https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

CEDDAR v1.0.2: Bridging physics and generative modelling for regional precipitation with controllable diffusion-based downscaling

Thea Quistgaard[1, 2], Tanja Denager[2], Raphael J. M. Schneider[2], Jesper R. Christiansen[3], Simon Stisen[2], and Peter L. Langen[1]

[1]Department of Environmental Science, iClimate, Aarhus University, Roskilde, Denmark
[2]Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark
[3]Forest and Landscape Ecology, Department of Geoscience and Nature Management, Copenhagen University, Denmark

Correspondence: Thea Quistgaard (tquistgaard@envs.au.dk)

Abstract. Understanding local impacts of extreme weather, from e.g. precipitation or temperature, requires high-resolution downscaling of those atmospheric variables. Deterministic and classical statistical methods fail to adequately represent variability and extremes, motivating the use of probabilistic generative methods. Recent advances in generative deep learning show promise for spatial realism and ensemble generation, but their physical fidelity and limitations remain insufficiently understood.

We present CEDDAR v1.0.2 (Controllable Ensemble Diffusion Downscaling for Atmospheric Rainfall), a diffusion-based generative model for daily precipitation downscaling, conditioned on large-scale ($\sim$30 $\times$ 30 km) ERA5 precipitation fields, to produce kilometre-scale (2.5 $\times$ 2.5 km) precipitation over Denmark, using the DANish ReAnalysis product (DANRA) as reference. The modelling framework is based on an Elucidated Diffusion Model (EDM) backbone and incorporates soft physics-guidance through seasonal (Day-of-Year) and geographic conditioning with a specific land-focused design. We further introduce regularisation terms through a Signed-Distance Function weighted loss, and an auxiliary RainGate head for shaping wet-dry occurrence statistics and improving ensemble calibration.

By design, the model is probabilistic and capable of producing ensembles. We further introduce an inference-time control parameter, σ^*, that allows for interpretable adjustment of stochastic scale sensitivity without retraining. For model-behaviour assessment, we introduce a multi-perspective evaluation protocol with emphasis on trade-offs across probabilistic, spatial, climatological, and scale-dependent diagnostics rather than single-metric optimisation.

Our results show strong spatial and spectral realism with generally well-calibrated ensembles, but also demonstrate that visually realistic fine-scale structure does not guarantee climatological fidelity. We also find that explicitly introducing σ^* reveals systematic, though non-linear, trade-offs between large-scale coherence, fine-scale variability, and ensemble spread. This highlights the potential and the limitations of controllable diffusion-based downscaling.

In conclusion, this work presents a diffusion-based downscaling framework and evaluation suite that works as a diagnostic testbed for understanding how generative models behave in physically constrained settings, and offers guidance on their appropriate use and common failure modes relevant to future operational solutions and climate-impact applications.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

1 Introduction

The impacts of climate change are often felt on a very local scale, particularly within ecosystems, hydrology, and other fast-reacting systems (IPCC, 2022; Seneviratne et al., 2012; Donat et al., 2013). Local climate impacts often depend on climate *extremes* and not only on the *mean* (Seneviratne et al., 2012; Donat et al., 2013). Currently, Global Climate Models (GCMs) and Regional Climate Models (RCMs) are generally good at capturing the means of the climate system, but less good at capturing the extremes of distributions due to their relatively coarse spatial grid spacing (Maraun et al., 2010; Wilby et al., 2004; Kendon et al., 2014) - especially when considering non-normally distributed variables like precipitation. An example is the case for local hydrology, where extreme precipitation events can have very localised and sudden effects on hydrology. Well-resolved and bias-corrected precipitation estimates are necessary in e.g. flood hazard assessment and extreme event management (Li et al., 2019; Meresa et al., 2022), and are also highly relevant for applications such as water-table depth variability in peatlands for greenhouse gas emission estimates (Denager et al., 2026).

Generally, the issue of low-resolution global models is solved through dynamical or statistical downscaling (Maraun et al., 2010; Wilby et al., 2004; Baño-Medina et al., 2020) which, in recent years, has adopted a myriad of methods and models from computer science (Baño-Medina et al., 2020; Vandal et al., 2017), and especially from Machine Learning (ML). Deterministic and simpler statistical ML downscaling methods are generally much faster than running high-resolution GCMs but struggle with extreme value representation and physical consistency (Karpatne et al., 2017; Beucler et al., 2021; Rasp and Thuerey, 2021). Many recent approaches to statistical downscaling are inspired by methods from computer vision and deep learning, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and more recently diffusion-based models (Accarino et al., 2021; Harris et al., 2022; Watt and Mansfield, 2024). These methods generally show promising results, but their performance and limitations vary widely depending on problem formulation and conditioning strategy (Rampal et al., 2024; Saha and Ravela, 2024). Often, a key challenge to adopting and applying generative models to physical systems is the lack of explicit physical constraints, which can lead to physically implausible outcomes (Karpatne et al., 2017; Beucler et al., 2021).

In this study, we develop a generative diffusion-based model that can capture stochastic variability and extremes while remaining anchored to large-scale atmospheric physics. Rather than treating the generative model as a static black box, we explicitly expose and interrogate its stochastic behaviour, providing a transparent handle for experimenting with scale-sensitivity and uncertainty at inference (sample generation after training) time.

Beyond proposing a specific downscaling model, we aim to critically assess the strengths, limitations, and trade-offs of diffusion-based generative models for climate downscaling. Instead of optimizing a single metric, we emphasise a systematic evaluation across four complementary evaluation pillars: probabilistic, spatial, climatological, and scale-dependent diagnostics. We aim to understand where generative realism aligns with - and diverges from - physical and climatological fidelity.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

1.1 Our contributions: a controllable diffusion-based downscaler

We present a diffusion-based generative downscaling framework, CEDDAR v1.0.2 (Controllable Ensemble Diffusion Downscaling for Atmospheric Rainfall), for daily precipitation that is explicitly designed to balance physical consistency, ensemble realism, and interpretability (Figure 1). Using an Elucidated Diffusion Model (EDM, Karras et al. (2022)), we downscale ERA5 (30 km scale) precipitation fields (Hersbach et al. (2020)) to kilometre-scale DANRA resolution (DANish regional atmospheric ReAnalysis (DANRA) Yang et al. (2025b)). From this model, we generate physically plausible high-resolution ensembles conditioned on large-scale atmospheric forcing.

Our main contribution with this study is a comprehensive multi-perspective evaluation of diffusion-based generative downscaling, demonstrating that strong spatial and spectral realism does not automatically translate into accurate climatology or extreme-value behaviour. Through a structured component study, we show that architectural and conditioning choices interact nonlinearly, and that commonly used strategies such as residual prediction or exponential moving averages may degrade performance once conditioning is sufficient. We further introduce the model's stochastic scale sensitivity via the inference-time parameter σ^* enabling controlled interrogation of the trade-off between large-scale coherence, fine-scale variability, and ensemble spread. All analyses build on a physics-guided EDM architecture that serves as a testbed for these analyses rather than as a single optimised solution.

The following sections detail the model architecture (Section 2), experimental setup and evaluation metrics (Section 3), and results demonstrating ensemble realism and calibration, multi-scale fidelity, and scale-controllability (Section 4).

2 Methods and architectural design

2.1 EDM: Model overview and problem formulation

In recent years, diffusion-based generative models have emerged as a powerful tool for problems with high-dimensional data and for enhancing image resolution through super-resolution (Ho et al., 2020; Song et al., 2021; Karras et al., 2022). In the context of downscaling atmospheric variable fields, generative diffusion-based models offer a key advantage compared to purely deterministic models: they model stochasticity explicitly and can thus represent multiple physically plausible fine-scale realisations consistent with a low-resolution large-scale forcing. This makes diffusion-based generative models appealing, especially for impact-analysis and applications, where variability and extremes are often more relevant than a single deterministic estimate.

The concrete problem is to map one or more low-resolution (LR) atmospheric fields to corresponding high-resolution (HR) precipitation, x_{HR}, at a daily resolution, with a target resolution of DANRA (2.5 $\times$ 2.5 km) (Yang et al., 2025b). See more on the data in Section 3.1 and in Supplementary Information Section S2. This problem is analogous to mapping from a low-resolution Global Climate Model (GCM) to higher-resolution Regional Climate Model (RCM), and the methods presented here are thus not limited to use with reanalysis data.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

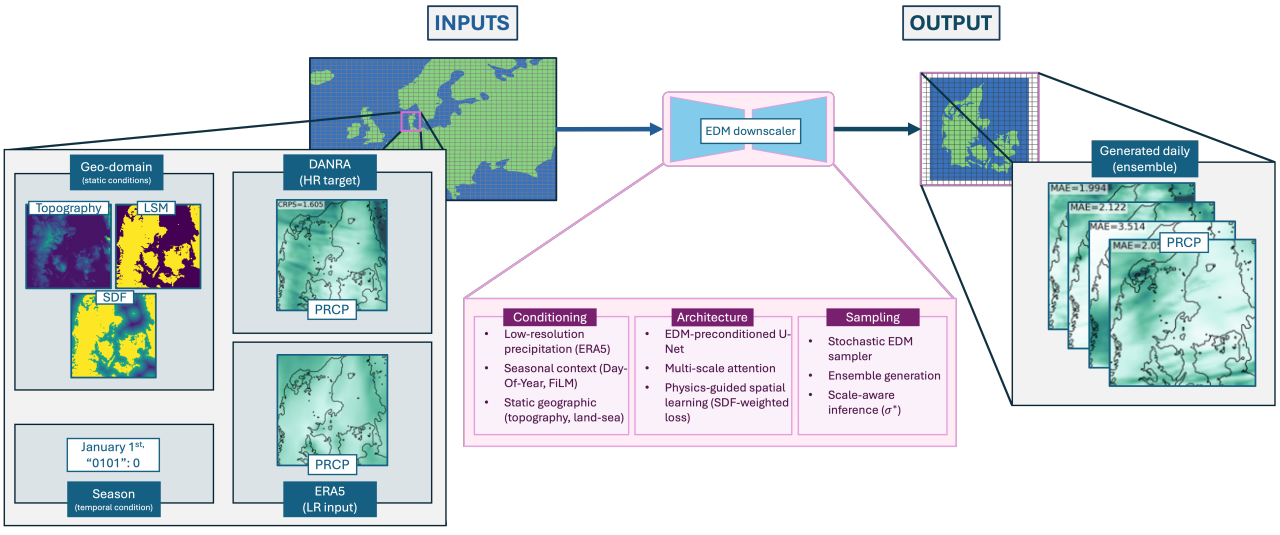

Figure 1. *Schematic overview of the Elucidated Diffusion Model (EDM) based downscaling framework*. The model is conditioned on low-resolution (LR) ERA5 precipitation (PRCP), seasonal context, and static geographical information (topography and land-sea mask (LSM)). It is trained to generated high-resolution (HR) DANRA-like precipitation (PRCP) using an EDM-preconditioned U-Net with multi-scale attention and physics-guided spatial learning through a land-sea aware (SDF-weighted) loss function. At inference, stochastic EDM sampling produces an ensemble of daily precipitation fields, with scale-aware control of spatial variability via the parameter σ^*.

