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

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

Response Sheet for GMD manuscript

Reviewer 1

This manuscript presents a timely and meaningful intercomparison of three “most used” gen AI models for precipitation super-resolution downscaling. The study addresses an important problem at the interface of atmospheric science and machine learning, and the effort to compare multiple model classes within a unified framework is valuable. At the same time, several aspects of the manuscript require substantial clarification and strengthening before the conclusions can be fully supported. Addressing these issues would improve the scientific rigor and impact of the paper.

Response: We sincerely thank the reviewer for the careful and insightful evaluation of our manuscript. We appreciate the recognition of the importance of the problem and the value of a unified comparison framework. We have carefully revised the manuscript to address all comments raised. Detailed responses to each comment are provided below.

Major Comments

Comment1

The authors state that the 10-member ensemble is generated from 10 independently trained models initialized with different random seeds. This procedure primarily reflects epistemic uncertainty associated with parameter estimation and training variability. However, the central theoretical advantage of conditional generative models is that for a given low-resolution input, they can generate a distribution of plausible high-resolution outputs through stochastic sampling. At present, the manuscript uses one prediction from each independently trained model and interprets the resulting spread as ensemble uncertainty, which is not equivalent to sampling the conditional output distribution of a single trained generative model. The authors must separate these two uncertainty sources explicitly. In addition to the current analysis, they should report results from repeated stochastic sampling using a single trained model, preferably the best-performing checkpoint, and compare that spread with the spread arising from different training seeds. This distinction is essential for a correct interpretation of the ensemble results.

Response: Thank you for your suggestion. We ignored this aspect to avoid confusion between these two types of stochasticity in our original version. We agree that since we are comparing generative models, we should also explore the stochastic generalization capabilities of these generative models. We have included some of the results (figures) and discussion around them in the revised manuscript.

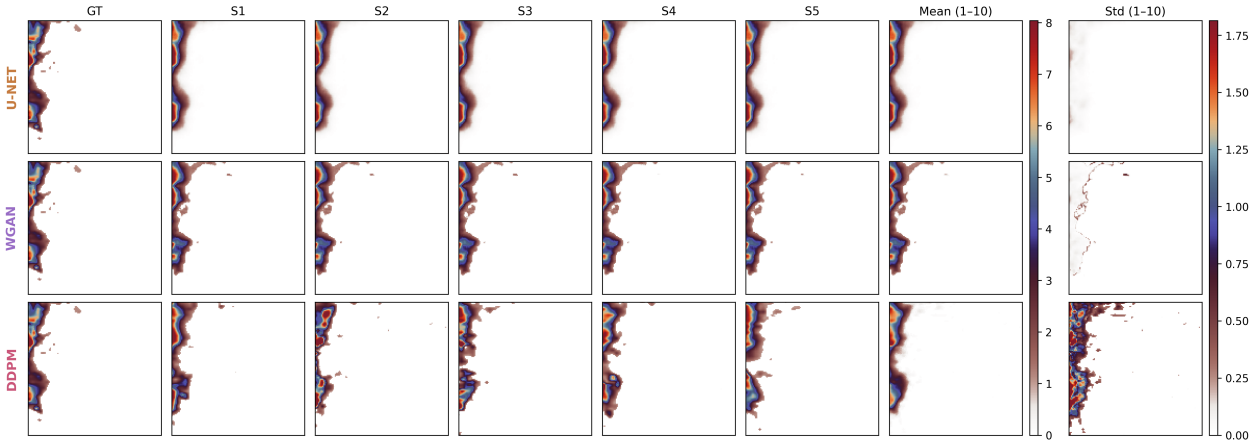

Figure 1a: Representative stochastic realizations and ensemble statistics for a randomly selected event. Each row shows results from U-Net, WGAN, and DDPM, respectively. Columns show the ground truth (GT), five stochastic samples (S1–S5), the ensemble mean across 10 realizations, and the corresponding ensemble standard deviation. The figure highlights differences in spatial realism, stochastic variability, and uncertainty structure across the three generative approaches.

