Comprehensive Inter-comparison of Generative AI Models for Super-Resolution Precipitation Downscaling Across Hydroclimatic Regimes

Shivam Singh*[1,2], Simon Michael Papalexiou[3,4], Hebatallah M. Abdelmoaty[5,6], Tom Hartvigsen[7], Antonios Mamalakis[2,7]

[1]Environmental Institute, University of Virginia, Charlottesville, VA, USA;
[2]Department of Environmental Sciences, University of Virginia, Charlottesville, VA, USA;
[3]Institute for Global Water Security, Hamburg University of Technology, Germany.
[4]Faculty of Environmental Sciences, Czech University of Life Sciences Prague;
[5]Department of Civil Engineering, Schulich School of Engineering, University of Calgary, Canada;
[6]Irrigation and Hydraulics Department, Faculty of Engineering, Cairo University, Egypt
[7]School of Data Science, University of Virginia, Charlottesville, VA, USA;
* Corresponding authors; Shivam Singh (wpa8me@virginia.edu), Antonios Mamalakis (npa4tg@virginia.edu)

Abstract

High-resolution precipitation information is essential for hydrologic modeling, flood forecasting, and climate-risk assessment, yet global weather and climate models operate at spatial resolutions too coarse to resolve storm structure, intermittency, and extremes. Deep-learning-based statistical downscaling provides a computationally efficient alternative to dynamical downscaling, but deterministic convolutional neural networks often yield overly smooth predictions and underestimate fine-scale variability and extreme events. Generative deep-learning models, including generative adversarial networks and diffusion models, offer a promising alternative by enabling stochastic downscaling and explicit representation of uncertainty. This study presents a systematic intercomparison of three representative deep-learning architectures for precipitation super-resolution, namely a Convolutional U-Net as baseline, a conditional Wasserstein GAN (WGAN), and a conditional Denoising Diffusion Probabilistic Model (DDPM). Using a perfect-model experimental design based on ERA5-Land precipitation fields over climatologically distinct regions of the United States, models are trained over the Central Plains and Northwest domains and evaluated over an independent Northeast test domain under 8× and 16× downscaling factors, providing a stringent test of cross-regional generalization. Evaluation diagnostics span precipitation distributions, wet–dry occurrence, extremes, spatial autocorrelation, spectral structure, and ensemble-based uncertainty quantification. All three models preserve large-scale precipitation organization, with differences emerging primarily at fine spatial scales and in the representation of extremes and spatial dependence. U-Net provides stable and computationally efficient predictions but consistently smooths fine-scale variability and suppresses extreme precipitation. WGAN improves distributional fidelity and heavy-tail behavior with comparatively modest computational overhead. DDPM yields the most physically coherent spatial structure and natural ensemble diversity for uncertainty quantification, at a substantially higher computational cost. Analysis of seed-based variability further reveals that training uncertainty dominates over stochastic generation variability, underscoring the need for multi-seed evaluation in generative downscaling systems.

Keywords: Precipitation downscaling; deep learning; generative models; super-resolution; hydrologic extremes; uncertainty quantification.

1. Introduction

Global climate models are fundamental tools for projecting future hydroclimate, yet their typical horizontal spatial resolution (often on the order of ∼100–200 km) remains too coarse to represent the mesoscale and storm-scale processes that govern precipitation intermittency and extremes (Feser et al., 2011; Palmer, 2014; Schär, 2019). In contrast, hydrologic impact modeling, flood-risk assessment, and climate adaptation planning commonly require precipitation information at finer resolution (∼10 km), where localized gradients, orographic forcing, land–sea contrasts, and convective organization must be adequately represented (Lucas-Picher et al., 2021; Nishant et al., 2023; Piani et al., 2010; Stephens, 2017). As a result, the direct application of coarse-resolution model output is inadequate for hydrologic impact assessment, flood risk estimation, and climate-risk analysis, especially in situations characterized by strong spatial intermittency and extremes (Giorgi & Gutowski, 2015; Tabari et al., 2021; Wood et al., 2004). To address this scale mismatch and provide high-resolution information required for hydrologic and climate-impact applications, downscaling techniques are employed to infer fine-scale fields from coarse-resolution model output. Dynamical downscaling, which relies on physics-based regional climate models (RCMs), has been widely used to improve the representation of mesoscale processes and precipitation extremes (Coppola et al., 2020; Giorgi & Gutowski, 2015; Giorgi & Mearns, 1991; Maraun et al., 2010). However, the substantial computational cost of RCMs severely constrains ensemble size, limits the exploration of uncertainty, and restricts their applicability for large multi-model or multi-scenario studies (Deser et al., 2012; Gao et al., 2012; Tomasi et al., 2025). These limitations have motivated growing interest in empirical downscaling approaches, including statistical and machine-learning-based methods, which offer orders-of-magnitude reductions in computational cost (Baño-Medina et al., 2020; Hobeichi et al., 2023; Lange, 2019; Mamalakis et al., 2017; Vrac et al., 2007).

Recent advances in deep learning have substantially reshaped empirical precipitation downscaling. Convolutional neural networks (CNNs), particularly end-to-end architectures such as the U-NET, have demonstrated strong skill in reproducing mean precipitation patterns, spatial continuity, and wet–dry occurrence with stable training and fast inference, making them attractive for large-scale and operational applications (Baño-Medina et al., 2020; Höhlein et al., 2020; Vandal et al., 2017). However, deterministic CNNs are typically optimized using pixel-wise loss functions that favor conditional mean solutions, resulting in overly smooth precipitation fields, reduced small-scale variability, and systematic underrepresentation of extremes, especially at high spatial resolutions and large downscaling factors (Abdelmoaty et al., 2025; Ravuri et al., 2021; Singh et al., 2026). To address these limitations, stochastic generative models have been increasingly explored to represent the inherent non-uniqueness of fine-scale precipitation conditioned on coarse inputs (Rampal et al., 2024). Generative adversarial networks (GANs), including Wasserstein GANs, have shown improved representation of fine-scale structure and extreme intensities relative to deterministic models, but their training is sensitive to hyperparameter choices and can suffer from instabilities and mode collapse, complicating robustness and calibration (Arjovsky et al., 2017; Gulrajani et al., 2017; Harris et al., 2022). More recently, diffusion-based generative models have emerged as an alternative stochastic framework, offering improved training stability and flexible uncertainty representation through a forward–reverse diffusion process, with latent diffusion variants improving computational efficiency by operating in a compressed feature space (Khader et al., 2023; Lyu et al., 2024; Tomasi et al., 2025). Early applications suggest that diffusion models can outperform both deterministic CNNs and GANs in capturing multiscale variability and spatial organization, though at substantially higher inference cost and with sensitivity to diffusion configuration (Mardani et al., 2025; X. Wang et al., 2025). Despite these advances, existing studies typically develop and evaluate deterministic, adversarial, and diffusion-based approaches in isolation, using case-specific designs and single target resolutions, leaving key questions unanswered regarding their relative performance, uncertainty representation, and computational scalability for hydrologically relevant diagnostics.

In this study, we address these gaps through a comprehensive, hydrologically oriented comparison of three representative deep-learning frameworks for daily precipitation super-resolution downscaling: a U-Net trained with pixel-wise mean squared error (MSE) loss, a Wasserstein generative adversarial network (WGAN), and a denoising diffusion probabilistic model (DDPM) in a conditional setting. Using a common training and evaluation framework, we assess their ability to reproduce precipitation occurrence, spatial structure, temporal consistency, precipitation distributions, extreme-event behavior, and uncertainty characteristics. We further investigate multi-seed and stochastic realizations to quantify within-seed and across-seed variability, providing insights into the relative contributions of model initialization and stochastic sampling to prediction uncertainty. By comparing regression-based, adversarial, and diffusion-based approaches under a consistent experimental design and emphasizing physically meaningful diagnostics, this work aims to clarify the strengths and limitations of contemporary deep-learning downscaling methods and provide guidance for hydrologic and climate-risk applications requiring reliable representation of precipitation extremes, spatial organization, temporal persistence, and predictive uncertainty.

2. Data and Methodology

2.1 Data

We use precipitation fields from the ERA5-Land reanalysis dataset (Muñoz-Sabater et al., 2021), which provides hourly accumulated precipitation at 0.1° (~9 km) spatial resolution over global land surfaces. Hourly precipitation is aggregated to daily totals and extracted over the contiguous United States (CONUS) for the period 1980–2014. This 35-year record is sufficiently long to sample a wide range of storm types, hydroclimatic variability, and interannual fluctuations. ERA5-Land is selected because it provides physically consistent land-surface precipitation estimates that are widely used in hydrologic modeling and downscaling evaluation.

To assess model behavior across distinct precipitation climatologies, we define three non-overlapping 128 × 128 grid domains (~1150 × 1150 km at mid-latitudes) representing major U.S. hydroclimatic regimes (Figure 1): (i) Central Plains (30.2°N–43.0°N, 105.4°W–92.6°W), dominated by warm-season convective precipitation; (ii) Northwest (36.6°N–49.4°N, 121.8°W–109.0°W), characterized by wintertime orographic enhancement and strong seasonal contrast; and (iii) Northeast (36.6°N–49.4°N, 90.4°W–77.6°W), influenced by synoptic-scale cyclones and frontal systems.

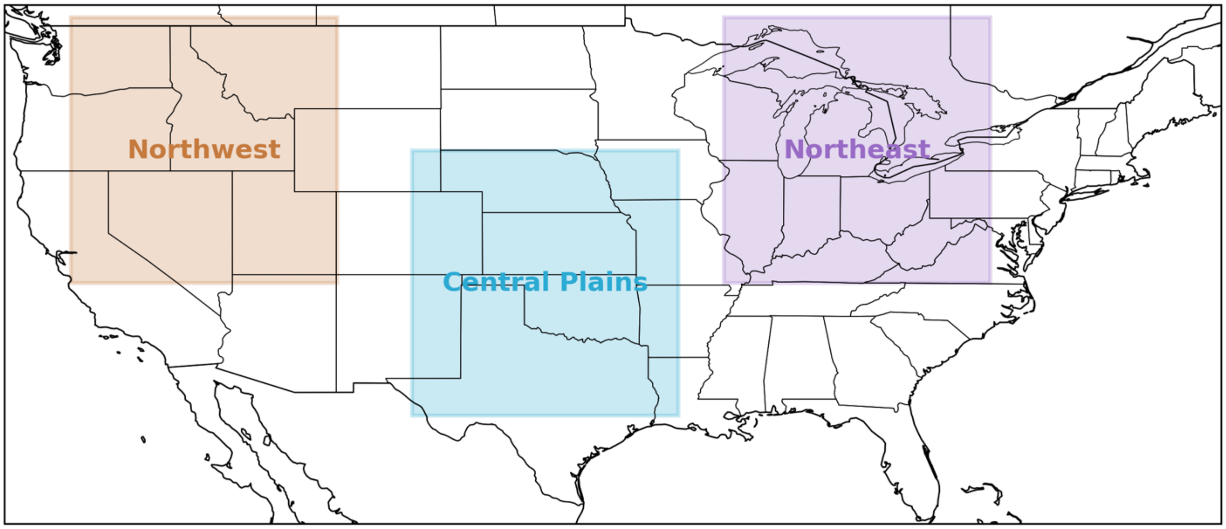

Figure 1. Geographic domains used in this study, showing the three 128 × 128 grid regions over the contiguous United States representing distinct hydroclimatic regimes: the Northwest, Central Plains, and Northeast.

A cross-regional evaluation strategy is adopted to test generalization across hydroclimatic regimes. Models are trained using samples from the Central Plains and the Northwest. The validation samples are drawn partially from the Central Plains to guide model selection and early stopping, while the Northeast samples are reserved for independent model evaluation and testing regional generalization. This setup ensures that the Northeast region is not used for parameter learning during training to keep it as a strictly independent test set (Rampal et al., 2024; Vandal et al., 2017).

Daily precipitation values below 1 mm/day are treated as dry and set to zero, following commonly used wet-day thresholds in hydro-climatological analyses (Teutschbein & Seibert, 2012; Trenberth et al., 2015). To ensure that learning is driven by meaningful spatial rainfall structure rather than near-empty scenes, days with fewer than 1% wet pixels (<164 wet pixels in a 128 × 128 domain) are excluded. This filtering removes 13.76%, 8.11%, and 3.41% of samples in the Central Plains, Northwest, and Northeast regions, respectively, with the highest removal occurring during winter (DJF) and autumn (SON), and minimal removal during summer (JJA). This indicates that the filtering primarily excludes dry and weak precipitation events. While this step improves training stability, it may lead to an underrepresentation of light rainfall and a slight reduction in wet-day occurrence frequency. After preprocessing, 11,025 samples are retained for the Central Plains, 11,747 for the Northwest, and 12,348 for the Northeast. The dataset is then split into 19,200 training samples, 3,304 validation samples, and 12,348 test samples.

Low-resolution inputs are generated through block averaging, producing 8× and 16× aggregated precipitation fields while conserving storm-total rainfall volume. Block averaging is preferred over interpolation because it preserves physical mass consistency and avoids introducing artificial spatial correlations (Hsu et al., 2024; Kumar et al., 2023; Stengel et al., 2020). For the 8× configuration, 128 × 128 fields (~9 km) are aggregated to 16 × 16 (~72 km) and models are trained to reconstruct the corresponding 128 × 128 target. For the 16× configuration, targets are reconstructed from 8 × 8 (~144 km) inputs. This design defines a perfect-model super-resolution framework in which inputs and targets originate from the same dataset, enabling controlled evaluation of spatial refinement independent of predictor mismatch or bias-correction effects.

2.2 Models

In this study we used a deterministic convolutional U-NET as a baseline and two generative models WGAN and DDPM for downscaling precipitation. A schematic of these deep-learning architectures is presented in Figure 2 and detailed architectures including total numbers of earnable parameters are included in supplementary Figure S1 and Table S1.