Formally, we learn a conditional diffusion model that maps noisy high-resolution precipitation fields to denoised estimates, conditioned on low-resolution atmospheric predictors and auxiliary information. Full mathematical details are provided in Supplement S1.

Rather than employing conventional score-based diffusion models ("Vanilla" DMs) or adversarial methods (GANs), we implement the Elucidated Diffusion Model (EDM) scheme. Compared to a standard Diffusion Model, the EDM framework introduces an improved noise schedule through a noise preconditioning, along with a log-normal sampling distribution over noise levels. These modifications improve gradient conditioning and training stability, while also accelerating convergence, and making the method less sensitive to hyperparameter tuning (Dhariwal and Nichol, 2021). To support reproducibility and accessibility for readers less familiar with EDMs, Figure 2 provides a schematic overview of the generic EDM training and sampling procedure used throughout this study.

We model the conditional distribution of high-resolution daily precipitation fields from the reanalysis DANRA, x_{HR}, given a number of conditions, c (low-resolution atmospheric reanalysis predictors c_{ERA5} from ERA5; geographical conditions c_{geo}), where we denote the corresponding LR precipitation field as x_{LR}. Both x_{HR} and x_{LR} are normalised using their own model-specific global statistics (computed on the training split), i.e. DANRA fields are scaled using DANRA statistics, and ERA5 fields are scaled using ERA5 statistics. This avoids implicit distribution misalignment and empirically yields the most stable and physically consistent results, see Supplement Section S4 for more details. The network is trained to denoise noisy HR

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

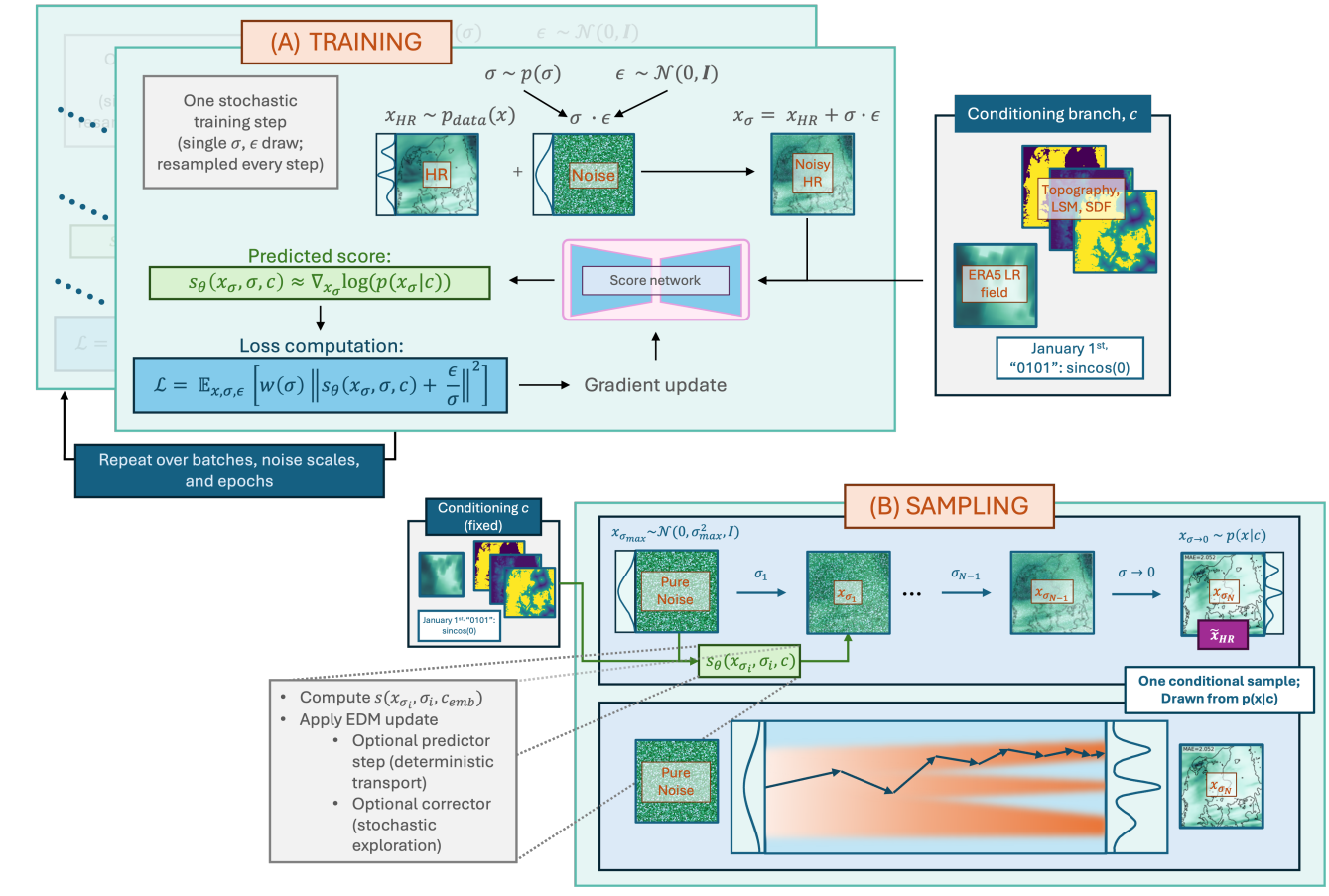

Figure 2. *Illustration of the Elucidated Diffusion Model (EDM) downscaler, training and sampling*. This figure provides a generic EDM overview to support readers unfamiliar with the framework. **(A) Training**: A clean high-resolution (HR) field $x_{\mathrm{HR}} \sim p_{\mathrm{data}}(x)$ is perturbed by additive Gaussian noise at a continuously sampled noise level σ, yielding $x_\sigma = x_{\mathrm{HR}} + \sigma\epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$ and σ drawn from a log-normal distribution (as proposed in EDM). The score network is trained using EDM preconditioning: the network receives a noised HR sample, the corresponding noise-scale, and any given conditions, here topography, land-sea mask (LSM) (x_σ, σ, c). It then predicts a denoised estimate (or equivalently a preconditioned score) with σ-dependent scaling, improving gradient conditioning across noise scales. The objective is a σ-weighted denoising loss, $\mathcal{L}$, where $w(\sigma)$ emphasises the appropriate noise regimes. In our work, we also allow for a reweighting of the loss through a signed-distance function (SDF). **(B) Sampling**: Starting from $x_{\sigma_{\max}} \sim \mathcal{N}(0, \sigma_{\max}^2 I)$, the model iteratively moves samples through a decreasing noise schedule $\sigma_1 > \cdots > \sigma_N \to 0$. Each step applies the EDM update using the trained network (optionally with *churn*, temporary noise inflation for stochastic exploration) followed by a deterministic update to the next noise level. Conditioning information c (e.g., low-resolution ERA5 input, geographic fields, and seasonal embeddings) is held fixed during sampling and guides the reverse process to yield conditional realisations $x_{\mathrm{gen}} \sim p(x \mid c)$.

samples, $x_t = x_{\mathrm{HR}} + \sigma_t\epsilon$, conditioned on c. The learned conditional denoiser then produces ensemble samples by varying the initial noise ϵ.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

2.2 Conditional EDM model and U-Net backbone

The entire model we present here is implemented in PyTorch, with the score network, f_θ, based on an EDM-preconditioned U-Net backbone with hierarchical skip connections, following the EDM parameterisation of Karras et al. (2022).

Noise levels σ_t are embedded using a learned projection of $\log \sigma_t$, following EDM conventions, and injected throughout the network. Seasonal information (Day-of-Year) is injected through a Feature-wise Linear Modulation-style modulation (Perez et al., 2018) in intermediate blocks at all U-Net resolutions. This design allows the model to modulate precipitation statistics coherently across spatial scales while keeping stochastic noise conditioning and physical seasonality disentangled.

Field-like conditioning variables, like topography, low-resolution precipitation, and land-sea mask, are injected to the network via channel concatenation with the noisy HR input, at the native input resolution and passed through the encoder. This minimal conditioning strategy avoids hard structural priors and yielded the most stable results in the controlled component experiments.

2.3 Physics-informed design choices

The model is fully data-driven, but several design choices have been made to embed physical and atmospheric dynamical knowledge and geographical realism into both the learning process and the model's spatial focus. These are soft physics-informed constraints that help guide the generative process towards physically plausible precipitation patterns over land - without imposing any hard conservation rules. The designs and features presented below are described in detail with mathematical formulations in Supplement S1.6.

Local precipitation characteristics relevant for impact modelling are predominantly shaped by land-surface interactions, including topography, coastlines, and surface roughness, which gives rise to highly localised and intermittent precipitation structures that are far less common over the open ocean (Roe, 2005; Vogel et al., 2020; Stevens and Bony, 2013). For Denmark, this distinction is particularly important, as coastal geometry and mesoscale land-sea contrasts strongly influence precipitation organisation. To enable the model to represent these processes, we provide static geographical information - topography and a land-sea mask - as additional conditioning inputs along with the LR dynamic fields. These static fields are broadcast to the HR grid and concatenated with the dynamic inputs, which thereby allows the network to learn regime-dependent mappings associated with coastal and orographic features. Additionally, we introduce a spatially varying loss weighting that emphasises predictive accuracy over land and near-coastal regions while still training on the full land-sea domain. This design biases learning towards physically relevant small-scale precipitation structures over land. Full details of the distance-based weighting formulation are provided in the Supplement.

Precipitation regimes and large-scale atmospheric dynamics vary strongly across seasons. To account for this, we incorporate explicit seasonal conditioning using Feature-wise Linear Modulation (FiLM; Perez et al. (2018)) based on Day-of-Year (DOY) information. The DOY is encoded using a continuous cyclic representation to reflect the annual cycle and to avoid artificial steep transitions and discontinuities between December and January. The seasonal signal is embedded into the same latent space as the diffusion noise level conditioning and combined with it to form a joint vector. At each resolution layer in the U-Net, this combined conditioning is used to modulate intermediate feature maps through learned channel-wise scaling and shifting

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

operations. This allows the network to smoothly reweight features in a season-dependent way, thereby enabling seasonal learning, such as enhanced representation of convective structures in summer and more stratiform patterns in winter. In practice, this results in a model with improved seasonal variability and realism and supports generalisation beyond mean climatology.

