generated high resolution precipitation fields. As such, ϵ is sampled from a uniform distribution between 0 and 1.

Conceptually, the Wasserstein loss functions aim for the critic to minimize its output for generated samples while maximizing it for real samples, effectively distinguishing between the two. The generator is trained to fool the critic. The gradient penalty in the critic's loss enforces constraints on the critic's output, limiting the likelihood of mode collapse and/or diverging solutions. For more details in training the WGAN, the reader is referred to Gulrajani et al., 2017; Leinonen et al., 2021). We train the WGAN on the training data (40,000 instances) for 20 epochs without applying early stopping, using a batch size of 32. In each training iteration, the critic's parameters are updated three times (using three batches), while the generator's parameters are updated once (typical practice; see Gulrajani et al., 2017; Leinonen et al., 2021).

Note that all model architectures, hyperparameters, optimizer settings, and regularization parameters used in this study are fully documented in the accompanying Python code, which is openly available at https://github.com/amamalak/downscaling_biascorrection.

4. Assessing the neural network downscaled storms

4.1. Visual inspection

Each of the four trained NNs were applied to downscale 5,000 aggregated storm fields from a coarse 6×6 grid to a finer 60×60 grid (Fig. 3 and Animation S1 in Supplementary Material). Note that in these visualizations, all generated negative values were set to zero, as none of the models produced exact zero values. Probably the NNs do not reproduce exact zeros because the output activations are continuous, making it statistically unlikely for the networks to generate a precise zero. Since precipitation is either zero or positive, this serves as an initial, straightforward post-processing step. The initial visual inspection of the fields generated by the NNs, compared to the original fields, demonstrates that all models successfully captured both low- and high-intensity regions. However, LNet, FCNet, and UNet exhibited a tendency to oversmooth storm cells, particularly along the boundaries between low- and high-intensity regions, as shown by the color scale in Fig. 3. These regions appear smoothed into elliptical shapes, indicating the anisotropic nature of the storms. Additionally, several high-intensity spots in the original fields (highlighted in purple) appeared with lower intensity, suggesting that these NNs might reduce extreme intensities. Among these three NNs, LNet produced noisier results, generating dry regions, albeit to a smaller extent compared to the original fields; FCNet exhibited excessive smoothing but appeared to more accurately reproduce the dry regions, while in contrast, UNet did not generate any dry regions. On the other hand, the WGAN model produced spatial patterns that closely resembled the original fields. The boundaries between high- and low-intensity regions and their irregular patterns were faithfully reproduced by WGAN, and the extreme intensities appear to agree with the benchmark fields. However, a key limitation of WGAN is its failure to generate true dry regions, instead producing areas of very low intensity.

The initial visual inspection provides a preliminary assessment, revealing noticeable differences among the NNs. However, a more thorough analysis is necessary to evaluate their performance in detail, focusing on key statistical properties of the marginal distribution and of the spatiotemporal dependence. In this analysis, we apply several statistics and use two error metrics to compare the observed and NN-downscaled fields, that is, the Bias and Root Mean Square Error

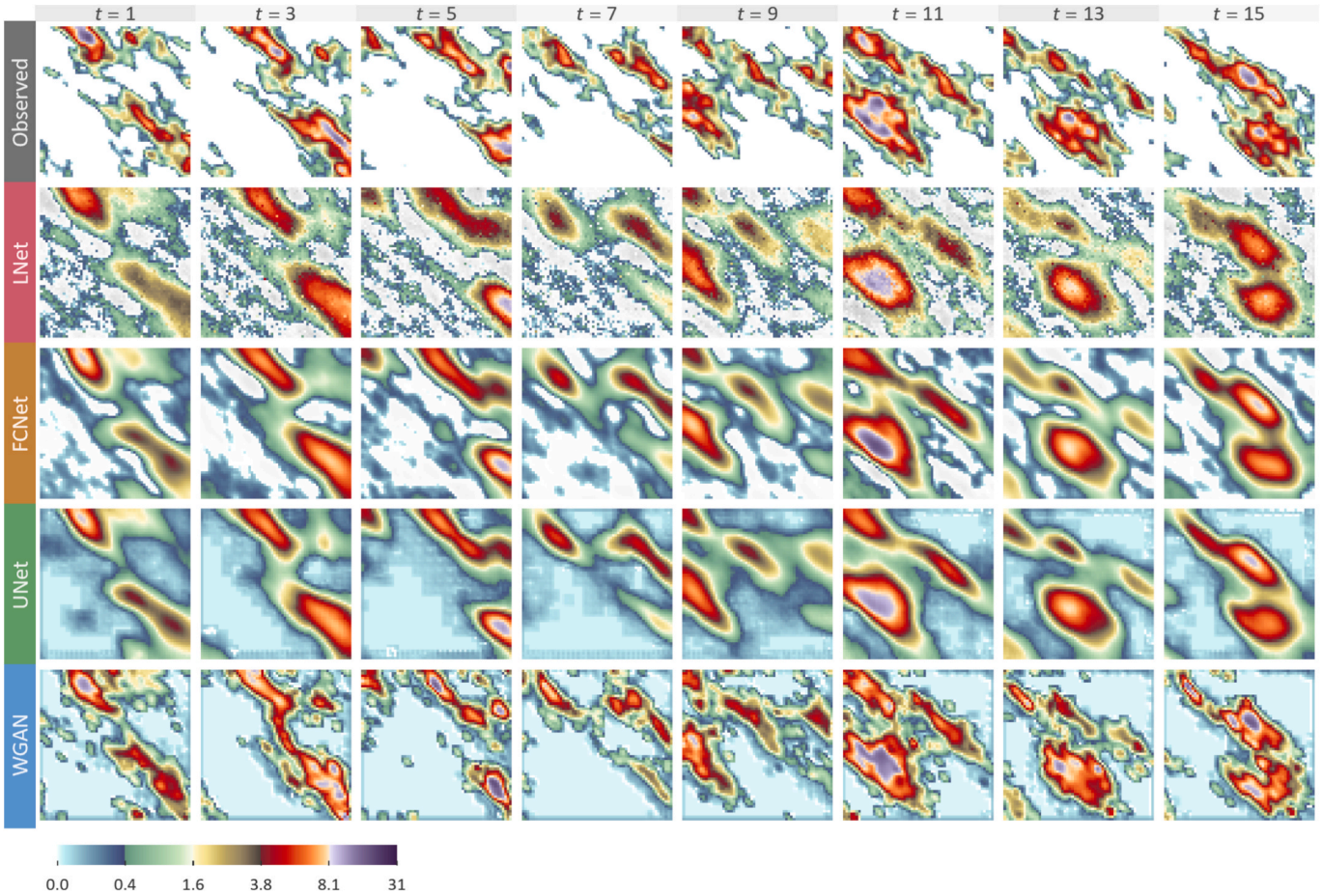


Fig. 3. Downscaled storm fields from LNet, FCNet, UNet, and WGAN from 6× 6 to 60 × 60 resolution. The first row shows benchmark high-resolution fields, with subsequent rows displaying the downscaled fields from the four NNs. The color scale represents storm intensity. See also Animation S1.

(RMSE). The Bias highlights systematic deviations (whether the model consistently over- or underestimates), while the RMSE provides a measure of error magnitude by capturing both positive and negative deviations. The formulas for these error metrics are as follows:

$$\text{Bias} = \frac{1}{n}\sum_{t=1}^{n}\left(\widehat{S}_s(t) - \widehat{S}_o(t)\right) \tag{10}$$

$$\text{RMSE} = \left(\frac{1}{n}\sum_{t=1}^{n}\left(\widehat{S}_s(t) - \widehat{S}_o(t)\right)^2\right)^{1/2} \tag{11}$$

where $\widehat{S}_s(t)$ and $\widehat{S}_o(t)$ represent any of the statistics assessed here and are estimated from each simulated and observed field values, respectively, from time $t = 1$ to $n = 5,000$.

4.2. Negative values and probability of zero

The analysis of the individual fields showed that all four NNs produced small negative values, though these values were generally close to zero. The average percentage of downscaled fields containing negative values was: LNet (32.9 %), FCNet (31.3 %), UNet (0.3 %), and WGAN (1.9 %). Notably, none of the networks generated exact zero values. To estimate the probability of zero (p_0) for each field, we used the ratio of values less than or equal to zero out of the total 3600 values per field. Thus, the average p_0 is equivalent to the percentage of negative values reported for each model, while the true average p_0 of the benchmark fields is 70 %. Based on this definition, the analysis revealed a significant negative bias and high RMSE across all models, as shown in the scatter plots of Fig. 4, where the simulated points fall below the equilibrium diagonal. Each scatter plot and its corresponding density region were constructed from 5000 points (one from each field), representing the estimated p_0 values from the simulated versus observed field data.

4.3. L-moments summary statistics

We estimated the mean, second L-moment (a measure of dispersion similar to the standard deviation), L-skewness (indicating asymmetry), and L-kurtosis (reflecting the heaviness of the tails) for the positive values in each simulated field (for more details on L-moments see Greenwood et al., 1979; Hosking, 1990; Sillitto, 1951). These statistics were then compared to those from the corresponding observed fields across the four NNs. To minimize sample variability, we calculated these statistics using only fields with more than 30 positive values. As previously noted, the NNs struggled to capture zero values, resulting in many simulated values close to zero, which can impact the summary statistics. This limitation may also misrepresent dry periods as slightly wet, leading to overestimation of soil moisture or premature runoff generation, thereby influencing flood and flash-flood modeling. The mean was consistently underestimated by all NNs, as evidenced by the clustering of points below the equilibrium diagonal in Fig. 4. LNet and FCNet performed similarly, with biases of − 0.79 and − 0.89, respectively. UNet showed the poorest performance, with a bias of − 1.16, while WGAN performed slightly better with a bias of − 1.05. The conclusions drawn for the mean also apply to the second L-moment. All NNs exhibit a clear underestimation of the second L-moment, as indicated by the negative bias, with all models performing similarly (Fig. 4).

