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]

Reviewer 2

We thank reviewer 2 for carefully reading our manuscript and providing some critical comments to further enhance the quality of our work. Some of the comments overlapped with those of Reviewer 1, and we have addressed all comments raised by both reviewers and will be revising the updated version accordingly.

1. General Comments

The study compares three deep learning frameworks (U-NET, WGAN, and DDPM) for precipitation super-resolution across different hydroclimatic regimes. While the comparison is timely, there are significant critical concerns regarding the experimental design and the technical execution that need to be addressed before the paper can be considered for publication.

2. Major Concerns

Training Stability and Overfitting (Critical Concern)

A fundamental issue exists regarding the training convergence and generalization of the generative models, particularly the DDPM. In Section 3 (Lines 398–400), the authors describe the dataset splitting into training and testing sets, but no validation set is mentioned. However, validation curves are provided in the Supplemental Material (Figure S2), at least for the UNET and the DDPM models. They are missing for the WGAN model. Upon inspection of Figure S2, despite the compressed scale of the y-axis, the DDPM model appears to exhibit clear signs of overfitting after approximately epoch 10. The divergence between training and validation loss suggests that the model is no longer learning generalizable features of the precipitation fields but is instead memorizing the training samples. Since the remainder of the paper relies on the results derived from these trained weights, the validity of the inter-comparison and the subsequent conclusions regarding DDPM's performance are in question. Also, the training curves of the WGAN raise some concerns, and no validation curve is shown. The authors must:

- Clarify the lack of a validation set description in the main text and for WGAN in Figure S2
- Provide a detailed analysis of the loss curves with a more appropriate y-axis scale. Is the overfitting actually happening?

- Address how they ensured the "optimal" stopping point for training to prevent reporting results from an overfit model.
- In lines 430-434, the authors acknowledge the potential for overfitting but treat it in a too simplistic manner.

Response: Thank you for this comment. We would like to clarify that a validation dataset was indeed used during the original training procedure, although its role was not described clearly enough in the main manuscript. This is also reflected in the originally submitted Supplementary Figure S2, which already included training and validation loss curves for both U-Net and DDPM models for the 8× and 16× experiments. We have updated the caption for clarity. To avoid ambiguity, we have now revised the main text to explicitly describe the dataset split into training, validation, and independent test sets, and clarified that the validation set was used for model monitoring, model selection, and early stopping where applicable.