2.2.1 U-NET

A U-NET architecture is widely used in deep learning-based super-resolution downscaling experiments (Abdelmoaty et al., 2025; Papalexiou & Mamalakis, 2025; Wang et al., 2021). The model takes as input a coarse precipitation field x_{LR} of size 16×16 for the 8× downscaling task or of size 8×8 for the 16× downscaling task, together with a spatially uncorrelated noise tensor $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ of identical dimensions. The noise input is included solely to maintain architectural compatibility with the stochastic generator used in the WGAN framework, but since the network is trained independently using a mean-squared error (MSE) loss, it converges to the conditional mean of the high-resolution target given the coarse input and therefore exhibits deterministic behavior during inference (Lakshminarayanan et al., 2017; Yan et al., 2019). The MSE loss is expressed as:

$$\mathcal{L}_{\mathrm{MSE}} = \mathbb{E}[\| G_\theta(x_{\mathrm{LR}}, \mathbf{z}) - x_{\mathrm{HR}} \|_2^2] \quad (1)$$

where G_θ denotes the U-NET mapping parameterized by θ, and x_{HR} represents the corresponding high-resolution target field.

Architecturally, the encoder consists of two down-sampling stages with 32 and 64 filters, each containing pairs of 3×3 convolutional layers with ReLU activations, followed by 2×2 max-pooling layers. The bottleneck includes two convolutional layers with 128 filters. The decoder employs transposed convolution up-sampling and incorporates skip connections from the encoder to preserve fine-scale spatial organization. To reach the target 128×128 resolution from the coarse input, three progressive up-sampling stages are applied (32 → 64 → 128), using 2×2 transposed convolutions with Leaky ReLU activations, followed by refinement convolutions at full resolution. The output layer uses a linear activation to produce precipitation intensity in mm/day. The model is trained end-to-end using the Adam optimizer (learning rate 1×10^{-4}, beta_1 = 0.9, beta_2 = 0.999, batch size 32) with early stopping when validation loss fails to improve for ten consecutive epochs (patience = 10). No weight decay is used, and the best checkpoint based on validation set MSE is retained for evaluation.

2.2.2 Wasserstein GAN (WGAN)

To enable stochastic and spatially realistic precipitation downscaling, we implement a conditional Wasserstein GAN (WGAN), following (Arjovsky et al., 2017; Gulrajani et al., 2017; Papalexiou & Mamalakis, 2025). The generator, which is the same noise-conditional U-NET described above, maps a coarse precipitation field x_{LR} together with a spatial noise tensor $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to a high-resolution field $\hat{x}_{\mathrm{HR}}$ on the 128×128 grid. Unlike the MSE-trained U-NET, which converges to the conditional mean, adversarial optimization enables sampling from the conditional distribution of plausible high-resolution rainfall structures conditioned on the same coarse input. The critic C_ψ serves as a conditional discriminator and evaluates the realism of the generated field given the coarse rainfall context. It consists of two pathways: an encoder that reduces the 128×128 field through strided convolutions, and an embedding of the low-resolution input into a matching spatial feature representation. These feature streams are fused and reduced to a scalar score, enabling the critic to assess global storm morphology while remaining aware of the large-scale meteorological state.

Let $x_{\mathrm{HR}} \sim P_r$ denote real high-resolution samples and $\hat{x}_{\mathrm{HR}} = G_\theta(x_{\mathrm{LR}}, \mathbf{z}) \sim P_g$ denote generated samples. The conditional Wasserstein critic loss is:

$$\mathcal{L}_{\mathrm{C}} = \mathbb{E}_{\hat{x}_{\mathrm{HR}} \sim P_g}\left[C_\psi(\hat{x}_{\mathrm{HR}}, x_{\mathrm{LR}})\right] - \mathbb{E}_{x_{\mathrm{HR}} \sim P_r}\left[C_\psi(x_{\mathrm{HR}}, x_{\mathrm{LR}})\right] + \lambda\, \mathbb{E}_{\tilde{x} \sim P_{\tilde{x}}}\left[\left(\| \nabla_{\tilde{x}} C_\psi(\tilde{x}, x_{\mathrm{LR}}) \|_2 - 1\right)^2\right] \quad (2)$$

where, interpolated sample $(\tilde{x})$ is defined as,

$$\tilde{x} = \in x_{\mathrm{HR}} + (1 - \in)\hat{x}_{\mathrm{HR}}, \in \sim U(0, 1) \quad (3)$$

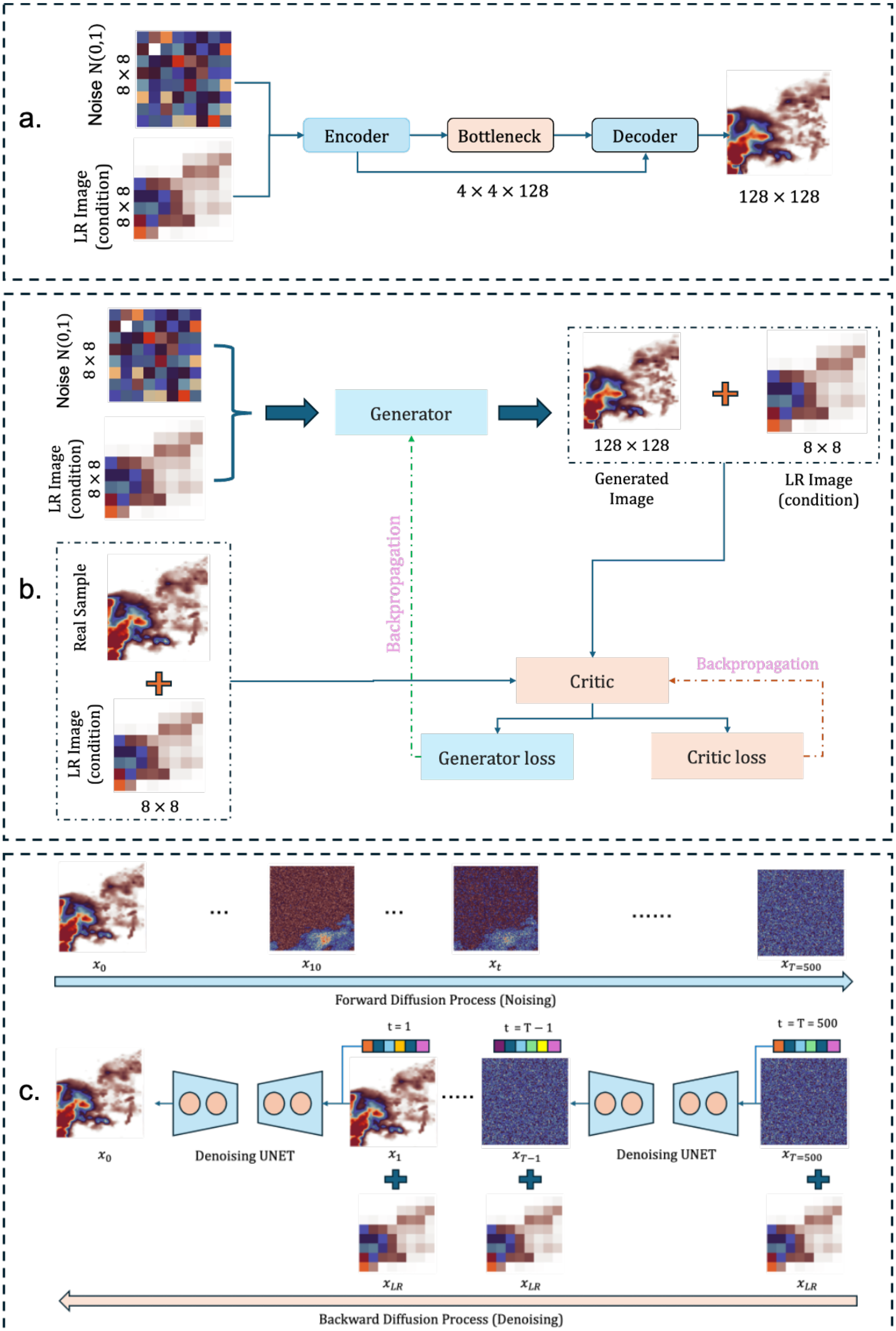
a.
Noise N(0,1)
8 × 8
LR Image (condition)
8 × 8
Encoder
Bottleneck
Decoder
4 × 4 × 128
128 × 128
b.
Noise N(0,1)
8 × 8
LR Image (condition)
8 × 8
Generator
128 × 128
8 × 8
Generated Image
LR Image (condition)
Real Sample
LR Image (condition)
8 × 8
Backpropagation
Critic
Backpropagation
Generator loss
Critic loss
c.
x_0
x_{10}
x_t
$x_{T=500}$
Forward Diffusion Process (Noising)
t = 1
t = T − 1
t = T = 500
x_0
Denoising UNET
x_1
x_{T-1}
Denoising UNET
$x_{T=500}$
x_{LR}
x_{LR}
x_{LR}
Backward Diffusion Process (Denoising)

Figure 2: Schematic of the deep-learning architectures used for precipitation super-resolution downscaling (16×). (a) Deterministic U-NET with an encoder–decoder structure and skip connections producing a single high-resolution output. (b) Conditional Wasserstein GAN (WGAN), where a stochastic generator produces high-resolution fields conditioned on low-resolution input and noise and is trained adversarially against a critic. (c) Conditional denoising diffusion probabilistic model (DDPM), which reconstructs high-resolution precipitation through an iterative, noise-to-signal denoising process conditioned on the coarse input.

The generator is trained to maximize the critic score for the generated samples, corresponding to:

$$\mathcal{L}_{\mathrm{G}} = -\mathbb{E}_{\hat{x}_{\mathrm{HR}} \sim P_g}\left[C_\psi(\hat{x}_{\mathrm{HR}}, x_{\mathrm{LR}})\right] \tag{4}$$

The model is trained using a gradient penalty weight $\lambda = 10$, updating the critic three times for each generator update. Both networks use the Adam optimizer with learning rate 1×10^{-4}, $\beta_1 = 0.0$, and $\beta_2 = 0.9$, with no weight decay. To characterize epistemic variability, ten WGAN models independently initialized with different random seeds are trained for 200 epochs, producing ensembles of generated samples for each coarse precipitation input.

2.2.3 Denoising Diffusion Probabilistic Model (DDPM)

To exploit recent advances in likelihood-based generative modeling for high-resolution precipitation reconstruction, we implement a conditional DDPM following the framework proposed by Ho et al. (2020). In their framework, the model learns to reverse a fixed forward diffusion process that progressively perturbs high-resolution (HR) precipitation fields with Gaussian noise over $\mathrm{T} = 500$ steps. During training, the network predicts the added noise at a randomly sampled timestep t, conditioned on the corresponding low-resolution (LR) precipitation field (either 16×16 or 8×8). At inference, samples are obtained by iteratively denoising pure Gaussian noise while conditioning on the LR field. The forward process is defined as:

$$\mathrm{q}(x_{1:T} \mid x_0) = \prod_{t=1}^{T} \mathrm{q}(x_t \mid x_{t-1}) \tag{5}$$

$$\mathrm{q}(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathrm{I}\right) \tag{6}$$

where $\{\beta_t\}_{t=1}^{T}$ is the noise schedule. We adopt the cosine noise schedule (Song & Dhariwal, 2023) to improve sample smoothness and reduce denoising artifacts. The cumulative signal retention is given by

$$\bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s) \tag{7}$$

$$\tilde{\alpha}_t = \frac{\cos^2\left(\frac{T/t + s}{1 + s} \cdot \frac{\pi}{2}\right)}{\cos^2\left(\frac{s}{1 + s} \cdot \frac{\pi}{2}\right)} \tag{8}$$

with a small offset $s = 0.008$ to avoid extreme noise ratios. At training time, the model predicts the noise ϵ added to x_0 at timestep t, conditioned on the LR field x_{LR}:

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon,t}[||\epsilon - \epsilon_\theta(x_t, t, x_{\mathrm{LR}})||_2^2] \tag{9}$$

where $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$. The network backbone is a conditional U-NET augmented with sinusoidal time embeddings and Feature-wise Linear Modulation (FiLM; (Perez et al., 2018) layers to inject timestep context and LR conditioning into intermediate feature representations. Time is encoded using 256-dimensional sinusoidal embeddings,

$$\gamma(\mathrm{t}) = \left[\sin\left(\frac{t}{10000^{2i/256}}\right), \cos\left(\frac{t}{10000^{2i/256}}\right)\right]_{i=0}^{127} \tag{10}$$

and refined using a multilayer perceptron to produce a contextualized embedding $\mathrm{t}_{\mathrm{emb}}$. Within each convolutional block, FiLM modulates internal activations according to the timestep and LR conditioning:

$$\mathrm{FiLM}(h, t_{emb}) = \gamma(t_{emb}) \odot h + \beta(t_{emb}) \tag{11}$$

This allows the model to adapt its representations to different stages of the denoising trajectory. To condition DDPM, the LR field is up-sampled to 128×128 via bilinear interpolation and concatenated with the noisy HR input, ensuring that the coarse-scale spatial context informs the fine-scale reconstruction. Unlike the U-Net and WGAN models, which were trained without an additional target normalization step, DDPM training employed a log(1+x) transformation followed by min-max normalization. This normalization was introduced to improve numerical stability during the iterative diffusion and denoising process. During training, we minimize the noise-prediction loss using the AdamW optimizer (learning rate 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 1×10^{-4}). During inference, HR precipitation samples are generated by iteratively applying the reverse diffusion update,

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(x_t, t, c)\right) + \sigma_t\mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(0, \mathrm{I}) \tag{12}$$

for all t from T to 1, injecting noise at each step except the final one. This additional noise makes the denoising process more stochastic and for same LR input, keeping all other parameters same, we get different downscaled output.