The high percentage of values near zero, which led to

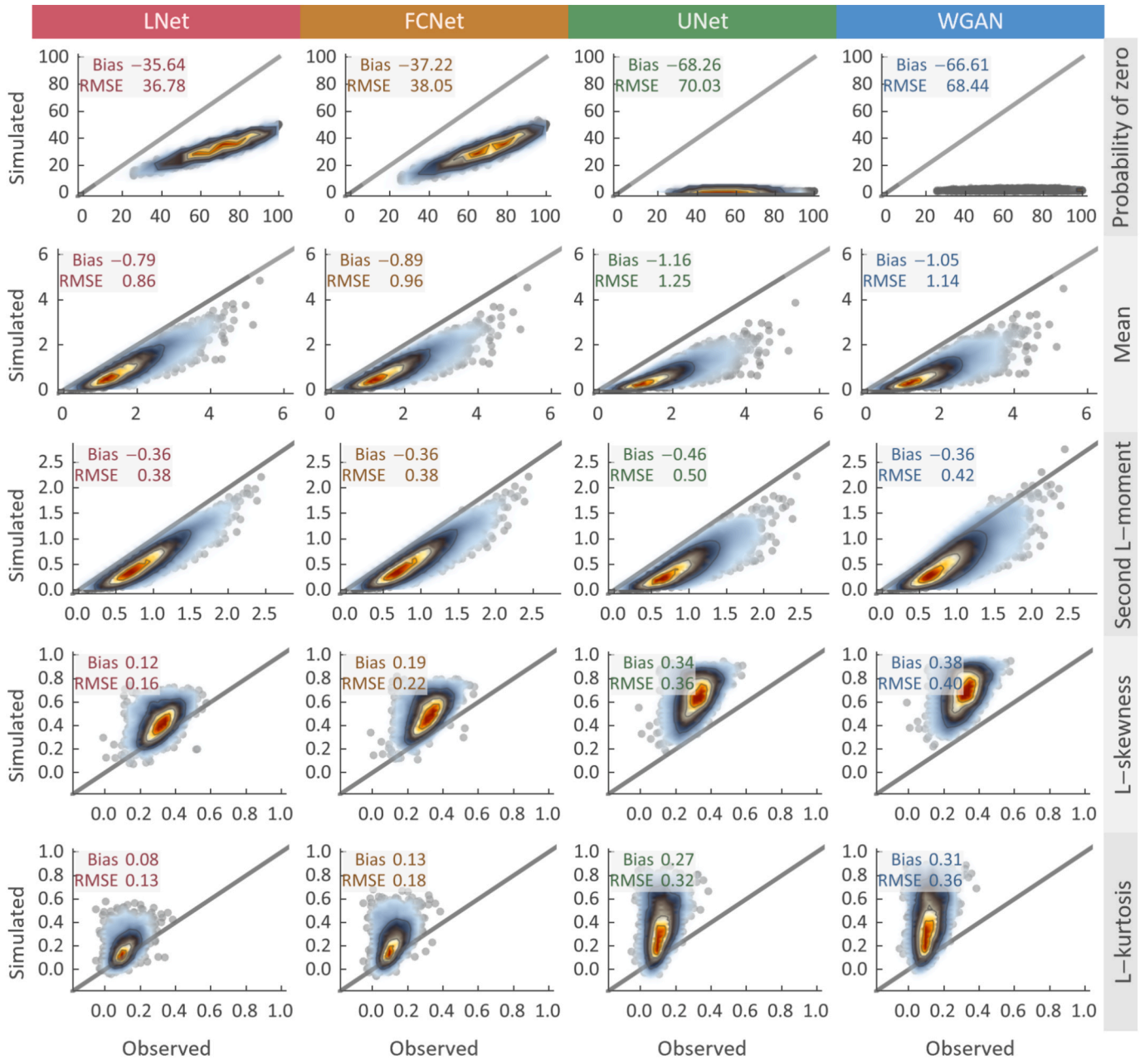


Fig. 4. Scatter density plots comparing the probability of zero, mean, second L-moment, L-skewness, and L-kurtosis in observed and downscaled storm fields from LNet, FCNet, UNet, and WGAN. Red regions indicate higher point density.

underestimation of the mean and dispersion in the NN-downscaled fields, had the opposite effect on L-skewness and L-kurtosis. These shape metrics, which assess asymmetry and tail heaviness more robustly than the classical skewness and kurtosis (Hosking, 1992; Vogel et al., 2024), were significantly overestimated. This is due to the high frequency of values near zero, which increases asymmetry and makes simulated extremes to appear less frequent, artificially inflating thus the heaviness of the right tail which is reflected in L-kurtosis values. Interestingly, the NNs performed differently, with LNet showing a lower positive bias, while WGAN exhibited the worst bias. However, it is important to note that the performance of these shape metrics is heavily influenced by the model's ability to accurately capture the proportion of positive values.

4.4. Temporal autocorrelation

Here, we assess the ability of the NN-downscaled storms to reproduce the temporal autocorrelation structure (ACS). The ACS is a key metric as it governs the persistence of values in the sequences of fields, leading to the clustering of high or low intensities—an important factor in estimating the risk of hazards such as flash flooding. To analyze the ACS, we construct time series for each grid cell, treating the sequence of the 5000 fields as 3600 individual time series (corresponding to the 60×60 grid cells). To clarify, let $x_{i,j}(t)$ represent the value at the (i,j)-th grid cell at time t, and let $\{x_{i,j}(t)\}$ for $t = 1, \cdots, 5,000$ denote the time series at this grid cell. For notational simplicity, hereafter we also use $x_{i,j}(t)$ to represent the time series, relying on context to indicate whether it refers to a single value or the entire series. The temporal correlation for the (i, j)-th grid time series is then defined as $\mathrm{Cor}\big(x_{i,j}(t), x_{i,j}(t-\tau)\big)$ where τ is the temporal lag. For each time series, we compute the autocorrelation

coefficient for $\tau = 1, \cdots, 5$.

The box plots of the estimated ACS (each derived from 3600 estimates) reveal a clear overestimation of the ACS compared to that estimated for the observed fields (Fig. 5a). Nevertheless, the WGAN consistently outperforms the other three NNs, exhibiting significantly lower positive bias across all temporal lags. This highlights WGAN's superior ability to capture the temporal evolution of storm events. Interestingly, despite its better performance in terms of bias, WGAN shows greater variability in the estimated ACS, as indicated by the wider range in the box plots (with whiskers representing the 95 % empirical confidence intervals). This suggests that while WGAN captures the average ACS more accurately, it introduces more variability in its predictions, potentially reflecting a broader spread of storm characteristics across different grid cells. Moreover, WGAN achieves both a lower bias (Fig. 5b) and reduced RMSE (Fig. 5c) across all temporal lags, further confirming its advantage over LNet, FCNet, and UNet.

4.5. Spatial correlation

We evaluate the spatial correlation within each individual field by correlating the field values with those of the field shifted by a specific direction θ and distance δ. To clarify, let $x_{i,j}(t)$ denote the value at the (i,j)-th grid cell at time t, and let $x_{i+k,j+l}(t)$ represent the shifted value located k rows below and l columns to the right of $x_{i,j}(t)$. The direction (angle) from $x_{i,j}(t)$ to $x_{i+k,j+l}(t)$ is given by $\phi = \tan^{-1}(l/k)$, and the distance is $\delta = \sqrt{k^2 + l^2}$. Using this rationale and notation, we denote the 2D grid of the field values at time t as $\boldsymbol{x}(t)$ and the shifted grid as $\boldsymbol{x}^{(k,l)}(t)$.

The directional correlation of the field is then expressed as $\mathrm{Cor}\left(\boldsymbol{x}(t), \boldsymbol{x}^{(k,l)}(t)\right)$ where the direction θ and spatial distance δ are as previously defined. Since the generated storm fields are anisotropic (see Section 2) the strength of the spatial dependence varies not only with distance but also with direction. Here we estimate the directional correlation along the -45° (for $k = d$, $l = d$) and 45° (for $k = -d$, $l = d$) diagonals, for $d = 1,3,5,8,12$. This results in spatial distances $\delta = \sqrt{2}, 3\sqrt{2}, 5\sqrt{2}, 8\sqrt{2}, 12\sqrt{2}$. We selected these two directions because the scaling factors ($\kappa_x = 2.5$, $\kappa_y = 1$) and the rotation ($\omega = -45^\circ$) that define the affine anisotropy introduced during the storm simulations (see Section 2) result in the strongest spatial dependence along this direction, while the weakest dependence occurs along the perpendicular 45° diagonal. To ensure robust estimates of directional correlation, we consider only fields with more than 20 % of positive values, resulting in over 3678 valid estimates.