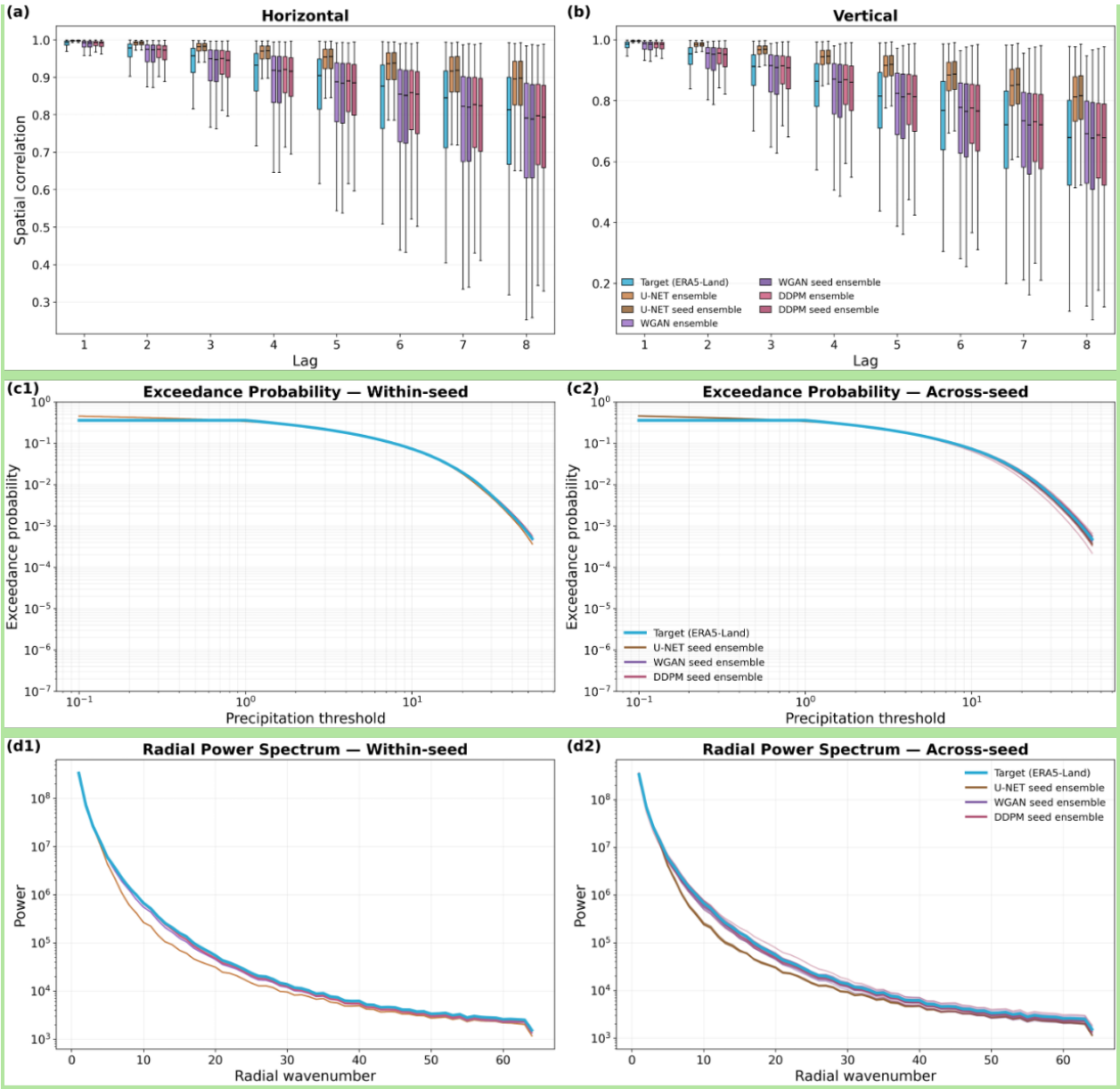

Figure 1b: Comparison of within-seed stochastic variability and across-seed training variability in spatial structure and precipitation statistics for generative downscaling models. Panels (a) and (b) show horizontal and vertical lagged spatial autocorrelation, respectively, summarized as boxplots across smaller test sample (1000) from U-Net, WGAN and DDPM against target precipitation (ERA5-Land), within-seed ensembles (multiple stochastic realizations from a fixed trained model), and across-seed ensembles (single realizations from independently trained models). Panels (c1) and (c2) present exceedance probability curves for precipitation intensity for within-seed and across-seed variability, respectively. Panels (d1) and (d2) show the corresponding radial power spectra. In exceedance and spectral panels, colored lines represent ensemble members, while the cyan line denotes observations. Together, these diagnostics distinguish uncertainty arising from stochastic sampling within a trained model from variability introduced by different random training initializations.

Across the three models, distinct responses to within-seed and across-seed variability are evident. U-Net exhibits the most compact spread in both within-seed and across-seed settings, indicating comparatively stable behavior under repeated stochastic sampling and retraining. Its lagged autocorrelation boxplots remain relatively narrow, and its exceedance probability and radial power spectrum curves show limited divergence, suggesting that the deterministic backbone of the architecture constrains variability. WGAN shows larger spread than U-Net, particularly in the across-seed autocorrelation and spectral diagnostics, implying stronger sensitivity of learned spatial texture to training initialization. DDPM generally exhibits the largest variability across seeds, with broader autocorrelation distributions and larger deviations in the power spectrum, indicating that diffusion-based generation is more sensitive to optimization pathway and model realization. However, within-seed variability for DDPM remains more controlled than its across-seed variability, implying that retraining brings additional uncertainty compared to stochastic sampling alone.

Considering all models collectively, a consistent contrast emerges between within-seed and across-seed behavior. Within-seed variability is generally smaller and more structured, meaning that once a model is trained, multiple stochastic realizations tend to preserve similar spatial coherence, intensity distributions, and multiscale variance. Across-seed variability is systematically larger, especially in lagged autocorrelation and radial power spectrum, demonstrating that training initialization alters the learned spatial organization more strongly than inference-time randomness. The exceedance probability curves remain comparatively stable in both cases, indicating that rainfall intensity statistics are more reproducible than spatial structure. Overall, the results suggest that uncertainty in generative precipitation downscaling is dominated less by stochastic sampling from a trained model and more by differences among independently trained model realizations, with this effect weakest for U-Net and strongest for DDPM.

Comment 2

The manuscript refers in several places to ERA5-Land as “observation”. This terminology is not correct. ERA5-Land is a reanalysis-based product, not a direct observational dataset. Since the study does not use in situ station obs, radar, satellite retrievals, or soundings as reference truth, the manuscript should consistently refer to ERA5-Land as a reanalysis or reanalysis-based target, not as observation. You could read this paper to obtain the detailed reason.

https://doi.org/10.1175/BAMS-D-14-00226.1

Response: Thank you for pointing this out. We agree ERA5-Land is a reanalysis data not observed. We intended to say it observed on behalf of our models to refer the target data but to avoid inconsistency and clarity, we will make sure to correct this in revised manuscript to Target (ERA5-Land) from Observed as suggested.

Comment 3

The use of min-max normalization may help stabilize training, but it raises an important concern for precipitation, especially for extreme events. Min-max scaling bounds the normalized target by the range seen in the training data, which may hinder robust extrapolation to unprecedented values. This issue is especially relevant for climate-related downscaling and extreme precipitation, where out-of-sample events may exceed the historical training maximum. This issue may also be relevant to the behavior shown in Figure 2, where DDPM with T=100 approaches the upper bound and cannot grow freely. The authors should discuss this limitation explicitly and test at least one alternative normalization strategy such as quantile normalization or z-score normalization over wet pixels only, reporting whether the normalization choice materially changes the extreme-value results.