In addition to the diffusion model, we optionally include a lightweight auxiliary *RainGate* head that predicts per-pixel wet/dry occurrence. The RainGate consists of a shallow convolutional network trained with a binary classification loss, where pixels exceeding 0.1 mm day^{-1} are considered wet. It takes as input the same large-scale and static conditioning fields as the diffusion model. The loss is used only during training and does not directly modify the diffusion loss through pixel-wise reweighting in the final configuration. Instead, it provides complementary supervision during training that helps the shared representation to better distinguish between wet and dry regimes. Results from the component study show that this auxiliary task acts as a calibration and occurrence-and-intensity shaper, see Section 4.1.

We tested a number of additional design choices, namely: (i) residual learning with an upsampled LR baseline, and (ii) dual representations of the LR precipitation field (normalised with LR statistics, and normalised with HR statistics). While these approaches can improve numerical stability in under-conditioned settings, they degraded performance for precipitation, once the EDM hyperparameters were tuned, and seasonal and geographic conditions are included. We therefore do not retain residual prediction or dual-LR conditioning in the final model. More information on hypotheses and limitations can be seen in Supplement Sections S1.4 and S1.5.

2.4 Scale-aware inference

The model architecture and design choices are made to promote consistency between large-scale forcing and fine-scale variability. We here also introduce a scale-aware parameter, a multiplicative factor σ^*, to enable control of the stochasticity during sampling. This factor is applied only to the injected noise term in the EDM sampler (see full formulation in Supplement S1).

Conceptually, σ^* rescales the EDM noise schedule at *inference only*, which means that the network - otherwise trained on a fixed σ-schedule - is asked to denoise along a slightly modified trajectory. Small adjustments, like the ones we implement, can be thought of as akin to how "temperature" can be modified in VAEs/GANs (Zhou et al., 2024), and moderate deviations yield plausible samples, whereas large deviations push the sampler off-manifold and thus into regimes not covered by the training distribution. We therefore treat σ^* as a calibrated inference knob for exploring the model's uncertainty and behaviour within a restricted, validated range. We implement this control through a late-stage rescaling of the EDM noise schedule, applied only at small noise levels. The full mathematical formulation is provided in Supplement Section S1.7.

For EDMs, the late-stage denoising (small σ) is where the fine-scale features are created. When we reduce σ in this regime, $\sigma^* < 1$, we increase the denoiser's confidence and amplify small-scale gradients, which in turn produces sharper and more textured fields. On the other hand, an increase in σ with $\sigma^* > 1$ moves the sampler into a higher-noise regime, where the denoiser becomes more conservative, with smaller updates, effectively resulting in smoother and more LR-coherent fields. Empirically, decreasing σ^* in late steps can increase high-k power due to sampler entering the low-σ regime earlier/more strongly. This regime is where detail is formed, after the more broad stochastic search at earlier steps, and thus decreasing σ^*

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

here leads to sharper (but potentially artifact-prone) small-scale structure. On the other hand, increasing σ^* delays or suppresses that transition, which in contrast yields smoother, more LR-coherent fields.

This scaling perturbation preserves the trained denoiser but alters the inference trajectory. We therefore treat σ^* as a calibrated control within a validated area.

For σ^* to be physically meaningful, we analyse the power spectral density (PSD) of generated precipitation fields for a range of σ^* values and compare it with the PSD of the HR DANRA samples (Lovejoy and Schertzer, 2013; Gires et al., 2012; Hess et al., 2025). This provides us with an analysis of the scale-awareness parameter, which can help us ensure that the generated fields reproduce spatial scaling properties while preserving LR anchoring, but with a range of σ^* for selection.

3 Experimental configuration and implementation

This section contains a description of the data used for training the model, including preprocessing to prepare data for training, full training setup and selected most-important hyperparameters and architectural choices, with the rest in Supplement S3, description of selected benchmarks for comparison, evaluation metrics, and experimental setup and design of the controlled architectural component study.

3.1 Data sources, study region and pipeline

The data used in this study is geographically located over Denmark and using:

- A high-resolution target (HR) of the DANish regional ReAnalysis, DANRA, (Yang et al., 2025b) daily precipitation on a 180×180 pixel subdomain at a 2.5 km $\times$ 2.5 km spatial resolution. Training samples are random 128×128 cutouts within this region (shuffled across dates for training stability).
- A low-resolution condition (LR) of ERA5 (Hersbach et al., 2020) daily precipitation for the same period, bilinearly regridded to HR resolution. All conditioning cutouts are co-located with the HR target cutouts.

In Figure 3, a few examples of precipitation maps of DANRA and ERA5 are shown, along with the difference between them, displaying the importance of high-resolution regional outputs, where extremes and localised events are more resolved. Figure 3 also shows DANRA and ERA5 data as seasonal distributions (on a logarithmic y-scale), showing clearly that ERA5 has a difficulty in capturing the extreme precipitation events at the tail of the distribution, especially in summer months. In this study, the high-resolution reanalysis product DANRA functions as the reference dataset used for both training and evaluation.

Unless noted, analyses and plots focus on land pixels only (coastline-buffered mask). Details of the end-to-end preprocessing (regridding, cropping, split generation and conversion to Zarr file-types for data handling (Zarr Developers, 2026)) appear in Supplementary Information Section S2 along with a schematic of the full data preprocessing pipeline (Figure S1).

All data are stored as day-wise arrays in .zarr archives. From the full geographical domain (589×789 pixels), we define the 180×180 pixel-area of interest, and draw the randomly located 128×128 pixel-area from within this area during data-loading.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

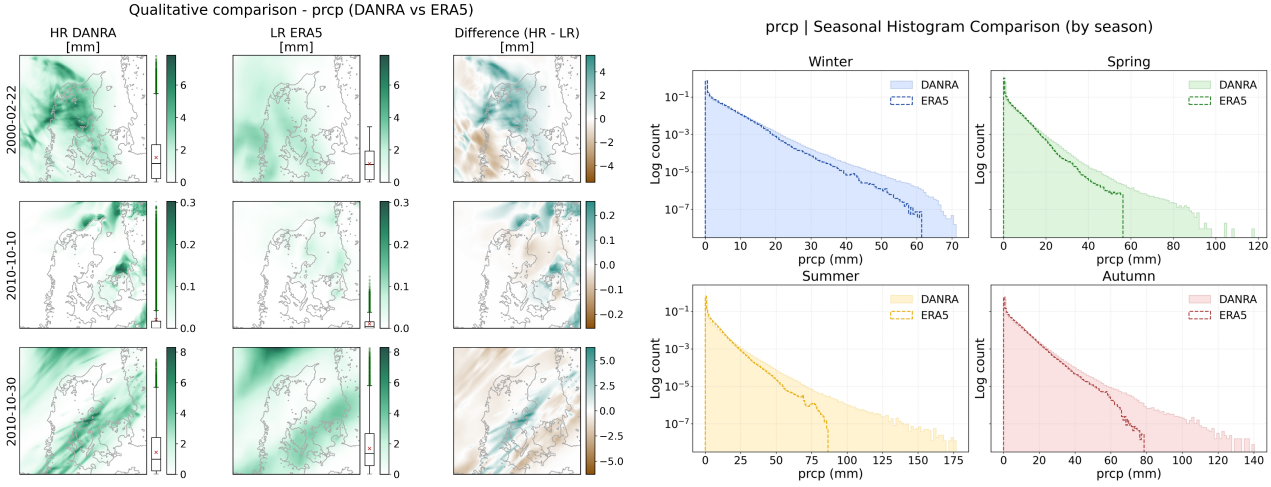

Figure 3. *Comparison of precipitation (daily fields and distributions) from DANRA and ERA5.* **Left**: Example of daily maps from DANRA (2.5 km × 2.5 km) and ERA5 (0.25°, ∼ 30 × 30 km, regridded to DANRA projection and resolution). The difference highlights the gain in spatial detail and representation of localised extremes in high-resolution data compared to the coarser ERA5 fields. **Right**: Seasonal histograms of ERA5 and DANRA showing that ERA5 underrepresents the upper tails of precipitation intensity, especially during summer (June-July-August). The comparison illustrates the need for downscaling to capture the full distribution of heavy precipitation events.

We apply a logarithmic Z-score normalisation to precipitation, combining a log transform with standardisation to zero mean and unit variance. We apply a small ϵ to avoid $\log(0)$ (here $\epsilon = 0.01$ mm day^{-1}). The model is trained to predict in this log-z-score scaled space and outputs are back-transformed for plotting and evaluation. The ERA5 precipitation is normalised using ERA5 (LR) training statistics, while the DANRA target is normalised using DANRA (HR) training statistics. This preserves the native distributions of each dataset.

To support reproducibility, we use fixed random seeds, with a seed of 504 for model training and generation and a separate seed of 1234 for ensemble subsampling where applicable. For the architectural component study, we use up to three seeds to test seed-noise and -importance. Paths and environment variables controlling data access, checkpoints, logging, and generation and evaluation outputs are fully specified through configuration files and job scripts. All experiments were executed on HPC GPU infrastructure, with full hardware and runtime details provided in Supplement S3. All code, configuration files, and experiment settings are version-controlled and publicly available (see Data and Code Availability). A summary of the main architectural, training, and data-handling settings for this representative configuration is presented in Table 1.

3.2 Validation and evaluation protocol

Our experimental workflow separates architectural model selection from sampler hyperparameter optimisation. This separation is intentional, as sampler parameters in diffusion models interact with the learned denoiser and can otherwise muddle results of architectural comparisons. We first define a fixed, stable EDM sampler configuration, which is unchanged for all architectural

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

Table 1. *Main architectural, training, and data handling configuration for the representative EDM-based model used throughout this study*. Unless explicitly stated otherwise, results in the main text correspond to this configuration. Deviations for architectural component experiments are described separately or in the Supplement.

COMPONENT	CONFIGURATION / SETTINGS
Architecture and Inference (EDM)	**Sampler**: edm_sampler with 56 steps; sinusoidal time embedding (dim = 256); last feature map with 512 channels; multi-head attention (4 heads); block layers $[2, 2, 2, 2]$.
	EDM schedule: $P_{\text{mean}} = -1.5$, $P_{\text{std}} = 1.2$, $\sigma_{\min} = 2\times10^{-3}$, $\sigma_{\max} = 80$, $\rho = 7$; modest churn ($S_{\text{churn}} = 2$, $S_{\min} = 40$, $S_{\max} = 80$, $S_{\text{noise}} = 1$).
	Decoder: transposed convolution (no resize-conv), group normalisation (8 groups), SiLU activation.
Training	**Optimiser**: AdamW, learning rate 2×10^{-4} with `ReduceLROnPlateau` (factor = 0.5, patience = 15, $\text{lr}_{\min} = 10^{-6}$); weight decay 10^{-6}.
	Batch size: 16; 400 **epochs**; early stopping **patience** = 75.
	Mixed precision: off. Exponential Moving Average (**EMA**) available but disabled for the basic run (used in architectural component experiments).
	Loss: diffusion loss in scaled space. **SDF-weighted loss** is enabled where stated (in specific component study experiments and final configuration), otherwise the diffusion loss is unweighted.
	Random seed = 504 for reproducibility; fixed date lists used for consistent quicklook generations.
Data handling	Cutouts of **size** 128×128 are sampled on-the-fly in the data loader. **Caching** is disabled in base runs. Dataloader **workers** are set from the LUMI job environment.
	Temporal split: *train:* 1991–2015, *validation:* 2016–2018, *test:* 2019–2020.

component experiments. This reference sampler is not assumed to be optimal, but is chosen to produce physically plausible samples across all tested architectures.

