https://doi.org/10.5194/egusphere-2026-861
Preprint. Discussion started: 27 February 2026

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, hydrologically oriented comparison of three representative deep-learning frameworks for precipitation super-resolution: a convolutional U-NET, a conditional Wasserstein GAN (WGAN), and a conditional denoising diffusion probabilistic model (DDPM). Using a perfect-model experimental design based on ERA5-Land precipitation over distinct hydroclimatic regions of the United States, we evaluate performance under 8-times (8×) and 16-times (16×) downscaling tasks within a unified training and evaluation framework. Models are evaluated using diagnostics that examine precipitation distributions, wet–dry occurrence, extremes, spatial structure, storm morphology, mass consistency, ensemble variability, and computational cost. All three models preserve aggregate rainfall mass despite the absence of explicit physical constraints. Differences arise primarily at fine spatial scales and in the representation of extremes, spatial dependence, and uncertainty. U-NET provides stable and computationally efficient predictions but smooths small-scale variability. WGAN improves fine-scale structure and heavy-tail behavior at the expense of increased noise. The DDPM yields physically coherent ensemble members and an explicit representation of uncertainty, at a substantially higher computational cost.

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). 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 unresolved 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 deterministic U-NET, a conditional Wasserstein GAN (WGAN), and a conditional denoising diffusion probabilistic model (DDPM). Using a common training and evaluation framework, we assess not only mean predictive accuracy but also wet–dry occurrence, storm morphology, spatial dependence, extreme precipitation behavior, and rainfall mass consistency across scales. We further exploit multi-seed realizations of the generative models to quantify ensemble variability and compare it with deterministic behavior, while explicitly evaluating computational requirements for training and inference. By placing deterministic and generative approaches on equal footing and emphasizing physically meaningful diagnostics, this work aims to clarify the relative strengths and limitations of contemporary deep-learning downscaling frameworks and to provide actionable guidance for hydrologic and climate-risk applications that require reliable extremes, coherent spatial structure, and interpretable uncertainty under increasing resolution demands.

https://doi.org/10.5194/egusphere-2026-861
Preprint. Discussion started: 27 February 2026

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) Pacific 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.

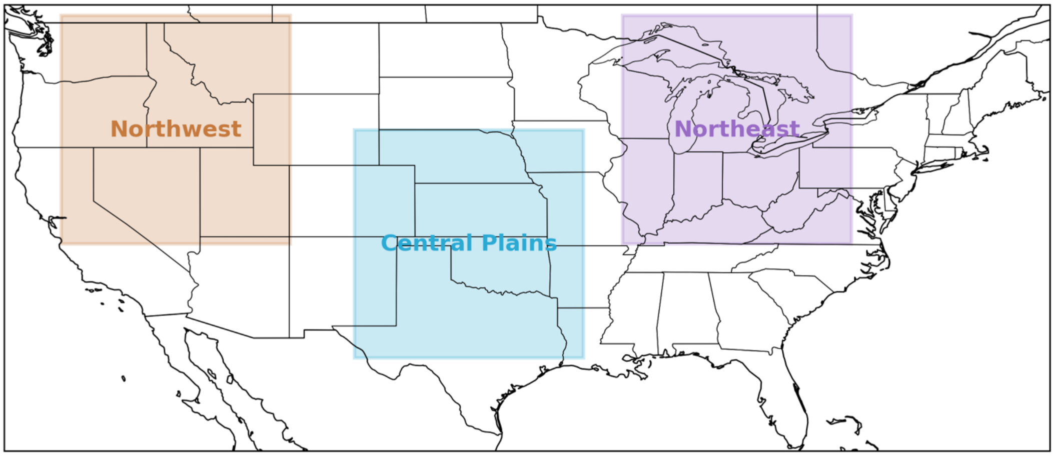


Figure 1. Geographic domains used in this study, showing the three 128 × 128 grid regions over the contiguous United States representing distinct hydroclimatic regimes: the Pacific Northwest, Central Plains, and Northeast

A cross-regional evaluation strategy is adopted to test generalization across hydroclimatic regimes. Models are trained using samples from the Central Plains and Pacific Northwest. Validation samples are drawn from the Central Plains and a subset of the Northeast region to guide model selection and early stopping under distribution shift, while the remaining Northeast samples are reserved for independent testing. This setup ensures that the Northeast region is not used for parameter learning during training, while enabling evaluation in an out-of-training-regime context (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 hydroclimatological analyses and reporting conventions (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. After filtering, 11,025 daily samples are retained for the Central Plains, 11,747 for the Pacific Northwest, and 12,348 for the Northeast.

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 data. A schematic of these deep-learning architectures is presented in Figure 2 and detailed architectures are included in supplementary information, Figure S1.

2.2.1 U-NET

A U-NET architecture is widely used in the 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 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). As demonstrated in prior studies, minimization of squared error causes the network to ignore injected noise and produce nearly identical outputs across different noise realizations (Abdelmoaty et al., 2025; Papalexiou & Mamalakis, 2025). The MSE loss is expressed as:

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

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}, batch size 32) with early stopping when validation loss fails to improve for ten consecutive epochs (patience =10), 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 a 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:

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

https://doi.org/10.5194/egusphere-2026-861
Preprint. Discussion started: 27 February 2026