Response: Thank you for raising this important point. We agree that normalization can influence the behavior of generative models, particularly for heavy-tailed variables such as precipitation and for the representation of extreme events. Our treatment of normalization differed across model classes because of differences in training dynamics and objective functions. For the U-Net and WGAN frameworks, models were trained directly on the original transformed precipitation field without an additional bounded min–max target scaling step. In contrast, DDPM training involves repeatedly adding and removing Gaussian noise across many diffusion steps, where maintaining a standardized target scale substantially improves numerical stability, noise scheduling consistency, and convergence. For this reason, normalized targets are commonly used in diffusion-based image generation workflows.

We also agree that min–max normalization may, in principle, constrain extrapolation beyond the historical training range. To examine whether our conclusions depended strongly on this choice, we performed an additional sensitivity experiment as suggested by the reviewer using an alternative z-score normalization strategy and retrained the DDPM across 10 independent seeds. We then compared these results with the original 10-seed min–max normalized DDPM ensemble (Figure 2). The two normalization strategies produced broadly consistent behavior across spatial correlation structure, exceedance probability, and radial power spectra, with only modest quantitative differences. In particular, the overall conclusions regarding model variability and stochastic behavior remained unchanged.

These additional experiments suggest that, for the present dataset, the normalization choice does not materially alter the key findings, although it can influence some aspects of tail behavior and fine-scale variance. We have now added discussion in the revised manuscript noting that normalization remains an important design choice for diffusion-based precipitation downscaling, especially for applications targeting unprecedented future extremes, where adaptive or tail-aware transformations (e.g., quantile-based or wet-pixel standardization approaches) may be valuable directions for future work.

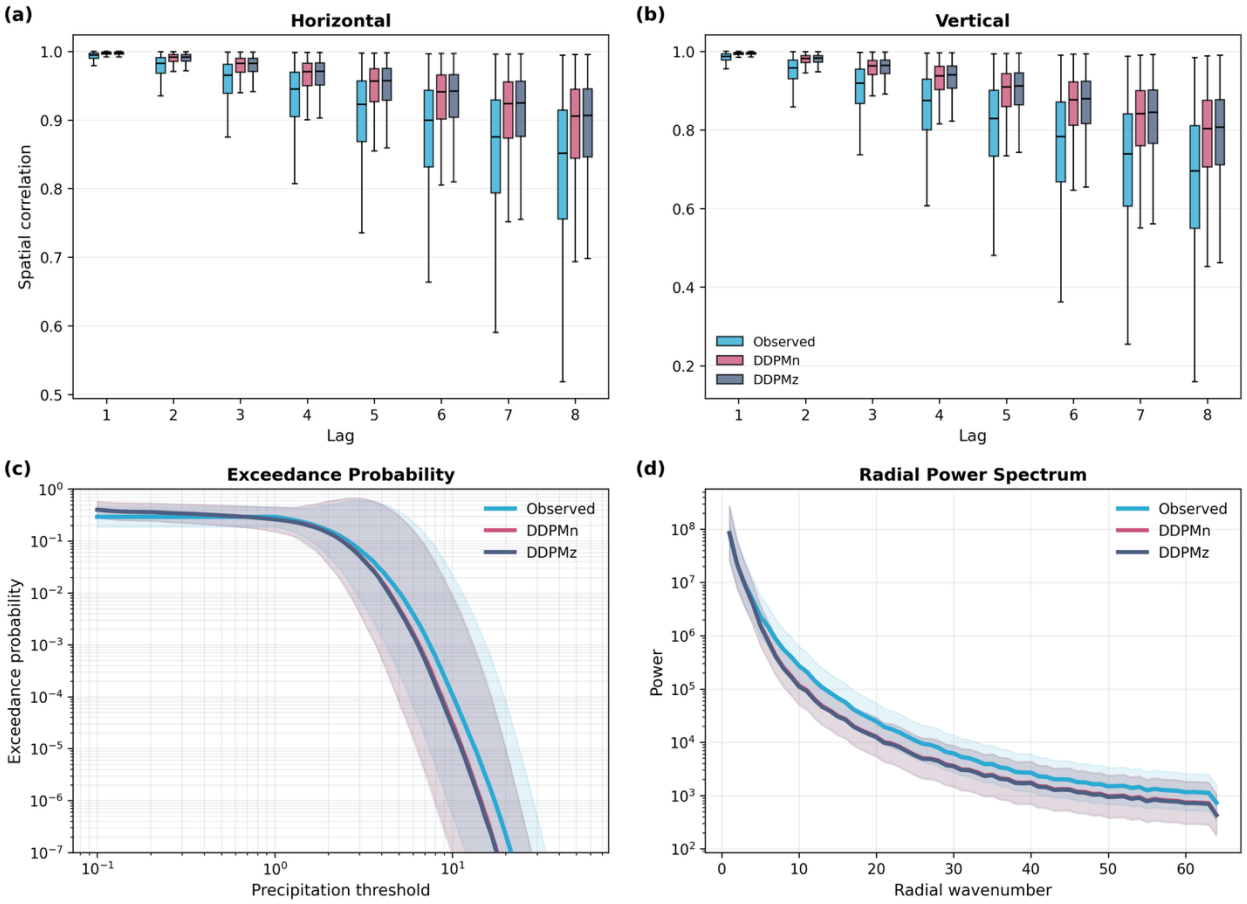

Figure 2: Comparison of spatial dependence and distributional characteristics between ERA5-Land target fields and DDPM predictions trained with min–max normalization (DDPMn) and z-score normalization (DDPMz). (a) Horizontal lagged spatial correlation and (b) vertical lagged spatial correlation shown as boxplots across evaluation samples for lags of 1–8 grid cells. ERA5-Land target fields are shown in cyan, DDPMn in pink, and DDPMz in dark blue. Both DDPM models reproduce the general decay of spatial correlation with increasing lag. (c) Exceedance probability curves showing the probability of precipitation intensity exceeding a given threshold on logarithmic axes. Both models capture the overall tail behaviour. Shaded regions denote spread across seeds, represented as mean ± 0.5 standard deviation. (d) Radial power spectrum as a function of radial wavenumber, illustrating the distribution of variance across spatial scales.