All three models (U-NET, WGAN, and DDPM) were trained using 10 different random seeds, yielding 10 independently initialized and trained realizations of each architecture. For a given low-resolution input, one prediction was generated from each trained instance, forming a 10-member ensemble. This ensemble primarily reflects epistemic uncertainty arising from random weight initialization during training. To ensure a fair comparison across downscaling factors, the core network architectures were kept fixed across experiments. For the deterministic U-NET and the WGAN, the same generator and critic architectures were used for both 8× and 16× configurations. The increased difficulty of the 16× case was introduced solely by providing coarser inputs and by bilinearly interpolating 8×8 fields to 16×16 before being passed to the networks, while keeping the target resolution at 128×128.

For the DDPM, the same denoising U-NET architecture was used for both scaling configurations. In the 8× setup, conditioning fields were bilinearly interpolated from 16×16 to 128×128 through three successive 2× interpolations, consistent with the diffusion model's multi-scale refinement process. In the more challenging 16× setup, conditioning fields originated from 8×8 inputs and were similarly interpolated to 128×128 through repeated (four-times) 2× bilinear interpolation steps before being concatenated with the noisy target field at each diffusion timestep. This design ensured that differences in performance across scaling factors reflect the increased information gap in the input, rather than changes in model capacity or architecture. By holding network architectures fixed and varying only the effective resolution of the conditioning input, this experimental setup enables a controlled and equitable comparison of deterministic and generative models across downscaling factors. This implementation allows direct comparison with U-NET and WGAN models under identical input configurations and downscaling ratios (8× and 16×) isolating differences attributable to generative formulation rather than architectural capacity or preprocessing design.

3. Performance Evaluation

To evaluate the statistical fidelity, spatial realism, and hydrologically relevant characteristics of the downscaled precipitation fields produced by the U-NET, WGAN, and DDPM models, we employ a set of complementary evaluation measures. These measures examine distributional consistency, representation of extremes and dry occurrence, rainfall mass conservation across spatial scales, spatial organization and storm morphology, and binary precipitation detection skill. All analyses are conducted on a test set of 12,348 daily precipitation fields from the unseen Northeast hydroclimatic region. For each modeling framework, results are computed across a 10-member ensemble consisting of independently trained models initialized with different random seeds.

3.1 Distributional Consistency

3.1.1 Quantile–Quantile (Q–Q) Analysis

Quantile behavior is assessed by comparing the p-quantile of model predictions (Q_p^{Model}) with the corresponding target (ERA5-land) quantile (Q_p^{Target}).

$$Q_p^{\text{Model}} = F_{\text{Model}}^{-1}(p) \tag{13}$$

$$Q_p^{\text{Target}} = F_{\text{Target}}^{-1}(p) \tag{14}$$

where $p \in [0,1]$ and F^{-1} denotes the empirical inverse of the cumulative distribution function (CDF). When plotted together, alignment with the 1:1 line indicates agreement in distributional shape, while deviations at high p quantify errors in extreme precipitation intensity representation.

3.1.2 Exceedance Probability

Extreme rainfall representation is examined using the complementary CDF and evaluated across intensity thresholds x. This emphasizes tail performance (e.g., $>$ 10 mm/day), which is critical for hydrologic risk estimation. The probability that exceeds the threshold (x) is expressed as:

$$\mathbb{P}(X^{\text{Model}} > x) = 1 - F_{\text{Model}}(x) \tag{15}$$

where F(x) is CDF of rainfall intensity.

3.1.3 Probability of Zero Precipitation (P_0)

Dry–wet occurrence skill is evaluated using the probability of zero (or near-zero) precipitation, defined as:

$$P_0^{\text{Model}} = \frac{1}{N} \sum_{i=1}^{N} 1(X_i^{\text{Model}} \leq x_{\text{th}}) \tag{16}$$

where N is the total number of evaluated grid cells (pixels), X_i is pixel intensity, $x_{\text{th}} =$ 1 mm/day. This metric quantifies how often a model predicts dry conditions and is essential for diagnosing dry bias.

3.1.4 Higher-Order Statistical Moments

To characterize the statistical properties of precipitation distributions, we compute the first four L-moments at each grid cell using precipitation values exceeding 1 mm/day. Specifically, we analyze the mean (first L-moment), L-dispersion (second L-moment), L-skewness, and L-kurtosis, which describe the central tendency, variability, asymmetry, and tail behavior of the precipitation distribution, respectively (Papalexiou & Mamalakis, 2025). These statistics are calculated independently for the Target (ERA5-Land) and model-generated precipitation fields and compared using pixel-wise scatter plots, bias, and RMSE metrics.

3.2 Mass Conservation

3.2.1 Cumulative Mean Rainfall Depth

We evaluate whether precipitation depth is preserved across spatial aggregation scales during downscaling. For an aggregation scale s, the precipitation field is partitioned into non-overlapping $s \times s$ blocks. The mean precipitation depth within block j is computed as

$$X_{s,j} = \frac{1}{s^2} \sum_{i \in j} P_i \tag{17}$$

where P_i denotes the precipitation value at grid cell i, and $X_{s,j}$ represents the mean precipitation depth within block j at aggregation scale s. This quantity is computed independently for both the Target (ERA5-Land) and model-generated precipitation fields.

The resulting block-mean precipitation depths are compared across aggregation scales using the Pearson correlation coefficient (r), mean bias, and root-mean-square error (RMSE). High correlation together with low bias and RMSE indicates that the downscaling model preserves precipitation depth consistently across spatial scales, thereby maintaining the large-scale rainfall characteristics of the target dataset.

3.3 Spatial Structure and Storm Morphology

3.3.1 Lagged Spatial Autocorrelation

Spatial dependence in precipitation fields is evaluated using lagged spatial autocorrelation, which quantifies the similarity of precipitation values separated by a given spatial lag (Papalexiou et al., 2021). For a specified lag distance d, the lagged autocorrelation is computed as the Pearson correlation coefficient between paired precipitation values sampled at locations separated by d. Specifically, the statistic is defined as

$$r(d) = \frac{\sum_{i=1}^{N} (X(s_i) - \bar{X})\,(X(s_i + d) - \bar{X})}{\sqrt{\sum_{i=1}^{N} (X(s_i) - \bar{X})^2 \, \sum_{i=1}^{N} (X(s_i + d) - \bar{X})^2}}, \qquad (18)$$

where X denotes either X^{Target} or X^{Model}, $\bar{X}$ is the sample mean of the precipitation field, s_i denotes a spatial location, d is the lag, and N is the number of valid grid-level pairs. The statistic is computed for a range of lag distances and averaged over all valid pairs at each lag.

3.3.2 Fraction Skill Score (FSS)

Spatial coherence and storm organization are assessed using the Fraction Skill Score (FSS), a window-based metric that compares the fractional rainfall coverage in predictions and targets (Gilleland et al., 2009; Roberts & Lean, 2008). For a window of size w, the score is defined as

$$\text{FSS}(w) = 1 - \frac{\mathbb{E}\left[(P_{\text{Model}} - P_{\text{Target}})^2\right]}{\mathbb{E}\left[P_{\text{Model}}^2 + P_{\text{Target}}^2\right] + \varepsilon} \qquad (19)$$

where P_{Model} and P_{Target} denote the fractional rainfall coverage within windows of size w computed from the model prediction and the reference target precipitation, respectively. FSS ranges from 0 (no skill) to 1 (perfect spatial agreement), making it particularly effective for diagnosing displacement errors, spatial smoothing, and the realism of storm geometry.

3.3.3 Radial Power Spectrum

Scale-dependent spatial variability in downscaled predictions is evaluated using the radially averaged Fourier power spectrum, which characterizes how energy is distributed across spatial wavenumbers (Bednarz & Cherukuri, 2023; Harrison et al., 2025; Skamarock, 2004). The 2-D discrete Fourier transform of the rainfall field,

$$F(k_x, k_y) = \mathcal{F}\{X(x, y)\} \qquad (20)$$

where $\mathcal{F}\{\cdot\}$ denotes the 2-D discrete Fourier transform, yields the spectral power,

$$P(k_x, k_y) = |\, F(k_x, k_y) \,|^2 \qquad (21)$$

which is then azimuthally averaged to obtain the one-dimensional spectrum $P(k)$. Agreement between predicted and targeted spectra indicates that the model accurately captures multiscale storm structure from large synoptic gradients to mesoscale organization and fine-scale convective patterns. Whereas deviations reveal scale-specific biases such as excessive smoothing, noise amplification, or loss of small-scale variability.

3.3.5 ROC Curve and AUC

Binary precipitation-occurrence skill is assessed using Receiver Operating Characteristic (ROC) analysis based on a random threshold (here, wet/dry threshold of $x = 1$ mm/day) (Harris et al., 2022). The Area Under the Curve (AUC) satisfies Eq. (22).

$$0.5 \leq \text{AUC} \leq 1 \tag{22}$$

where AUC $= 1$, denotes perfect discrimination between wet and dry events, and AUC $= 0.5$ indicates no discriminative ability (random chance). This metric evaluates ability of models to correctly identify rainfall occurrence independently of intensity, making it particularly useful for diagnosing dry–wet classification bias.

3.4 Temporal correlation

To evaluate temporal consistency of precipitation fields, we construct time series at each grid cell by considering the sequence of spatial fields as a temporal evolution. Let $P_t(i,j)$ denote precipitation at grid cell (i,j) and time t. The temporal correlation at each grid cell is defined as:

$$r_{i,j}(\tau) = \text{corr}(P_t(i,j), P_{t+\tau}(i,j)) \tag{23}$$

where τ is the temporal lag. This results in a distribution of correlation values across all spatial locations for each lag.

4. Computational Requirements

All model trainings were conducted on a single NVIDIA A100 GPU (80 GB memory). Each architecture was trained for 200 epochs using 19,200 training samples and tested on 12,348 test samples. Substantial differences in computational demand were observed across the three model classes (Table 1). The U-NET baseline completed training in approximately 16 minutes, reflecting the efficiency of direct supervised optimization with a single forward–backward pass per batch. In contrast, the WGAN required 1 hour and 12 minutes to train, driven primarily by the need to update the critic multiple times per generator update and to compute the gradient-penalty term that enforces the 1-Lipschitz constraint. The DDPM exhibited the highest training cost, requiring 2 hours and 29 minutes, since each optimization step involves predicting injected noise over a sequence of 500 diffusion timesteps.

Table 1: Comparison of training and inference times for U-NET, WGAN, and DDPM models on a single NVIDIA A100 (80 GB) GPU

Resource	Model	Training	Inference (12,348 samples)	Per-Sample Time	Remarks
A100-80GB GPU (1)	U-NET	~16 min	~1.33 sec	~0.0001sec	Fastest, single forward pass
	WGAN (Generator)	~1h 12min	~1.32 sec	~0.0001sec	Multiple critic steps
	DDPM (T=500)	~2h 29min	~1h 38 min	~0.5 sec	500 denoising steps per image
	DDPM (T=100)	~2h 23min	~20 min 59 sec	~0.1 sec	100 denoising steps per image
	DDPM (T=50)	~2h 21min	~10 min 27 sec	~0.05 sec	50 denoising steps per image

Inference performance showed an even stronger divergence. Because the U-NET and WGAN generators share identical architectures and an equal number of trainable parameters (differing only in training objectives), their inference times were nearly identical, requiring 1.33 seconds and 1.32 seconds, respectively, to process the full test set (12,348 samples). In contrast, the DDPM required 1 hour and 38 minutes for inference, since high-resolution precipitation fields are generated through an iterative reverse-diffusion sampling procedure that sequentially refines noise over 500 denoising steps. These differences underscore the computational trade-offs between deterministic convolutional downscaling and likelihood-based generative modeling. While WGAN offers stochastic outputs with moderate added computational burden, DDPM provides calibrated ensemble diversity at substantially higher computational cost, a factor that may strongly influence operational deployment, ensemble forecasting, and climate-model downscaling at scale.

5. Results and Discussion

5.1 Training behavior

The three downscaling architectures exhibit distinct optimization characteristics that directly reflect their learning objectives and model structures. The deterministic U-NET converges most rapidly under both the 8× and 16× super-resolution configurations, with training and validation losses decreasing sharply and stabilizing within the first few epochs (supplementary Figure S2a). The near-perfect overlap between training and validation curves indicates negligible overfitting and strong generalization, consistent with optimization under a mean-squared-error objective that drives the network toward a conditional mean solution. Final loss values are systematically higher for the 16× configuration, reflecting the larger information gap between the coarse 8×8 inputs and the high-resolution targets.

The WGAN displays the characteristic oscillatory dynamics associated with adversarial training. For the 8× case, the critic loss stabilizes within a narrow negative range while the generator loss fluctuates around a stationary mean, indicating a sustained adversarial equilibrium and effective enforcement of the 1-Lipschitz constraint through gradient penalty regularization (supplementary Figure S2b). Under the more challenging 16× configuration, critic and generator losses exhibit increased variability, reflecting the greater difficulty of discriminating realistic structure when the conditioning input contains extremely limited spatial information. Nevertheless, the loss trajectories remain bounded across all seeds, with no evidence of training instability or mode collapse. It is important to note that adversarial loss values are not directly comparable to conventional supervised validation losses and are often only weakly correlated with sample quality. Therefore, WGAN convergence was assessed not only through adversarial losses but also through independent diagnostics, including SSIM, exceedance probability characteristics, spatial correlation statistics, and qualitative sample realism. Collectively, these results indicate stable adversarial training across both super-resolution configurations.

The DDPM shows the most monotonic and stable optimization behavior among the three models. The noise-prediction loss decreases smoothly for both super-resolution factors, and training and validation curves remain nearly indistinguishable throughout training (supplementary Figure S2c), indicating strong generalization. Convergence is slower than U-NET and WGAN, reflecting the iterative denoising process and timestep conditioning intrinsic to diffusion models. As expected, loss values are slightly higher for the more challenging 16× configuration, where substantially less information is available in the coarse-resolution input. Nevertheless, training remains stable across all seeds, underscoring the robustness of likelihood-based diffusion models for learning structured precipitation fields under severe information constraints.