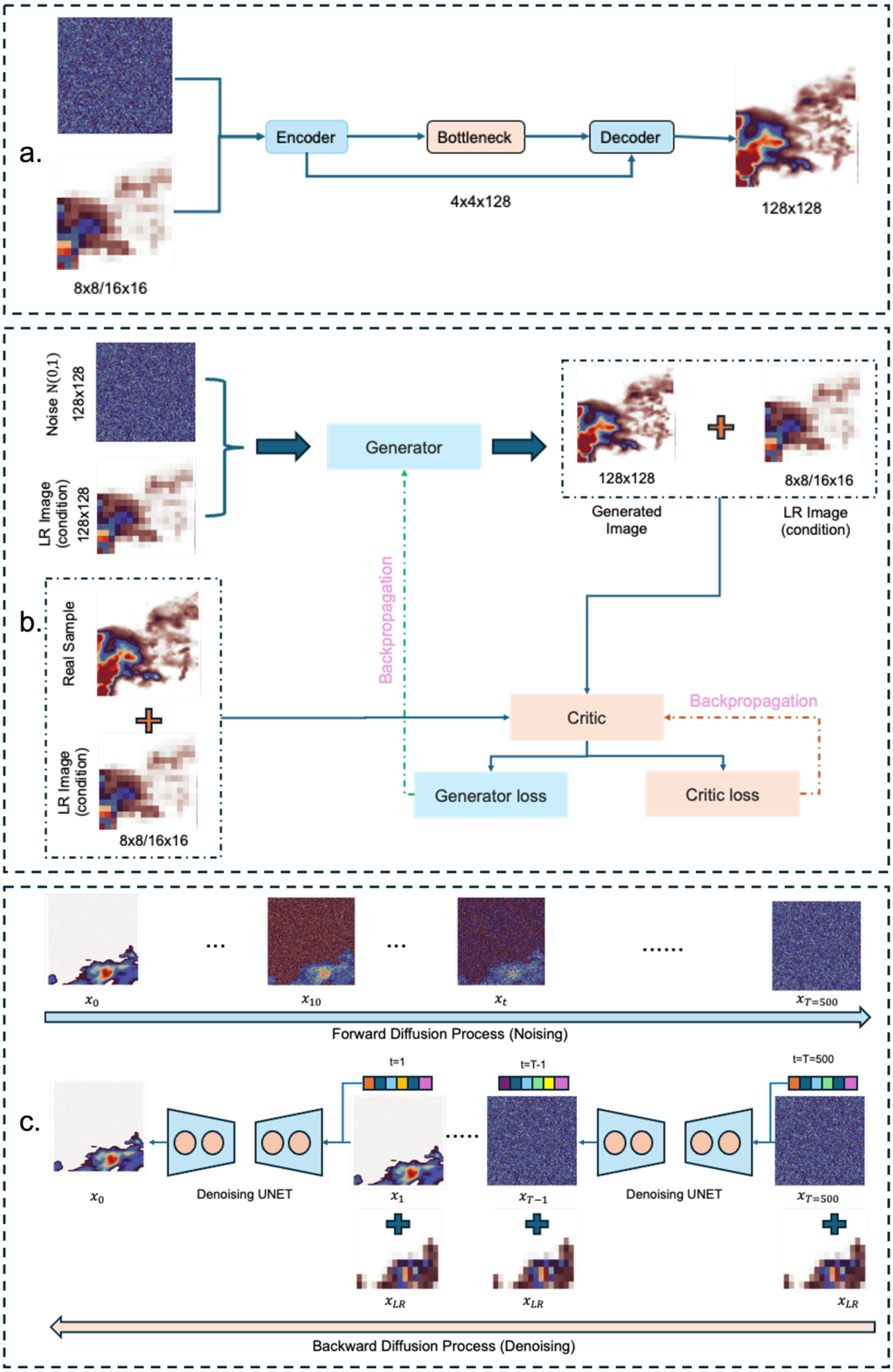


Figure 2: Schematic of the deep-learning architectures used for precipitation super-resolution downscaling. (a) Deterministic U-NET with an encoder–decoder structure and skip connections producing a single high-resolution output. (b) Conditional Wasserstein GAN (WGAN), where a stochastic generator produces 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, corresponding to:

$$\mathcal{L}_{\mathrm{G}} = -\mathbb{E}_{\hat{x}_{\mathrm{HR}} \sim P_g}\left[C_{\psi}(\hat{x}_{\mathrm{HR}}, x_{\mathrm{LR}})\right] \tag{3}$$

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 $\beta_2 = 0.9$. To characterize stochastic variability, ten WGAN models independently initialized with a random seed 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 T(500) steps. During training, the network predicts the added noise at a randomly sampled timestep, 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:

$$\mathrm{q}(x_{1:T} \mid x_0) = \prod_{t=1}^{T} \mathrm{q}(x_t \mid x_{t-1}) \tag{4}$$

$$\mathrm{q}(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathrm{I}\right) \tag{5}$$

where $\{\beta_t\}_{t=1}^{T}$ is the variance 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

$$\bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s) \tag{6}$$

$$\tilde{\alpha}_t = \frac{\cos^2\left(\frac{T/t+s}{1+s}\cdot\frac{\pi}{2}\right)}{\cos^2\left(\frac{s}{1+s}\cdot\frac{\pi}{2}\right)} \tag{7}$$

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}:

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, t, x_{\mathrm{LR}})\|_2^2] \tag{8}$$

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,

$$\gamma(\mathrm{t}) = \left[\sin\left(\frac{t}{10000^{2i/256}}\right), \cos\left(\frac{t}{10000^{2i/256}}\right)\right]_{i=0}^{127} \tag{9}$$

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:

$$\mathrm{FiLM}(h, t_{emb}) = \gamma(t_{emb}) \odot h + \beta(t_{emb}) \tag{10}$$

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. All precipitation fields undergo a $\log(1+x)$ transformation followed by min–max normalization to stabilize training under heavy-tailed rainfall distributions. Training minimizes the noise-prediction loss using the AdamW optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-4}). During inference, HR precipitation samples are generated by iteratively applying the reverse diffusion update,

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, c)\right) + \sigma_t\mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(0, \mathrm{I}) \tag{11}$$

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 and stochastic optimization 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, obtained by bilinearly interpolating 16 × 16 fields to 8 × 8 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 12,616 daily precipitation fields from an 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^{mod}) with the corresponding observational quantile (Q_p^{obs}).

$$Q_p^{\text{mod}} = F_{\text{mod}}^{-1}(p) \qquad (12)$$

$$Q_p^{\text{obs}} = F_{\text{obs}}^{-1}(p) \qquad (13)$$

where $p \in [0,1]$ and F^{-1} denotes the empirical inverse 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:

$$\mathbb{P}(X^{\text{mod}} > x) = 1 - F_{\text{mod}}(x) \qquad (14)$$

where F(x) is cumulative distribution function 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:

$$P_0^{\text{mod}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(X_i^{\text{mod}} \le x_{\text{th}}) \qquad (15)$$

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 evaluate the distributional structure of rainfall intensities predicted from the models, we compute the first four statistical moments for all wet pixels ($X_i >$ 1 mm/day): the mean, variance, skewness, and kurtosis. These moments respectively describe the central tendency, spread, asymmetry, and tail-heaviness of the precipitation distribution properties. Model fidelity in capturing these distributional characteristics is assessed through mean bias and RMSE, which summarize systematic and random departures from the observed moments.

3.2 Mass Conservation

3.2.1 Cumulative Mean Rainfall Depth

We compute mean precipitation at multiple aggregation scales to assess whether rainfall volume is preserved during spatial refinement. For block size d, the observed and predicted mean depths are:

$$P_d^{\text{obs}} = \frac{1}{M}\sum_{j=1}^{M} X_{d,j}^{\text{obs}} \; \boldsymbol{and} \; P_d^{\text{mod}} = \frac{1}{M}\sum_{j=1}^{M} X_{d,j}^{\text{mod}} \tag{16}$$

where M is the number of non-overlapping blocks and denotes the mean precipitation within block j at scale d. Agreement across spatial scales indicates physically consistent redistribution of rainfall mass in downscaling precipitation.