Comment 4

Precipitation is not a typical continuous variable like temperature, pressure, or geopotential height. It is sparse, intermittent, highly skewed, and often better represented by zero-inflated or Tweedie-like distributions. For this reason, the loss-function choice deserves much more discussion than it currently receives. The authors should discuss why their selected losses are appropriate for precipitation specifically, and whether distribution-aware losses could improve tail behavior and wet-day occurrence. Recent studies suggest that distributional losses can be beneficial for

precipitation prediction and downscaling. At minimum, this should be discussed more clearly. Ideally, the authors would include a sensitivity test or ablation experiment.

https://arxiv.org/html/2509.08369

https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024GL111828

Response: We thank the referee for this thoughtful and important comment. We agree that precipitation differs fundamentally from approximately Gaussian continuous variables such as temperature or geopotential height, as it is intermittent, zero-inflated, highly skewed, and characterized by heavy tails. Consequently, the choice of training loss is particularly important for precipitation modeling. In the present study, our objective was to compare three widely used generative frameworks under their commonly adopted baseline training formulations: MSE for the U-Net, adversarial Wasserstein loss for the WGAN, and noise-prediction loss for the diffusion model. This provides a controlled comparison of representative deterministic and stochastic approaches. However, we agree that distribution-aware objectives (e.g., Tweedie deviance, quantile-based losses, or hybrid occurrence–intensity formulations) may further improve the representation of wet-day occurrence and extreme precipitation behavior.

We appreciate the suggestion and note that recent studies have highlighted the benefits of precipitation-specific or distribution-sensitive losses for sparse and skewed rainfall data (e.g., Hunt, 2025; Rastogi et al., 2025). We will expand the revised manuscript to discuss these considerations more explicitly. While a full sensitivity or ablation study of alternative loss functions is beyond the current scope, we agree this is an important direction for future work.

Comment 5

Figure 5 appears to compare model predictions against an upsampled low-resolution field rather than the native high-resolution ERA5-Land target, given the clustering of identical reference values on the x-axis. If so, the comparison is not appropriate and the figure needs to be redone using the actual high-resolution target field. If this interpretation is incorrect, the authors should clarify exactly how the reference field in Figure 5 was constructed.

Response: Thank you for the comment. For each model class (U-NET, WGAN, and DDPM), predictions from 10 independently initialized and trained seeds were evaluated against the same high-resolution ERA5-Land test targets, which is why the same reference statistic appears repeatedly along the x-axis. This repeated use of the same target values across multiple seed realizations leads to visible clustering in the scatter plots. Therefore, the clustering does **not** indicate that an upsampled low-resolution field was used as reference.

To avoid ambiguity, we will revise the manuscript to clarify explicitly in the figure caption: "Comparison of precipitation statistics from high-resolution ERA5-Land reanalysis and model predictions for the 8× 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."

Comment 6

The spatial lag analysis in Fig. 6 is not the most informative way to evaluate scale-dependent structure for precipitation super-resolution. A spatial power spectrum would be more standard and more physically interpretable. The authors should add a spectral analysis to the main paper. The current spatial lag figure could be moved to the supplement.

Response: We thank the reviewer for this suggestion. We agree that spectral analysis provides a more physically interpretable evaluation of scale-dependent structure. We would like to clarify that a radial power spectrum analysis was already included in the original submission (Figure 8a), where power (dB) is evaluated as a function of spatial wavelength. This analysis shows that all models reproduce large-scale energy well, while differences emerge at smaller spatial scales (in a range of few 100-km), with U-NET exhibiting reduced power due to smoothing and the generative models retaining more small-scale variability.

At the same time, we retain the lagged spatial autocorrelation analysis (Figure 6) because it provides a complementary and intuitive measure of spatial dependence and structural decay. In particular, U-NET shows higher lagged autocorrelation, indicating over-smoothing, whereas WGAN and DDPM better match the decay pattern in the target (ERA5-land) test data. To address the reviewer's concern, we have revised the manuscript to more clearly emphasize the spectral analysis in the main discussion and to explicitly state the complementary roles of these two diagnostics.

Comment 7

All three models condition only on coarse-resolution precipitation. Precipitation is not a self-contained variable. For instance, topography is a key control on high-resolution precipitation structure, especially in regions where orographic effects are important. The manuscript does not sufficiently discuss the implications of omitting terrain height or other static geographic information as conditioning variables. The smoothness seen in the deterministic baseline may partly reflect the lack of physically informative conditioning, rather than only the architecture itself. This point is also relevant for the generative models. One of the strengths of conditional DDPM is the flexibility with which conditioning information can be incorporated, including modulation-based conditioning (like FiLM used here and AdaGN etc.). The authors should discuss more directly whether including terrain or other physically meaningful covariates could materially change the conclusions.

Response: Thank you for your suggestion. We agree that after giving additional high-resolution topography (static predictor), UNET could perform better as it has been reported in some previous studies (Rastogi et al., 2023, Liu et al., 2020, Reddy et al., 2023). The main motivation of the study was to fairly compare all three model. In present case, we provide same data to all three

models; even UNET and WGAN shares the same generator architecture. So, the intention was to compare the performance of these 3 widely used AI models for super-resolution downscaling in terms of statistical properties, extremes, computational efficiency etc. We will take this suggestion as a future direction to explore further.