Using the fixed sampler, we perform a structured architectural component study on the validation period and select a final model architecture, denoted F. Only after the final model F is selected do we perform a dedicated sampler hyperparameter sweep for optimisation, thereby tuning the EDM sampling parameters (ρ, S_{churn}, σ) on the validation period. All results on the independent test split are reported using this final architecture and its tuned sampler configuration.

We evaluate our model on two years of unseen data to test the model's overall generalizability and capability. It would be possible to do an evaluation on the full 28-year training-validation dataset to generate a full ensemble dataset that can be useful for impact modellers. The focus in this paper, though, is rather on the model's general generative performance especially for future use, and therefore we choose to focus our evaluation on an unseen data-split. From the test years, we generate an ensemble of $N = 32$ samples per day, and also provide evaluation of a Probability Matched Mean (PMM) presented as the

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

best estimate for selecting a single field of the ensemble that is still relatively representable for the entire distribution of the ensemble. The evaluation, results, and discussions are focused on ensembles rather than single deterministic maps to better do the stochasticity of the model justice. We compare the model performance against two simple benchmarks with the same data pipeline and splits (Table 2).

A preliminary sweep of the σ^* parameter was performed ($\sigma^* \in [0.70, 0.85, 1.00, 1.15, 1.30]$) to identify a stable and physically admissible range (full details in Supplement S4, and Figures S3 and S6). This sweep is used only to delimit the region in which σ^* produces physically consistent downscaling behaviour and does *not* imply any operational interpretation beyond the final suggested calibrated range. The final σ^* evaluations are tested in a tightened grid of $\sigma^* \in [0.90, 0.95, 1.00, 1.05, 1.10, 1.15]$ which brackets the visually and physically realistic regime. This sweep is used only when assessing inference-time scale-awareness (Section 4.5), only evaluated on the final selected model. Unless noted, presented results use $\sigma^* = 1.0$

Table 2. *The two benchmarks used for comparison of model performance*. Two simple benchmarks are selected for general performance comparison, a bilinearly upsampled ERA5, and a Quantile mapping (QM) fitted on the training dataset and tested on the test set.

Benchmark	Type / Reference	Inputs (core)	Settings / Notes
Bilinear ERA5	Deterministic upsampling	ERA5 precipitation (daily), bilinearly upsampled to HR grid	No trainable parameters; establishes LR→HR fidelity of large-scale patterns and climatology.
Quantile Mapping (QM)	Statistical bias correction (Themeßl et al., 2011)	Bilinear ERA5 (as above)	Fit on *train* set CDFs; apply to *test*. Quantiles: {0.00, 0.01, 0.02, 0.05, 0.10, 0.20, 0.40, 0.60, 0.80, 0.90, 0.95, 0.98, 0.99, 0.995}; wet-day threshold 0.5 mm day^{-1}

To comprehensively assess a generative downscaling model, we evaluate across four metrics pillars, investigating (A) the *statistical and climatological fidelity* (mean and physical behaviour over time), the (B) *spatial and scale-dependent structure* (realism of spatial organisation and scale-dependent behaviour), (C) the *extreme values and tails* (rare and high-impact events), and (D) the *probabilistic and ensemble skill* (ensemble calibration). Finally, to investigate σ^* behaviour and controllability, we also evaluate on (D) the controllable scale-awareness of the model through σ^*. The details of the five groups of metrics are presented in Table 3, and described in more detail in Supplement S5.1.

Where relevant, we also compute auxiliary diagnostics such as autocorrelation, wet/dry spell length distributions, spread–skill relationships, reliability diagrams, and variograms (see configuration parameters and code in the public repository) (Gneiting et al., 2005). The main text focuses on CRPS, ISS, PSD, pixel distributions, extremes, and climatology as these together capture the core generative and physical realism. Complementary results of the additional metrics are presented in S5.

To evaluate the different architectural components, we performed a structured component study starting from a minimal stable EDM reference configuration. During early development we also explored residual-aware prediction and dual-normalised representation of LR precipitation as possible ways to stabilise optimisation and bias alignment. While these mechanisms did

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

Table 3. Summary of the multi-pillar evaluation framework used to assess generative downscaling performance. Each pillar targets a distinct aspect of model realism and addresses complementary failure modes.

Pillar	What it tests	Core diagnostics used in this study	Typical failure modes
A. Statistical and climatological fidelity	Mean behavior and marginal distributions	Pixel-value distributions (all days); annual mean and sum maps (2019–2020); wet-day frequency	Correct means but biased variability or truncated tails
B. Spatial and scale-dependent structure	Realism of spatial organisation across scales	Intensity–Scale Skill (ISS) across neighbourhoods (5–40 km) and thresholds (1–50 mm day^{-1}); isotropic Power Spectral Density (PSD) and mesoscale slope (5–20 km)	Over-smoothing; missing fine-scale variance; scale leakage
C. Extremes and tails	Representation of rare, high-impact events	Upper tail quantiles (P95–P99.99); wet-day occurrence vs intensity; tail-focused diagnostics	Underestimated extremes; incorrect occurrence–intensity balance
D. Probabilistic and ensemble skill	Reliability and calibration of predictive uncertainty	Continuous Ranked Probability Score (CRPS); ensemble MAE; Probability Integral Transform (PIT) histogram	Under-dispersed or overconfident ensembles
E. Controllable scale-awareness (σ^*)	Sensitivity of spatial structure and uncertainty to inference scale	σ^*-sweeps of CRPS, low-pass LR–GEN correlation, PSD slope; identification of physically grounded operating regimes	Uncontrolled texture amplification or loss of spatial coherence

ease convergence in preliminary runs, they were not beneficial once the conditioning strategy and auxiliary objectives were fully functional (see Supplement S4 for more). Therefore they were omitted in the final model configuration. All reported experimental configurations are defined relative to the selected baseline model, and we evaluate each addition using the eight selected metrics seen in Table 4, which each helps cover one of the four aspects of scale awareness (PSD slope and ISS), probabilistic skill (CRPS), extreme-value performance (P99, P99.9), and climatological capacity (wet-day frequency, yearly sum spatial mean and standard deviation). In addition to single-feature additions, selected combined architecture experiments were conducted to assess interactions.

The following section reports results from benchmarks and architectural experiments relative to the baseline EDM, as well as results for the final selected model, quantifying ensemble realism, multiscale fidelity, and the practical range of controllability afforded by σ^*.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

4 Results

The following section first presents results from the architectural component experiments and then focuses on the final model performance, presenting results from ensemble and probabilistic skill, spatial structure and scale-dependence, statistical and climatological fidelity, and finally the effects of introducing scale-awareness at inference time, through σ^*.

4.1 Controlled architectural component study

We evaluate how different architectural components affect different aspects of model behaviour relative to a common reference configuration, recognizing that improvements in one metric may come at the expense of others. We investigate the change to a suite of metrics given different additions to the baseline model. The results can be seen in Table 4 with colours indicating performance closer to or further from the DANRA target.

All architectural experiments reported in this section are performed using a fixed reference EDM sampler configuration, $[\rho = 7.0, S_{\text{churn}} = 2.0, \sigma_{\text{data}} = 1.0]$. The goal of these experiments is not to maximise absolute performance for each individual configuration but to assess the relative impact and interactions of architectural components under identical sampling dynamics. We evaluate performance changes in spatial structure (PSD slope and intercept), scale-dependent skill (ISS at 20 km for intensities of ≥ 10 mm day^{-1}), probabilistic performance (CRPS), extreme-value statistics, wet-day frequency, and mean and standard deviation of spatially aggregated yearly precipitation sums. STD in the table denotes the spatial standard deviation over Denmark of the annual-sum field (i.e., heterogeneity), not temporal variability. Metrics are presented as means across training seeds (number of runs in parenthesis next to the model name) with spread across seed-runs attached as standard deviation. Note that we distinguish variability across space (annual-sum heterogeneity, last column in table) from variability across training seeds (optimisation stochasticity). Here, 'seed' refers to the training random seed (network initialisation and data ordering), while the sampler seed is kept fixed across all experiments. All metrics are computed consistently over the same evaluation mask/domain across experiments. We report P99 and P99.9 as representative of frequent and rare extremes. Additional percentiles (P95 and P99.99) are provided in the supplement.

Based on Table 4, showing results from the architectural experiments, we see that seasonal conditioning (B0_S) mainly corrects mean-state (yearly sum) and wet-day bias, but leaves the annual-sum field overly smooth (low spatial variance) and tails conservative. Adding geographic conditioning to the baseline (B0_G) systematically increases scale skill (ISS), improves wet-day frequency, and, most importantly, restores spatial heterogeneity, as seen in the yearly-sum STD. Seasonal and geographic information are the primary correctors, where SDF-weighted loss (_SDF) and RainGate BCE-loss (_RGBCE) act as secondary refinements. In combination with geography and seasonal information, RGBCE improves probabilistic calibration (CRPS) and sometimes tails, while SDF further helps on spatial structure. An interesting result is that adding geographic information on top of seasonal DOY information (B1_G) does not automatically help. This configuration actually harms calibration and pushes the model toward a dry, conservative regime. But combining the two conditions with an additional constraint (SDF) and/or a calibration-oriented loss (RGBCE) allows the model to translate the added information into improved precipitation statistics (B1_ GSDF, B1_ G_ RGBCE, and B1_ GSDF_ RGBCE). A clear challenge for all configurations remains extreme

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

tail behaviour - and especially lower-end extremes (P99). While most configurations improve over the baseline in spatial structure and calibration, P99 remains systematically underestimated relative to DANRA across all models, indicating that tail deficiencies are persistent rather than configuration specific. The only configuration that overshoots on high-tail extremes (P99.9), B0_RGBCE, does so with larger seed variability, indicating that tail fidelity can come at the cost of robustness.