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

$$r(d) = \frac{\sum_{i=1}^{N}(X(s_i) - \bar{X})\,(X(s_i + d) - \bar{X})}{\sqrt{\sum_{i=1}^{N}(X(s_i) - \bar{X})^2 \; \sum_{i=1}^{N}(X(s_i + d) - \bar{X})^2}}, \tag{17}$$

where X denotes either X^{obs} or X^{mod}, $\bar{X}$ is the sample mean of the precipitation field, s_i denotes a spatial location, d is the lag vector, and N is the number of valid grid-point 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 observations (Gilleland et al., 2009; Roberts & Lean, 2008). For a window of size w, the score is defined as

$$\mathrm{FSS}(w) = 1 - \frac{\mathbb{E}[(P_{\mathrm{mod}} - P_{\mathrm{obs}})^2]}{\mathbb{E}\left[P_{\mathrm{mod}}^2 + P_{\mathrm{obs}}^2\right] + \varepsilon} \qquad (18)$$

where P_{mod} and P_{obs} denote the fractional rainfall coverage within windows of size $w$computed from the model prediction and the reference observation, 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,

$$F(k_x, k_y) = \mathcal{F}\{X(x, y)\} \qquad (19)$$

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

$$P(k_x, k_y) = |\ F(k_x, k_y)\ |^2 \qquad (20)$$

which is then azimuthally averaged to obtain the one-dimensional spectrum $P(k)$. Agreement between predicted and observed 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.

3.3.4 Structural Similarity Index (SSIM)

Structural realism is evaluated using the Structural Similarity Index (SSIM), which jointly measures agreement in luminance, contrast, and local spatial structure between the predicted and observed rainfall fields (Meghani et al., 2023; Singh & Goyal, 2023; Zhou Wang et al., 2004). Unlike pixel-wise metrics, SSIM emphasizes coherent patterns such as storm cores, gradients, and spatial organization. SSIM distributions are visualized using violin/box plots, enabling comparison of structural fidelity across models and highlighting variability in performance across the ensemble.

$$\mathrm{SSIM}(X^{\mathrm{mod}}, X^{\mathrm{obs}}) \in [-1,1] \qquad (21)$$

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. (25).

$$0.5 \le \mathrm{AUC} \le 1 \qquad (22)$$

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.

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 evaluated on 12,616 test samples. Substantial differences in computational demand were observed across the three model classes. 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.

Table1: 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	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 42 min	~0.5 sec	500 denoising steps per image
	DDPM (T=100)	~2h 23min	~20 min 53 sec	~0.1 sec	100 denoising steps per image
	DDPM (T=50)	~2h 21min	~10 min 30 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. In contrast, the DDPM required 1 hour and 42 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 (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 (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. Importantly, the loss trajectories remain bounded across all seeds, with no evidence of divergence or mode collapse, highlighting the stabilizing influence of conditional adversarial training even under severe super-resolution.

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 (Figure S2c), indicating strong generalization. Convergence is slower than for U-NET and WGAN, reflecting the iterative denoising process and timestep conditioning intrinsic to diffusion models. In the 16× configuration, a modest widening between training and validation losses emerges after prolonged training, suggesting that the model approaches representational limits imposed by the extremely coarse 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 (Figure 3; Figure S3). 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 attenuates 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, but with slightly smoother textures attributable to the progressive denoising mechanism. 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 U-NET, although fine-scale detail is reduced relative to the 8× case. WGAN continues to generate sharper, more intermittent structures, whereas DDPM maintains coherent storm morphology with comparatively smoother gradients. Overall, 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.

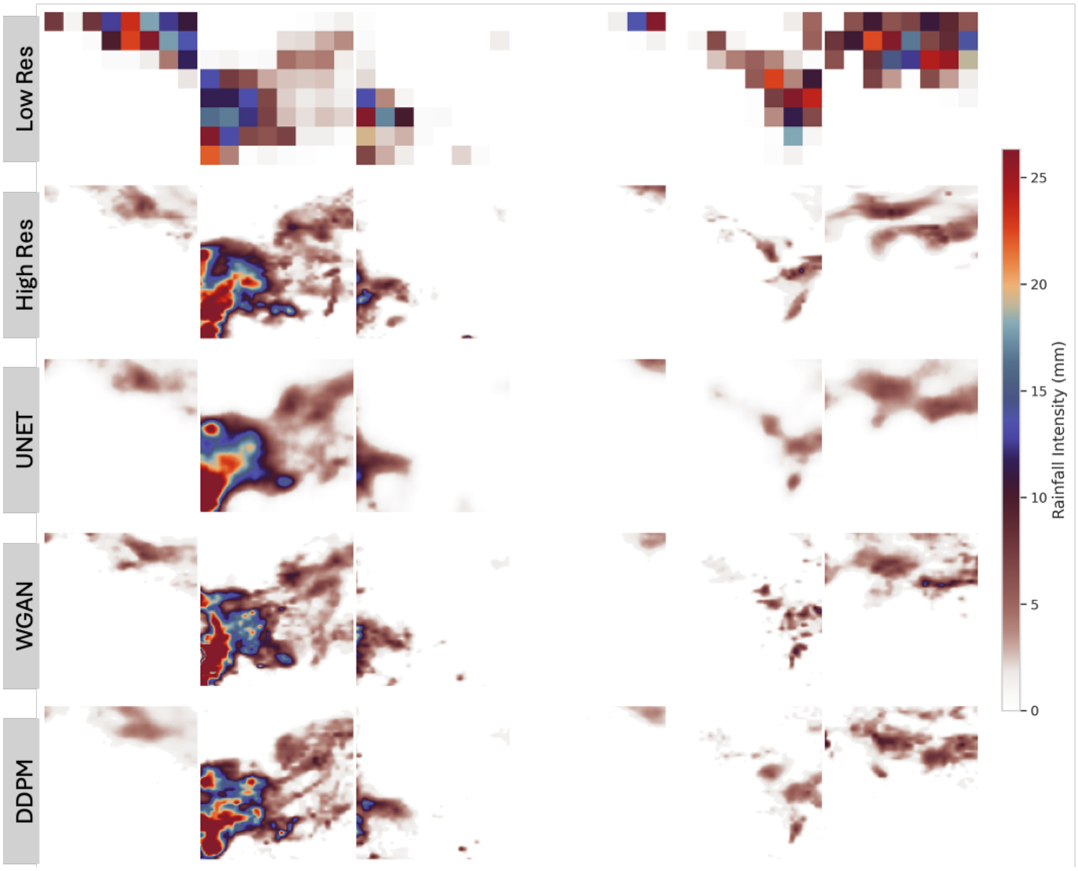


Figure 3. Reconstruction of high-resolution precipitation fields for the 16× downscaling task (8×8 → 128×128). Rows show the low-resolution input, high-resolution ground truth, and predictions from U-NET, WGAN, and DDPM respectively.