Comment 8

A plain UNet trained with a pointwise loss is well known to produce overly smooth outputs in super resolution tasks, so the trad contrast between UNet and generative models may partly reflect the baseline choice rather than an inherent limitation of deterministic approaches. The paper should justify this baseline more carefully or include at least one stronger deterministic baseline such as a sub-pixel convolution or PixelShuffle-based architecture.

https://arxiv.org/pdf/1609.05158

Response: We thank the reviewer for this important comment. We agree that a standard U-NET trained with a pointwise loss (e.g., MSE) is known to produce overly smooth outputs in super-resolution tasks. In this study, our objective is to provide a controlled comparison among widely used models for precipitation downscaling, and U-NET was selected as a popularly used and well-established baseline in the literature. Its tendency toward smooth predictions is consistent with prior work and serves as a useful reference for evaluating improvements in spatial variability and extremes. At the same time, we acknowledge that stronger deterministic architectures (e.g., sub-pixel convolution or PixelShuffle-based models) or even better loss function (distribution aware) or even additional predictors (HR topography input) could potentially reduce this gap. To address this, we have clarified the rationale for the baseline choice, explicitly discussed its limitations in the revised version of the manuscript. Given multiseed and multiple model training and evaluation framework, a full evaluation of more advanced deterministic architectures is beyond the scope of this study at this stage, but we now highlight this as an important direction for future work.

Comment 9

SSIM is a perceptual image metric designed for natural-image comparison based on luminance and contrast, and its physical meaning for sparse, intermittent precipitation fields is limited. SSIM should not be emphasized as a primary result and may be moved to the supplement. The Q-Q diagnostics currently in Figure S7 are more physically meaningful for a heavy-tailed intermittent variable and should be promoted to the main text.

Response: Thank you for your suggestion. We agree Q-Q diagnostics is more important than SSIM in hydrometeorological context and therefore, we have made that change in our revised version of the manuscript.

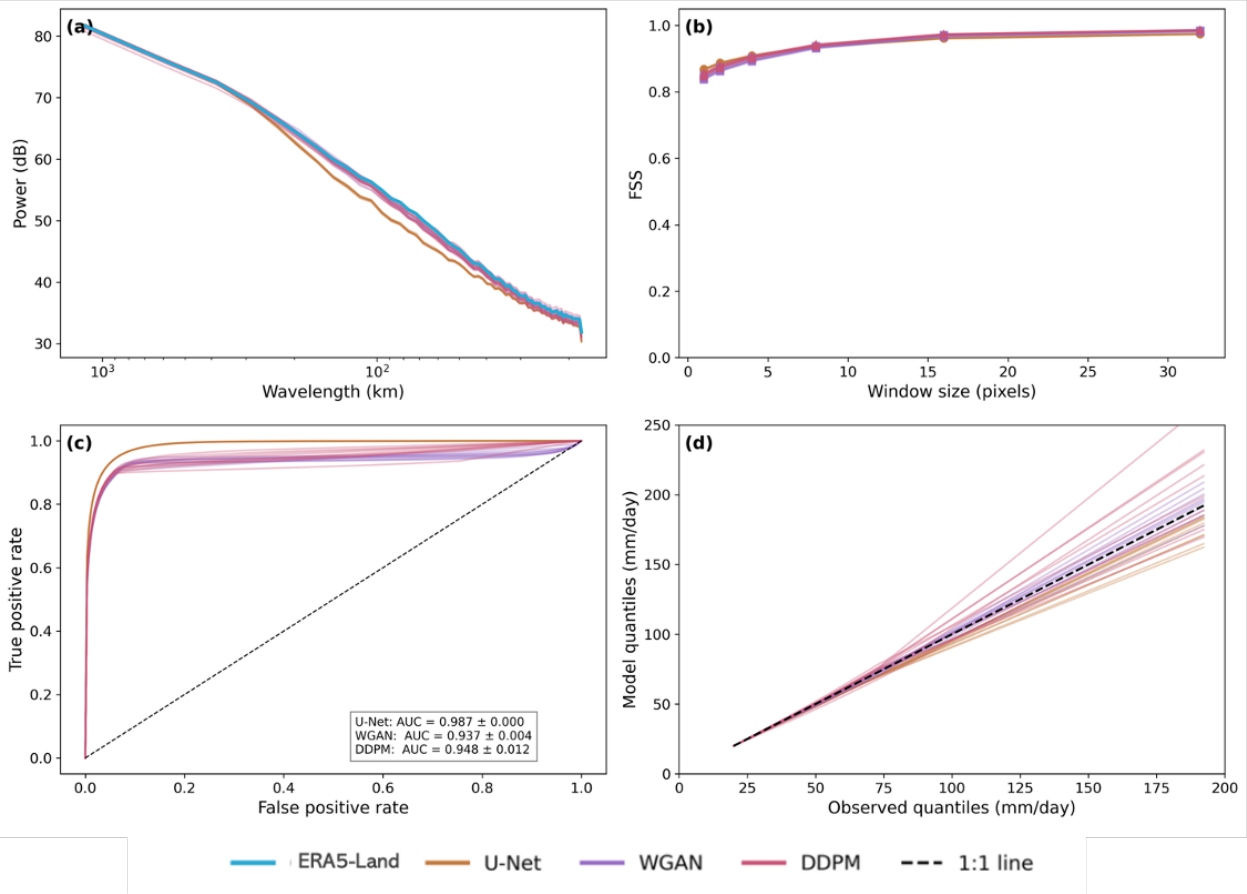

Figure 3. Comparative evaluation of target precipitation fields (ERA5-Land) and predictions from U-Net, WGAN, and DDPM models across multiple diagnostic metrics. (a) Power spectrum as a function of wavelength (km), illustrating the representation of spatial variance across scales. (b) Fractions Skill Score (FSS) as a function of neighborhood window size (pixels). (c) Receiver operating characteristic (ROC) curves for precipitation occurrence classification, with mean area under the curve (AUC ± standard deviation across seeds) reported in the inset. (d) Quantile–quantile comparison of model and observed precipitation intensities for upper-tail events. The dashed 1:1 line denotes perfect agreement. Cyan denotes ERA5-Land target fields, orange denotes U-Net, purple denotes WGAN, and pink denotes DDPM. Model spread across seeds is shown by semi-transparent lines, highlighting differences in the representation of high precipitation extremes.