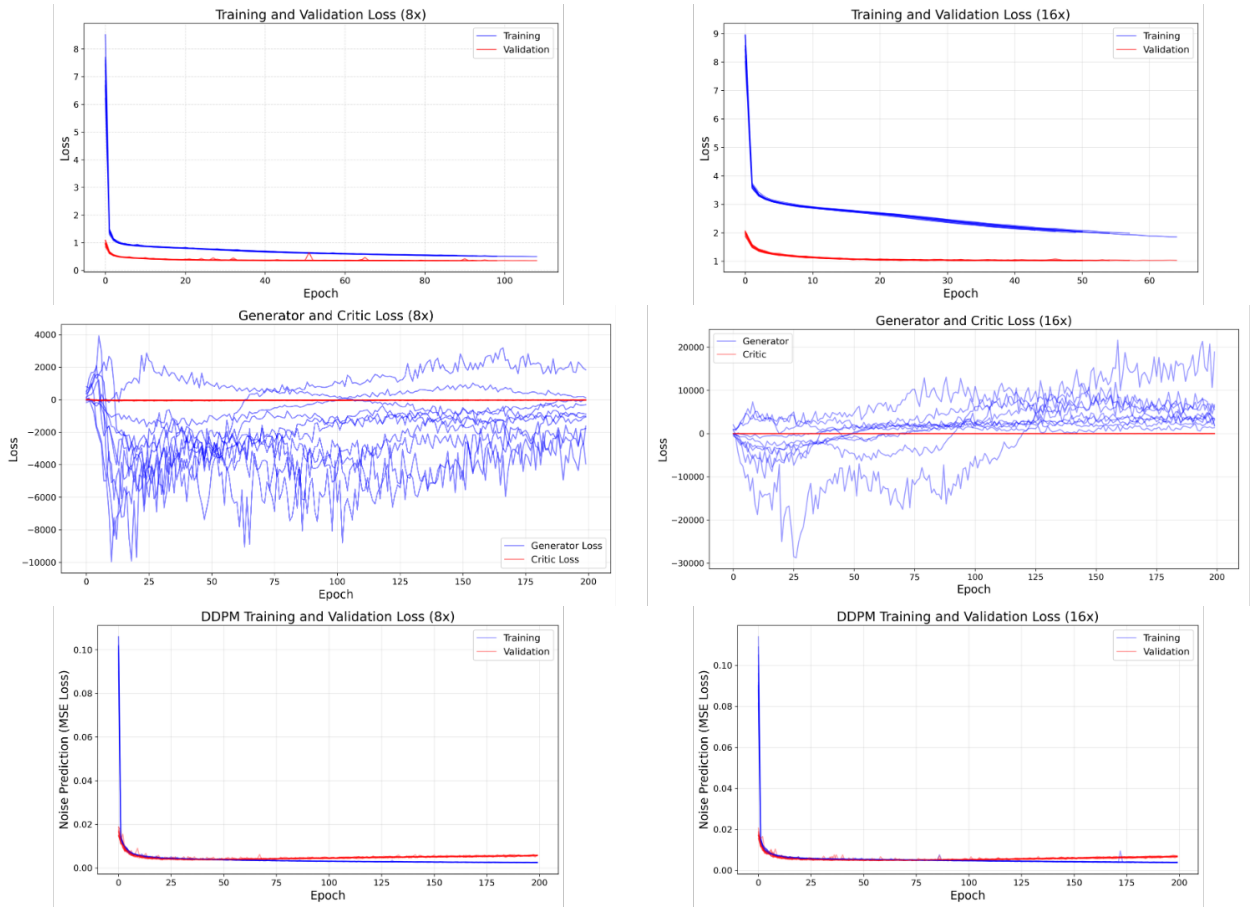

Figure 1. (Figure S2 from our original submission) Training and validation curves for the U-Net, WGAN, and DDPM models under 8× and 16× downscaling configurations. (top) Training and validation loss for U-Net; (middle) Generator and critic losses for WGAN; (bottom) Noise-prediction MSE loss for DDPM. Each subplot represents one of ten independent (random) initialization seeds, shown separately for 8× (left column) and 16× (right column). Blue line represents training loss behavior (generator loss in case of WGAN) whereas red line represents validation loss behavior with epochs (critic loss in case of WGAN).

Regarding the concern about overfitting in DDPM, we respectfully note that while the training loss continues to decrease after the initial epochs, the validation loss remains comparatively stable with only minor fluctuations rather than showing sustained divergence. This behavior is more consistent with convergence to a plateau than severe memorization. We have revised the supplementary figure presentation (clearer axis scaling and discussion) to make this interpretation more transparent.

For WGAN, we agree that a standard validation-loss curve was not shown previously. Because adversarial training does not optimize a direct reconstruction objective, validation assessment was instead based on multiple diagnostics, including generator loss, critic loss, gradient penalty behaviour, validation SSIM, Pearson correlation, exceedance probability, and visual inspection of generated samples across epochs. We now clarify this in both the manuscript and supplementary material.

Finally, we have expanded the discussion of overfitting and model generalization in the revised manuscript, particularly for the higher-capacity DDPM model.

Directly related to the previous point: Sample Size vs. Model Complexity (Critical)

The technical rigor of the study is challenged by the potential imbalance between the available data and the model's capacity. The authors utilize ~33,000 daily fields (with only ~19,200 samples for training) to train high-capacity architectures. Modern WGAN and DDPM implementations (including features mentioned like sinusoidal time embeddings and FiLM layers) often contain millions of trainable parameters.