5.2.2 Scale-Dependent Rainfall Depth Consistency (Mass Conservation)

The scatter plots of block-averaged precipitation depth demonstrate strong scale-dependent consistency between predicted and observed rainfall across both super-resolution configurations (Figure 4; Figures S4–S5). Despite the absence of any explicit mass-conservation constraint during training, all three models preserve storm-integrated rainfall depth remarkably well. At the native grid scale (1×1), discrepancies are largest, with RMSE dominated by pixel-scale intensity errors and spatial displacement. In the 8× case, U-NET exhibits the lowest RMSE (0.885 mm/day) and negligible bias, while WGAN and DDPM show comparatively larger errors (1.065 mm/day and 1.156 mm/day, respectively), reflecting increased small-scale variability. Nevertheless, even at this finest scale, correlations remain high, indicating that the total precipitation volume is broadly preserved despite local mismatches.

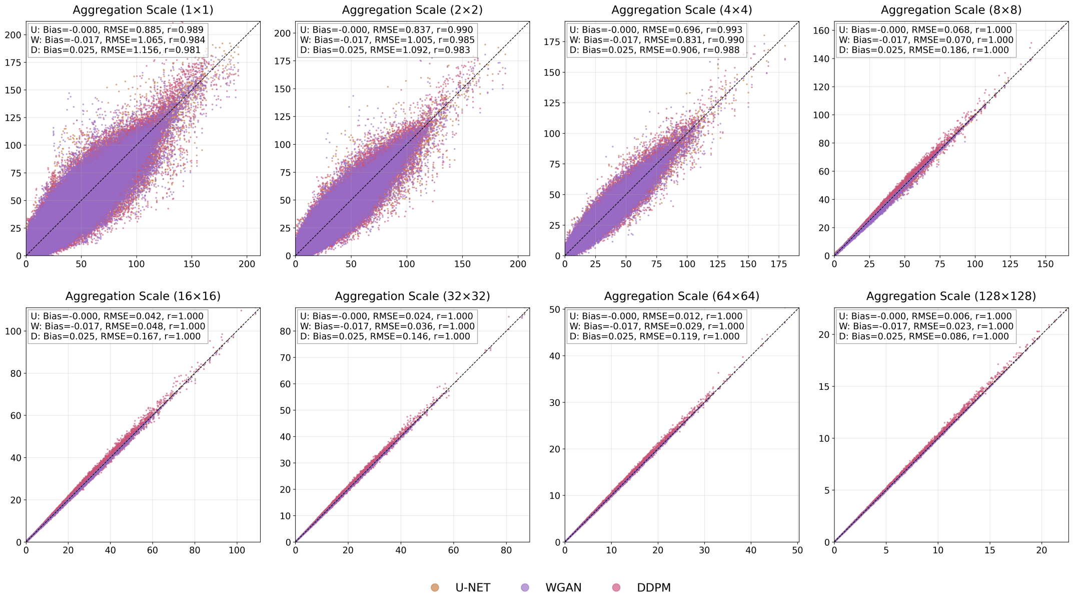


Figure 4. Mass (rainfall depth) consistency across spatial aggregation scales for 8× downscaling. Scatter plots compare block-averaged predicted and observed precipitation depths at aggregation scales of 1×1, 2×2, 4×4, 8×8, 16×16, 32×32, 64×64, and 128×128. Each point represents a spatially aggregated 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.

As aggregation scale increases, scatter collapses rapidly toward the 1:1 line and RMSE decreases sharply for all models. At coarser scales (≥16×16 blocks), biases approach zero and correlations approach unity, demonstrating that errors associated with fine-scale structure largely cancel under spatial averaging. These results indicate that while pixel-level fidelity remains challenging, particularly for generative models, the downscaled fields retain robust depth consistency at hydrologically relevant scales, supporting their applicability for basin-scale water-balance and impact analyses.

5.2.3 Statistical Distribution and Storm Morphology

To ensure consistent wet–dry classification, predicted precipitation values below 1 mm/day were treated as dry and set to zero. The statistical comparison plots show that U-NET, WGAN and DDPM reproduce key distributional properties of the precipitation fields under 8× and 16× downscaling (Figure 5 and Supplementary Information, Figure S6 respectively).