The results from the component study clearly highlight that information does not equal skill: though added conditioning can increase the hypothesis space, it does not guarantee better training dynamics or calibrated output. Regularisation of some kind is necessary to convert information into usable structure, here with SDF acting like a spatial constraint and regulariser, and RGBCE like a calibration/occurrence-and-intensity shaping term. Combining them prevents the model from collapsing into the conservative, dry mode seen in B1_G. This further underlines exactly why structural architectural studies are needed - effects are strongly non-linear and interaction driven. We therefore select B1_GSDF_RGBCE as the final model for full test-set evaluation as it provides the most balanced performance across diagnostics with comparatively low inter-seed variability, which indicates greater robustness than B1_G_RGBCE, which exhibits higher seed sensitivity, particularly in annual-sum and spatial STD. We also investigated additional architectural features (residual prediction, dual LR-scaled conditioning, Exponential Moving Average (EMA)), with results summarised in Supplement S4. We observed that residual prediction had small-to-no effect on results, and dual LR conditioning clearly degraded results after the model was properly conditioned and calibrated. The effect of EMA was more nuanced: while it altered seed-to-seed variability in several configurations, it was not uniformly stabilizing or consistently conservative - which might otherwise have been expected. Instead, EMA interacted strongly with conditioning and seed initialisation, in some cases improving tail and mean-state statistics, and in others showing only small changes. Due to this large variability and unpredictability, we do not interpret EMA as a guaranteed stabiliser in this setting.

After selecting the final architecture, F, based on the architectural experiments (run on the validation dataset), we perform a dedicated sampler hyperparameter sweep for this final model (Supplement Table S2). We scan over $(\rho, S_{\text{churn}}, \sigma)$ on the validation split and select the configuration that optimises probabilistic skill (CRPS), wet-day statistics and extremes, and ensemble reliability. This tuned sampler is used for all results reported on the test period. We select the sampler configuration ($\rho = 7.0$, $S_{\text{churn}} = 2.0$, $\sigma = 1.1$) as the final setting, as it provides the best balance between hydrologically relevant extremes, wet-day frequency, and spatial variability, while maintaining stable probabilistic skill and realistic small-scale structure. While some configurations marginally outperform individual metrics, this sampler consistently performs well across all key diagnostics relevant for precipitation downscaling. Specifically increasing σ from 1.0 to 1.1 improves tail behaviour and intermittency with only marginal impact on CRPS and PSD slope, and was therefore preferred for this paper's focus on impacts through extremes. For the full sampler-sweep results, see Table S2 in Supplement. The following results are from running the calibrated model on an unseen test dataset, covering the years 2019 and 2020.

4.2 Ensemble realism and probabilistic skill

Figure 4 shows four examples of different days in the test dataset, with the low-resolution conditioning ERA5, high-resolution DANRA target, three ensemble members (Ens-1 to -3), and the Probability Matched Mean (PMM) of the generated ensemble. These spatial maps illustrate the spatial diversity and non-deterministic nature of the diffusion sampling and provide insight

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

Table 4. *Overview of model performance given various architectural and conditioning choices.* Summary of results from the architectural component experiments for the precipitation downscaling EDM. Metrics are reported as mean ± standard deviation across training seeds, with a fixed sampler seed and configuration. Colours indicate performance compared to the reference dataset DANRA: green denotes improved agreement, orange denotes degraded agreement, and purple indicates metrics with unusually large inter-seed variability, signalling reduced robustness. The standard deviation of the annual precipitation sum represents spatial heterogeneity across Denmark, not temporal variability. We select the final configuration as the most robust across training seeds, prioritizing stable annual-sum statistics and tail behaviour over single-metric optima. † Selected final model based on robustness across seeds.

	PSD slope ($\lambda < 20$ km)	**ISS at 20 km** (≥ 10 mm day^{-1})	**CRPS** (mean)	**P99** [mm day^{-1}]	**P99.9** [mm day^{-1}]	**Wet-day frequency** [%]	**Yearly sum** (2017) [mm]	**STD on yearly sum** [mm]
DANRA	**-2.29**	**-**	**-**	**18.41**	**18.41**	**33.28**	**834.13**	**116.12**
B0: Baseline (3)	-2.59 ± 0.14	0.7849 ± 0.0043	0.803 ± 0.055	15.61 ± 0.30	34.28 ± 2.43	29.09 ± 2.89	652.95 ± 61.95	48.25 ± 8.63
B0_S (2) B0 + seasonal	-2.36 ± 0.00	0.7860 ± 0.0064	0.735 ± 0.007	15.70 ± 0.47	32.44 ± 1.80	31.25 ± 0.87	700.46 ± 2.65	54.45 ± 1.00
B0_G (2) B0 + geographic	-2.45 ± 0.00	0.7924 ± 0.0012	0.758 ± 0.013	16.41 ± 0.35	34.59 ± 0.54	30.85 ± 0.13	707.89 ± 17.28	74.848 ± 1.39
B0_GSDF (2) B0 + geographic and SDF	-2.48 ± 0.03	0.7903 ± 0.0071	0.755 ± 0.002	16.14 ± 0.25	33.47 ± 0.02	31.26 ± 0.23	714.67 ± 10.48	78.212 ± 4.57
B0_RGBCE (2) B0 + RainGate BCE	-2.49 ± 0.18	0.7901 ± 0.0018	0.765 ± 0.007	16.42 ± 1.10	35.19 ± 5.53	30.77 ± 1.52	707.69 ± 4.45	52.16 ± 1.61
B1 = B0_S New baseline								
B1_G (2) B1 + Geographic	-2.38 ± 0.05	0.7819 ± 0.0018	0.761 ± 0.033	14.42 ± 1.04	28.85 ± 1.59	29.79 ± 1.39	632.16 ± 50.60	73.65 ± 7.32
B1_GSDF (2) B1 + Geographic and SDF	-2.34 ± 0.02	0.7896 ± 0.0012	0.738 ± 0.026	14.96 ± 0.86	29.72 ± 1.062	32.74 ± 0.37	697.42 ± 39.18	74.08 ± 2.07
B1_RGBCE (2) B1 + RainGate BCE	-2.39 ± 0.08	0.7919 ± 0.0005	0.741 ± 0.014	15.24 ± 0.35	31.05 ± 0.89	31.60 ± 0.58	686.41 ± 19.86	60.73 ± 2.17
B1_G_RGBCE (2) B1 + Geo. + RainGate BCE	-2.33 ± 0.10	0.7937 ± 0.0024	0.718 ± 0.032	15.96 ± 1.06	31.05 ± 1.07	32.37 ± 1.89	731.08 ± 78.78	81.24 ± 16.48
B1_GSDF_RGBCE† (2) B1 + Geo., SDF + RainGate	-2.38 ± 0.00	0.7874 ± 0.0004	0.732 ± 0.005	15.65 ± 0.04	32.16 ± 0.88	31.66 ± 0.77	704.22 ± 10.69	77.82 ± 3.74

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

into ensemble sharpness and calibration. The LR ERA5 is shown with the full field to show that the model is conditioned on the full atmospheric field, and HR, ensemble members and PMM are shown only for land, as we mask out the ocean for evaluation. Attached to each realisation is a boxplot that illustrates the (land-only) full pixel distributions for that day and sample. The colour bars are fixed across rows for easier comparison between LR, HR and generated samples. Complementary ensemble metrics (PIT histogram, spatial, temporal and seasonal CRPS, in Supplementary Information Section S5) show that the ensemble is biased towards dry samples and slightly under-dispersed. CRPS analysis also reveals higher CRPS in summer months, consistent with our knowledge on difficulty of mapping summer convective precipitation. These results illustrate that the model generally produces statistically coherent ensembles, but does exhibit a clear bias towards being drier than DANRA with a mild under-dispersion (Figures 6 and 7).

4.3 Spatial and scale-dependent structure

Figure 5 shows the isotropic Power Spectral Density (PSD) of precipitation for the mean of the individual EDM-generated ensemble members PSDs (GEN), the Probability-Matched Mean (PMM), the low-resolution conditioning input (LR, ERA5), a quantile-mapped baseline (QM), and the high-resolution reference (HR, DANRA). The PSD is plotted as a function of wavelength on a reversed logarithmic axis, such that large-scale variability (low wavenumber k, large λ) appears on the left and fine-scale variability (high wavenumber k, small λ) on the right. Vertical markers indicate the transition between low- and high-wavenumber regimes used for slope estimation. The EDM ensemble mean closely reproduces the reference PSD across scales, with high-k slopes that are nearly identical to DANRA, indicating a physically consistent reconstruction of the multiscale precipitation without excess small-scale noise. The consistent lower intensity on GEN than on HR indicates that there is a systematic dry-bias across scales. As expected, the PMM degrades skill relative to the ensemble mean by smoothing scale-dependent structure, despite preserving marginal extremes - by construction.

While the quantile-mapped baseline appears to perform well in terms of absolute PSD magnitude and slope, this agreement must be interpreted with caution: quantile mapping is a deterministic marginal transformation that adjusts grid-point distributions without explicitly modifying *spatial* dependence (Maraun, 2013). When applied across scales, such transformations can effectively inflate variability rather than generate physically consistent small-scale structure, which is otherwise what we are looking for. The result is an increase in high-wavenumber power and slight overshoot in the fine-scale regime. Thus, the apparent spectral agreement can be interpreted as a form of spectral overfitting, where most of the PSD is matched, but the underlying spatial coherence and event morphology can remain misrepresented. Similar phenomena are documented in other fields of generative modelling, where outputs reproduce the sought spectral statistics, yet systematically show high-frequency distortions (Durall et al., 2020).

These findings demonstrate that PSD-based metrics cannot stand alone when evaluating downscaling performance, as the QM achieves clear favourable spectral scores, while failing to reproduce physically meaningful spatial organisation. By contrast, the EDM maintains coherent multiscale structure while retaining consistency with the large-scale atmospheric signal.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