- The authors must provide a table explicitly summarizing the total number of trainable parameters for the U-NET, WGAN, and DDPM.
- Given the relatively small training set size and the high complexity of the models, the risk of overfitting is severe. The authors must discuss the generalization potential of these models in this context.

Conclusion on the first two points:

This evidence of potential overfitting is a major flaw that propagates through all results and discussions in the manuscript. Before any further analysis can be considered, the authors must demonstrate that the models (especially the DDPM) are not over-parameterized for the provided dataset and that the reported performance metrics are not the result of a model that has failed to generalize.

Response: Thank you for raising this important concern regarding model complexity relative to training sample size. We agree that reporting model capacity explicitly is necessary for a rigorous comparison, and we have revised the manuscript accordingly.

First, we have added a new table summarizing the total number of trainable parameters for all models used in the study:

Model	Trainable Parameters
U-Net	497,121
WGAN Generator	497,121
WGAN Critic	552,417
DDPM Denoising U-Net	14,567,649

The relatively larger parameter count of the DDPM arises from the conditional denoising U-Net backbone, which includes timestep-conditioning components (learnable time-embedding MLP and FiLM layers) in addition to convolutional encoder–decoder layers. The sinusoidal timestep encoding itself and the diffusion schedule are fixed and non-learnable.

Second, while the DDPM has substantially higher capacity than the U-Net and WGAN generator, model complexity alone does not imply overfitting. Generalization was assessed using an independent held-out test set that was not used during training or validation. In addition, validation data were used during training to monitor convergence and select model checkpoints. Across all three frameworks, performance was evaluated not only using pointwise metrics, but also using independent structural diagnostics including exceedance probability, radial power spectra, spatial autocorrelation, temporal autocorrelation, and reconstruction examples. The consistency of these diagnostics on unseen test samples suggests that the models learned transferable spatial precipitation features rather than memorizing training data.

Third, the downscaling task considered here maps relatively low-dimensional coarse-resolution inputs (8×8 or 16×16) to 128×128 outputs under strong conditioning by the input precipitation field. This supervised conditional setting is less unconstrained than unconditional image generation tasks, which helps regularize learning and reduces the effective risk of memorization.

Additional concerns

Perfect-Model Framework and Practical Usability (Lines 154–157): The authors describe their approach as a "perfect-model super-resolution framework" where both inputs and targets originate from the same dataset (ERA5-Land). While this allows for controlled evaluation, it ignores the critical real-world challenge of "predictor mismatch" or bias between different datasets (e.g., GCM output vs. regional simulations). This design choice significantly hinders the practical usability of the proposed models, as they are not tested against actual biased low-resolution information found in climate model outputs. This limitation must be explicitly stated in the **Abstract**, and discussed in the **Introduction**, and **Conclusion**. The authors should clarify how these models would perform when the "trace" of mesoscale information is truly absent or biased in the low-resolution input.

Response: Thank you for this thoughtful comment. We agree that the present study adopts a controlled perfect-model super-resolution framework**,** in which both the coarse-resolution inputs

and high-resolution targets are derived from the same ERA5-Land dataset. This setup does not fully represent operational climate downscaling applications, where coarse predictors often originate from external sources such as GCMs or regional climate models and may contain systematic biases, structural errors, or missing mesoscale information.

Our motivation for using this framework was to enable a fair and controlled comparison of the three deep-learning approaches (U-Net, WGAN, and DDPM) under identical input-output conditions. By minimizing predictor mismatch as a confounding factor, the study isolates differences attributable to the learning frameworks themselves and allows clearer assessment of their ability to reconstruct fine-scale precipitation structure, spatial dependence, extremes, and uncertainty characteristics. We agree that strong performance in a perfect-model setting does not automatically guarantee equivalent skill when applied to biased climate-model predictors. In practical settings, additional challenges such as bias correction, domain adaptation, predictor selection, and limited mesoscale information must be addressed.