The spatial correlation analysis, similar to the temporal correlation, demonstrates that the WGAN model excels in reproducing realistic spatial dependencies (Fig. 5). As expected from the earlier visual inspection of spatial patterns (see Fig. 3), WGAN shows superior performance in capturing spatial correlations in both directions. Specifically, the directional correlation estimates of WGAN, represented by box plots (Figs. 5d,g), closely mirror those of the observed fields. This is evident not only in the mean values but also in the variability, as indicated by the similar spread of the box plot whiskers. In contrast, LNet, FCNet, and UNet display a consistent overestimation of spatial dependence, which aligns with their visually smoother spatial patterns that deviate from the observed ones. WGAN also shows minimal positive bias in both spatial

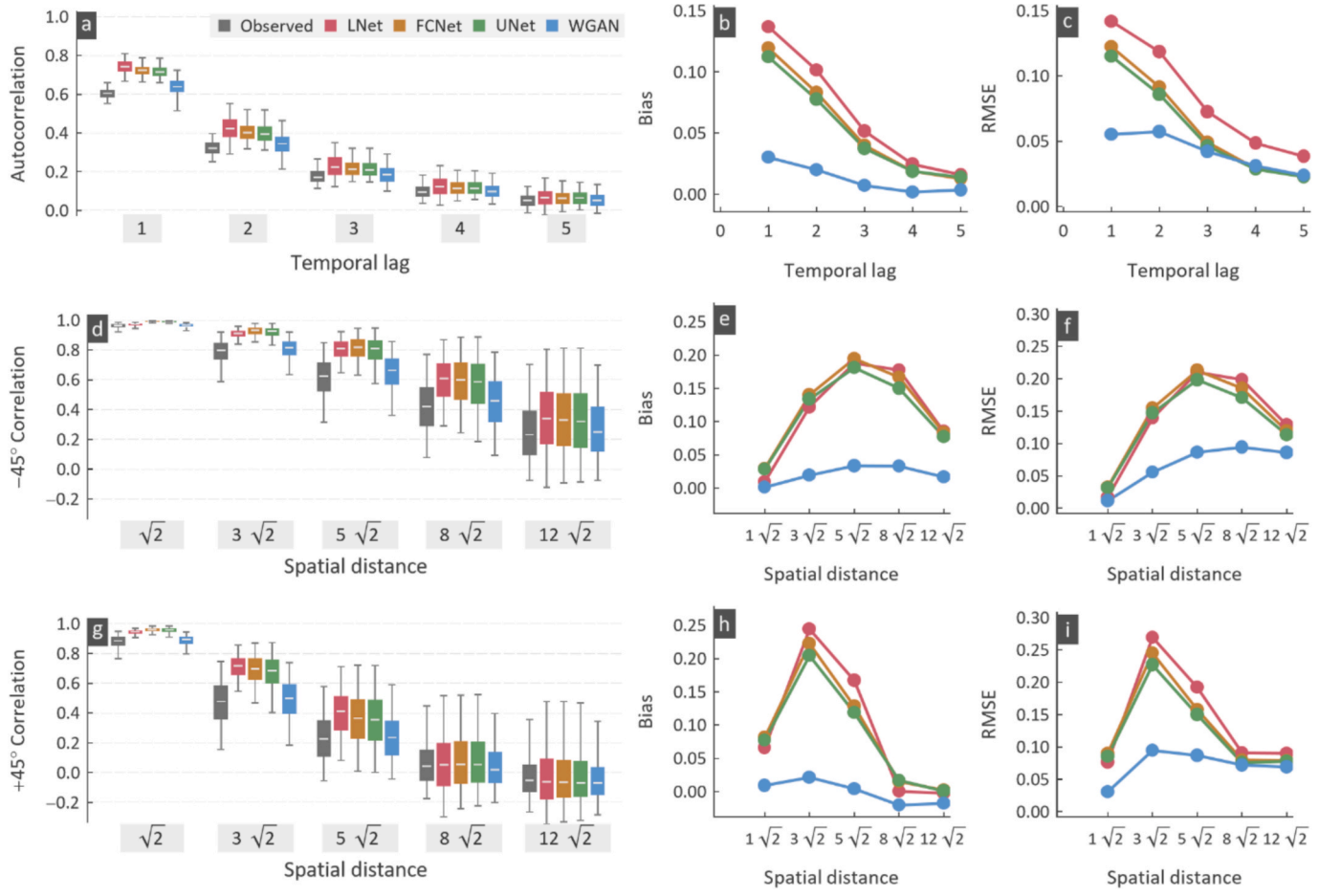


Fig. 5. Performance of downscaled fields from LNet, FCNet, UNet, and WGAN in replicating temporal autocorrelation and spatial dependence of the observed storm fields. (a-c) Temporal autocorrelation estimates, bias, and RMSE across lags. (d-f) -45° directional spatial correlation, bias, and RMSE across distances. (g-i) $+45^\circ$ directional spatial correlation, bias, and RMSE. Box plots show estimates for individual fields, with whiskers indicating 95 % empirical confidence intervals.

directions analyzed (Figs. 5e, h). For instance, in the -45° direction, the bias peaks at a distance of approximately $5\sqrt{2}$ with a value of around 0.03. In contrast, the other models have biases approximately seven times larger at the same distance. The RMSE results (Figs. 5f,i) also highlight WGAN's superior performance, with consistently lower and less variable errors across spatial distances. As a final comment, all models successfully captured the stronger spatial dependence along the -45° diagonal and the weaker dependence along the 45° diagonal, demonstrating that all NNs were able to reproduce anisotropy. However, only WGAN accurately reflected the true geometrical complexity of the wet-dry region borders, producing realistic spatial patterns, as noted earlier.

5. Bias correction of NN-downscaled storms

The analysis of NN-downscaled storms reveals several key issues. All models produced small negative values and failed to generate realistic dry regions, even when negative values were set to zero. A high percentage of values near zero distorted the summary statistics, underestimating the mean and second L-moment while inflating L-skewness and L-kurtosis. In terms of dependence structure, only WGAN, despite some positive bias, successfully reproduced both temporal and spatial correlations with realistic spatial patterns. These findings indicate that post-processing of raw outputs is clearly required before operational use. To address this, we explore and apply two bias correction (BC) methodologies: the first is a general linear bias correction (LBC), while the second is a fully nonlinear bias correction (NLBC) approach.

5.1. Linear bias correction (LBC) of downscaled storms

The most general linear transformation, that could be used for bias correction, includes translation and rescaling and can be applied to adjust the mean (μ) and standard deviation (σ) of a random variable. However, correcting μ and σ using a linear transformation is problematic for variables like precipitation, especially at fine scales, as it may result in negative or unrealistically low values. Thus, here we apply a linear bias correction (LBC) to adjust the probability of zero and the mean of positive values in the downscaled fields. First, we modify p_o by setting a proportion of positive values to zero to match the target p_o of the observed fields. This adjustment can result in minimum positive values above a certain threshold (α), which may seem unnatural for precipitation data. To mitigate this, we correct the lower limit by applying a translation, subtracting the threshold α. The translated variable is then rescaled to achieve the desired mean. Specifically, any simulated value $x_s := x_{s|i,j}(t)$ belonging to the (i,j)-th grid cell for the t-th NN–downscaled field, is corrected using the LBC as follows:

$$\widetilde{x}_s = \begin{cases} 0 & x_s \le \alpha \\ \dfrac{\widehat{\mu}_o}{\widehat{\mu}_{s,\alpha}}(x_s - \alpha) & x_s > \alpha \end{cases} \tag{12}$$

where $\alpha = \widehat{F}_s^{-1}(p_0)$ is the threshold set to match the observed p_0 and can be estimated by the empirical quantile function $\widehat{F}_s^{-1}(\cdot)$ of simulations; and $\widehat{\mu}_o$ and $\widehat{\mu}_{s,\alpha}$ are the estimated mean of the positive values in the observations and the translated by α simulations respectively. Note that while termed linear, this correction is linear only for values above the wet/dry threshold; adjusting the probability of zero makes the overall transformation nonlinear. In general, this bias correction is a special case of a distribution-mapping framework, where simulated values are transformed to match the reference distribution.

The estimated threshold values to match p_0, by setting $x_s \le \alpha$ to zero in the simulations for LNet, FCNet, UNet, and WGAN, are 0.46, 0.31, 0.22, and 0.14, respectively. These thresholds indicate significantly higher-than-zero values as the lower bounds, which justifies and necessitates the translation approach to bring the lower limit to zero, the natural lower bound for precipitation. The rescaling factors $\mu_o/\mu_{s,\alpha}$ for the four NNs are estimated as 1.27, 1.21, 1.19, and 0.95, respectively, indicating values significantly greater than 1 for all cases except for WGAN. Based on this approach, all linear bias corrected downscaled simulations (hereafter denoted as LBC fields) now match the probability of zero and the mean of positive values of the observed fields. To avoid any misunderstanding, the LBC is not applied to each individual field; instead, it corrects the p_0 and the mean of positive values for the entire simulation, equivalent to merging all field values. We stress that all correction parameters (the p_0 threshold and mean rescaling factor) were derived from the training/validation dataset to ensure that no information from the test set was used during bias correction, maintaining full independence between model training and evaluation.

The LBC fields (Fig. 6 and Animation S2 in Supplementary Material) show the expected improvement in capturing wet and dry regions compared to the uncorrected downscaled fields (Fig. 3), as the probability of zero is now matched which is reflected in the individual fields. All negative values have been eliminated and set to zero, effectively addressing this limitation in the uncorrected fields. Additionally, a visual inspection of several fields suggests that high- and low-intensity regions are represented more accurately relative to the observations. While the LBC method enhances certain features, the smooth field patterns seen in LNet, FCNet, and UNet remain, as LBC is not designed to modify spatial smoothness but rather to correct the probability of zero and the mean of positive values.

The linear bias correction had a profound impact on the statistical properties of the downscaled storms generated by the NNs. The scatter density plots (Fig. 7) show a marked improvement in the accuracy of key metrics, such as the probability of zero, mean, second L-moment, L-skewness, and L-kurtosis, compared to the uncorrected outputs (Fig. 4). Specifically, the probability of zero in the LBC-corrected fields is now accurately represented across all NNs, as evidenced by the alignment of scatter points along the equilibrium diagonal (Fig. 7). This contrasts with the uncorrected fields, where significant negative biases were present, particularly in WGAN and UNet (Fig. 4). The mean values of the LBC-corrected fields also now exhibit minimal bias, particularly for WGAN, which shows near-perfect alignment with the diagonal (bias of -0.01 and RMSE of 0.17). Similarly, the second L-moment is significantly improved, with WGAN demonstrating very low bias (0.01) and RMSE (0.10). LNet, FCNet, and UNet still show slight underestimations, with density regions positioned slightly below the equilibrium line. The LBC also drastically corrected L-skewness and L-kurtosis, which were artificially inflated in the uncorrected fields due to an overrepresentation of near-zero values. These metrics are now more aligned with the corresponding metrics of the observed fields, particularly in WGAN, which exhibits near-zero bias and low RMSE for both.

Regarding the dependence structure, the LBC fields (Fig. 8) display largely similar properties to the uncorrected ones (see Fig. 5), with minor improvements. This is expected, as linear transformations generally do not alter correlation properties. However, it is important to note that the component of the LBC transformation that adjusts negative values to zero to match the p_0 in the benchmark fields is not strictly linear. While its impact is not significant in this instance, it has the potential to affect the strength of spatiotemporal dependence (see Papalexiou, 2018 for a detailed discussion on the influence of nonlinear transformations on correlation).