Comment 10

Because the low-resolution inputs are constructed using a block-averaging operator, the mass conservation result in Section 5.2.2 is partly guaranteed by the experimental design and is less informative than the manuscript implies. This section should be shortened and some of the discussion moved to the supplement.

Response: Thank you for this comment. We agree that the block-averaging operator used to construct low-resolution inputs ensures mass consistency between the coarse inputs and high-resolution targets at the training scale by design. However, we respectfully argue that this does not trivially guarantee mass conservation in the model predictions across all spatial scales.

To clarify: the models are trained to reconstruct 128×128 fields from 16×16 (or 8×8) inputs, and no explicit conservation constraint is imposed during training. Critically, the models receive no information about intermediate spatial aggregations such as 1×1, 2×2, 4×4, or 8×8 grids during training or inference. Yet when the predicted fields are aggregated at these intermediate scales and compared against the corresponding ERA5-Land targets, all three models exhibit near-perfect mass consistency at scales of 8×8 and above ($r = 1.000$, Bias ≈ 0), with scatter decreasing systematically as aggregation scale increases (Figure 4, original manuscript). This emergent behavior at scales the models have no direct knowledge of is a non-trivial result. We acknowledge, however, that the full-domain aggregation result (128×128) is partly a consequence of the experimental design. In the revised manuscript, we will restructure this section to make this distinction explicit.

Comment 11

The manuscript compares a U-Net against generative models, but it does not include a simple interpolation baseline. That omission weakens the benchmark. At minimum, the authors should include one standard interpolation baseline so that readers can assess whether the deterministic neural model actually adds value beyond trivial reconstruction.

Response: We thank the reviewer for this suggestion. We agree that including a simple interpolation baseline provides a useful lower-bound reference for assessing model skill. In the revised manuscript, we have decided to include some results with **bilinear interpolation** as a baseline to add credibility to our UNET predictions in supplementary materials (Figure 4). This allows us to explicitly evaluate whether the U-NET model provides added value beyond trivial reconstruction.