In response to the reviewer's suggestion, we have revised the Abstract, Introduction, and Conclusion to explicitly acknowledge this limitation and clarify that the present study should be interpreted as a controlled benchmark rather than a direct demonstration of readiness for operational GCM downscaling. We also note that extending these methods to true climate-model predictors is an important direction for future work.

Data Normalization: The manuscript mentions a log(1+x) transformation and min-max normalization specifically for the DDPM (Line 258). It is unclear if any normalization or transformation was applied to the precipitation data for the U-NET and WGAN models. Given the heavy-tailed nature of precipitation, this is a critical detail. If no normalization was used for the other models, the authors must justify this choice or clarify the preprocessing steps for all architectures.

Response: 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 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 alter the key findings, although it can influence some aspects of tail behavior and fine-scale variance slightly. 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.

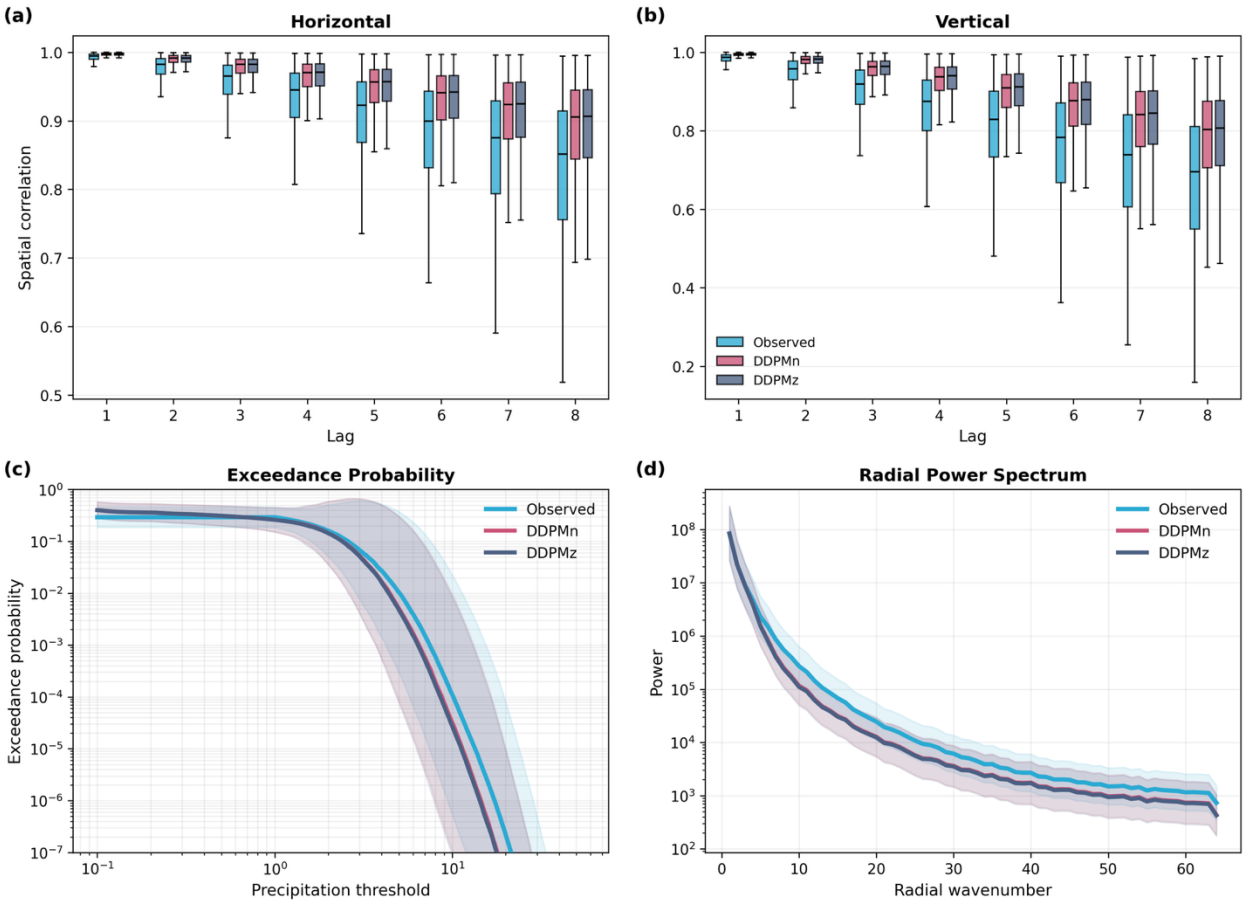