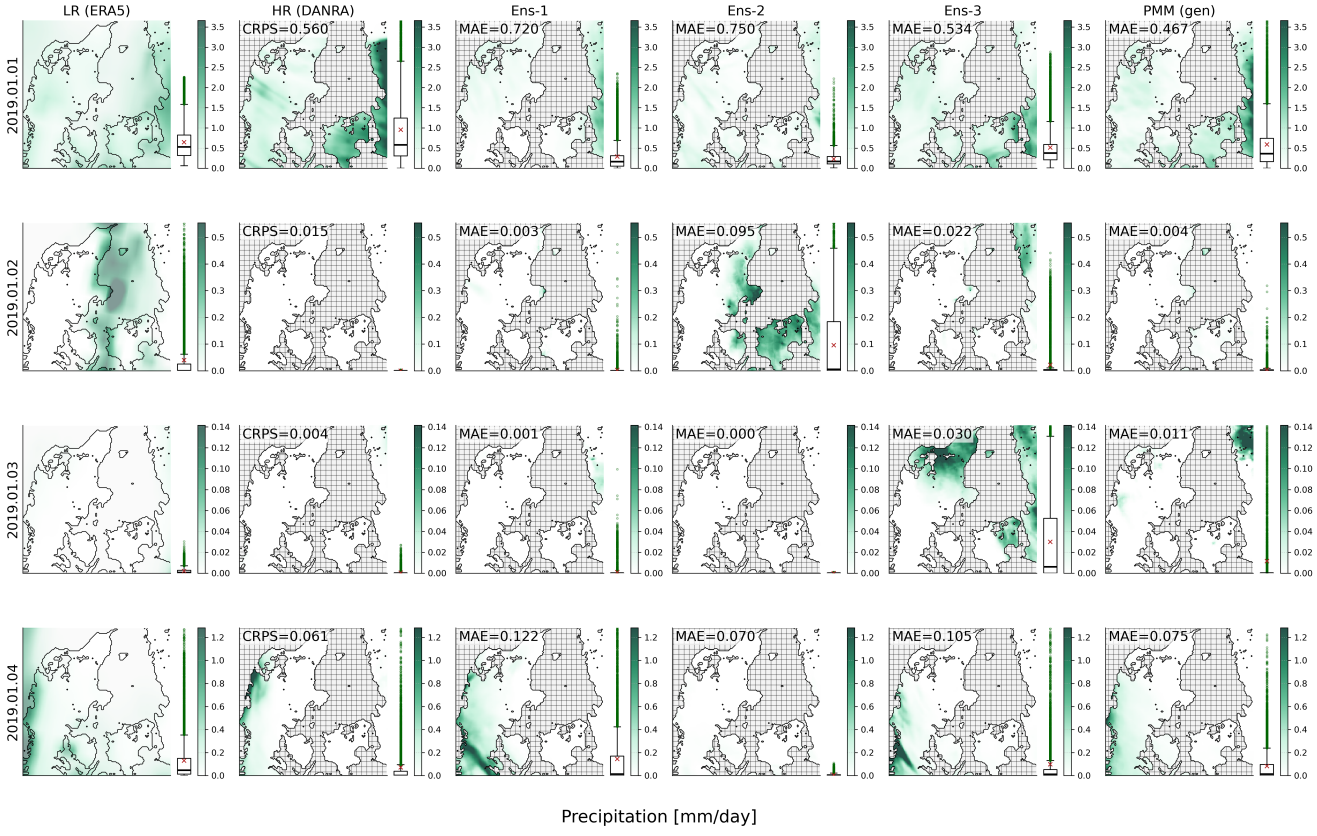

Figure 4. *Example model ensemble realisation and reference fields (HR DANRA and LR ERA5).* Examples comparing the EDM ensemble members, their probability-matched mean (PMM) and the reference DANRA and ERA5 inputs. The boxplots to the left of the colour bars illustrate the full pooled pixel-distribution for the respective sample (masked to land-only pixels). The examples illustrate the model's capability of representing uncertainty and variability in localised precipitation events. The continuous ranked probability score (CRPS) for the entire ensemble is attached to the HR DANRA reference, and each ensemble member and PMM has the mean-average error (MAE) attached.

4.4 Statistical and climatological fidelity

Figures 6-8 assess statistical and climatological fidelity from complementary perspectives, covering distributional shape, spatial aggregation, and extreme-event statistics. Together, these diagnostics distinguish agreement in marginal precipitation distributions from physically meaningful spatial and event-based realism. Figure 6 focuses on the shape and tail behaviour of pooled pixel-value distributions, separated by season. The EDM generally follows the distribution shape of HR, but with a slight underestimate from the mid-range and up to around P99.9. Looking at the separate seasons, it is evident that the

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

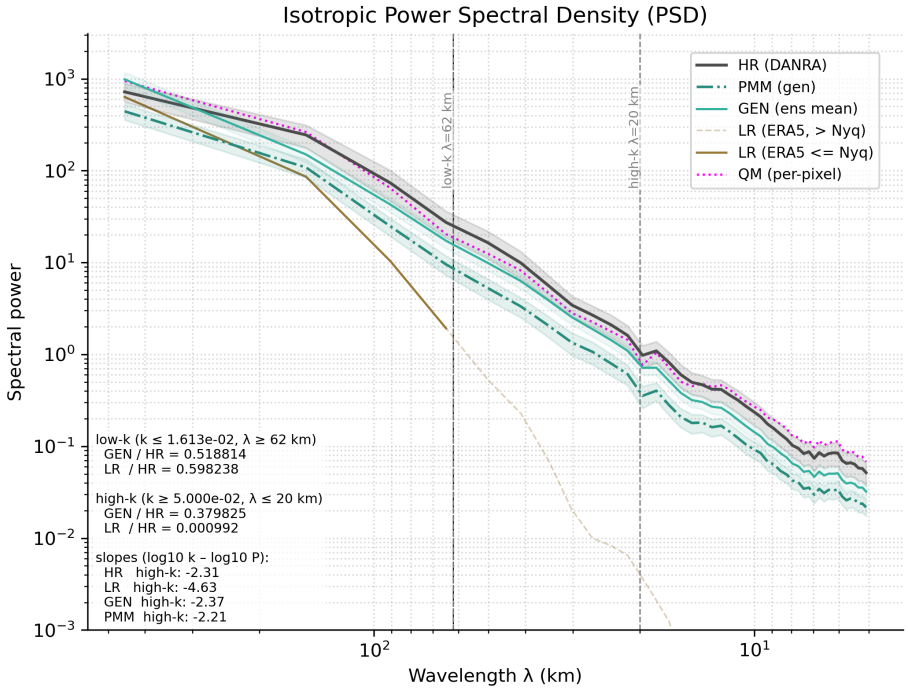

Figure 5. *Power spectral density (PSD) analysis of spatial scaling.* Isotropic PSD distributions are shown for the EDM-generated ensemble mean (GEN), probability-matched mean (PMM), reference LR and HR data, and a quantile-mapped (QM) benchmark. LR is represented with a solid line for $k \leq$ LR Nyquist frequency and with a dashed line for $k >$ LR Nyquist. While QM closely matches the reference PSD across scales, this agreement primarily reflects variance redistribution through point-wise bias correction, which results in close-matching PSDs but overly smoothed precipitation fields. In contrast, the EDM ensemble mean reproduces the observed scaling behaviour to a large degree while still retaining realistic fine-scale variability without spectral overfitting (excess high-k power).

model captures spring precipitation distributions best, but that it also seems to capture high-end extreme convective events in summertime well (>P99.9).

Figure 7 evaluates whether correct spatial climatology and large-scale precipitation gradients are preserved in annual aggregates. The ensemble mean reproduces the dominant west-east precipitation gradient and regional maxima over Denmark, also seen in DANRA. This is indication that the EDM preserves large-scale atmospheric controls on precipitation. The bilinearly upsampled ERA5 field (LR) appears spatially smooth, still with the west-east gradient, and with a slight wet bias, masking fine-scale variability despite comparable large-scale totals. These maps emphasise that correct spatial climatology requires both accurate temporal aggregates and realistic spatial heterogeneity.

Figure 8 isolates extreme-value behaviour and wet-day occurrence statistics, highlighting differences between intensity calibration and event detection skill. While QM and PMM achieve competitive performance for some tail percentiles, these gains do not consistently translate into improved wet-day frequency or hit-rate. The EDM ensemble balances intensity calibration

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

and event occurrence more consistently, though lower extreme intensities remain slightly underestimated. These results highlight that extreme-value agreement alone is insufficient to assess hydrological and impact-model relevant realism without both evaluation of occurrence and spatial structure.

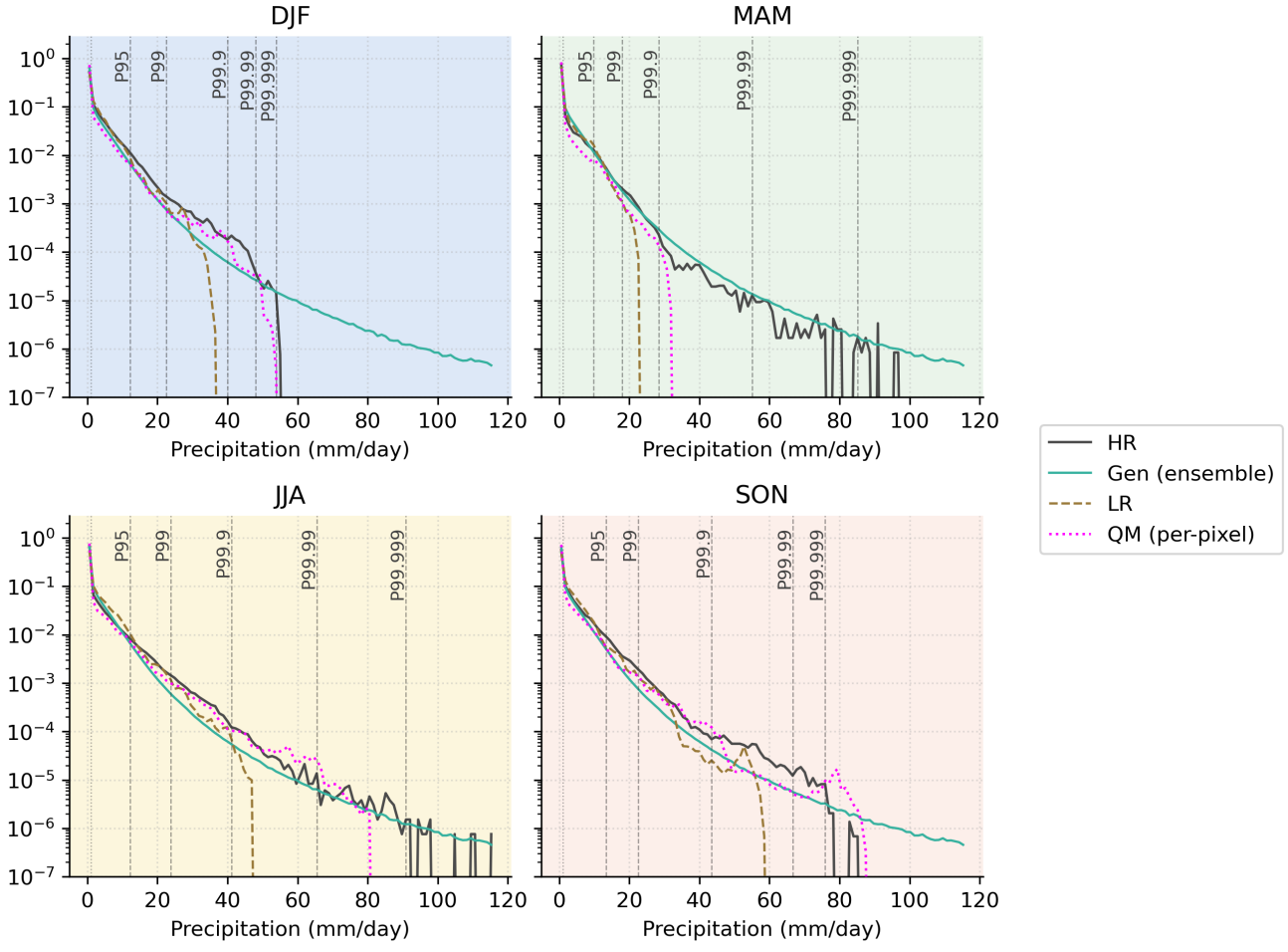

Figure 6. *Pooled pixel-value distributions of daily precipitation (log-plot), separated based on season.* Vertical dashed lines indicate the P95, P99, P99.9, and P99.99 percentiles of the reference (DANRA) distribution, providing fixed intensity thresholds for comparison across methods. Agreement in this figure therefore reflects similarity in marginal intensity distributions rather than spatial or event-level realism.