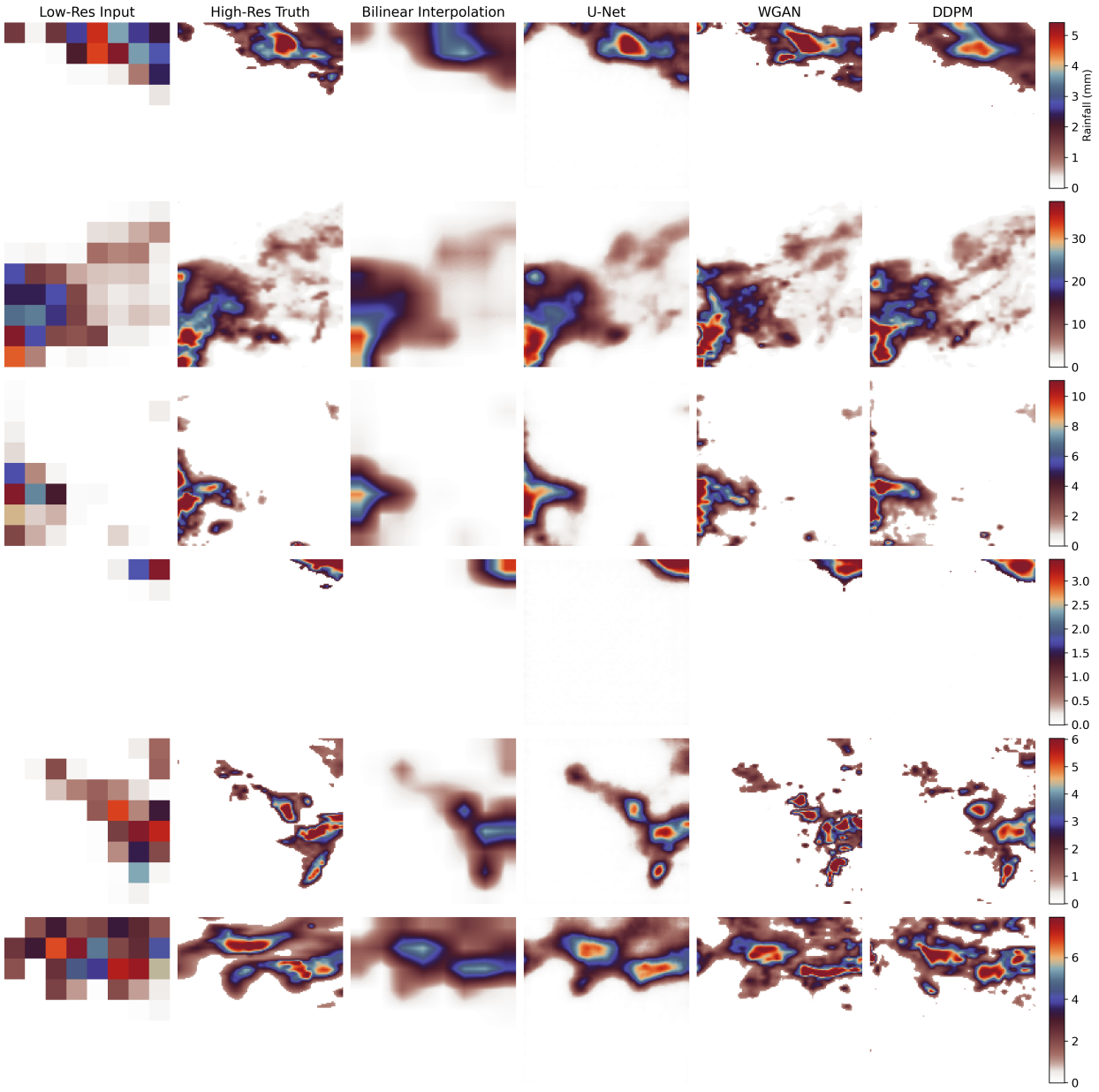

Figure 4. Visual comparison of precipitation downscaling methods for randomly selected test examples used to assess reconstruction skill from a given low-resolution (LR) input. Each row represents an independently selected sample, while columns show the LR input, ERA5-Land high-resolution target field, bilinear interpolation baseline, U-Net prediction, WGAN prediction, and DDPM prediction. Color shading indicates daily rainfall intensity (mm/ day), with a separate color scale for each row to emphasize within-case structural differences.

Comment 12

Precipitation has strong temporal autocorrelation because it is tied to evolving synoptic and mesoscale systems. Wet-spell duration, dry-spell duration, and multi-day persistence are among

the most hydrologically relevant properties of any downscaled product. A model may match daily spatial structure while still failing to reproduce realistic persistence across time. The manuscript does not evaluate this sufficiently. The authors should either include temporal diagnostics that directly assess persistence behavior or clearly state that the current evaluation is not enough to establish hydrological usefulness.

Response: Thank you for this valuable suggestion. We agree that evaluating temporal dependence is important for precipitation downscaling, particularly to assess whether generated fields preserve the persistence structure present in the reference data.

In response, we have added a new analysis and corresponding results (Figure 6) in the revised manuscript) examining the temporal autocorrelation structure of consecutive test samples. Specifically, we computed pixel-wise temporal lag correlations for lags 1–5 using the last 1000 consecutive test samples and compared the ERA5-Land target fields with model-generated outputs from U-Net, WGAN, and DDPM.

The new results show that all three models closely reproduce the observed temporal lag-correlation structure. The distributions of temporal correlations are similar to ERA5-Land, and both the mean bias and RMSE relative to the target are small across lags. This indicates that the models retain the short-range temporal persistence present in the reference dataset rather than generating temporally inconsistent fields.

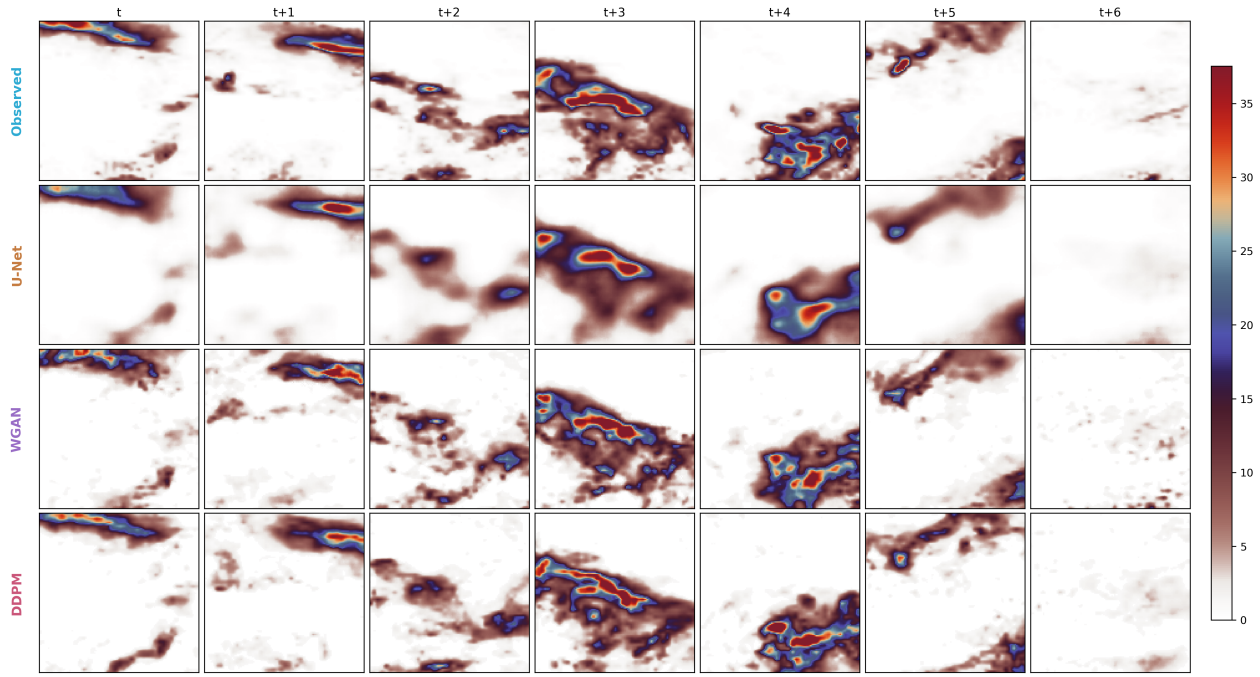

Figure 5. Reconstruction of seven randomly selected consecutive test samples by the U-Net, WGAN, and DDPM downscaling models. Columns correspond to six consecutive low-resolution input cases (t to $t+6$), while rows show the ERA5-Land target field (Observed) and the

corresponding model predictions from U-Net, WGAN, and DDPM. Color shading indicates daily rainfall intensity (mm/ day).