Figure 2: Comparison of spatial dependence and distributional characteristics between ERA5-Land target fields and DDPM predictions trained with min–max normalization (DDPMn) and z-score normalization (DDPMz). (a) Horizontal lagged spatial correlation and (b) vertical lagged spatial correlation shown 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.

Stochastic vs. Epistemic Uncertainty: The authors use 10 independently trained models to form an ensemble (Line 270), which they state reflects "epistemic uncertainty." However, the primary strength of generative models like WGAN and DDPM is their ability to produce multiple stochastic realizations from a single trained model. The authors should evaluate the generative potential/stochasticity of a single trained model rather than relying solely on an ensemble of differently trained models, as the latter is computationally expensive and less practical for end-users.

Response: Thank you for your suggestion. We have 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.

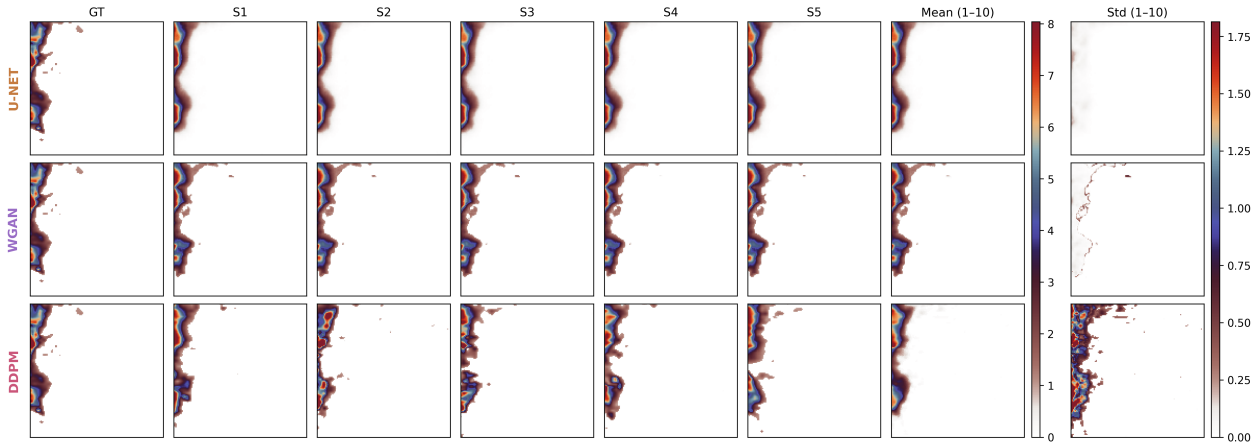

Figure 3a: *Representative stochastic realizations and ensemble statistics for a* randomly *selected event.* Each row shows results from U-Net**,** WGAN**,** and DDPM**,** respectively. Columns show the ground truth (GT)**,** five stochastic samples **(**S1–S5**),** the ensemble mean across 10 realizations, and the corresponding ensemble standard deviation. The figure highlights differences in spatial realism, stochastic variability, and uncertainty structure across the three generative approaches.