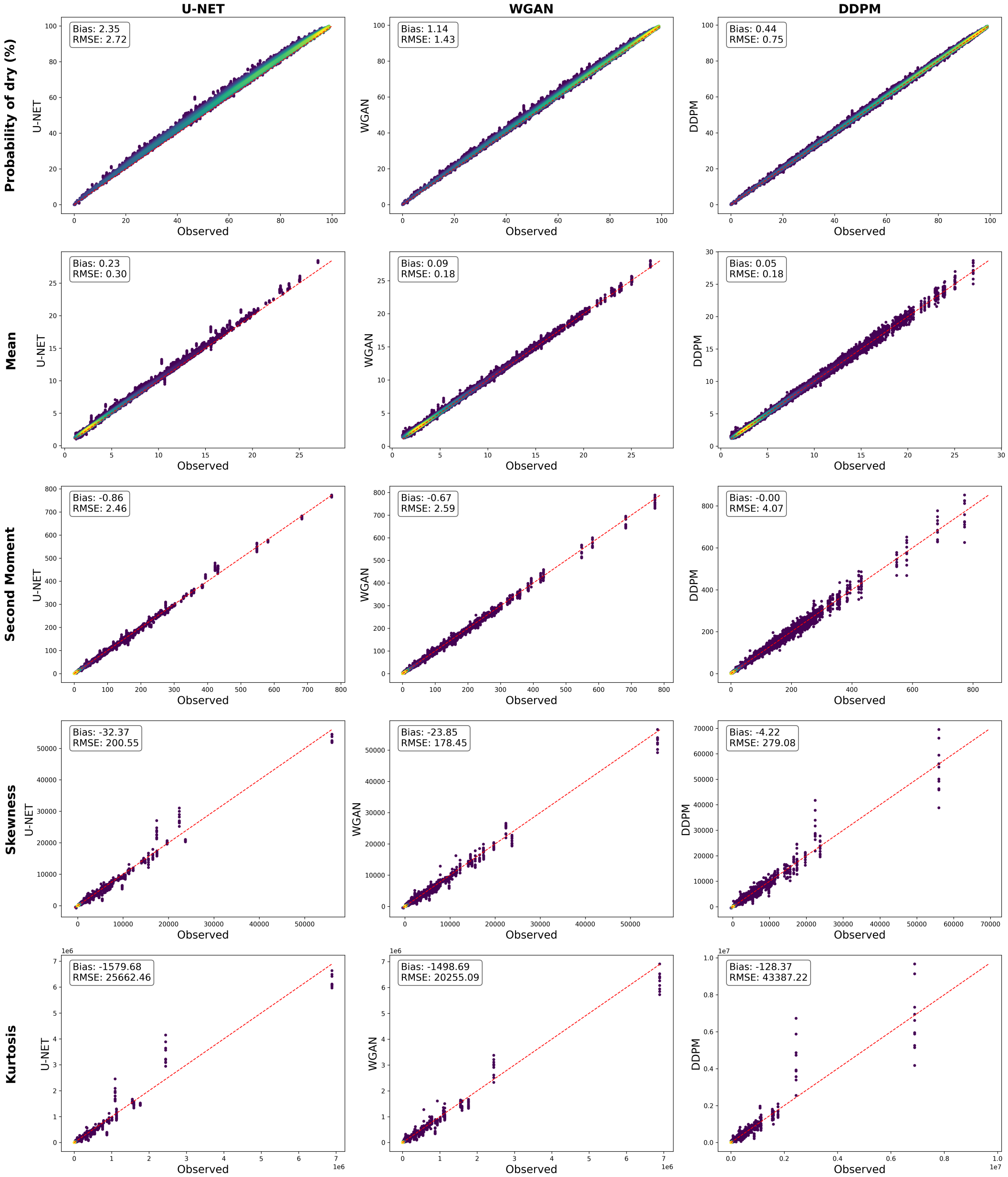


Figure 5. Comparison of observed versus predicted precipitation statistics for 8× downscaling across U-NET, WGAN, and DDPM. Each panel shows the relationship between

observed statistics (x-axis) and model predictions (y-axis) for dry-pixel probability (P_0), mean, second moment, skewness, and kurtosis. Bias and RMSE are reported for each metric.

All models align closely with the 1:1 reference line for dry-pixel probability and mean rainfall, indicating accurate reconstruction of occurrence frequency and overall storm magnitude. U-NET and WGAN exhibit particularly small deviations for these two metrics, while DDPM achieves the lowest RMSE for P_0 in the 16× case. We observe more pronounced differences for higher-order statistics. The second moment is captured reasonably well by all models, although DDPM shows a larger spread at higher values, especially in the 8× configuration. Skewness and kurtosis exhibit the largest errors across models, reflecting the difficulty of recovering tail behavior and extreme-event structure. WGAN generally shows smaller bias in skewness, whereas U-NET and DDPM display higher variance. Despite these challenges at higher-order moments, all models maintain consistent performance trends across both downscaling factors, with U-NET providing stable low-order statistics, WGAN delivering sharper distributional structure, and DDPM capturing overall variability while showing higher spread for extremes.

The spatial autocorrelation analysis illustrates how effectively each model preserves fine-scale structure as pixel lag increases. Under the 8× configuration, all three architectures reproduce the observed correlation decay well at short lags ($1-3$ pixels), indicating that they can recover local spatial gradients when the coarse input still contains partial storm structure (Figure 6). U-NET yields the highest correlations at the smallest lags, consistent with its tendency to generate smooth and spatially coherent fields. WGAN produces slightly lower short-lag correlations but aligns more closely with the observed decay pattern at intermediate lags ($4-6$ pixels), reflecting its capacity to introduce sharper gradients and finer texture. DDPM follows a similar trajectory to WGAN, with correlations slightly below WGAN at short lags but closer to observations across mid-range scales, suggesting a balanced reconstruction of local variability and structural detail. Ensemble spread is wider for WGAN and DDPM than for U-NET, reflecting the stochastic nature of generative sampling and the diversity of high-resolution realizations.

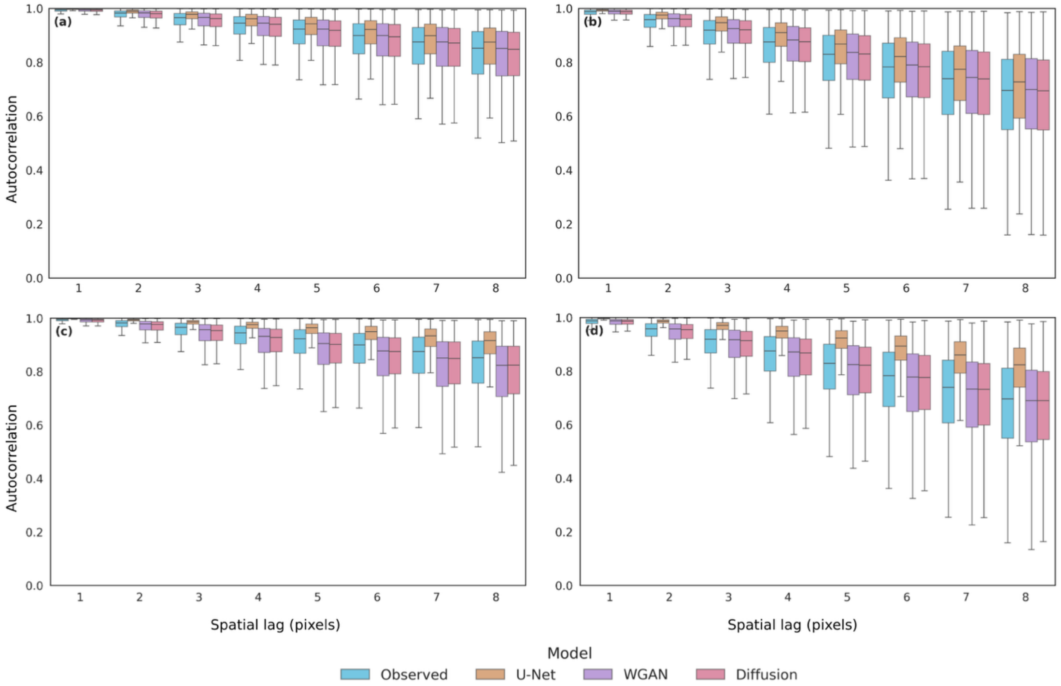


Figure 6. Spatial autocorrelation of observed and predicted precipitation fields for horizontal and vertical pixel lags (1–8 pixels) under 8× (top row) and 16× (bottom row) downscaling. Boxplots summarize ensemble variability across 10 seeds for each model (U-NET, WGAN, DDPM). Observed correlations are shown in green.