The overall improvements across all tested metrics demonstrate the robustness of this simple yet effective transformation. The field-by-field analysis confirms that the LBC method successfully corrected the distributional properties of the downscaled fields. However, when we use all positive values from the benchmark and LBC fields to construct empirical probability plots (Fig. 9) we observed tail deviations even after the LBC. Specifically, the empirical distribution of the positive values from the uncorrected downscaled fields (Fig. 9a) deviate significantly from that of the benchmark values as it was anticipated from the

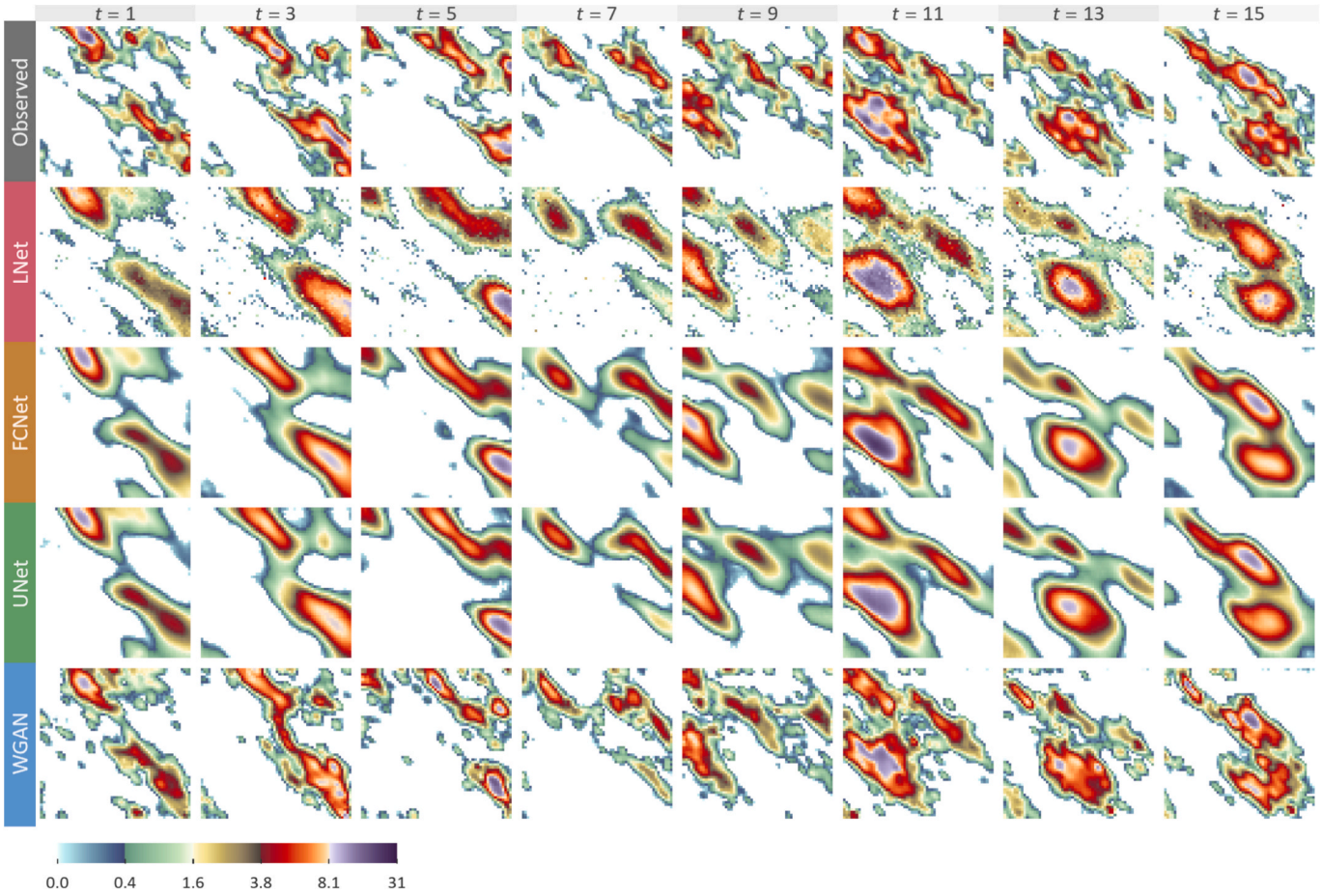


Fig. 6. Linear bias-corrected downscaled storm fields from LNet, FCNet, UNet, and WGAN from 6×6 to 60×60 resolution. The first row shows benchmark high-resolution fields, followed by the LBC downscaled fields from the four NNs. The color scale represents storm intensity. See also Animation S2.

summary statistics analysis (see Section 4). Specifically, LNet, FCNet, and UNet underestimate the exceedance probabilities, while WGAN overestimates the exceedance probabilities for extreme values. In contrast, the exceedance probabilities of the LBC fields (Fig. 9b) show substantial improvement. The right tails of the probabilities produced by LNet, FCNet, and UNet align closely with the observed tail, with a slight overestimation, which could be attributed to random fluctuations, as tails are sensitive to such variations. However, WGAN still overestimates extremes, producing a heavier tail. It is important to note that, by definition, a linear transformation cannot alter the asymptotic behavior of the tail (loosely speaking its curvature as visualized in probability plots). This indicates that a nonlinear transformation may be necessary to more accurately capture the behavior of extremes and requires further investigation.

5.2. Nonlinear bias correction (NLBC) of downscaled storms

The previous analysis and application of the LBC demonstrated that, despite significant improvements in the statistical properties of the simulated storm fields, all NNs did not reproduce the tail of the benchmark fields accurately. Especially WGAN, which was the most realistic NN in reproducing spatial patterns, tended to overestimate extremes. To address this, we explore the use of a nonlinear bias correction (NLBC) transformation, which corrects the entire distribution. This approach maps one mixed-type distribution to another. The transformation is nonlinear by design, targeting discrepancies in shape and distribution tails that linear transformations cannot address.

The NLBC process starts similarly to the LBC. To match the probability of zero p_0 in the observed fields, we determine a threshold α for each NN-downscaled field. We set all values below α to zero and subtract α from the remaining values (i.e., $x_s - \alpha$). Let $F_o(x)$ denote the fitted distribution to the positive values of the observed fields, and $F_{s,\alpha}(x)$ denote the fitted distribution to the translated values of the simulated fields. The nonlinearly bias-corrected values $\widetilde{x}_s$ are given by:

$$\widetilde{x}_s = \begin{cases} 0 & x_s \le \alpha \\ F_o^{-1}\left(F_{s,\alpha}(x_s - \alpha)\right) & x_s > \alpha \end{cases} \tag{13}$$

where $\alpha = \widehat{F}_s^{-1}(p_0)$ matches the probability of zero p_0 in observations, estimated using the empirical quantile function $\widehat{F}_s^{-1}(\cdot)$ of the NN-downscaled data. The function $F_o^{-1}(\cdot)$ is the quantile function of the fitted distribution to the observed values. Thus, the nonlinear bias correction is a generic distribution-mapping employing a quantile-based nonlinear transformation, conceptually related to approaches used in climate model bias correction (e.g., Rajulapati & Papalexiou, 2023).

Apparently, the thresholds to match the probability of zero in the NN-downscaled fields are the same as those estimated in the previous section for LBC. After subtracting the corresponding threshold from the NN-fields, all positive values are used to fit a continuous parametric distribution. As described in Section 2, the marginal distribution used to describe positive precipitation in the synthetic generation of the storms was the Generalized Exponential Type IV ($\mathscr{GE}4$). However, in practice, this information is often unknown. To apply the NLBC framework without relying on prior knowledge that is not available in practice, we instead fit the Generalized Exponential Type I ($\mathscr{GE}1$) distribution (Papalexiou, 2022) by minimizing the squared error norm between the theoretical quantiles and the positive values of observations and