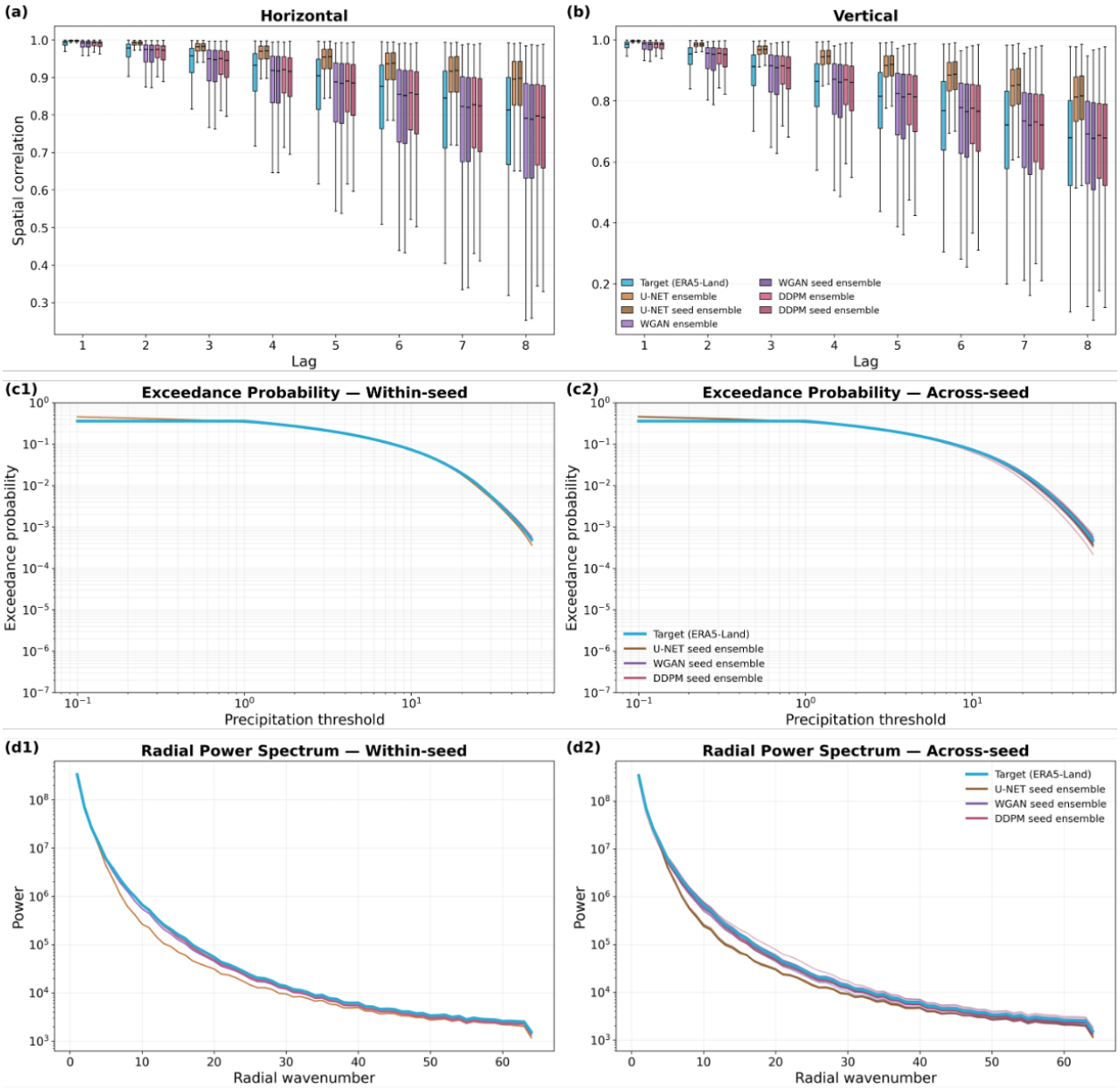

Figure 3b: Comparison of within-seed stochastic variability and across-seed training variability in spatial structure and precipitation statistics for generative downscaling models. Panels (a) and (b) show horizontal and vertical lagged spatial autocorrelation, respectively, summarized as boxplots across 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.