4.5 Scale-awareness and interpretability

Figure 9 illustrates the effect of increasing or decreasing σ^* in qualitative examples, where the top row shows three ensemble members sampled with $\sigma^* = 0.8$ at inference time. The bottom row shows an increased σ^* at 1.20. With lower σ^*, we see an increase in fine-scale variability, and, at low enough σ^*, even unphysical noise-like patterns. With higher σ^* on the other

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

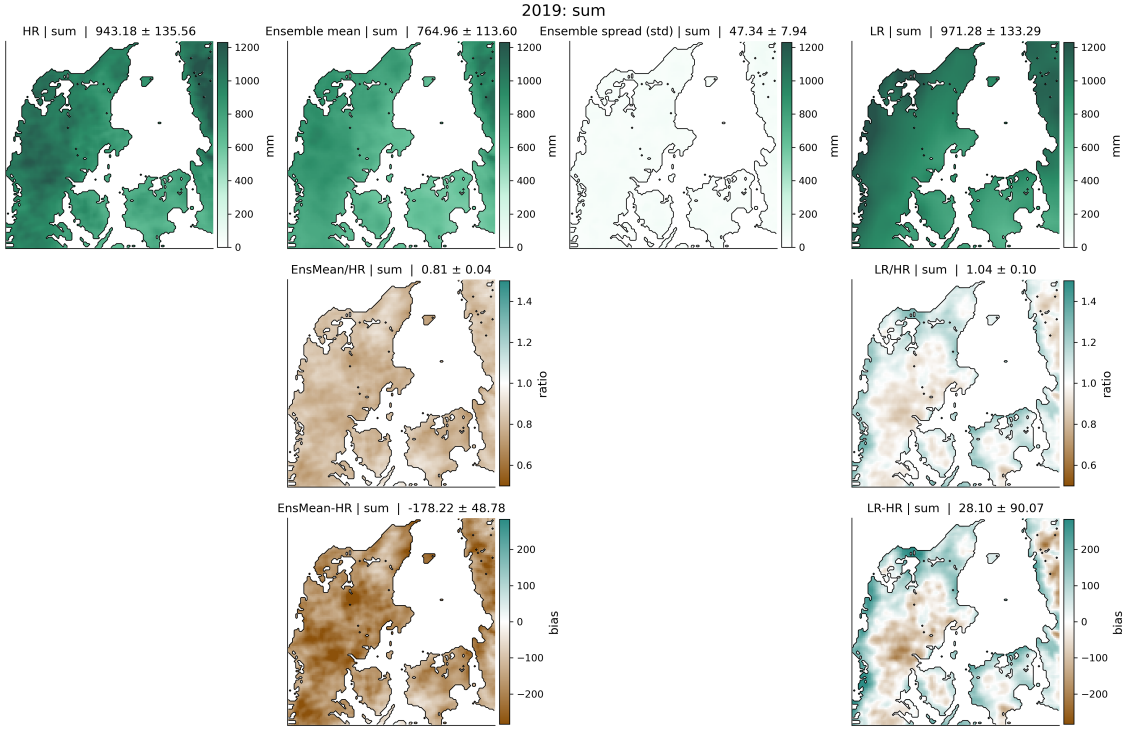

Figure 7. *Spatial annual sum of precipitation (2019).* Maps of DANRA, ensemble mean, and ERA5 (bilinearly upsampled) annual precipitation sum. The EDM reproduces the large-scale gradients and regional maxima, seen in the west-east precipitation contrast across Denmark, but exhibits a systematic dry-bias. ERA5 on the other hand shows a generally smoothed field with a wet-bias.

hand, we see a smoothing of the precipitation patterns, with more large-scale coherence. Notably, the direction of this effect is opposite to an obvious interpretation of stochastic scaling: in our EDM implementation, reducing σ^* increases high-frequency variance by perturbing the late denoising trajectory, whereas increasing σ^* suppresses small-scale structure.

The right panel in Figure 10 (Scale-aware correlation, Mesoscale PSD slope, and ensemble CRPS) also backs these interpretations up, by illustrating how σ^* influences spectral power, correlation between large-scale condition (LR) and generated samples (GEN), and ensemble spread (CRPS). Within a moderate range around $\sigma^* = 1$, changes in σ^* systematically affect spatial variability, large-scale coherence, and ensemble spread. However, pushing σ^* too far (e.g. $\sigma^* \simeq 0.8$) produces excess power at high wavenumbers and visually noisy artifacts. This is an indication that the samples have been moved outside the physical manifold representing the training data (red hatched zone above DANRA PSD in the left panel of Figure 10). We therefore treat these extreme deviations as diagnostic edge cases rather than physically meaningful outcomes. For the metrics analysis of the σ^*-sweep in Figure 10 (right), the error bars denote the standard error of the mean (SEM) across evaluation dates, rather than the inter-date standard deviation (STD). The STD is large since the daily metrics exhibit strong seasonal and synoptic variability, which can visually dominate small but systematic mean differences across σ^*. SEM therefore more directly communicates the

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

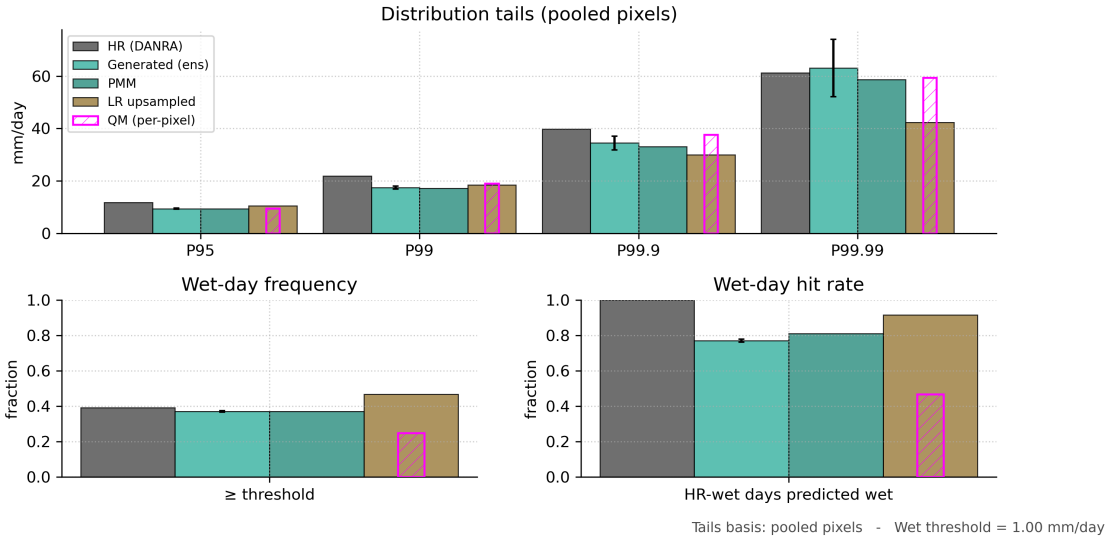

Figure 8. *Extreme values of pooled pixel distributions (P95, P99, P99.9, P99.99) and wet-day frequency and hit-rates*. **Top**: Bars showing the 95th, 99th, 99.9th and 99.99th percentile extreme values of the pooled distributions' tails. **Lower-left**: Wet-day frequency, i.e. percentage of rainy days (rain corresponding to more than 1 mm day^{-1}). **Lower-right**: Wet-day hit rate, corresponding to the ratio of correctly estimated rainy days (i.e. true positives). Bars show ensemble-based, probability matched mean (PMM), LR upsampled and quantile-mapped (QM) estimates of tail percentiles and wet-day statistics. Error bars indicate variability across ensemble members.

uncertainty in the estimated mean response to changing σ^*. We also provide the corresponding STD-based version, reflecting the full day-to-day spread, in the Supplement Figure S5 for transparency.

5 Discussion

5.1 Advantages of our EDM approach

As our results demonstrate, the EDM approach grants us a number of advantages compared to simpler, conventional downscaling methods. Compared to both a bilinear upsampled LR and the quantile-mapped LR, our model improves spatial structure, distributional realism, and wet-day statistics, while still underestimating some aspects of climatology and extremes. In our model, physical fidelity is achieved through a combination of tuned stochastic diffusion dynamics, seasonal and geographical informed soft constraints, and large-scale LR conditioning. These designs together preserve large-scale atmospheric structure while allowing realistic fine-scale variability to emerge, as evident in the PSD and ISS analyses (Figures 5 and S11). The stochastic nature of the EDM provides us with probabilistic realism, emphasised by sharp and reliable ensembles, as is supported by our visual examples with CRPS (Figure 4, and Figure S10 and PIT histogram in Supplementary Information,