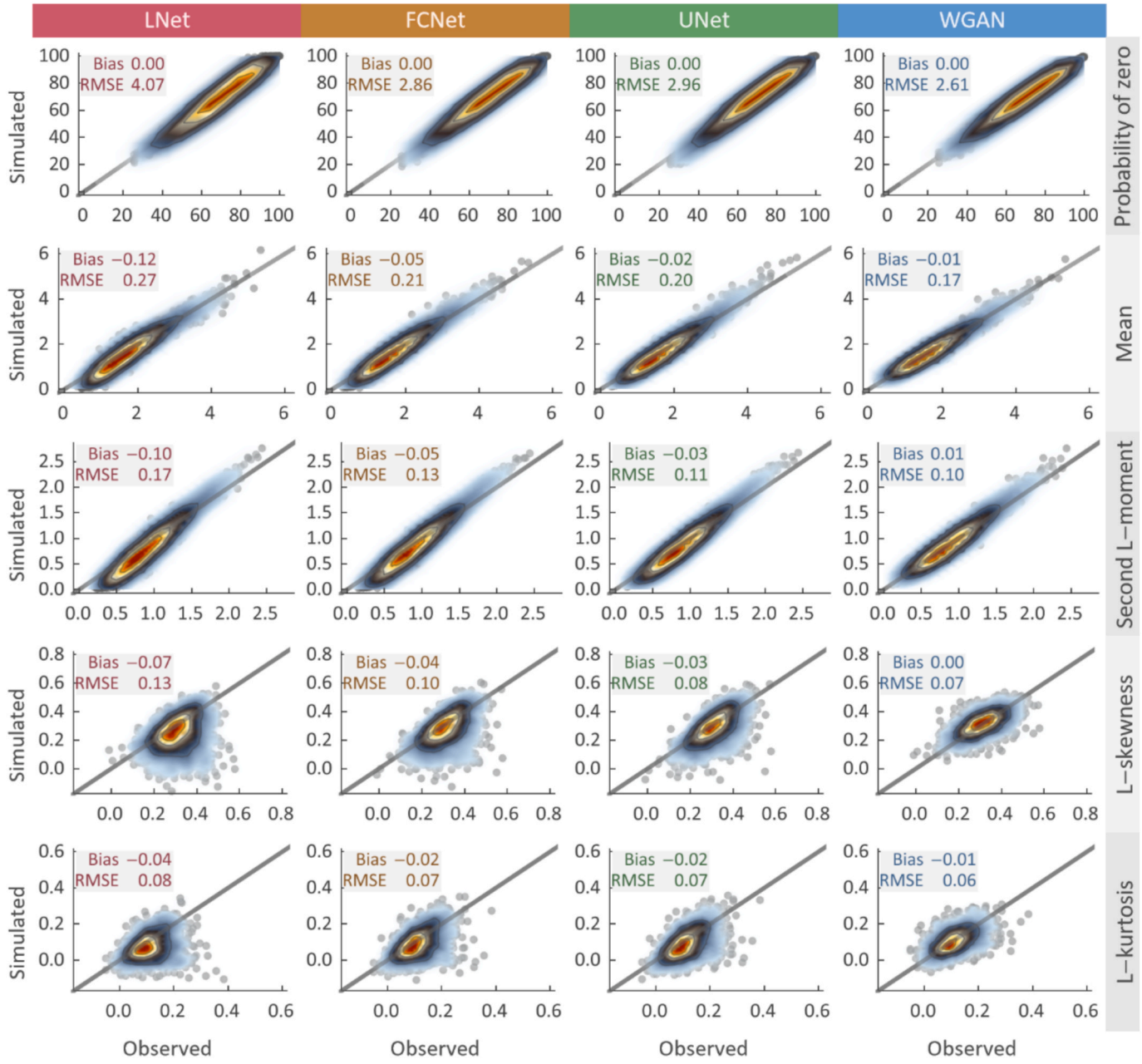


Fig. 7. Scatter density plots comparing the probability of zero, mean, second L-moment, L-skewness, and L-kurtosis in observed and LBC downscaled storm fields from LNet, FCNet, UNet, and WGAN. Red regions indicate higher point density.

simulations. The $\mathscr{GE}1$ distribution function is given by:

$$F_{\mathscr{GE}1}(x;\beta,\gamma_1,\gamma_2) = 1 - \exp\left(1 - \left(\gamma_2\left(\frac{x}{\beta}\right)^{\gamma_1} + 1\right)^{\frac{1}{\gamma_2}}\right) \tag{14}$$

where $\beta > 0$ is a scale parameter and $\gamma_1 > 0$ and $\gamma_2 > 0$ are shape parameters. Note that preliminary Monte Carlo tests confirmed that the least-squares quantile fitting procedure applied yields stable and reliable parameter estimates. Clearly, any other fitting method can be used.

The fitted $\mathscr{GE}1$ distributions (solid lines in Fig. 10a) consistently describe the empirical distributions of positive values (dots in Fig. 10a) for both the benchmark and the NN-downscaled fields. After applying the NLBC, as outlined in Equation (13), we estimated the empirical distribution function of the NLBC fields. As shown in Fig. 10b, the empirical distributions now closely overlap with the distribution of the benchmark fields, demonstrating the effectiveness of the NLBC in correcting the distribution tail.

Similar to LBC, NLBC is applied collectively to all positive values of the NN-downscaled fields rather than on a field-by-field basis. Consequently, it is important to examine its effects on individual fields, as previously done. Visually, the NLBC fields (Fig. A2 and Animation S3 in Supplementary Material) resemble the LBC fields shown in Fig. 6; both methods produce identical wet/dry regions, with only subtle differences in high intensities noticeable upon careful inspection. As expected, the NLBC fields preserve the probability of zero well across all networks (Fig. A3). Interestingly, NLBC introduces a slightly larger mean bias for LNet, FCNet, and UNet, while WGAN achieves minimal bias and near-perfect alignment (Fig. A3). The second L-moment indicates that WGAN performs best under both methods, showing minimal bias, whereas LNet, FCNet, and UNet exhibit larger negative biases compared to LBC. In terms of L-skewness, NLBC improves WGAN's alignment with observed values, with scatter point densities better aligned along the

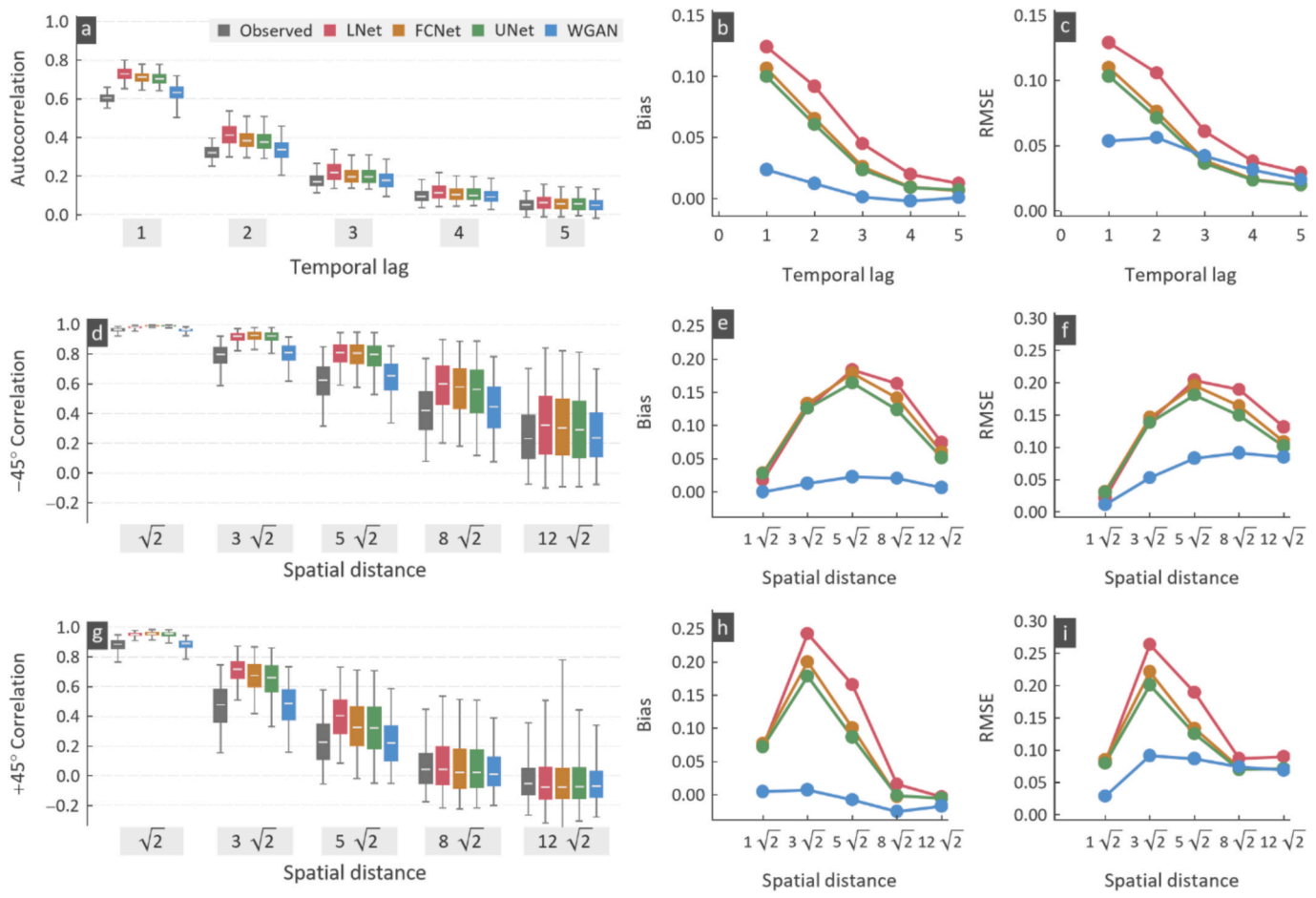


Fig. 8. Performance of LBC downscaled fields from LNet, FCNet, UNet, and WGAN in replicating temporal autocorrelation and spatial dependence of the observed storm fields. (a-c) Temporal autocorrelation estimates, bias, and RMSE across lags. (d-f) − 45° directional spatial correlation, bias, and RMSE across distances. (g-i) + 45° directional spatial correlation, bias, and RMSE. Box plots show estimates for individual fields, with whiskers indicating 95 % empirical confidence intervals.

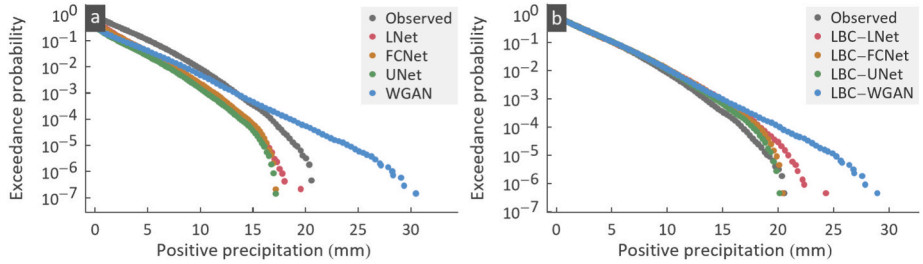


Fig. 9. Empirical exceedance probability distribution based on positive values from the benchmark and NN-downscaled fields. (a) For uncorrected fields. (b) For LBC fields.