Static Variables: Were any static variables (e.g., elevation, land cover, distance to coast) included as auxiliary inputs to the models? Orographic forcing is mentioned as a relevant driver (especially for one of the domains/regimes line 127), but looks like the models were trained without providing them with this information. Having access to the underlying topography can assist in the downscaling process: why did the authors choose not to include any static covariants in the training?

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.

3. Specific Comments and Technical Corrections

Hydrologic Relevance: The authors mention "hydrologic modeling" as a motivation many times in the text. Could the authors clarify the specific "hydrologic" metrics or assessments used beyond standard meteorological diagnostics?

Response: Thank you for this helpful comment. We agree that the phrase “hydrologic relevance” should be clarified more precisely in the manuscript. In the present study, we did not directly couple the downscaled precipitation outputs with a hydrologic model such as a rainfall–runoff or streamflow simulation model. Instead, we use the term to refer to precipitation characteristics that are highly relevant for hydrologic applications and water-resources assessments.

Beyond standard meteorological diagnostics, our evaluation includes several precipitation metrics with clear hydrologic significance. These include the probability of zero precipitation (wet/dry occurrence), mean and variance, which are relevant for water balance and rainfall variability; and skewness and kurtosis, which characterize asymmetry and heavy-tailed behavior associated with extreme rainfall events.

We also assess exceedance probability curves, which are directly relevant for heavy-rainfall frequency and flood-generating events. In addition, spatial correlation diagnostics are important for understanding storm organization and basin-scale runoff coherence, while temporal autocorrelation helps characterize multi-day persistence, storm sequencing, and antecedent wetness conditions that strongly influence hydrologic response.

Target Data Selection (Lines 115–122): The authors use ERA5-Land as the target data. Please provide a brief justification for using a reanalysis product as "Ground Truth" (GT) rather than observational-based data (e.g., high-resolution radar products).

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.

DDPM Hyperparameters (Section 2.2.3): training details for the DDPM are missing. Specifically, please provide the optimizer used (e.g., Adam, AdamW) and the specific learning rate or learning rate schedule employed during training.

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}.

Figure 3 Clarification: What do the columns in Figure 3 represent? Please label them clearly in the figure or the caption. The low-resolution input in the last column does not appear to match the

high-resolution Ground Truth (GT). Please verify if the block-averaging or interpolation used to create the low-resolution inputs is consistent across the visualization.

Response: Thank you for the comment. The columns represent independent precipitation events (randomly selected test samples), while the rows correspond to the coarse-resolution input, high-resolution target (ERA5-Land), and the corresponding reconstructions from U-Net, WGAN, and DDPM. We have revised the figure caption to explicitly state that each column corresponds to one randomly selected test case, while rows indicate the different input, target, and model outputs.

Regarding the reviewer's concern about the apparent mismatch between the low-resolution input in the final column and the corresponding high-resolution target, we appreciate this comment and carefully rechecked the plotting pipeline. The low-resolution inputs were consistently generated from the corresponding high-resolution target fields using the same preprocessing procedure (block averaging) which we used for training models. However, because coarse-resolution aggregation strongly smooths and compresses sparse or localized precipitation structures, some low-resolution panels can appear visually weak or spatially simplified relative to the associated high-resolution target, particularly for intermittent or fragmented rainfall events.

We also note that the low-resolution panels were displayed on their native coarse grid (8x8), whereas the high-resolution targets and model outputs are shown on the fine grid (128x128). This difference in spatial resolution can further exaggerate the perceived mismatch in side-by-side visual comparison.

To avoid confusion, we have reverified sample alignment across all columns, confirmed that each low-resolution input corresponds to the same precipitation event as its target field, and revised the caption to clearly describe the preprocessing and visualization procedure. Where necessary, we have also updated the figure to improve clarity and consistency.