5.2 Model Performance Evaluation

5.2.1 Visual Reconstruction of Precipitation Fields

Visual inspection of reconstructed precipitation fields reveals clear performance differences between the 8× and 16× downscaling tasks (supplementary Figure S3; Figure 3). Under the

8× configuration, where coarse inputs retain recognizable storm-scale organization, all three models successfully recover the dominant spatial structure of precipitation events. The U-NET produces smooth, spatially coherent fields but systematically reduces sharp gradients and localized convective maxima. In contrast, WGAN outputs exhibit sharper boundaries and more textured rainfall patterns, consistent with adversarial training encouraging the reconstruction of high-frequency spatial variability. DDPM reconstructions closely resemble those of WGAN in terms of storm morphology and spatial extent.

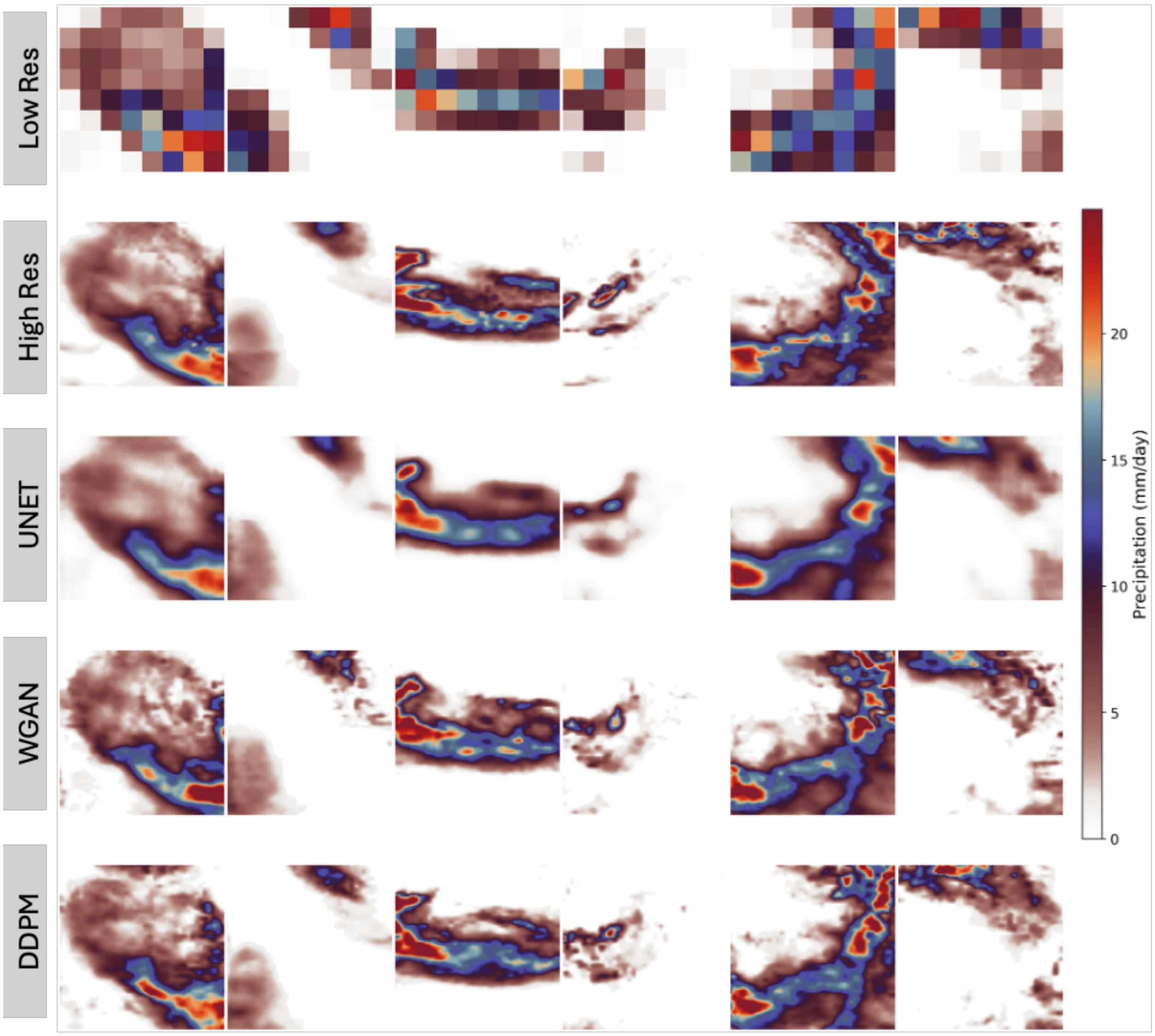

Figure 3. Reconstruction of high-resolution precipitation fields for the 16× downscaling task (8×8 → 128×128). Columns show six randomly selected test samples, while rows show the low-resolution input, high-resolution target (ERA5-land), and predictions from U-NET, WGAN, and DDPM respectively.

Performance degradation becomes evident for all models under the 16× configuration, reflecting the extremely limited information content of the 8×8 conditioning inputs. In this case, U-NET predictions increasingly collapse toward overly smoothed and diffuse rainfall fields, frequently failing to recover localized storm features. Both generative models retain substantially better spatial organization than the U-NET, although fine-scale detail is reduced relative to the 8× case. Both WGAN and DDPM recover fine-scale precipitation structures while maintaining spatially coherent storm morphology that closely matches the target precipitation patterns. These visual results highlight the growing limitations of deterministic regression under extreme downscaling and demonstrate the advantage of generative models in reconstructing physically plausible precipitation patterns when conditioning information becomes severely constrained. For completeness, a qualitative comparison with a bilinear interpolation baseline is provided in the supplementary Figure S4 of the Supplementary Material, illustrating the added value of learned super-resolution models beyond conventional interpolation.

5.2.2 Scale-Dependent Rainfall Depth Consistency (Mass Conservation)

To assess whether the generated precipitation fields preserve precipitation depth across spatial scales, we compared block-mean precipitation depth between the target (ERA5-Land) and the generated fields at aggregation scales ranging from 1×1 to 128×128 pixels for both the 8× and 16× downscaling experiments (Figure 4 and supplementary Figure S5, respectively). For each aggregation scale, precipitation was averaged within non-overlapping blocks, and the resulting block-mean values from the generated fields were compared against the corresponding target values using bias, root-mean-square error (RMSE), and Pearson correlation coefficient. At the native resolution (1×1), all models exhibit noticeable scatter around the one-to-one line, reflecting differences in local precipitation intensity and spatial placement. These discrepancies are more pronounced for the 16× downscaling task, which involves reconstructing substantially finer-scale structure from a coarser input field. However, as the aggregation scale increases, the scatter progressively decreases, correlations approach unity, and RMSE is substantially reduced for all models. This behavior indicates that a large fraction of the disagreement originates from fine-scale spatial variability, whereas precipitation depth is reproduced much more accurately at mesoscale and regional scales.

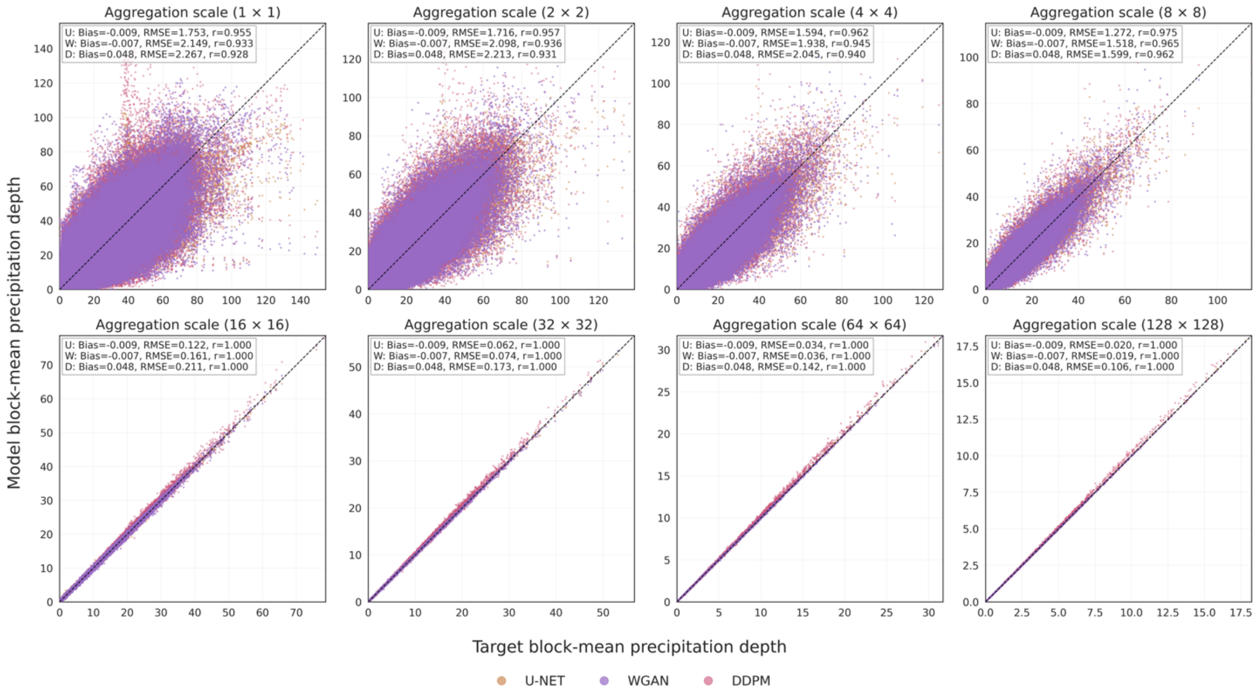

Figure 4. Mass (rainfall depth) consistency across spatial aggregation scales for 16× downscaling. Scatter plots compare block-averaged predicted and target (ERA5-land) precipitation depths at aggregation scales of 1×1, 2×2, 4×4, 8×8, 16×16, 32×32, 64×64, and 128×128. Each point represents a spatially aggregated precipitation value computed over the corresponding block size. Results are shown for single seed initialized U-NET (orange), WGAN (purple), and DDPM (pink). The dashed black line denotes perfect 1:1 agreement. Panel annotations report model-specific mean bias, root-mean-square error (RMSE), and Pearson correlation coefficient (r) at each scale. The analysis was performed using 2000 randomly selected samples from the test set (Northeast region) and one representative model initialization for each generative framework due to the computational expense associated with evaluating the full test dataset and multiple model initializations.

The bias remains nearly constant across aggregation levels for each model. This is expected because spatial aggregation preserves the domain-averaged precipitation amount and therefore has limited influence on systematic precipitation-depth biases. Consequently, the reduction in RMSE with increasing aggregation scale primarily reflects the smoothing of local-scale errors rather than changes in the overall precipitation budget. By aggregation scales of approximately 8×8 pixels and larger, all three models exhibit near-perfect correspondence with the target precipitation depth, with correlations approaching 1.0 and only minimal residual errors. These results demonstrate that the U-Net, WGAN, and DDPM largely conserve precipitation depth across spatial scales, while most prediction errors are associated with the representation of fine-scale precipitation structure rather than systematic biases in the total precipitation amount. Consistent behavior across both the 8× and 16× experiments further indicates that the models maintain physically coherent precipitation totals despite the increased difficulty of higher downscaling.

5.2.3 Statistical Distribution and Storm Morphology

Accurate precipitation downscaling requires preserving not only precipitation occurrence and intensity but also the higher-order statistical characteristics of precipitation fields. To evaluate the statistical fidelity of the generated fields, the probability of dry pixels (P_0), mean precipitation, second L-moment, L-skewness, and L-kurtosis were computed for each generated 128×128 precipitation field and compared against the corresponding ERA5-Land target field. Statistics were evaluated across all generated realizations obtained from the 10 independently initialized model seeds.

All three models reproduced the probability of dry pixels with high fidelity, indicating accurate representation of precipitation occurrence and dry-wet transitions. For the 8× experiment (supplementary Figure S6), DDPM achieved the lowest bias and RMSE (0.35% and 0.67%), followed by WGAN (1.17% and 1.47%) and U-Net (2.39% and 2.77%). A similar ranking was observed for the more challenging 16× experiment (Figure 5), where DDPM again produced the smallest errors (bias = 0.58%, RMSE = 1.82%). Mean precipitation and second L-moment were also reproduced accurately across all models, with small biases and low RMSE values, demonstrating strong preservation of precipitation intensity and variability.