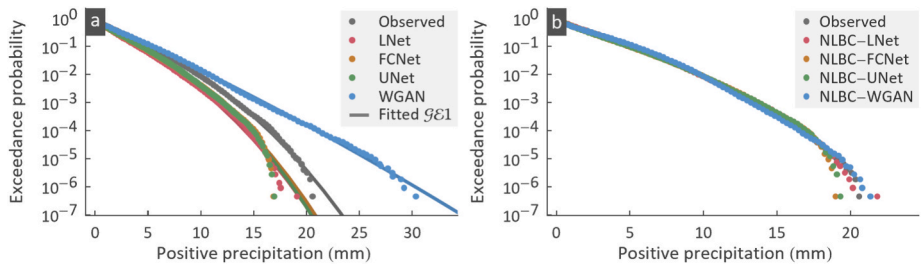


Fig. 10. (a) Empirical exceedance probability distributions of positive values for benchmark (observed) and uncorrected downscaled fields, along with the corresponding fitted $\mathcal{GE}1$ distributions. (b) Empirical exceedance probability distributions for NLBC fields.

diagonal, although the other networks display similar biases under both methods (Fig. A3). L-kurtosis follows a similar pattern, with WGAN exhibiting low negative bias and superior performance. Overall, NLBC—particularly for WGAN, the only model with strong operational potential—not only corrects tail behavior by accurately matching extremes but also preserves individual field properties with high precision.

Regarding autocorrelation, the differences between the LBC (Fig. 8) and NLBC (Fig. A4) downscaled fields are minimal. NLBC seems to slightly improve the bias of WGAN while slightly worsening the bias of the other neural networks. For spatial correlation, WGAN shows subtle improvements under NLBC, exhibiting nearly zero bias in both -45° and $+45^{\circ}$ directional spatial correlations (Fig. A4). Overall, although the differences between LBC and NLBC are not substantial across all models, NLBC slightly enhances WGAN's performance, making it the most promising model for replicating both temporal and spatial dependencies with minimal bias and error. We stress that while both LBC and NLBC target marginal distribution correction, changes in dependence structures (e.g., temporal and spatial correlations) arise indirectly as a by-product through the bias correction transformation.

6. Discussion

It is important to acknowledge that the performance of the NNs in this study may be influenced by specific parameter settings used during training. While we employed standard implementations representative of each NN architecture, variations in hyperparameters, learning rates, or training epochs could potentially enhance or reduce the performance before any bias correction is applied. Therefore, the results presented here should be seen as indicative of typical NN behavior, but not necessarily the optimal configuration. Future research could investigate alternative parameter settings to further improve performance and explore models that may eliminate the need for bias corrections altogether.

However, a critical question remains: what exactly are the NNs learning? The straightforward answer is that we do not know with certainty what these models have learned. Future research should focus more on exploring downscaling schemes that integrate interpretability and explainability tools (Holzinger et al., 2022; Mamalakis et al., 2022a, Mamalakis et al., 2022b; Mamalakis et al., 2023; McGovern et al., 2019; Rampal et al., 2022, 2024; Toms et al., 2020). Despite the theoretical results that some of the models used herein are based on (Arjovsky et al., 2017; Gulrajani et al., 2017), the distributional properties and the spatiotemporal structure produced by these models were shown to be critically biased. In this context, post-processing the NN outputs serves as a form of "machine unlearning," applied to ensure that we have knowledge, at the very least, of the theoretical properties of the distribution describing intensities. We note here that "machine unlearning" is a known term in computer science, that is used to describe (among others) the post-training removal of biases and inaccuracies that neural networks might learn during training (Bourtoule et al., 2021; Li et al., 2024). In the broader ML literature, it also refers to methods designed to remove or alter previously learned information from trained models, often for reasons of privacy, security, or fairness (see e.g., Shaik et al., 2025). The bias correction methods adopted in this study are purely statistical and do not re-estimate the parameters of the network (as done in typical "unlearning" methods), we nevertheless emphasize that such post-processing is necessary. Nonlinear bias correction provides full control over the final distribution, including its desired tail behavior, which is crucial. The tail governs the frequency and magnitude of extremes, and thus, having control over it ensures more accurate risk estimates and informed decision-making.

A key question and challenge that must be discussed is how we can ensure that bias correction will have the desired effect. The linear bias correction, as applied here, by definition, corrects the probability of zero and the mean of positive rainfall. However, nonlinear bias correction requires more caution, as it is technically more complex and may produce unintended effects. If the distribution fitted to the raw outputs of the neural networks is poorly chosen and does not represent the data well, its mapping to the data will not result in a uniform distribution of probabilities, leading next to inaccurate quantile mapping (see Equation (13)). This also applies to the distribution fitted to the observations, which must accurately describe the data; otherwise, the bias correction will adjust the NN outputs incorrectly. Therefore, nonlinear quantile mapping can be particularly challenging, especially when dealing with real-world data that often have quality issues, leading to poorly fitted distributions. This challenge also explains why the NLBC caused slightly higher bias in the mean and second L-moment in the first three NNs (Fig. A3). Their empirical distribution was not as well captured by the fitted $\mathscr{GE}1$ distribution leading to increased bias in these two metrics.

Among the NNs tested, WGAN stands out as the only model with operational potential. The primary limitation of LNet, FCNet, and UNet in reproducing storms was their tendency to generate overly smoothed spatial patterns, failing to capture the fine-scale structure of storms, particularly at wet/dry boundaries. This issue cannot be easily corrected through post-processing. In contrast, WGAN showed a clear advantage in accurately capturing the complex geometry of wet/dry borders and intensity fluctuations within storm cells, as demonstrated both by visual inspection of the downscaled fields and its superior performance in reproducing spatiotemporal dependence metrics, as outlined in the analysis. From a technical perspective, WGAN's use of the Wasserstein distance provides a more robust and meaningful measure of differences between real and generated data compared to the traditional loss functions used in the other networks. This allows WGAN to preserve the finer details necessary to replicate the complex spatial patterns of storms. The misrepresentation of the probability distribution by WGAN, including the probability of zero and tail behavior, is not a concern for operational use, as we have shown that post-processing, in the form of linear or nonlinear bias correction, can effectively resolve this issue.

Our controlled experiments indicate that generative models like WGAN can reproduce temporal and spatial dependence with comparatively low bias, yet residual discrepancies often persist in higher-order distributional statistics and in the occurrence of exact zeros. This suggests a principled two-stage workflow: representation first (learning realistic structures), followed by calibration (post-processing "machine unlearning") to enforce marginal properties—including zeros and tail behavior—needed for hydrologic use. We stress that this is not an ad hoc fix but a standard, interpretable step analogous to bias correction widely applied to climate-model outputs. Even after several generations of global and regional climate models, their raw precipitation fields still do not reproduce statistics with full accuracy and therefore require systematic bias correction before use. Following a similar approach for neural networks is both practical and necessary until loss functions and architectures explicitly encode marginal constraints (e.g., zero inflation, L-moment, or tail targets). Recent studies have begun to explore extreme-aware loss formulations within GAN frameworks (e.g., Lee & Park, 2025) representing a promising complementary direction. In contrast, our controlled experiments demonstrate that statistical bias correction can be systematically coupled with neural networks to achieve reliable, operationally usable results—an aspect not yet addressed in purely data-driven implementations. Controlled synthetic benchmarks therefore remain essential to reveal strengths and weaknesses that may be obscured in observational evaluations.

Overall, this study offers not only a technical evaluation but also a conceptual message: even the most advanced neural networks require systematic post-processing to achieve reliable, bias-free outputs. Until architectures or loss functions inherently encode physical and statistical constraints, coupling neural networks with robust bias-correction methods remains an operationally sound and transparent approach—mirroring decades of practice in climate modeling. We hope this perspective encourages a more critical and systematic assessment of neural-network-based downscaling frameworks in hydrologic and climate applications.

We stress that our results are derived from controlled synthetic storms where the true marginal and dependence properties are known. This design has the advantage of isolating what the networks learn and consequently what post-processing can fix. Translating to operational application using observations will introduce challenges such as measurement noise and detection limits (affecting p_0), data sparsity for extremes in short records, orographic and representativeness effects in many regions, and potentially climate nonstationarity. For operational use, we propose specific steps that should be considered: (1) quality-control of inputs (outlier, flags, etc.) and define a wet/dry threshold consistent with instrument limits and current practice; (2) retrain the network on historical reference fields, for example, using radar fields or fine scale regional climate model outputs; (3) estimate all bias-correction parameters from the training set only as we applied in this study; (4) assess in detail outputs using both distributional and dependence diagnostics; and (5) monitor and assess potential drifts and recalibrate as the climatology changes. Such a framework we deem preserves the workflow of using NNs coupled with statistical bias correction and could address observational uncertainties and nonstationarity.

For future research, training WGAN on larger spatial domains presents both a challenge and an opportunity, particularly because storms can move at varying velocities and exhibit spatially varying anisotropies, which could affect the model's ability to accurately capture storm dynamics. Incorporating a binary layer specifically designed to assess the geometry of wet/dry regions could improve the representation of these critical boundaries or of other physical constraints (Harder et al., 2023). Future efforts could also focus on leveraging information from multiple spatiotemporal scales to better capture both small- and large-scale storm features. Additionally, integrating exogenous variables, such as temperature, could enhance the potential to assess changes in storm behavior under varying climate conditions. Downscaling coarse fields to fine-scale storms remains challenging for all these reasons; however, the promising results suggest that other important hydrometeorological variables, such as wind or temperature, could also be effectively downscaled using WGAN, creating a more versatile framework for broader climate and risk applications.

7. Conclusions

Advanced hydrologic modeling relies on precipitation data at fine spatiotemporal scales. Despite advancements in radar technology and regional climate models, most observed and simulated precipitation data remain at coarse scales, limiting their utility for risk assessment and decision-making. Machine learning methods offer a promising solution by transforming coarse-scale data into the fine-scale information necessary for accurate hydrologic assessments. However, neural network outputs have limitations that must be rigorously evaluated before operational use. Their "black-box" nature complicates understanding specific performance aspects, and strong results on cases similar to the training set may not reliably extend to unseen or extreme scenarios.

In this study, we conducted a controlled experiment to evaluate the potential of machine learning methods for downscaling storms to fine spatial resolutions. Using CoSMoS, we generated synthetic storm fields emulating real-world storms with skewed distributions, strong spatiotemporal correlations, advection, and anisotropy. We aggregated these fine-scale fields to coarse scales to represent typical observational data and trained four neural networks (LNet, FCNet, UNet, and WGAN) to assess their performance in downscaling back to high-resolution outputs.