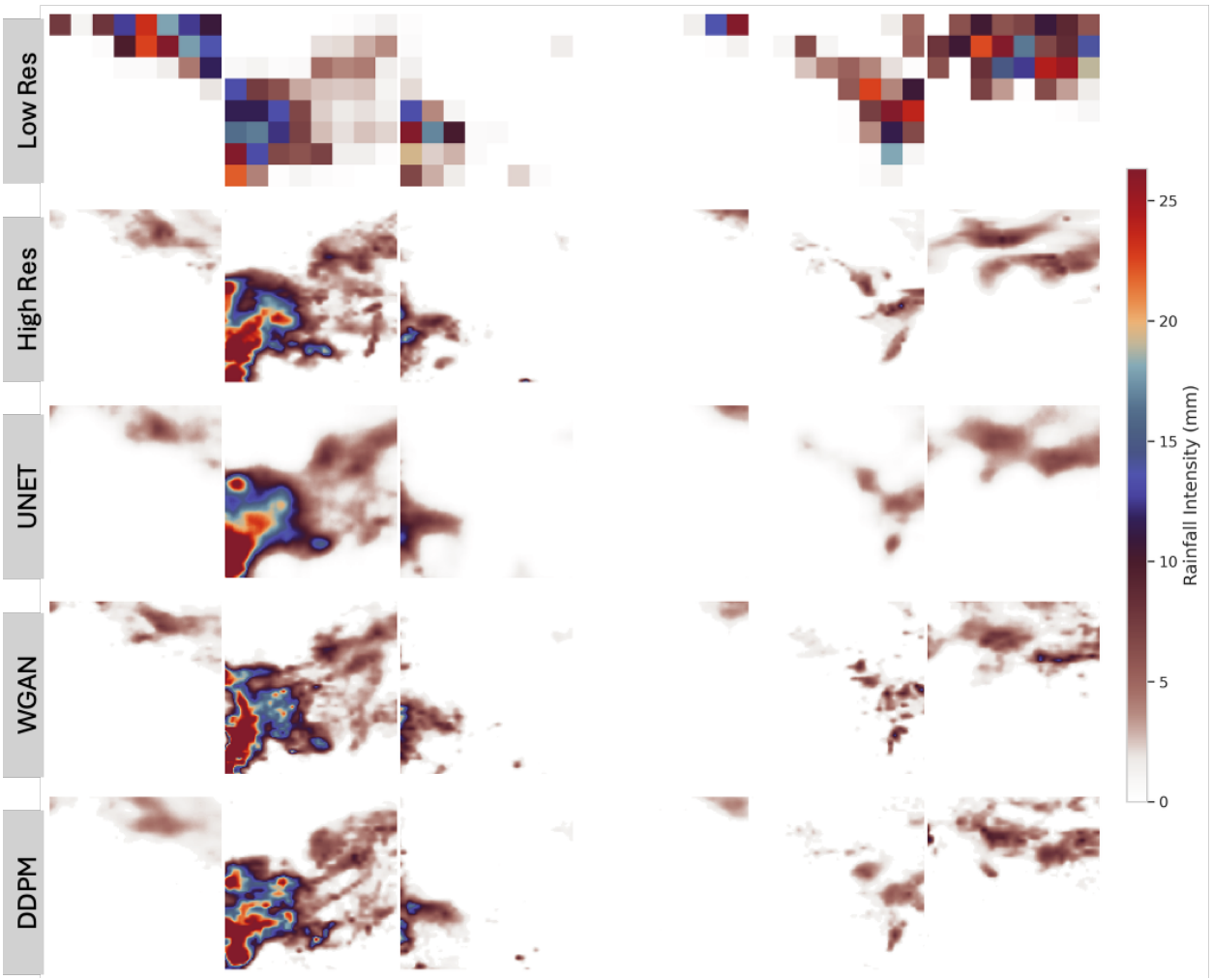

Figure 4. Reconstruction of high-resolution precipitation fields for the 16× downscaling task (8×8 → 128×128). Rows show the low-resolution input, high-resolution ground truth, and predictions from U-NET, WGAN, and DDPM respectively. The columns represent randomly picked LR samples (first row) from test dataset to visualize the reconstruction by the trained models (row 3 to 5) and compare with respective target (ERA5-land) HR samples (second row)

References

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

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