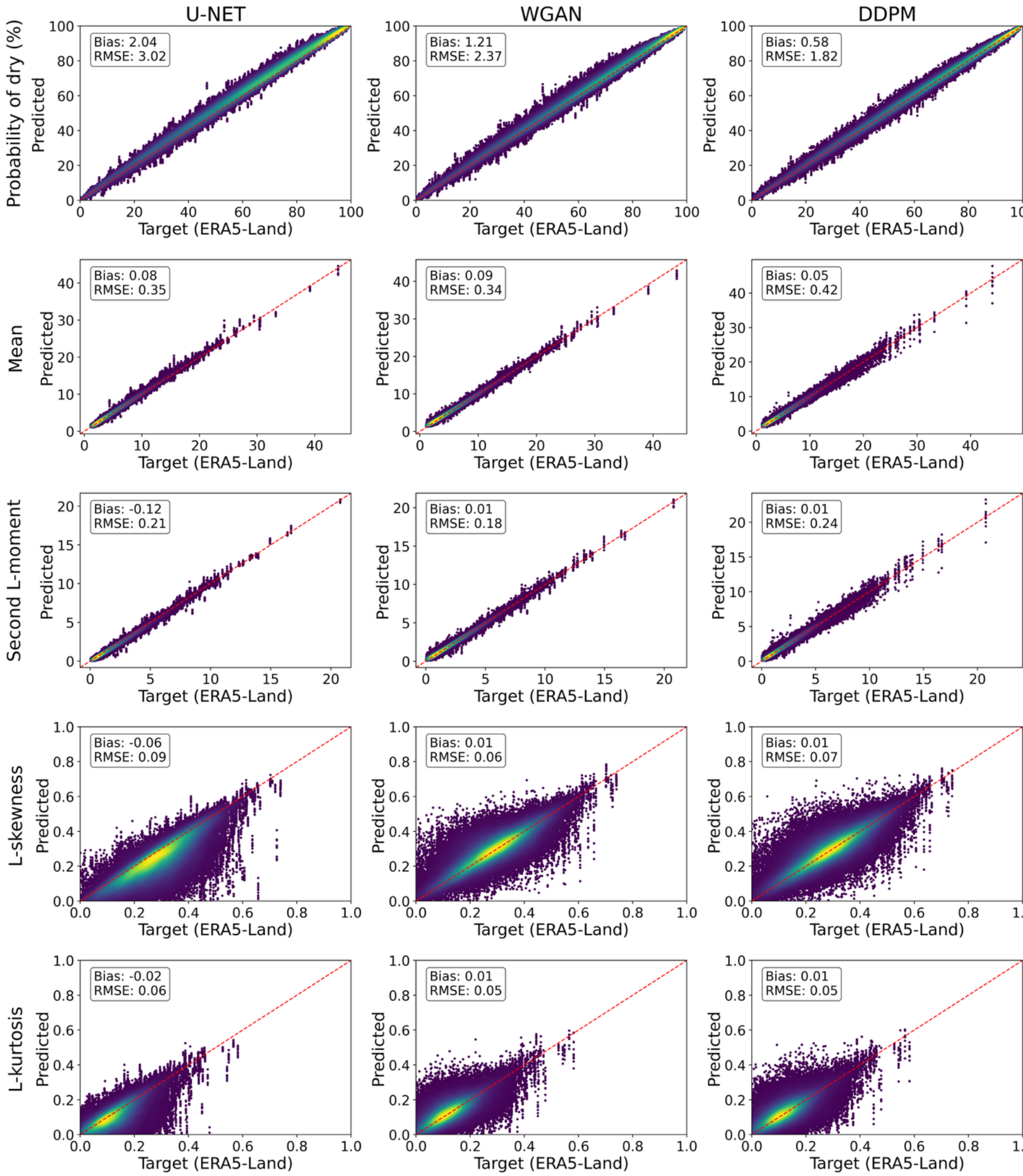

Figure 5. Comparison of precipitation statistics from high-resolution ERA5-Land reanalysis and model predictions for the 16× downscaling experiment across U-NET, WGAN, and DDPM. Each panel compares reference statistics computed from native high-resolution ERA5-Land fields (x-axis) with corresponding model predictions (y-axis), including dry-pixel probability (P_0), mean, and second to fourth central moments. Multiple seed-based predictions are evaluated against the same reference samples, leading to repeated x-axis values. Bias and RMSE are shown in each panel.

Greater differences among the models emerged for the higher-order L-moments. Although L-skewness and L-kurtosis exhibited larger scatter than the lower-order statistics due to their sensitivity to localized extremes and tail behavior, biases remained close to zero and RMSE values remained relatively small in both experiments. WGAN and DDPM generally produced lower errors than U-Net, particularly for the 16× downscaling task, indicating a modest improvement in reproducing precipitation distribution asymmetry and tail characteristics. Consistent with expectations, agreement was generally stronger for the 8× experiment than for the more challenging 16× experiment, as reflected by the tighter clustering around the one-to-one line and lower RMSE values across all statistics. These results demonstrate that all three models successfully preserve the key statistical properties of precipitation fields at the native high-resolution scale. While higher-order distributional characteristics remain more challenging to reproduce than lower-order moments, particularly for 16× downscaling, WGAN and DDPM provide slightly improved representation of precipitation occurrence, asymmetry, and tail behavior compared to U-Net.

Spatial autocorrelation was evaluated in both the horizontal and vertical directions for lags ranging from 1 to 8 pixels in order to assess the ability of the models to reproduce the spatial organization of precipitation fields. Autocorrelation distributions were computed across all test samples and compared against the corresponding ERA5-Land target fields for both the 8× and 16× downscaling experiments. All models reproduced the overall decline in autocorrelation with increasing spatial lag observed in the target precipitation data, indicating preservation of the dominant spatial dependence structure. However, systematic differences emerged in the rate at which autocorrelation decayed with distance. U-Net consistently exhibited higher autocorrelation values than ERA5-Land across intermediate and larger lags, indicating a slower decay of spatial dependence. This behavior is consistent with the smoother precipitation structures commonly produced by MSE-based optimization, which tend to increase spatial coherence and suppress fine-scale variability.

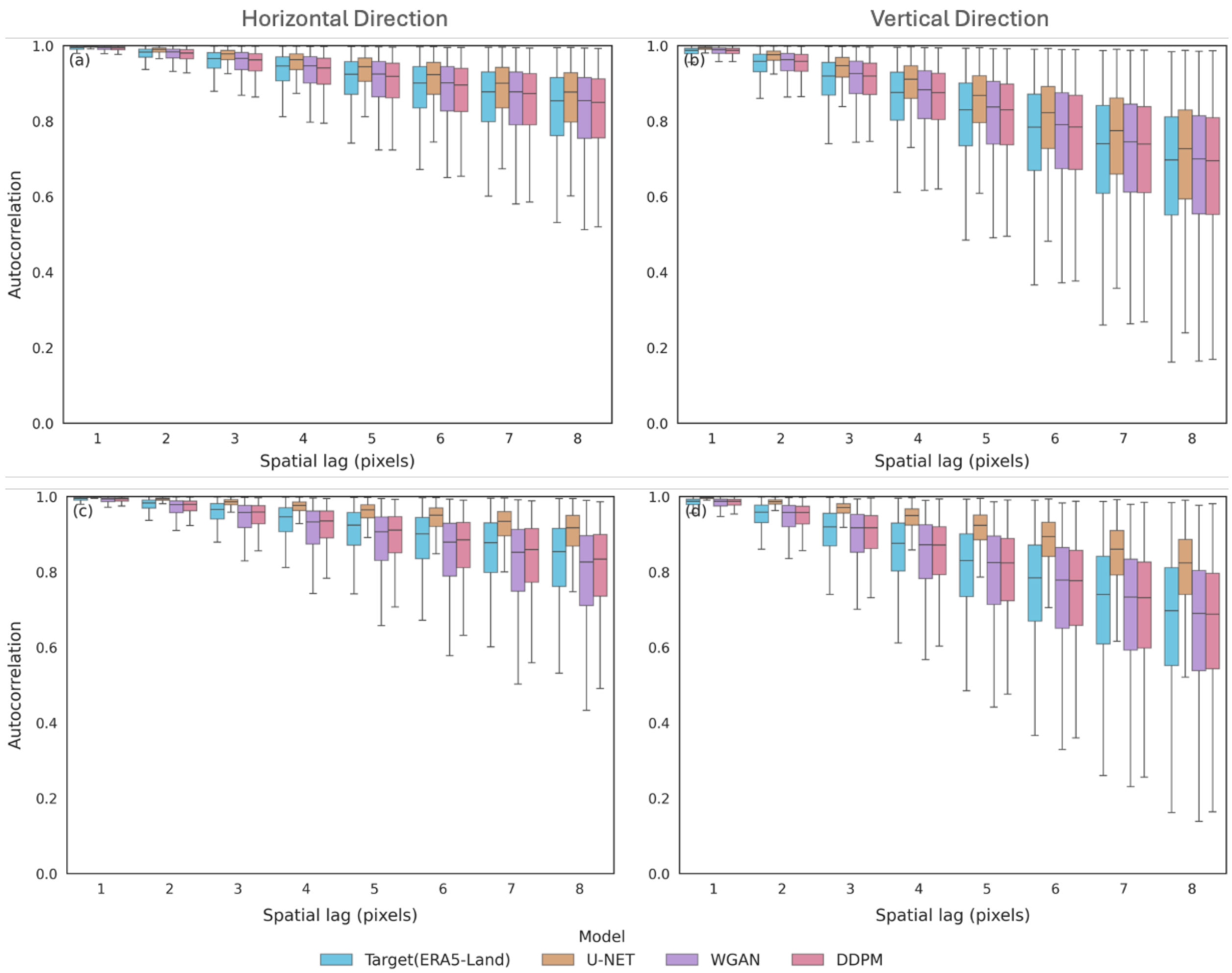

Figure 6. Distributions of horizontal and vertical spatial autocorrelation for lags 1–8 pixels for the ERA5-Land target fields and precipitation fields generated by U-Net, WGAN, and DDPM. Results are aggregated across all test samples and across the 10 independently trained model seeds for each method. Panels (a, b) correspond to the 8× downscaling experiment, while panels (c, d) correspond to the 16× downscaling experiment.

In contrast, WGAN and DDPM more closely reproduced the autocorrelation distributions of ERA5-Land in both the horizontal and vertical directions. The agreement was particularly evident at intermediate and larger lags, where the observed decay of spatial dependence was better captured than by U-Net. This suggests that adversarial and diffusion-based training more effectively preserve realistic precipitation variability and storm morphology by generating sharper and less spatially smoothed precipitation structures. This behavior was consistent between the horizontal and vertical directions, although autocorrelation decreased slightly more rapidly in the vertical direction. Similar patterns were observed for both downscaling factors, with somewhat larger departures from the ERA5-Land distributions in the more challenging 16× experiment (especially in the case of the U-Net). Nevertheless, all models reproduced the dominant spatial organization of precipitation fields, while WGAN and DDPM provided a closer representation of the observed autocorrelation decay and associated storm-scale structure.

5.2.4 Extreme Precipitation and Tail Behavior

Extreme precipitation behavior was evaluated using exceedance probability distributions computed from all test samples across all 10 independently initialized model realizations. The exceedance curves reveal that all models reproduce the bulk precipitation distribution reasonably well, with deviations becoming increasingly apparent toward the upper tail. For the 8× downscaling experiment, U-Net, WGAN, and DDPM closely follow the ERA5-Land exceedance distribution over most precipitation intensities. Differences emerge primarily for rare events exceeding approximately 150 mm/day, where U-Net exhibits a slightly lighter tail than ERA5-Land, indicating a tendency to suppress the most intense precipitation extremes. WGAN more closely reproduces the target tail behavior, while DDPM displays greater variability among realizations and occasionally generates events exceeding the maximum intensities observed in the target dataset.

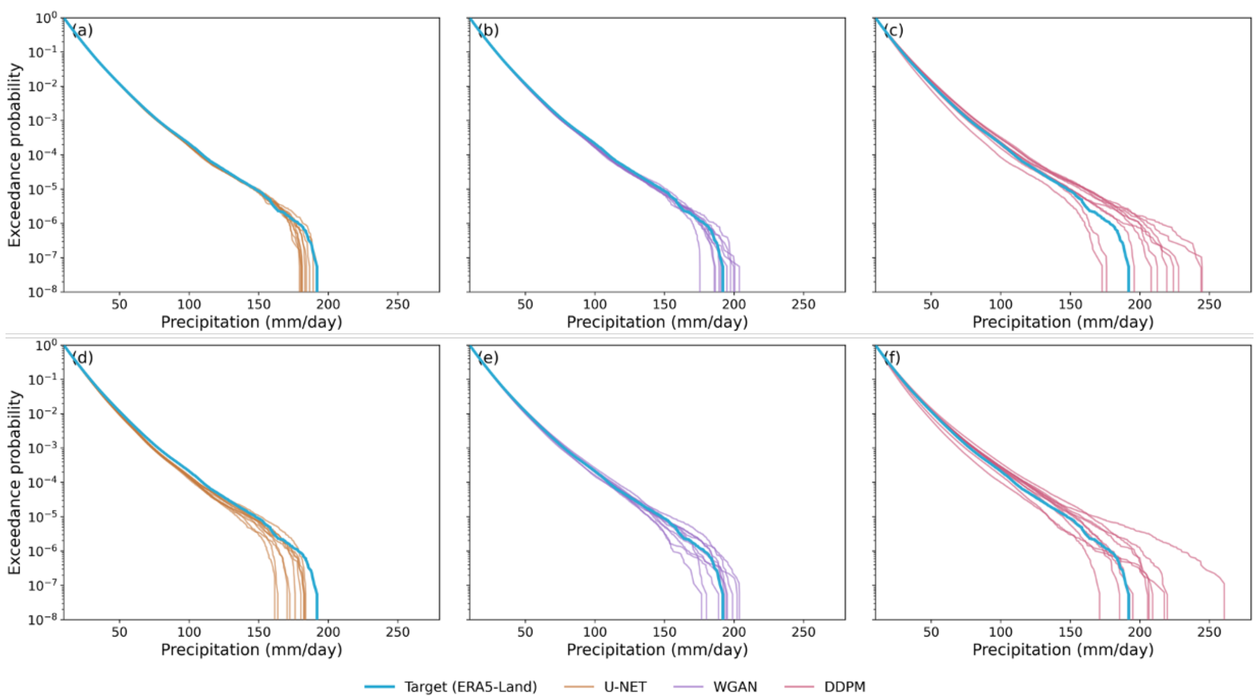

Figure 7. Exceedance probability curves for precipitation generated by U-Net, WGAN, and DDPM across 10 independent seed initializations for the 8× (a–c) (top row) and 16× (d–f) (bottom row) downscaling experiments. The Target (ERA5-Land) distribution is shown in blue, while other colored lines represent individual seed realizations. Exceedance probability is displayed on a logarithmic scale to highlight differences in the upper tail of the precipitation distribution.