Under the more demanding 16× configuration, all models show weakened spatial structure, with reduced correlations and greater spread across lags (Figure 6). U-NET continues to produce the highest lagged autocorrelations and fails to reproduce the observed rate of decorrelation (decay in autocorrelation), indicating over smoothing when limited information is available in the 8 × 8 inputs. WGAN and DDPM better capture the observed decline in autocorrelation with increasing lag, although WGAN exhibit slightly broader ensemble variability due to uncertainty introduced at this extreme downscaling ratio in vertical direction. WGAN maintains the closest agreement with observed correlations at intermediate lags, preserving sharper spatial transitions, while DDPM yields slightly lower correlations but relatively consistent behavior across lags. Overall, for 16× downscaling, deterministic models over-smooth the fields, while generative models capture the observed spatial dependence more accurately but with higher ensemble dispersion due to increased uncertainty.

https://doi.org/10.5194/egusphere-2026-861
Preprint. Discussion started: 27 February 2026

5.2.4 Extreme Precipitation and Tail Behavior

The exceedance probability analysis evaluates the ability of each model to reproduce the upper tail of the precipitation distribution (Figure 7). Under the 8× configuration, all models broadly follow the observed exceedance behavior at moderate intensities, with clear divergence emerging at the highest values. U-NET systematically underestimates extreme precipitation, consistent with regression-induced smoothing. WGAN exhibits the closest agreement with the observed upper tail, maintaining stronger alignment at high intensities, while DDPM performs comparably at low to intermediate values but displays greater ensemble spread at the most extreme quantiles, reflecting increased sampling variability for rare events.

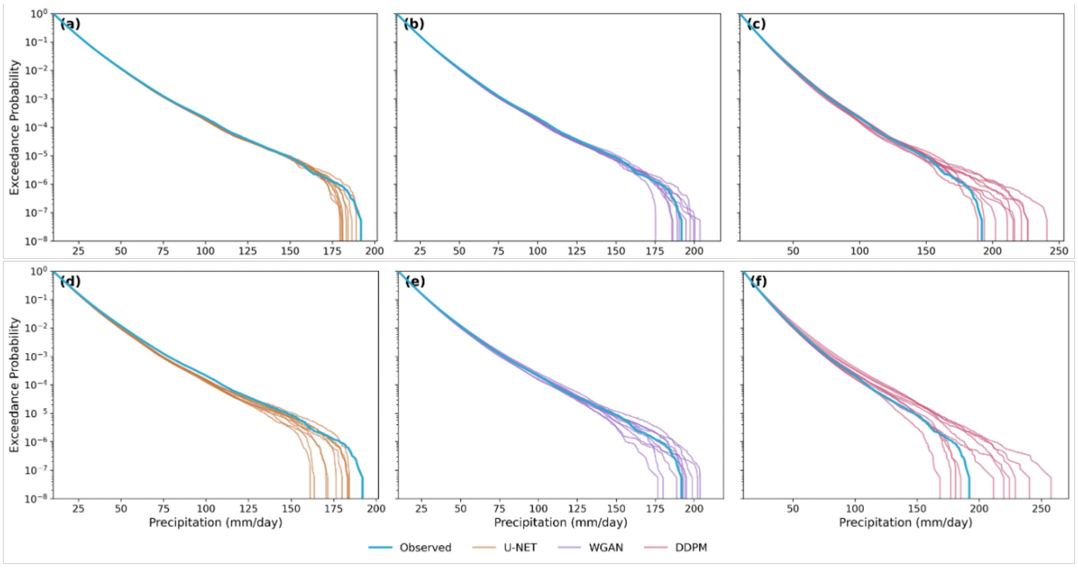


Figure 7. Exceedance probability (1–CDF) curves of daily precipitation for observations and model predictions at 8× (top row) and 16× (bottom row) spatial downscaling. Panels show results for (a, d) U-NET, (b, e) WGAN, and (c, f) DDPM. The observed precipitation is shown by the thick blue curve, while thin colored curves represent 10 ensemble members for each generative model. All curves are computed using the same test samples and plotted on a logarithmic exceedance scale, highlighting differences in the representation of extreme precipitation tails across downscaling approaches and spatial scales.

Under the more challenging 16× configuration, deviations from observations become more pronounced across all models due to the severely limited information content of the coarse inputs. U-NET further suppresses high-intensity events, whereas WGAN retains the closest correspondence to the observed tail despite increased ensemble dispersion. DDPM captures the overall tail shape but exhibits the largest spread among realizations, indicating increased uncertainty in reconstructing extremes at this downscaling ratio. Overall, generative models, particularly WGAN, better preserve the heavy-tailed nature of precipitation, while deterministic regression systematically attenuates extremes.

The corresponding Q–Q diagnostics (Figures S7–S8) support these findings, showing consistent underestimation of upper quantiles by U-NET, closer alignment by WGAN at high intensities, and broader dispersion for DDPM in the upper tail. Together, these results indicate that differences among models are dominated by extreme-value behavior: generative approaches better preserve tail behavior but exhibit increased ensemble variability, particularly under severe super-resolution, whereas deterministic methods favor stability at the expense of extreme-value fidelity.