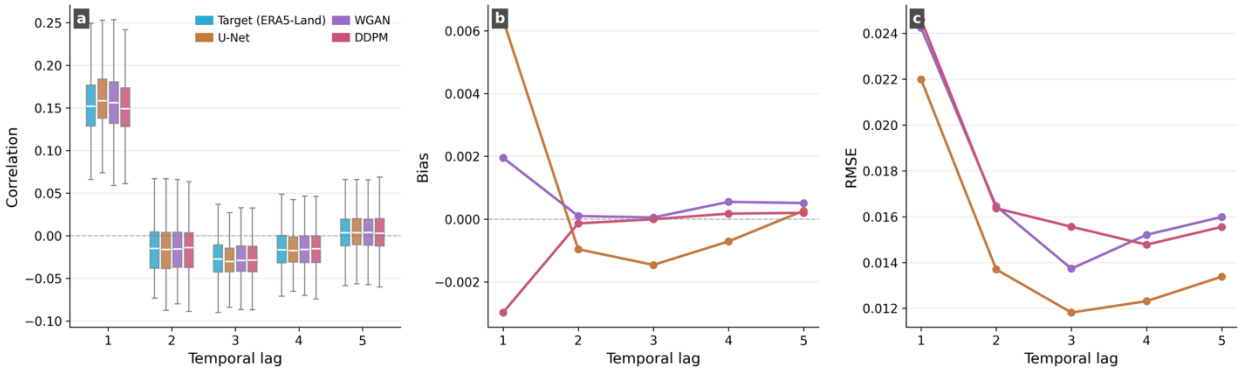

Figure 6: Temporal correlation structure of observed and model-generated precipitation fields over the test set. **(a)** Distribution of pixel-wise temporal correlation coefficients at lags 1–5 for the observed data, and one realization of U-Net, WGAN, and DDPM. **(b)** Mean bias in temporal autocorrelation relative to the observed data as a function of lag. **(c)** Root-mean-square error (RMSE) in temporal autocorrelation relative to the observed data as a function of lag. These results are computed over the last 1000 consecutive test samples.

Minor Comments:

Comment 1

Figures 3 and S3 should state the timestamp or time period shown. The figure captions should clearly indicate which date or sample is being plotted.

Response: Thank you for this comment. The examples in Figures 3 and S3 were selected randomly from the test set and are used only as representative reconstruction examples from low-resolution inputs to high-resolution outputs. Since the models were trained and evaluated on randomly paired low-resolution and high-resolution precipitation fields, without using timestamp information as an input or conditioning variable, the exact timestamp was not used in the modelling framework. We will revise the captions of Figures 3 and S3 to explicitly state that the panels show randomly selected test samples and to indicate the corresponding sample indices where applicable.

Comment 2

At line 126, the manuscript refers to one region as the “Pacific Northwest”. Based on the domain shown, this terminology appears inaccurate, since Utah and western Nevada are not usually considered part of the Pacific Northwest. It would be better to use “Northwest” unless the domain is redefined.

Response: Thank you for your suggestion, we have replaced the region Pacific Northwest with Northwest throughout the manuscript.

Comment 3

The manuscript states that the models are trained on the Central Plains and Northwest, while validation uses the Central Plains plus a subset of the Northeast, and the remaining Northeast samples are used for independent testing. This is understandable, but the exact fractions or sample counts should be stated explicitly in the main text.

Response: 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 12,616 test samples.

Comment 4

The manuscript sets daily precipitation below 1 mm per day to zero and excludes days with fewer than 1 percent wet pixels. These choices may be reasonable, but the authors should report how many samples are removed by region and season and briefly discuss the potential impact on light-rain statistics and wet-day occurrence.

Response: Thank you for this comment. We have now quantified the number of samples removed by the filtering criteria for each region and season and added this information to the revised manuscript.

Overall, the filtering removes:

- 13.76% of samples in Central Plains
- 8.11% in Northwest
- 3.41% in Northeast

A clear seasonal dependence is observed, with the largest removal rates occurring during winter (DJF) and autumn (SON), particularly in the Central Plains (24.82% and 21.10%, respectively), while removal during summer (JJA) is minimal across all regions (≤5.25%).

Added in main text, “Daily precipitation values below 1 mm day^{-1} 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.”

Comment 5

The manuscript states that the models are trained on the Central Plains and Northwest, while validation uses the Central Plains plus a subset of the Northeast, and the remaining Northeast samples are used for independent testing. This is understandable, but the exact fractions or sample counts should be stated explicitly in the main text.

Response: As addressed in our response to Minor Comment 3 above**, "**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 12,616 test samples.

Comment 6

The U-Net training description omits weight decay and the Adam beta values; since the WGAN uses non-default beta1=0.0 and beta2=0.9, any deviation from defaults for other models must also be explicitly reported. The DDPM section does not specify the optimizer, initial learning rate, or weight decay.

Response: Thank you for pointing this out. We have revised the manuscript to explicitly report optimizer hyperparameters for all models. The U-Net uses Adam with default $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and no weight decay. The WGAN already used non-default values ($\beta_1 = 0.0$, $\beta_2 = 0.9$), which are now consistently reported alongside other models. The DDPM section has been updated to specify the use of AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1×10^{-4}.

References

Hunt, K. M. R. (2025). *Stop using root-mean-square error as a precipitation target!* arXiv preprint arXiv:2509.08369.

Liu, Y., Ganguly, A. R., & Dy, J. (2020). Climate downscaling using YNet: A deep convolutional network with skip connections and fusion. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 3145–3153).

Rastogi, D., Niu, H., Passarella, L., Mahajan, S., Kao, S.-C., Vahmani, P., & Jones, A. D. (2025). Complementing dynamical downscaling with super-resolution convolutional neural networks. Geophysical Research Letters, 52, e2024GL111828. https://doi.org/10.1029/2024GL111828

Reddy, P. J., Matear, R., Taylor, J., Thatcher, M., & Grose, M. (2023). A precipitation downscaling method using a super-resolution deconvolution neural network with step orography. Environmental Data Science, 2, e17. https://doi.org/10.1017/eds.2023.18