The differences become more pronounced in the 16× experiment. The U-Net consistently underestimates the most intense events and exhibits the most rapid decline in the precipitation intensities at the lowest exceedance probability. WGAN better preserves the upper-tail structure and generally remains closer to the ERA5-Land exceedance distribution. DDPM produces the broadest range of tail behaviors and the largest inter-seed variability, with several realizations generating substantially higher precipitation intensities than those present in the target data. This indicates an enhanced ability to generate rare extremes, although it may also lead to occasional overestimation of upper-tail probabilities. These results demonstrate a trade-off between smoothness and extreme-event representation. U-Net provides the most conservative tail estimates, whereas WGAN and particularly DDPM better reproduce the variability and intensity of rare precipitation events, with the differences becoming increasingly evident in the more challenging 16× downscaling experiment.

5.2.5 Composite Diagnostics

The composite diagnostics provide a complementary evaluation of scale-dependent structure, spatial agreement, event detection, and perceptual similarity using ensemble statistics derived from ten independently trained realizations for each model (Figure 8). These results, too, indicate that all three models successfully reproduce the key spatial and statistical characteristics of precipitation over the independent Northeast test region, although notable differences emerge in their representation of fine-scale variability and precipitation extremes. The close agreement between model and target power spectra (Figure 8a), high FSS values across spatial scales (Figure 8b), and strong ROC performance (Figure 8c) demonstrate that all models capture the dominant spatial organization and occurrence patterns of precipitation. Furthermore, the limited spread among the individual seed realizations across these diagnostics suggests that the results are robust to model initialization.

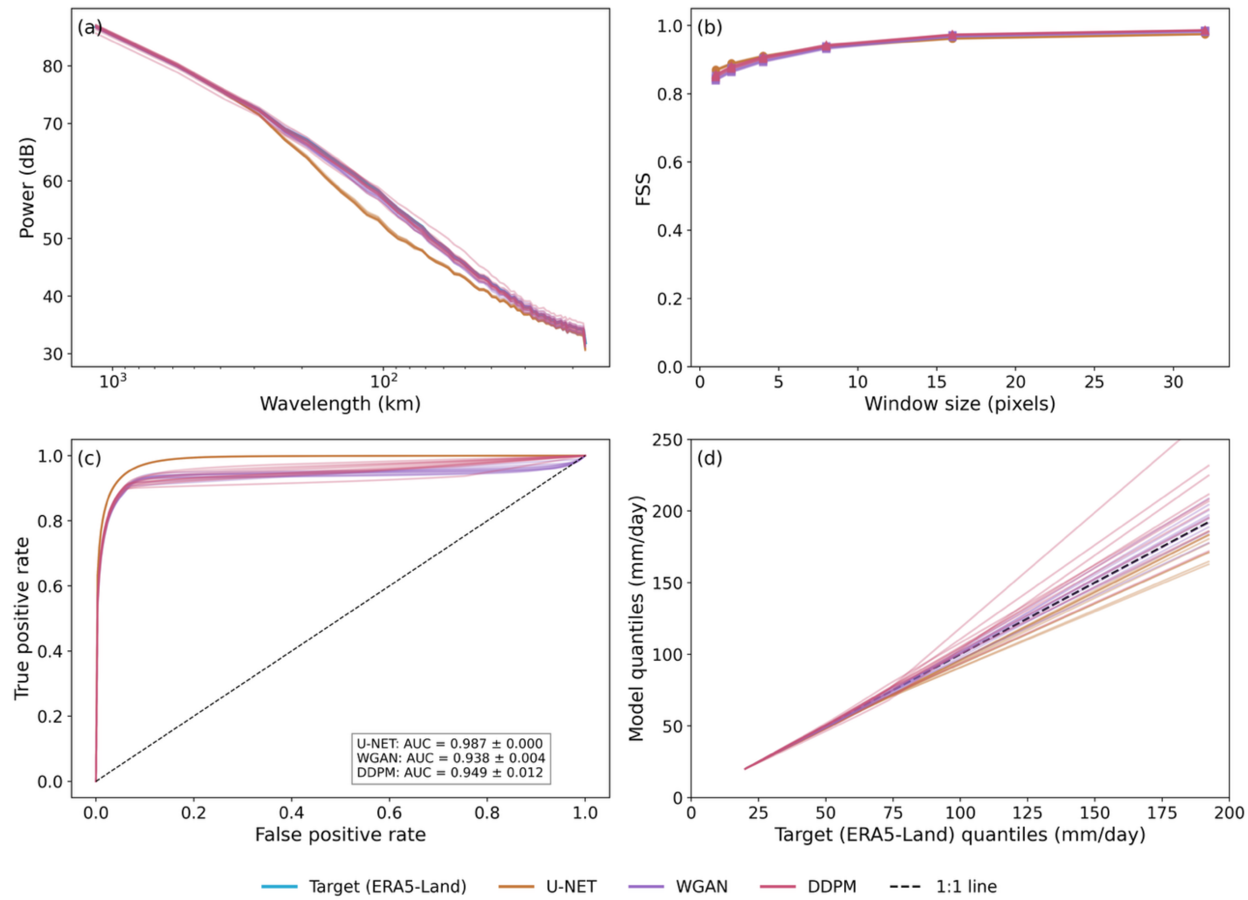

Figure 8. Comparison of U-Net, WGAN, and DDPM performance over the independent Northeast test region at 16× super-resolution. (a) Radial power spectra showing the distribution of spatial variance across wavelengths relative to the target (ERA5-Land) fields. (b) Fractions Skill Score (FSS) as a function of spatial window size. (c) Receiver Operating Characteristic (ROC) curves for wet-pixel classification, with inset values showing the mean ± standard deviation of the Area Under the Curve (AUC) across ten independent model initializations. (d) Quantile–Quantile (Q–Q) comparison between model and target precipitation quantiles for precipitation amounts exceeding 20 mm/day. Colored lines represent individual model initializations, and the dashed line in (d) denotes perfect agreement (1:1 relationship).

Despite these overall similarities, systematic differences become evident when examining the representation of spatial variability and precipitation intensity distributions. As shown in Figure 8a, U-Net exhibits a reduction in spectral power at smaller wavelengths, indicating smoother precipitation structures and a loss of fine-scale variability relative to the target fields. In contrast, WGAN and DDPM retain more high-frequency variance and better reproduce small-scale precipitation features. The superior wet-pixel classification skill of U-Net, reflected by its higher AUC values in Figure 8c, is consistent with its optimization toward pixel-wise accuracy. However, this advantage does not necessarily translate into a more realistic representation of precipitation intensities. The Q–Q analysis (Figure 8d) highlights the largest differences among the models. For precipitation amounts exceeding 20 mm/day, U-Net systematically underestimates the upper quantiles, suggesting a tendency to smooth intense precipitation events. In comparison, WGAN and DDPM remain closer to the 1:1 reference line across much of the distribution and produce heavier tails, indicating an improved representation of moderate-to-extreme precipitation intensities. At the highest quantiles, both models exhibit increased spread among seed realizations, reflecting greater uncertainty in the simulation of rare extreme events. All models exhibit high sample-wise structural similarity index (SSIM) values, particularly for the 8× downscaling experiment. SSIM decreases for the more challenging 16× downscaling task, although median values remain above 0.95 for all model classes (supplementary Figure S7). Overall, these results further emphasize that WGAN and DDPM achieve a more realistic representation of fine-scale spatial variability and high-intensity precipitation.

5.2.6 Stochastic Variability and Uncertainty

The comparison between within-seed and across-seed ensembles shows that both sources of variability affect the simulated precipitation distribution, but across-seed variability is consistently larger. This distinction is most evident in the exceedance probability curves (Figure 9b1–b2). Within-seed realizations remain relatively clustered at moderate thresholds but begin to diverge in the upper tail, particularly beyond approximately 150 mm/day. In contrast, across-seed ensembles show a broader spread even at lower precipitation thresholds, and this spread increases further toward the most extreme events. This indicates that stochastic sampling within a fixed trained model contributes to tail uncertainty, but differences among independently trained model initializations produce a stronger and more persistent variability. The lagged spatial autocorrelation characteristics were highly consistent across independent model initializations (Supplementary Figure S8), indicating that the learned spatial dependence structure is robust to training variability. Although the spread among realizations increased slightly at larger lags, particularly for DDPM, the median correlations remained stable across seeds and closely followed the target spatial autocorrelation decay.

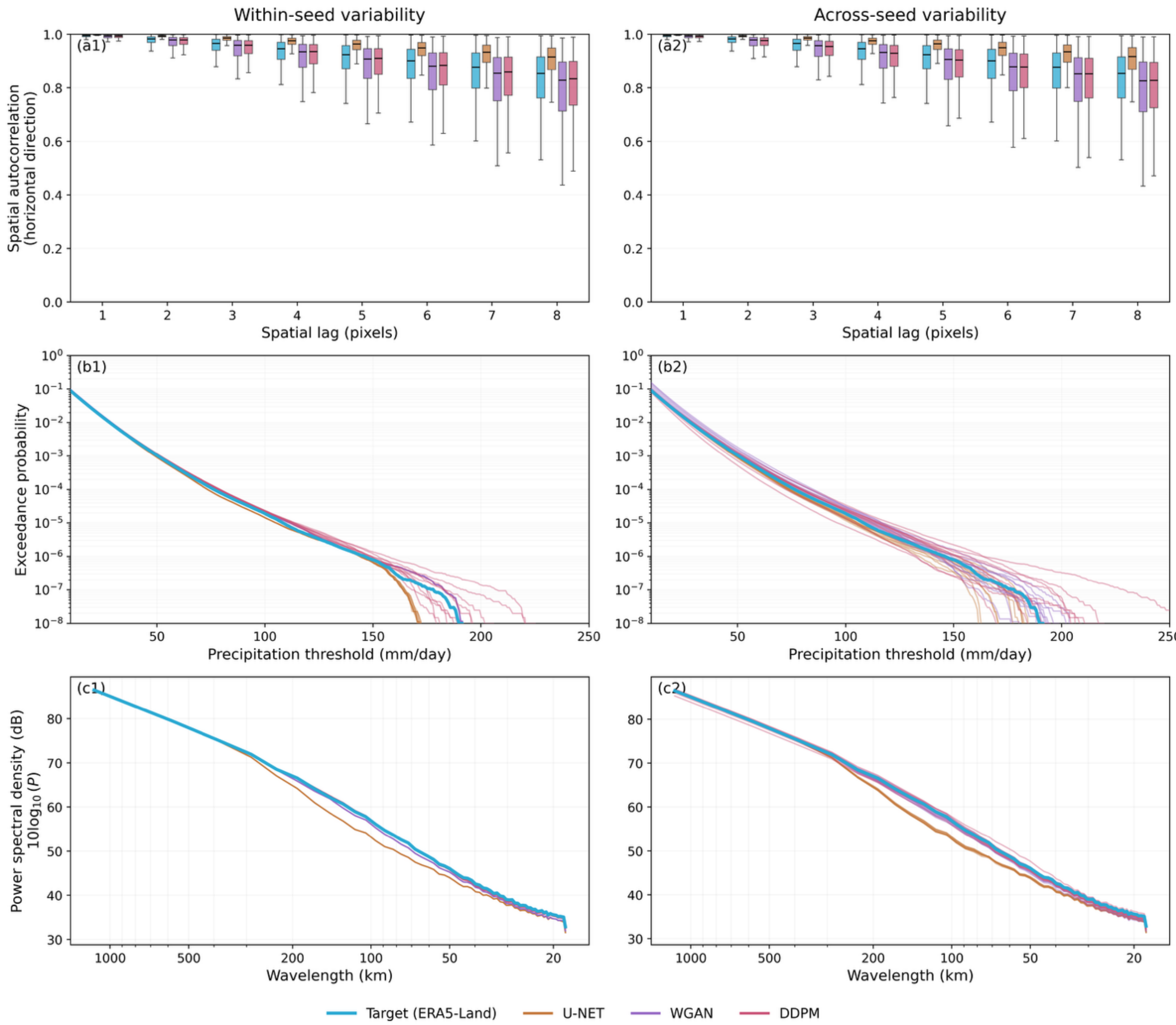

Figure 9: Within-seed and across-seed variability of U-Net, WGAN, and DDPM predictions over the independent Northeast test region at 16× super-resolution. Panels (a1–a2) show boxplots of horizontal spatial autocorrelation at lags 1–8 pixels, where within-seed variability is estimated from multiple stochastic realizations generated by a single trained model and across-seed variability is estimated from independently trained model initializations. Panels (b1–b2) present exceedance probability curves as a function of precipitation threshold (mm/day), illustrating variability in the representation of moderate-to-extreme precipitation intensities. Panels (c1–c2) show radial power spectra describing the distribution of spatial variance across wavelengths. Left-column panels correspond to within-seed variability and right-column panels correspond to across-seed variability.

For spatial autocorrelation (Figure 9a1–a2) and radial power spectra (Figure 9c1–c2), the contrast between within-seed and across-seed behavior is less pronounced. Both ensemble configurations show broadly similar spatial dependence and spectral characteristics, suggesting that the learned spatial organization is relatively stable. Nevertheless, the across-seed boxplots show slightly larger dispersion at longer spatial lags, indicating that model initialization also affects spatial dependence, although less strongly than it affects precipitation extremes. These results suggest that uncertainty in generated precipitation fields is metric dependent. Spatial structure is comparatively robust to both stochastic sampling and independent model initialization, whereas distributional tail behavior is more sensitive, with across-seed variability representing the dominant source of uncertainty.