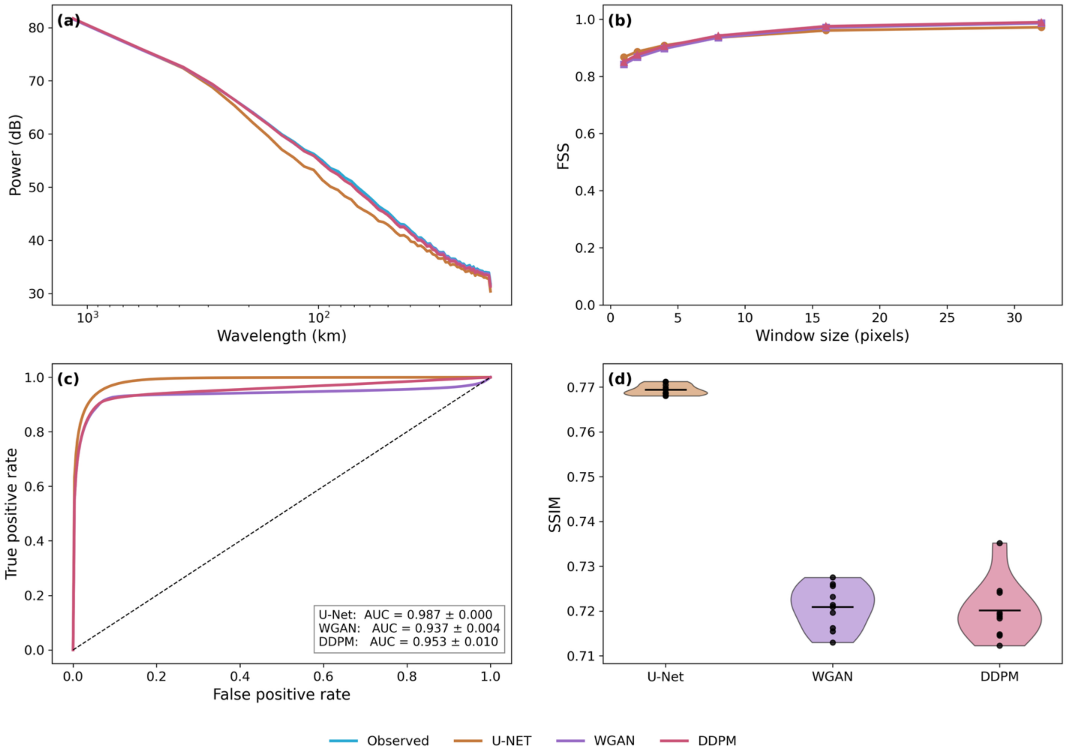
5.2.5 Composite Diagnostics

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). The radial power spectra indicate that all three models accurately reproduce the observed large-scale energy content, with close agreement at wavelengths exceeding several hundred kilometers, demonstrating robust preservation of synoptic-scale precipitation organization. At smaller spatial scales, however, U-NET exhibits a pronounced loss of spectral power, reflecting excessive smoothing and suppression of fine-scale variability, whereas WGAN and DDPM retain substantially more energy and remain closer to the observed spectrum, indicating improved representation of small-scale spatial intermittency and storm texture.

These scale-dependent differences are consistent with the Fractions Skill Score (Figure 8b), which increases monotonically with window size for all models, reflecting reduced sensitivity to displacement errors at larger spatial scales. U-NET shows marginally higher skill at the smallest windows due to its smoother fields, while WGAN and DDPM converge rapidly and achieve comparable skill at moderate and large window sizes, indicating that enhanced small-scale variability does not compromise spatial agreement at physically meaningful scales. The ROC curves demonstrate strong precipitation occurrence discrimination for all models, with high AUC values indicating reliable separation between wet and dry pixels; U-NET attains the highest AUC, followed by DDPM and WGAN, consistent with the more conservative nature of deterministic predictions versus the sharper, more variable fields produced by generative models (Figure 8c). Finally, the SSIM distributions summarize structural similarity across inference ensembles, showing that U-NET yields the highest median SSIM with minimal spread, while WGAN and DDPM exhibit lower medians and broader distributions, reflecting increased structural diversity and stochasticity. Overall, the diagnostics indicate that while all models perform comparably in capturing large-scale precipitation characteristics, the generative approaches better reproduce fine-scale spatial variability. At the 16× downscaling factor, DDPM achieves a more physically coherent

https://doi.org/10.5194/egusphere-2026-861
Preprint. Discussion started: 27 February 2026

balance between spatial organization and ensemble diversity, whereas WGAN emphasizes sharper extremes with comparatively higher structural variability.



U-Net: AUC = 0.987 ± 0.000
WGAN: AUC = 0.937 ± 0.004
DDPM: AUC = 0.953 ± 0.010

Figure 8. Multi-metric evaluation of 16× precipitation downscaling performance across 10 independently trained models for each architecture. Panels show (a) ensemble-mean radial power spectra, (b) mean Fractions Skill Score (FSS) as a function of spatial window size, (c) mean ROC curves for precipitation occurrence with AUC reported as mean ± standard deviation across models, and (d) distributions of mean SSIM values summarizing structural similarity to observations.

5.3 Sensitivity of DDPM to Diffusion Length

We also investigated the sensitivity of DDPM performance to different diffusion lengths (T = 500, 100, 50) by comparing key spatial and distributional characteristics. A primary motivation for this analysis was the substantial difference in inference cost between these models. While U-NET and WGAN generate high-resolution precipitation fields in a single forward pass, DDPM relies on iterative sampling, with inference time scaling approximately linearly with the number of diffusion steps. To assess whether shorter diffusion schedules could offer a practical compromise between computational efficiency and predictive skill, we

https://doi.org/10.5194/egusphere-2026-861
Preprint. Discussion started: 27 February 2026

compared DDPM ensembles trained with T = 500 and T = 100 across 10 independent seeds (Figure 9).

The lagged spatial correlation analysis shows that both diffusion configurations reproduce the physically consistent decay of spatial dependence with increasing lag in both the horizontal and vertical directions. However, the T = 500 configuration maintains systematically higher correlations and reduced inter-seed variability at intermediate and larger lags, indicating stronger spatial coherence. In contrast, the T = 100 configuration exhibits modestly lower correlations and a broader spread, reflecting increased variability when fewer diffusion steps are used. Despite these differences, both configurations preserve the dominant spatial dependence structure of the observed precipitation fields, indicating that the overall spatial realism of DDPM is retained even when inference cost is substantially reduced.

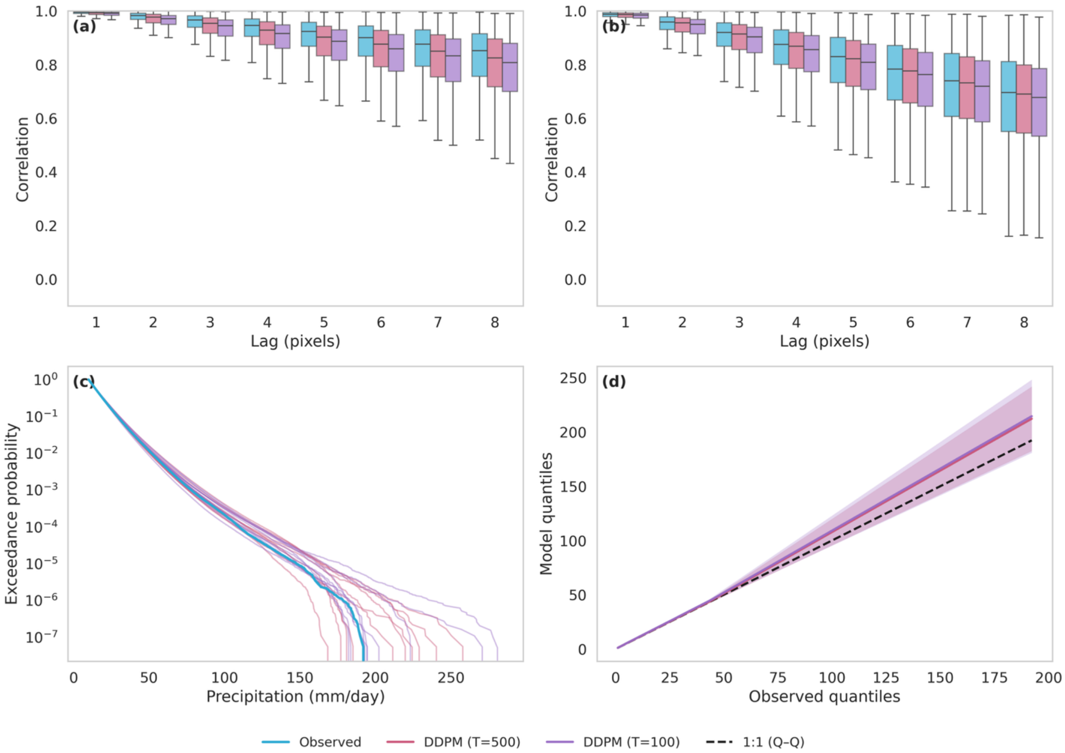


Figure 9. Comparison of DDPM downscaling performance (16×) across 10 random seeds for two diffusion lengths, T = 500 and T = 100. Panels (a–d) summarize complementary aspects of spatial structure and extremes: (a) and (b) show lagged spatial correlation (lags 1–8 pixels) in the horizontal and vertical directions, respectively, with boxplots summarizing inter-seed variability and the observed correlations shown for reference. (c) shows