- **Raw Downscaled Outputs Had Limitations:** In general, all models reproduced low- and high-intensity regions. However, LNet, FCNet, and UNet oversmoothed storm cells, resulting in unrealistically smooth intensity boundaries. All models produced small negative values and did not match the observed probability of zero resulting in inaccurately representing wet and dry regions. The high frequency of near-zero values skewed summary statistics for positive precipitation: the mean and second L-moment were underestimated, while L-skewness and L-kurtosis were overestimated. Additionally, LNet, FCNet, and UNet overestimated temporal and spatial correlations. In contrast, WGAN accurately captured spatiotemporal dependencies.
- **Linear Bias Correction (LBC) Improved Performance:** We applied LBC to adjust the probability of zero and the mean of positive values. This correction eliminated negative values and led to accurate representation of wet and dry boundaries. Additionally, LBC significantly improved summary statistics: the mean and second L-moment were close to observed values, and the severe overestimations in L-skewness and L-kurtosis were almost eliminated. LBC also resulted in minor improvements in spatiotemporal dependence structures due to the correction of the probability of zero in the downscaled fields. WGAN continued to exhibit superior performance, but discrepancies in tail behavior suggested the need for exploring nonlinear bias correction methods to better capture extreme values.
- **Nonlinear Bias Correction (NLBC) Improved Extremes:** We applied a nonlinear bias correction (NLBC) to adjust the entire distribution and correct the tail behavior that LBC could not address. NLBC effectively aligned the empirical distributions with benchmark data, significantly improving the representation of extreme values. The representation of wet/dry regions was similar to that achieved with LBC, and summary statistics were also well reproduced. Specifically, for WGAN, NLBC resulted in minimal mean bias, better alignment in L-skewness and L-kurtosis, and additional slight enhancements in spatiotemporal dependencies, exhibiting nearly zero bias in spatial correlations.

Generative models such as WGAN show great potential in capturing the complex spatiotemporal variability of storms, offering valuable tools for high-resolution hydrologic modeling. However, our study underscores the need for post-processing as a "machine unlearning" step to refine results, particularly for accurately representing extremes that can otherwise be missed. The two bias correction methods presented here—linear and nonlinear—drastically improved model outputs, enhancing WGAN's ability to replicate both distributional properties and fine-scale dependencies critical for climate resilience and risk management.

In practical terms, a reliable application requires a two-step process: first, use a generative model to reproduce the fine-scale spatial and temporal structure, and then apply bias correction to ensure realistic zeros and accurate distribution tails. This approach follows the long-established practice in climate modeling, where outputs from global and regional models are routinely bias-corrected before use. Following a similar philosophy helps reduce residual biases and makes the results suitable for operational hydrologic applications, while future work can focus on models that better capture these properties directly.

CRediT authorship contribution statement

Simon Michael Papalexiou: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Antonios Mamalakis:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The CoSMoS R package (Papalexiou et al., 2021a, Papalexiou et al., 2021b), originally resealed in April 2019, is available at CRAN (R Core Team, 2021); see also https://cran.r-project.org/web/packages/CoSMoS/vignettes/vignette.html.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2025.134689.

Data availability

The authors used only synthetic data generated by the models and methods described in the main text. All model architectures, hyperparameters, optimizer settings, and regularization parameters used in this study are fully documented in the accompanying Python code, which is openly available at https://github.com/amamalak/downscaling_biascorrection.

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN. arXiv. Retrieved from http://arxiv.org/abs/1701.07875.

Bacchi, B., Kottegoda, N.T., 1995. Identification and calibration of spatial correlation patterns of rainfall. J. Hydrol. 165 (1), 311–348. https://doi.org/10.1016/0022-1694(94)02590-8.

Bardossy, A., Plate, E.J., 1992. Space-time model for daily rainfall using atmospheric circulation patterns. Water Resour. Res. 28 (5), 1247–1259. https://doi.org/10.1029/91WR02589.

Benestad, R.E., 2004. Empirical-statistical downscaling in climate modeling. Eos Trans. AGU 85 (42), 417–422. https://doi.org/10.1029/2004EO420002.

Benestad, R.E., 2010. Downscaling precipitation extremes. Theor. Appl. Climatol. 100 (1), 1–21. https://doi.org/10.1007/s00704-009-0158-1.

Berne, A., Krajewski, W.F., 2013. Radar for hydrology: unfulfilled promise or unrecognized potential? Adv. Water Resour. 51, 357–366. https://doi.org/10.1016/j.advwatres.2012.05.005.

Boulaguiem, Y., Zscheischler, J., Vignotto, E., Van Der Wiel, K., Engelke, S., 2022. Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. Environ. Data Sci. 1, e5.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., et al., 2021. Machine unlearning: 42nd IEEE Symposium on Security and Privacy, SP 2021. *Proceedings - 2021 IEEE Symposium on Security and Privacy, SP 2021*, 141–159. https://doi.org/10.1109/SP40001.2021.00019.

Cavanaugh, N.R., Gershunov, A., Panorska, A.K., Kozubowski, T.J., 2015. The probability distribution of intense daily precipitation. Geophys. Res. Lett. 42 (5), 1560–1567. https://doi.org/10.1002/2015GL063238.

Chen, J., Janke, T., Steinke, F., Lerch, S., 2024. Generative machine learning methods for multivariate ensemble post-processing. Ann. Appl. Stat. 18 (1). https://doi.org/10.1214/23-AOAS1784.

Chilès, J.-P., Delfiner, P., 2009. *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons.

Duncan, J., Subramanian, S., & Harrington, P. (2022, October 22). Generative Modeling of High-resolution Global Precipitation Forecasts. arXiv. Retrieved from http://arxiv.org/abs/2210.12504.

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., Taylor, K.E., 2016. Overview of the coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. Geosci. Model Dev. 9 (5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016.

Glawion, L., Polz, J., Kunstmann, H., Fersch, B., Chwala, C., 2023. spateGAN: spatio-temporal downscaling of rainfall fields using a cGAN approach. Earth Space Sci. 10 (10), e2023EA002906. https://doi.org/10.1029/2023EA002906.

Glawion, L., Polz, J., Kunstmann, H. G., Fersch, B., & Chwala, C. (2023b, February 20). spateGAN: Spatio-Temporal Downscaling of Rainfall Fields using a cGAN Approach. Preprints. https://doi.org/10.22541/essoar.167690003.33629126/v1.

Gneiting, T., 2002. Nonseparable, stationary covariance functions for space–time data. J. Am. Stat. Assoc. 97 (458), 590–600. https://doi.org/10.1198/016214502760047113.

González-Abad, J., Baño-Medina, J., & Cachá, I. H. (2023, April 27). On the use of Deep Generative Models for Perfect Prognosis Climate Downscaling. arXiv. Retrieved from http://arxiv.org/abs/2305.00974.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C., Wallis, J.R., 1979. Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. Water Resour. Res. 15 (5), 1049–1054.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017, December 25). Improved Training of Wasserstein GANs. arXiv. Retrieved from http://arxiv.org/abs/1704.00028.

Harder, P., Hernandez-Garcia, A., Ramesh, V., Yang, Q., Sattegeri, P., Szwarcman, D., et al. (n.d.). Hard-Constrained Deep Learning for Climate Downscaling.

Harris, L., McRae, A.T.T., Chantry, M., Dueben, P.D., Palmer, T.N., 2022. A generative deep learning approach to stochastic downscaling of precipitation forecasts. J. Adv. Model. Earth Syst. 14 (10), e2022MS003120. https://doi.org/10.1029/2022MS003120.

Hobeichi, S., Nishant, N., Shao, Y., Abramowitz, G., Pitman, A., Sherwood, S., et al., 2023. Using machine learning to cut the cost of dynamical downscaling. Earth's Future 11 (3), e2022EF003291. https://doi.org/10.1029/2022EF003291.

Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., & Samek, W. (Eds.). (2022). *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (Vol. 13200). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2.

Hosking, J.R.M., 1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. J. Roy. Stat. Soc.: Ser. B (Methodol.) 52 (1), 105–124. https://doi.org/10.1111/j.2517-6161.1990.tb01775.x.

Hosking, J.R.M., 1992. Moments or L moments? an example comparing two measures of distributional shape. Am. Stat. 46 (3), 186–189. https://doi.org/10.2307/2685210.

Houze, R.A., 2018. 100 years of research on mesoscale convective systems. Meteorol. Monogr. 59, 17.1-17.54. https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0001.1.

Hristopulos, D.T., 2020. *Random Fields for Spatial Data Modeling: a Primer for scientists and Engineers*. Springer Nature.

Leblois, E., Creutin, J.-D., 2013. Space-time simulation of intermittent rainfall with prescribed advection field: Adaptation of the turning band method. Water Resour. Res. 49 (6), 3375–3387. https://doi.org/10.1002/wrcr.20190.

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

Lee, J., Park, S.Y., 2025. WGAN-GP-based conditional GAN (cGAN) with extreme critic for precipitation downscaling in a key agricultural region of the Northeastern U.S. IEEE Access 13, 46030–46041. https://doi.org/10.1109/ACCESS.2025.3549443.

Leinonen, J., Nerini, D., Berne, A., 2021. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. IEEE Trans. Geosci. Remote Sens. 59 (9), 7211–7223. https://doi.org/10.1109/TGRS.2020.3032790.

Li, C., Jiang, H., Chen, J., Zhao, Y., Fu, S., Jing, F., Guo, Y., 2024. An overview of machine unlearning. High-Confid. Comput. 100254. https://doi.org/10.1016/j.hcc.2024.100254.

Liu, G., Zhang, R., Hang, R., Ge, L., Shi, C., Liu, Q., 2023. Statistical downscaling of temperature distributions in southwest China by using terrain-guided attention network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 1678–1690. https://doi.org/10.1109/JSTARS.2023.3239109.

Lucas-Picher, P., Argüeso, D., Brisson, E., Tramblay, Y., Berg, P., Lemonsu, A., et al., 2021. Convection-permitting modeling with regional climate models: latest developments and next steps. WIREs Clim. Change 12 (6), e731.