5.2.7 Temporal Correlation

Although the downscaling models were trained independently on daily precipitation fields without explicit temporal constraints, they reproduce the observed temporal dependence structure reasonably well. As shown in Figure 10a, all three models capture the overall distribution of pixel-wise temporal correlation across lags 1–5 days, with median values closely matching those of the target (ERA5-Land). The strongest persistence is observed at lag 1, followed by a rapid reduction in autocorrelation at longer lags, consistent with the temporal behavior of the target precipitation fields.

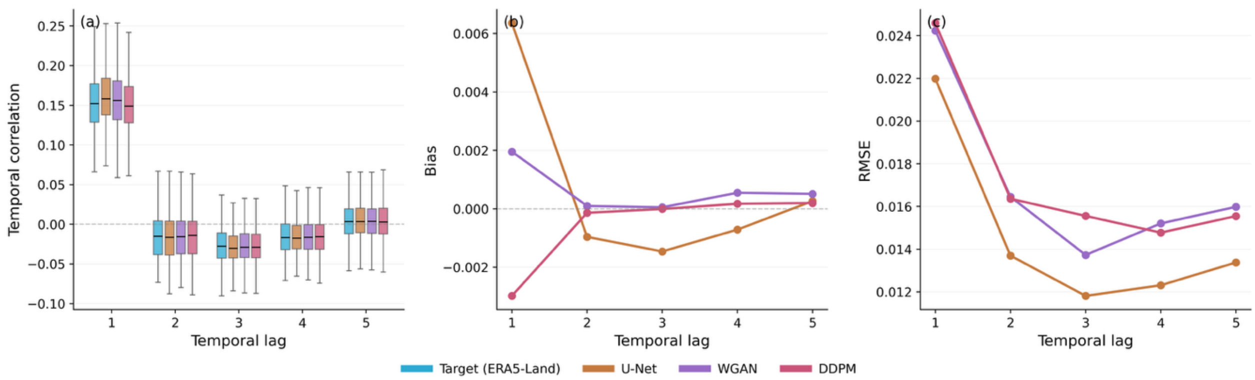

Figure 10: Temporal correlation characteristics of U-Net, WGAN, and DDPM evaluated using 1000 consecutive precipitation samples from the Northeast test region without wet-day filtering. (a) Distribution of pixel-wise temporal correlation coefficients at lags 1–5 days for the target (ERA5-Land) and model-generated precipitation fields. (b) Bias of temporal autocorrelation relative to the target as a function of temporal lag. (c) RMSE of temporal correlation relative to the target as a function of temporal lag. Temporal correlation was computed independently at each grid cell using lagged correlations between daily precipitation time series.

Differences among the models become more apparent in the bias and RMSE statistics (Figure10b–c). WGAN and DDPM exhibit near-zero bias across most temporal lags, indicating that they reproduce the mean temporal correlation structure of the target data with high fidelity. In contrast, U-Net slightly overestimates persistence at lag 1 and marginally underestimates correlations at longer lags. Despite this bias pattern, U-Net achieves the lowest RMSE across all lags, indicating the closest overall agreement with the target temporal correlations. This may reflect the more deterministic and spatially smoother nature of U-Net predictions, whereas the additional variability introduced by WGAN and DDPM to better represent spatial structure and extremes results in slightly larger temporal-correlation errors while remaining nearly unbiased. Nevertheless, these differences are modest, and all three models successfully reproduce the observed decay of temporal correlation with increasing lag. The RMSE values decrease substantially from lag 1 to lag 3 before stabilizing at longer lags (Figure 10c), indicating that model differences are most pronounced in the representation of short-term persistence. These findings suggest that the spatial downscaling models preserve much of the temporal correlation structure present in the target data, despite not being explicitly trained to model temporal evolution. This provides additional evidence that the generated precipitation fields remain physically consistent not only in their spatial characteristics but also in their short-term temporal dependence.

5.3 Sensitivity of DDPM to Diffusion Length

To evaluate the computational efficiency of the DDPM framework, we examined the sensitivity of model performance to the number of reverse-diffusion inference steps. In addition to the baseline configuration using T=500 diffusion steps, inference was performed using reduced diffusion lengths of T=100 and T=50 without retraining the model (supplementary Figure S9-S10). The resulting precipitation fields were evaluated using spatial autocorrelation, exceedance probability distributions, quantile–quantile (Q–Q) relationships, and radial averaged power spectra (Figure 11 and supplementary Figure S11).

Reducing the diffusion length from T=500 to T=100 produces only minor changes in the generated precipitation fields. Spatial autocorrelation distributions remain nearly identical, indicating that large-scale spatial organization is preserved despite the shorter reverse-diffusion process. Exceedance probability distributions and Q–Q relationships also show substantial overlap, with only modest increases in variability among seeds at the highest precipitation intensities. Similarly, the power spectra closely match those obtained using T=500, demonstrating that the representation of spatial variability across scales is largely unaffected by this reduction in inference steps.

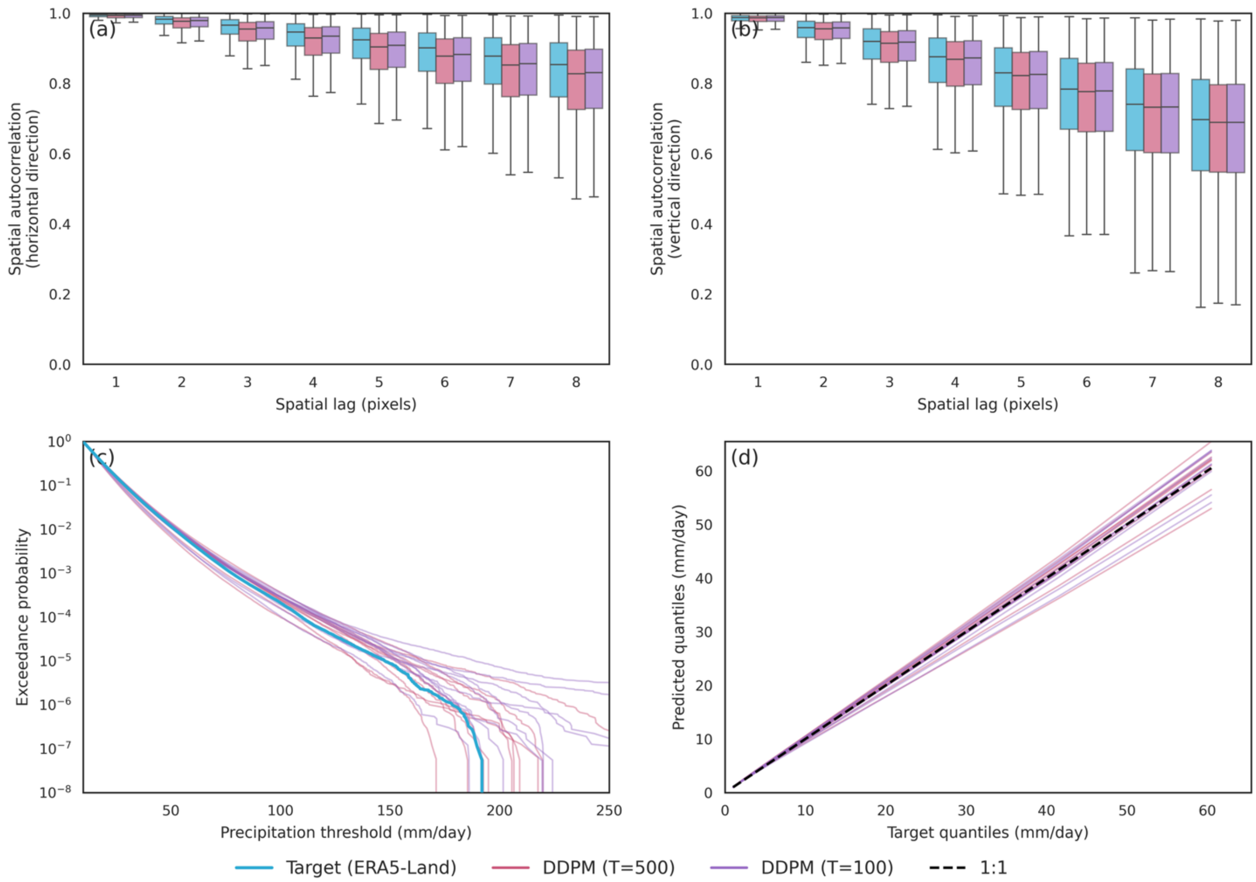

Figure 11. Comparison of DDPM predictions generated using 500 and 100 denoising steps for 16× precipitation downscaling across ten independently trained seeds. Panels show (a) horizontal spatial autocorrelation, (b) vertical spatial autocorrelation, (c) exceedance probability for precipitation intensities exceeding 10 mm/day, and (d) quantile–quantile (Q–Q) relationships relative to Target (ERA5-Land). Boxplots in (a–b) summarize the distribution of spatial correlations across all test samples, while panels (c–d) display individual seed realizations as separate lines. The Target (ERA5-Land) distribution is shown in cyan, DDPM T=500 in pink, and DDPM T=100 in purple. The comparison assesses the sensitivity of spatial dependence and precipitation statistics to the number of diffusion inference steps.

In contrast, a further reduction from T=100 to T=50 results in noticeably larger deviations from the Target (ERA5-Land) statistics. While the spatial autocorrelation and spectral characteristics remain broadly similar, the exceedance probability distributions exhibit substantially greater spread among seed realizations and increasingly divergent behavior in the upper tail. The Q–Q relationships likewise show larger departures from the one-to-one line at high quantiles, indicating reduced stability in the representation of extreme precipitation events. Several T=50 realizations generate substantially higher precipitation intensities than those observed in the target dataset, suggesting an increased tendency to overestimate rare extremes. These results indicate that DDPM performance is relatively insensitive to moderate reductions in diffusion length, with T=100 retaining most of the statistical and spatial characteristics obtained using T=500 while reducing computational cost by approximately a factor of five. However, further reduction to T=50 leads to measurable degradation in the representation of precipitation extremes.

6. Conclusions

This study presented a comprehensive intercomparison of three deep learning architectures, namely a convolutional U-Net, a WGAN, and a DDPM, for statistical precipitation downscaling at 8× and 16× super-resolution factors. Models were trained over the Central Plains and Northwest regions and evaluated over an independent Northeast test domain, providing a rigorous assessment of cross-regional generalization under realistic out-of-sample conditions. All three models successfully captured the large-scale spatial organization and bulk precipitation statistics of the ERA5-Land target fields, demonstrating that deep learning approaches can effectively transfer learned downscaling relationships to unseen geographic domains. However, systematic differences emerged in their ability to represent fine-scale variability, precipitation extremes, spatial dependence structure, and predictive uncertainty, all of which are critical dimensions for downstream climate impact applications.

The U-Net, the baseline model in this study, consistently produced the smoothest precipitation fields and achieved strong performance in wet–dry discrimination and precipitation occurrence. These characteristics, however, came at the cost of drastically underestimated spatial variability, suppressed precipitation extremes, and overestimated spatial autocorrelation. These limitations are well-documented in deterministic regression-based downscaling frameworks and reflect the inherent smoothing tendency of mean squared error optimization. The two generative frameworks, WGAN and DDPM, substantially outperformed U-Net in reproducing precipitation distributions, higher-order statistical moments, and spatial structure. The WGAN demonstrated particularly strong fidelity in capturing exceedance probabilities and the full precipitation intensity spectrum, while maintaining a favorable computational cost relative to the DDPM. The DDPM, by contrast, most faithfully reproduced the target spatial autocorrelation decay and spectral energy distribution across spatial scales and uniquely provided stochastic ensemble realizations that quantify uncertainty in the downscaling process, a capability absent in both U-Net and WGAN.

A key finding of this study concerns the sources of uncertainty in generative downscaling systems. Analysis of within-seed and across-seed variability revealed that differences arising from independent model initializations consistently exceeded variability associated with repeated stochastic generations from a single trained model. This result has important practical implications. Single-seed model assessments risk underestimating true model uncertainty, and robust evaluation of generative downscaling systems should incorporate multi-seed training ensembles alongside stochastic sampling. This finding also suggests that the stochastic diversity of DDPM realizations, while valuable, does not fully capture the epistemic uncertainty inherent in the training process itself.

Sensitivity experiments demonstrated that DDPM performance is relatively robust to moderate reductions in the number of reverse-diffusion steps. Reducing the step count from T=500 to T=100 produced only minor degradation in precipitation statistics and spatial structure, whereas further reduction to T=50 led to more noticeable deterioration in extreme precipitation representation and increased inter-seed variability. These results identify T=100 as an effective operational compromise between physical fidelity and computational efficiency, a finding with practical relevance for ensemble-based downscaling applications requiring large numbers of realizations.

Overall, the results of this study demonstrate that the choice of downscaling architecture involves non-trivial trade-offs among statistical accuracy, spatial realism, uncertainty quantification, and computational cost. Both WGAN and DDPM substantially outperform the deterministic U-Net across many of the evaluated diagnostics and exhibit broadly comparable performance in reproducing precipitation statistics, spatial dependence, and higher-order distributional characteristics. While DDPM occasionally provides modest improvements in spatial structure and offers a native framework for probabilistic ensemble generation, these gains are generally small relative to its substantially greater computational cost. Consequently, WGAN represents an attractive balance between performance and efficiency, whereas DDPM may be preferable when explicit uncertainty quantification and ensemble-based probabilistic predictions are primary objectives. These results suggest that the optimal model choice depends on the specific application requirements and the relative importance of computational efficiency versus probabilistic information.