exceedance probability (1–CDF) curves for precipitation intensities above 10 mm/day, where colored lines represent individual seeds for each diffusion setting. (d) presents Q–Q plots of precipitation quantiles above 10 mm/day, with shaded bands indicating inter-seed variability.

Differences between diffusion lengths are also impacting the representation of extreme precipitation. The exceedance probability curves show that both configurations capture the general shape of the upper tail, but the T = 100 ensemble exhibits a slight overestimation at high intensities, with some realizations producing heavier tails. This behavior is also shown in the joint Q − Q plots (Figure 9). These results indicate that reducing diffusion length slightly amplifies small-scale noisiness and extreme-value sampling, rather than altering the central tendency or overall structure of the precipitation distribution.

Additional experiments using an even shorter diffusion length (T = 50), presented in Figure S9 of the Supplementary Information, further illustrate the trade-off between inference cost and predictive uncertainty. While the T = 50 configuration continues to reproduce the overall spatial organization and distributional shape of precipitation, the bias in the upper tail becomes more pronounced. Across diffusion lengths (T = 500, 100, and 50), DDPM consistently captures the characteristic decay of spatial dependence and the general structure of the precipitation distribution; however, larger diffusion lengths yield systematically closer agreement with observations and reduced ensemble spread, particularly for the 16× downscaling case. These results indicate that diffusion depth primarily controls the strength and coherence of spatial structure rather than altering the fundamental behavior of the model. This analysis underlines a practical trade-off in diffusion-based downscaling that longer diffusion schedules provide more constrained and spatially coherent realizations at higher computational cost, whereas shorter schedules substantially reduce inference time at the expense of increased ensemble variability, and overestimation of extreme precipitation.

6. Conclusions

This study presents a systematic comparison of deterministic and generative deep-learning approaches for precipitation super-resolution under challenging 8× and 16× downscaling tasks. Using a consistent experimental framework, we evaluated U-Net, WGAN, and DDPM across complementary diagnostics of statistical fidelity, spatial structure, and extreme-value behavior. All three models preserve aggregate rainfall mass despite the absence of explicit conservation constraints, with spatial aggregation showing rapid convergence toward near-perfect agreement at coarse scales. Model differences emerge primarily at fine spatial scales, particularly in the representation of extremes and spatial dependence. Importantly, the cross-regional design provides insight into model robustness across distinct hydroclimatic regimes, highlighting performance under distribution shift between training and evaluation domains.

U-NET demonstrates strong stability, structural consistency, and precipitation-occurrence skill, yielding high SSIM and ROC performance and smooth spatial fields that perform well at short spatial lags. However, this deterministic smoothing suppresses small-scale variance, attenuates spectral power, and systematically underestimates extremes, particularly under 16× downscaling. The generative models provide complementary strengths. WGAN more effectively captures fine-scale variability, spatial dependence, and upper-tail behavior, producing sharper precipitation structures and improved extreme-value statistics, though with increased variability and reduced structural consistency at small scales. DDPM offers a balanced alternative, maintaining coherent multi-scale storm morphology while explicitly representing conditional uncertainty through stochastic sampling. This uncertainty manifests as increased ensemble spread in extreme-value diagnostics, especially at high downscaling factors.

Sensitivity experiments further show that diffusion length primarily governs a trade-off between computational efficiency and spatial coherence: reducing the number of diffusion steps increases variance and weakens fine-scale structure while preserving the overall distributional and morphological characteristics of precipitation. These results highlight fundamental trade-offs among determinism, spatial realism, uncertainty representation, and computational cost. No single approach is uniformly optimal; instead, model selection should be guided by application-specific priorities. U-NET is well suited for tasks emphasizing stability, occurrence detection, and computational efficiency, whereas generative models are preferable when accurate representation of spatial variability and extremes is critical. Among the generative approaches, WGAN most closely reproduces observed spatial correlations and extreme-value behavior, while DDPM provides physically coherent ensemble realizations with explicit uncertainty representation at higher computational cost.

Several limitations necessitate future investigation. The analysis is conducted under a perfect-model framework and focuses on a single region and resolution pair. Diffusion-based inference remains computationally expensive, and further work is needed to assess scalability for operational applications. Future research should explore conditioning on additional physical predictors, hybrid deterministic–generative architectures, explicit physical constraints, and calibration strategies to better control ensemble spread. Overall, this study demonstrates that generative downscaling methods offer clear advantages for representing fine-scale spatial structure and extremes, while deterministic approaches remain valuable for stable and efficient precipitation reconstruction under increasing resolution demands.

https://doi.org/10.5194/egusphere-2026-861
Preprint. Discussion started: 27 February 2026

Data Availability

ERA5-Land precipitation data used in this study are publicly available from the Copernicus Climate Change Service (C3S) Climate Data Store. ERA5-Land provides global land-surface variables at approximately 9-km spatial resolution and hourly temporal resolution. The data can be accessed through the C3S Climate Data Store (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land?tab=download) (Muñoz Sabater, 2019). All preprocessing steps applied in this study are described in the Methods section.

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.

Acknowledgement

We gratefully acknowledge the financial support provide by the Environmental Institute, University of Virginia to carry out this research work.

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

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

Meghani, S., Singh, S., Kumar, N., & Goyal, M. K. (2023). Predicting the spatiotemporal characteristics of atmospheric rivers: A novel data-driven approach. *Global and Planetary Change*, *231*, 104295. https://doi.org/10.1016/j.gloplacha.2023.104295

Muñoz Sabater, J. (2019). *ERA5-Land hourly data from 1950 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.e2161bac

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

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*.

Singh, S., & Goyal, M. K. (2023). An innovative approach to predict atmospheric rivers: Exploring convolutional autoencoder. *Atmospheric Research*, *289*, 106754.

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

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