Mamalakis, A., Langousis, A., Deidda, R., Marrocu, M., 2017. A parametric approach for simultaneous bias correction and high-resolution downscaling of climate model rainfall. Water Resour. Res. 53 (3), 2149–2170. https://doi.org/10.1002/2016WR019578.

Mamalakis, A., Barnes, E.A., Ebert-Uphoff, I., 2022a. Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. Artif. Intelligence Earth Syst. 1 (4), e220012. https://doi.org/10.1175/AIES-D-22-0012.1.

Mamalakis, A., Ebert-Uphoff, I., Barnes, E.A., 2022b. Neural network attribution methods for problems in geoscience: a novel synthetic benchmark dataset. Environ. Data Sci. 1, e8.

Mamalakis, A., Barnes, E.A., Hurrell, J.W., 2023. Using explainable artificial intelligence to quantify "climate distinguishability" after stratospheric aerosol injection. Geophys. Res. Lett. 50 (20), e2023GL106137. https://doi.org/10.1029/2023GL106137.

Maraun, D., Wetterhall, F., Ireson, A.M., Chandler, R.E., Kendon, E.J., Widmann, M., et al., 2010. Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. Rev. Geophys. 48 (3), RG3003. https://doi.org/10.1029/2009RG000314.

Maraun, D., Widmann, M., Gutiérrez, J.M., Kotlarski, S., Chandler, R.E., Hertig, E., et al., 2015. VALUE : a framework to validate downscaling approaches for climate change studies. Earth's Future 3 (1), 1–14. https://doi.org/10.1002/2014EF000259.

Marra, F., Amponsah, W., Papalexiou, S.M., 2023. Non-asymptotic Weibull tails explain the statistics of extreme daily precipitation. Adv. Water Resour. 173, 104388. https://doi.org/10.1016/j.advwatres.2023.104388.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G.E., Elmore, K.L., Homeyer, C. R., Smith, T., 2019. Making the black box more transparent: understanding the physical implications of machine learning. Bull. Am. Meteorol. Soc. 100 (11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1.

Moszkowicz, S., 2000. Small-scale structure of rain field — preliminary results basing on a digital gauge network and on MRL-5 legionowo radar. Phys. Chem. Earth Part B 25 (10), 933–938. https://doi.org/10.1016/S1464-1909(00)00128-3.

Moustakis, Y., Papalexiou, S.M., Onof, C.J., Paschalis, A., 2021. Seasonality, intensity, and duration of rainfall extremes change in a warmer climate. Earth's Future 9 (3), e2020EF001824. https://doi.org/10.1029/2020EF001824.

Nerantzaki, S.D., Papalexiou, S.M., 2019. Tails of extremes: advancing a graphical method and harnessing big data to assess precipitation extremes. Adv. Water Resour. 134, 103448. https://doi.org/10.1016/j.advwatres.2019.103448.

Niemi, T.J., Kokkonen, T., Seed, A.W., 2014. A simple and effective method for quantifying spatial anisotropy of time series of precipitation fields. Water Resour. Res. 50 (7), 5906–5925. https://doi.org/10.1002/2013WR015190.
Nishant, N., Hobeichi, S., Sherwood, S., Abramowitz, G., Shao, Y., Bishop, C., Pitman, A., 2023. Comparison of a novel machine learning approach with dynamical downscaling for Australian precipitation. Environ. Res. Lett. 18 (9), 094006. https://doi.org/10.1088/1748-9326/ace463.
Ochoa-Rodriguez, S., Wang, L.-P., Willems, P., Onof, C., 2019. A review of radar-rain gauge data merging methods and their potential for urban hydrological applications. Water Resour. Res. 55 (8), 6356–6391. https://doi.org/10.1029/2018WR023332.
Papalexiou, S.M., 2018. Unified theory for stochastic modelling of hydroclimatic processes: preserving marginal distributions, correlation structures, and intermittency. Adv. Water Resour. 115, 234–252. https://doi.org/10.1016/j.advwatres.2018.02.013.
Papalexiou, S.M., 2022. Rainfall generation revisited: introducing CoSMoS-2s and advancing copula-based intermittent time series modeling. Water Resour. Res. 58 (6), e2021WR031641. https://doi.org/10.1029/2021WR031641.
Papalexiou, S.M., Serinaldi, F., 2020. Random fields simplified: preserving marginal distributions, correlations, and intermittency, with applications from rainfall to humidity. Water Resour. Res. 56 (2), e2019WR026331. https://doi.org/10.1029/2019WR026331.
Papalexiou, S.M., AghaKouchak, A., Foufoula-Georgiou, E., 2018. A diagnostic framework for understanding climatology of tails of hourly precipitation extremes in the United States. Water Resour. Res. 54 (9), 6725–6738. https://doi.org/10.1029/2018WR022732.
Papalexiou, S.M., Serinaldi, F., Porcu, E., 2021a. Advancing space-time simulation of random fields: from storms to cyclones and beyond. Water Resour. Res. 57 (8), e2020WR029466. https://doi.org/10.1029/2020WR029466.
Papalexiou, S. M., Serinaldi, F., Strnad, F., Markonis, Y., & Shook, K. (2021). CoSMoS: Complete Stochastic Modelling Solution. R package version 2.1.0. Retrieved from https://CRAN.R-project.org/package=CoSMoS.
Papalexiou, S.M., Serinaldi, F., Clark, M.P., 2023. Large-domain multisite precipitation generation: operational blueprint and demonstration for 1,000 sites. Water Resour. Res. 59 (3), e2022WR034094. https://doi.org/10.1029/2022WR034094.
Paschalis, A., Molnar, P., Fatichi, S., Burlando, P., 2013. A stochastic model for high-resolution space-time precipitation simulation. Water Resour. Res. 49 (12), 8400–8417. https://doi.org/10.1002/2013WR014437.
Pegram, G.G.S., Clothier, A.N., 2001. High resolution space–time modelling of rainfall: the "String of Beads" model. J. Hydrol. 241 (1), 26–41. https://doi.org/10.1016/S0022-1694(00)00373-5.
Peleg, N., Fatichi, S., Paschalis, A., Molnar, P., Burlando, P., 2017. An advanced stochastic weather generator for simulating 2-D high-resolution climate variables. J. Adv. Model. Earth Syst. 9 (3), 1595–1627. https://doi.org/10.1002/2016MS000854.
Porcu, E., Furrer, R., Nychka, D., 2020. 30 Years of space–time covariance functions. WIREs Comput. Stat. https://doi.org/10.1002/wics.1512.
Powers, J.G., Klemp, J.B., Skamarock, W.C., Davis, C.A., Dudhia, J., Gill, D.O., et al., 2017. The weather research and forecasting model: overview, system efforts, and future directions. Bull. Am. Meteorol. Soc. 98 (8), 1717–1737. https://doi.org/10.1175/BAMS-D-15-00308.1.
R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.
Rajulapati, C.R., Papalexiou, S.M., 2023. Precipitation bias correction: a novel semi-parametric quantile mapping method. Earth Space Sci. 10 (4), e2023EA002823. https://doi.org/10.1029/2023EA002823.
Rampal, N., Gibson, P.B., Sood, A., Stuart, S., Fauchereau, N.C., Brandolino, C., et al., 2022. High-resolution downscaling with interpretable deep learning: rainfall extremes over New Zealand. Weather Clim. Extremes 38, 100525. https://doi.org/10.1016/j.wace.2022.100525.
Rampal, N., Hobeichi, S., Gibson, P.B., Baño-Medina, J., Abramowitz, G., Beucler, T., et al., 2024. Enhancing regional climate downscaling through advances in machine learning. Artif. Intellig. Earth Syst. 3 (2), 230066. https://doi.org/10.1175/AIES-D-23-0066.1.
Rumelhart, D. E., Hintont, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors.
Serinaldi, F., Kilsby, C.G., 2014. Rainfall extremes: toward reconciliation after the battle of distributions. Water Resour. Res. 50 (1), 336–352.
Shaik, T., Tao, X., Xie, H., Li, L., Zhu, X., Li, Q., 2025. Exploring the landscape of machine unlearning: a comprehensive survey and taxonomy. IEEE Trans. Neural Networks Learn. Syst. 36 (7), 11676–11696. https://doi.org/10.1109/TNNLS.2024.3486109.
Sillitto, G.P., 1951. Interrelations between certain linear systematic statistics of samples from any continuous population. Biometrika 38 (3/4), 377–382. https://doi.org/10.2307/2332583.
Teutschbein, C., Seibert, J., 2013. Is bias correction of regional climate model (RCM) simulations possible for non-stationary conditions? Hydrol. Earth Syst. Sci. 17 (12), 5061–5077. https://doi.org/10.5194/hess-17-5061-2013.
Themeßl, M.J., Gobiet, A., Heinrich, G., 2012. Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal. Clim. Change 112 (2), 449–468. https://doi.org/10.1007/s10584-011-0224-4.
Toms, B.A., Barnes, E.A., Ebert-Uphoff, I., 2020. Physically interpretable neural networks for the geosciences: applications to earth system variability. J. Adv. Model. Earth Syst. 12 (9), e2019MS002002. https://doi.org/10.1029/2019MS002002.
Villarini, G., Krajewski, W.F., 2010. Review of the different sources of uncertainty in single polarization radar-based estimates of rainfall. Surv. Geophys. 31 (1), 107–129. https://doi.org/10.1007/s10712-009-9079-x.
Vogel, R.M., Papalexiou, S.M., Lamontagne, J.R., Dolan, F., 2024. When heavy tails disrupt statistical inference. Am. Statist. 1–42. https://doi.org/10.1080/00031305.2024.2402898.
Wilby, R.L., Wigley, T.M.L., Conway, D., Jones, P.D., Hewitson, B.C., Main, J., Wilks, D. S., 1998. Statistical downscaling of general circulation model output: a comparison of methods. Water Resour. Res. 34 (11), 2995–3008. https://doi.org/10.1029/98WR02577.