Some limitations should be acknowledged. The models were trained and evaluated using ERA5-Land as the high-resolution reference, which, although widely used, remains a reanalysis product rather than a direct observational dataset. The analysis was also restricted to precipitation, and the relative advantages of generative approaches remain to be shown in multivariate downscaling problems involving additional variables. Furthermore, the models did not explicitly incorporate physiographic information such as topography and land-surface characteristics, which may influence performance in regions where terrain strongly affects precipitation patterns. The architectures evaluated here should be viewed as representative implementations rather than fully optimized models. In particular, the baseline model U-Net employed a conventional MSE objective, whereas precipitation fields exhibit intermittency, skewness, and heavy-tailed behavior that may be better represented by distribution-aware loss functions (e.g., Tweedie, quantile-based, or probabilistic objectives; (Hunt, 2025; Rastogi et al., 2025)). Finally, all experiments were conducted under historical climate conditions, and the robustness of these approaches under future climate states and associated distributional shifts remains an important area for future investigation.

Future work should address these limitations progressively. Evaluating model performance against observation-based targets, incorporating physiographic conditioning variables, and extending the framework to multivariate downscaling problems are all important near-term directions. Further research should also explore the sensitivity of conclusions to architectural design choices, loss functions, data transformations, and conditioning strategies. Assessing robustness under future climate conditions will likely require complementary approaches such as domain adaptation, physically informed constraints, and hybrid frameworks that combine the flexibility of generative deep learning with the dynamical consistency of process-based climate models. Overall, the results demonstrate that generative approaches, particularly WGANs and DDPMs, offer substantial advantages over conventional deterministic downscaling for reproducing precipitation variability, extremes, and uncertainty, while also highlighting the importance of evaluating model performance across multiple complementary diagnostics rather than relying on a single metric.

Code availability

All scripts used for data downloading, preprocessing, model training, inference, and evaluation are openly available in the public GitHub repository: https://github.com/shivamsinghhada/precipitation-downscaling. The exact version of the code used in this study is permanently archived on Zenodo at: https://doi.org/10.5281/zenodo.20549613.

Data availability

The raw ERA5-Land precipitation data used in this study are publicly available from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS) (Muñoz Sabater, 2019). ERA5-Land provides global land-surface variables at approximately 9 km spatial resolution and hourly temporal resolution and can be accessed at: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=download. The processed dataset splits and selected trained model weights used for the main experiments and figure generation are archived separately on Zenodo at: https://doi.org/10.5281/zenodo.19324377. These archived materials include the processed 8× and 16× downscaling datasets, selected trained U-Net, WGAN, and DDPM model weights, and the associated inference scripts required to reproduce the main analyses presented in this manuscript.

Raw ERA5-Land data are not redistributed here because they are already publicly available from ECMWF / CDS. Instead, the full preprocessing workflow required to reproduce the derived datasets is provided in the archived code repository.

Author Contributions

SS led the conceptualization, methodology development, model implementation, experiments, analysis, and manuscript writing. SMP, TH, and AM supervised the research, contributed to experimental design, interpretation of results, and manuscript review and editing, and supported funding acquisition. HMA contributed to experimental design and assisted with manuscript review and editing. All authors approved the final manuscript.

Competing Interests

The author declares no competing interests.

Funding Statement

This work was supported by the UVA Environmental Institute through the Strategic Investment Fund**.**

References

Abdelmoaty, H. M., Papalexiou, S. M., Mamalakis, A., Singh, S., Coia, V., Hairabedian, M., Szeftel, P., & Grover, P. (2025). Generative Adversarial Networks for Downscaling Hourly Precipitation in the Canadian Prairies. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(4). https://doi.org/10.1029/2025JH000678

Arjovsky, M., Chintala, S., & Bottou, L. (2017). *Wasserstein GAN*. http://arxiv.org/abs/1701.07875

Baño-Medina, J., Manzanas, R., & Gutiérrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, *13*(4), 2109–2124. https://doi.org/10.5194/gmd-13-2109-2020

Bednarz, T., & Cherukuri, R. (2023). Use of Physics-Based AI for Simulations and Modeling in the Era of Digital Twins. *SIGGRAPH Asia 2023 Courses*, 1–45. https://doi.org/10.1145/3610538.3614627

Coppola, E., Sobolowski, S., Pichelli, E., Raffaele, F., Ahrens, B., Anders, I., Ban, N., Bastin, S., Belda, M., Belusic, D., Caldas-Alvarez, A., Cardoso, R. M., Davolio, S., Dobler, A., Fernandez, J., Fita, L., Fumiere, Q., Giorgi, F., Goergen, K., … Warrach-Sagi, K. (2020). A first-of-its-kind multi-model convection permitting ensemble for investigating convective phenomena over Europe and the Mediterranean. *Climate Dynamics*, *55*(1–2), 3–34. https://doi.org/10.1007/s00382-018-4521-8

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, *38*(3–4), 527–546. https://doi.org/10.1007/s00382-010-0977-x

Feser, F., Rockel, B., von Storch, H., Winterfeldt, J., & Zahn, M. (2011). Regional Climate Models Add Value to Global Model Data: A Review and Selected Examples. *Bulletin of the American Meteorological Society*, *92*(9), 1181–1192. https://doi.org/10.1175/2011BAMS3061.1

Gao, X. J., Shi, Y., Zhang, D., Wu, J., Giorgi, F., Ji, Z., & Wang, Y. (2012). Uncertainties in monsoon precipitation projections over China: Results from two high-resolution RCM simulations. *Clim Res*, *52*. https://doi.org/10.3354/cr0108

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., & Ebert, E. E. (2009). Intercomparison of Spatial Forecast Verification Methods. *Weather and Forecasting*, *24*(5), 1416–1430. https://doi.org/10.1175/2009WAF2222269.1

Giorgi, F., & Gutowski, W. J. (2015). Regional dynamical downscaling and the CORDEX initiative. *Annu. Rev. Environ. Resour.*, *40*. https://doi.org/10.1146/annurev-environ-102014-021217

Giorgi, F., & Mearns, L. O. (1991). Approaches to the simulation of regional climate change: A review. *Reviews of Geophysics*, *29*(2), 191–216. https://doi.org/10.1029/90RG02636

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. (2017). *Improved Training of Wasserstein GANs*.

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. *Journal of Advances in Modeling Earth Systems*, *14*(10). https://doi.org/10.1029/2022MS003120

Harrison, D. R., McGovern, A., Karstens, C. D., Bostrom, A., Demuth, J. L., Jirak, I. L., & Marsh, P. T. (2025). An Assessment of How Domain Experts Evaluate Machine Learning in Operational Meteorology. *Weather and Forecasting*, *40*(3), 393–410. https://doi.org/10.1175/WAF-D-24-0144.1

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information*

Processing Systems (Vol. 33, pp. 6840–6851). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

Hobeichi, S., Nishant, N., Shao, Y., Abramowitz, G., Pitman, A., Sherwood, S., Bishop, C., & Green, S. (2023). Using Machine Learning to Cut the Cost of Dynamical Downscaling. *Earth's Future*, *11*(3). https://doi.org/10.1029/2022EF003291

Höhlein, K., Kern, M., Hewson, T., & Westermann, R. (2020). A comparative study of convolutional neural network models for wind field downscaling. *Meteorological Applications*, *27*(6). https://doi.org/10.1002/met.1961

Hsu, L.-H., Chiang, C.-C., Lin, K.-L., Lin, H.-H., Chu, J.-L., Yu, Y.-C., & Fahn, C.-S. (2024). Downscaling Taiwan precipitation with a residual deep learning approach. *Geoscience Letters*, *11*(1), 23. https://doi.org/10.1186/s40562-024-00340-y

Hunt, K. M. R. (2025). *Stop using root-mean-square error as a precipitation target!* http://arxiv.org/abs/2509.08369

Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., Stegmaier, J., Kuhl, C., Nebelung, S., Kather, J. N., & Truhn, D. (2023). Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*, *13*(1), 7303. https://doi.org/10.1038/s41598-023-34341-2

Kumar, B., Atey, K., Singh, B. B., Chattopadhyay, R., Acharya, N., Singh, M., Nanjundiah, R. S., & Rao, S. A. (2023). On the modern deep learning approaches for precipitation downscaling. *Earth Science Informatics*, *16*(2), 1459–1472. https://doi.org/10.1007/s12145-023-00970-4

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

Lange, S. (2019). Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1.0). *Geoscientific Model Development*, *12*(7), 3055–3070. https://doi.org/10.5194/gmd-12-3055-2019

Lucas-Picher, P., Argüeso, D., Brisson, E., Tramblay, Y., Berg, P., Lemonsu, A., Kotlarski, S., & Caillaud, C. (2021). Convection-permitting modeling with regional climate models: Latest developments and next steps. *WIREs Climate Change*, *12*(6). https://doi.org/10.1002/wcc.731

Lyu, R., Wang, L., Sun, Y., Bai, H., & Lu, C.-T. (2024). Downscaling Precipitation with Bias-informed Conditional Diffusion Model. *2024 IEEE International Conference on Big Data (BigData)*, 8768–8770. https://doi.org/10.1109/BigData62323.2024.10825056

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 Resources Research*, *53*(3), 2149–2170. https://doi.org/10.1002/2016WR019578

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., Brienen, S., Rust, H. W., Sauter, T., Themeßl, M., Venema, V. K. C., Chun, K. P., Goodess, C. M., Jones, R. G., Onof, C., Vrac, M., & Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, *48*(3), RG3003. https://doi.org/10.1029/2009RG000314

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

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. *Environmental Research Letters*, *18*(9), 094006. https://doi.org/10.1088/1748-9326/ace463

Palmer, T. (2014). Build high-resolution global climate models. *Nature*, *515*. https://doi.org/10.1038/515338a

Papalexiou, S. M., & Mamalakis, A. (2025). Machine unlearning: bias correction in neural network downscaled storms. *Journal of Hydrology*, 134689. https://doi.org/10.1016/j.jhydrol.2025.134689

Papalexiou, S. M., Serinaldi, F., & Porcu, E. (2021). Advancing Space-Time Simulation of Random Fields: From Storms to Cyclones and Beyond. *Water Resources Research*, *57*(8). https://doi.org/10.1029/2020WR029466

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018). FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11671

Piani, C., Haerter, J. O., & Coppola, E. (2010). Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology*, *99*(1–2), 187–192. https://doi.org/10.1007/s00704-009-0134-9

Rampal, N., Hobeichi, S., Gibson, P. B., Baño-Medina, J., Abramowitz, G., Beucler, T., González-Abad, J., Chapman, W., Harder, P., & Gutiérrez, J. M. (2024). Enhancing Regional Climate Downscaling through Advances in Machine Learning. *Artificial Intelligence for the Earth Systems*, *3*(2). https://doi.org/10.1175/aies-d-23-0066.1

Rastogi, D., Niu, H., Passarella, L., Mahajan, S., Kao, S., Vahmani, P., & Jones, A. D. (2025). Complementing Dynamical Downscaling With Super-Resolution Convolutional Neural Networks. *Geophysical Research Letters*, *52*(4). https://doi.org/10.1029/2024GL111828

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., & Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. https://doi.org/10.1038/s41586-021-03854-z

Roberts, N. M., & Lean, H. W. (2008). Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Schär, C. (2019). Kilometer-scale climate models. Prospects and challenges article. *Am. Meteorol. Soc.*, *101*.

Skamarock, W. C. (2004). Evaluating Mesoscale NWP Models Using Kinetic Energy Spectra. *Monthly Weather Review*, *132*(12), 3019–3032. https://doi.org/10.1175/MWR2830.1

Song, Y., & Dhariwal, P. (2023). *Improved Techniques for Training Consistency Models*. http://arxiv.org/abs/2310.14189

Stengel, K., Glaws, A., Hettinger, D., & King, R. N. (2020). Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, *117*(29), 16805–16815. https://doi.org/10.1073/pnas.1918964117

Stephens, G. (2017). Challenges and Advances in Convection-Permitting Climate Modeling. *Bulletin of the American Meteorological Society*, *98*(5), 1027–1030. https://doi.org/10.1175/BAMS-D-16-0263.1

Tabari, H., Paz, S. M., Buekenhout, D., & Willems, P. (2021). Comparison of statistical downscaling methods for climate change impact analysis on precipitation-driven drought. *Hydrology and Earth System Sciences*, *25*(6), 3493–3517. https://doi.org/10.5194/hess-25-3493-2021

Teutschbein, C., & Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, *456–457*, 12–29. https://doi.org/10.1016/j.jhydrol.2012.05.052

Tomasi, E., Franch, G., & Cristoforetti, M. (2025). Can AI be enabled to perform dynamical downscaling? A latent diffusion model to mimic kilometer-scale COSMO5.0_CLM9 simulations. *Geoscientific Model Development*, *18*(6), 2051–2078. https://doi.org/10.5194/gmd-18-2051-2025

Trenberth, K. E., Fasullo, J. T., & Shepherd, T. G. (2015). Attribution of climate extreme events. *Nature Climate Change*, *5*(8), 725–730. https://doi.org/10.1038/nclimate2657

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., & Ganguly, A. R. (2017). DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1663–1672. https://doi.org/10.1145/3097983.3098004

Vrac, M., Stein, M. L., Hayhoe, K., & Liang, X. Z. (2007). A general method for validating statistical downscaling methods under future climate change. *Geophysical Research Letters*, *34*(18). https://doi.org/10.1029/2007GL030295

Wang, F., Tian, D., Lowe, L., Kalin, L., & Lehrter, J. (2021). Deep Learning for Daily Precipitation and Temperature Downscaling. *Water Resources Research*, *57*(4). https://doi.org/10.1029/2020WR029308

Wood, A. W., Leung, L. R., Sridhar, V., & Lettenmaier, D. P. (2004). Hydrologic Implications of Dynamical and Statistical Approaches to Downscaling Climate Model Outputs. *Climatic Change*, *62*(1–3), 189–216. https://doi.org/10.1023/B:CLIM.0000013685.99609.9e

Yan, W., Wang, Y., Gu, S., Huang, L., Yan, F., Xia, L., & Tao, Q. (2019). *The Domain Shift Problem of Medical Image Segmentation and Vendor-Adaptation by Unet-GAN* (pp. 623–631). https://doi.org/10.1007/978-3-030-32245-8_69