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

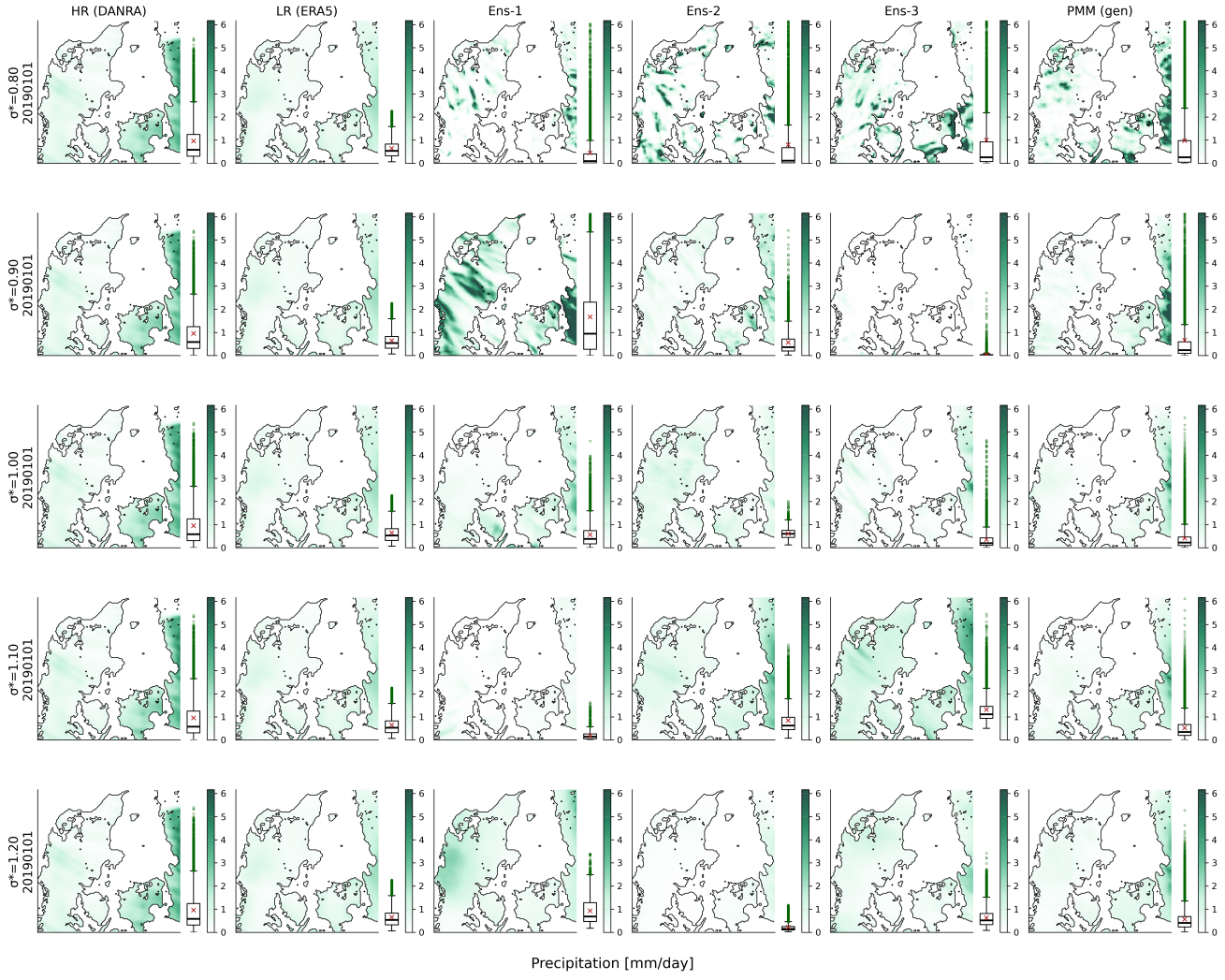

Figure 9. *Visual effect of the σ^* controllability parameter on ensemble samples*. Generated example fields for different σ^* values. High-resolution (HR) DANRA and low-resolution (LR) ERA5 are fixed and the same input for all rows. Higher σ^* yields smoother, large-scale coherent precipitation patterns, whereas lower σ^* increases fine-scale variability and ensemble spread.

Figure S7). This shows that the model clearly captures relevant physical variance rather than pure random noise. Together, ISS close to DANRA at 5-10 km and matched PSD slopes indicate realistic multiscale variability without spectral overfitting.

Another contribution of this work is the explicit demonstration of a generative model's stochastic scale-sensitivity through the inference-time parameter σ^*. This parameter allows for a direct way to explore the trade-off between large-scale coherence and fine-scale stochastic structure. In most other generative downscalers, this behaviour is kept implicit in sampler choices, but our work makes it explicit, interpretable and user-tunable.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

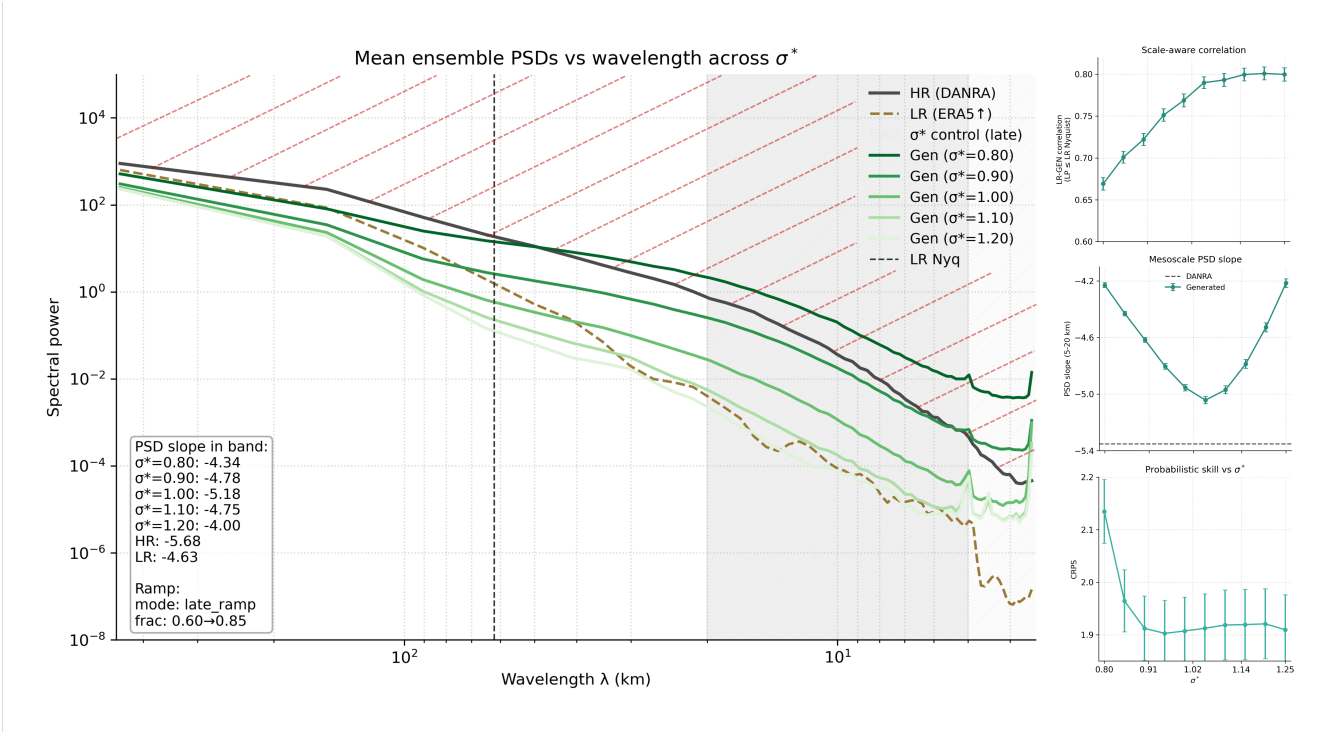

Figure 10. *Scale-aware inference analysis*. **Left**: Full isotropic PSD curves for selected σ^* values, showcasing how varying σ^* systematically shifts spectral energy at higher spatial frequencies, while preserving large-scale structure. Red hatched area correspond to spectral power outside the range of what is exhibited by the HR target DANRA. **Right**: Relationships between σ^* and (top) large-scale LR-Generated correlation, (middle) mesoscale PSD slope, and (bottom) ensemble CRPS. Error bars show Standard Error (SEM) across evaluation dates.

Our model helps bridge the gap between deterministic statistical models (often too smooth) and unconstrained generative ones (often physically inconsistent). Further, unlike other ML approaches like GANs and U-Nets, our EDM with scale-aware inference offers a predictable, monotonic control and avoids mode collapse (Dhariwal and Nichol, 2021; Accarino et al., 2021).

5.2 Interpretation, limitations, and implications

Our extensive evaluation indicates that good spectral-scale performance is not necessarily accompanied by strong performance on, e.g., extreme metrics or climatology. Achieving both typically requires careful tuning and validation focused on the downstream use case. Our current model exhibits a systematic dry bias, visible both in the climatology and in moderate-to-high precipitation intensities (P95-P99.99, Figures 7 and 8). This indicates that the model is conservative across much of the intensity range, not only during extreme events. Several mechanisms likely contribute to this behaviour: (i) the logarithmic Z-score transformation can have a distorting effect on the target distribution, (ii) the EDM's own log-normal denoising dynamics might favour smoother solutions at low noise levels, which can suppress localised high-intensity structure, and (iii) the RainGate and BCE reweighted losses have pixel weights ≥ 1, so we upweight where RainGate thinks "wet", which tends to encourage sharp

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

wet/dry decisions and focus on the most confidently wet pixels. This gives nice rain-structures, but might be too selective with where the rain happens, resulting in a dry bias in the main part of the distribution, but the "very wet" (confident) spots are still getting corrected. Potential mitigations for these contributions include: (i) testing alternative EDM schedules or customised low-σ ramps, (ii) experimenting with alternative scalings, such as Box-Cox, asinh, or mixture-based normalisation, and (iii) exploring modified noise priors or targeted high-σ sampling during training. These avenues are beyond scope of the present study, but they will be valuable for future development.

Although our model does create spatially physically grounded downscaling, it only downscales daily snapshots independently, but large-scale atmospheric dynamics are driven on more than just daily timescales. This means that any temporal persistence in our model is learned implicitly through the LR condition. We do get a reasonable autocorrelation due to the model being conditioned on output from the dynamical ERA5, Supplement Figure S13, and the same goes for wet- and dry-spell lengths, Supplement Figure S14. Because LR forcing carries persistence, explicit temporal modelling may be unnecessary for some use-cases. Explicit spatio-temporal diffusion or sequential conditioning is a logical next step, particularly for applications sensitive to multi-day persistence, such as droughts or compound events, but the necessity of this remains to be confirmed.

All experiments are performed on daily precipitation, and sub-daily convective variability, diurnal cycles and short-duration extremes (cloudbursts) are not directly resolved in the temporal resolution that we use the DANRA reanalysis in - but extreme events still show up to some extent as daily accumulated high-intensity events. While the EDM framework in principle is well-suited for sub-daily generative modelling, such an extension requires high-frequency HR and LR training data and careful treatment of temporal correlation. Investigating how our model generalises to sub-daily or even convection-resolving data is an important focus of future work.

Another physical limitation that we are aware of is the oversimplification of our atmospheric LR conditioning. For this study, we have only used LR ERA5 precipitation as the dynamic atmospheric driver. Other drivers (temperature, Convective Available Potential Energy (CAPE), equivalent potential temperature (θ_e), water vapour fluxes, and geopotential heights) are included in our model framework and design, but not fully explored in training, as the focus of this paper is more on architecture and design, along with controllability and interpretability of an EDM-based model. A natural extension is multi-variable conditioning using additional ERA5 predictors which the framework already supports and which is expected to strengthen physical coupling. Additionally, extending the LR conditioning domain (e.g., to the North Atlantic) may further stabilise large-scale circulation constraints and move the framework toward a full regional emulator.

As with any data-driven model, we are constrained by the availability and characteristics of our training data. In this case, we confine the training and evaluation to the Danish domain which is coastal and mid-latitude, and evaluate the model exclusively using high-resolution reanalysis data. This leaves questions of how the model will generalise to other regions or climates, e.g., to tropical convection or strongly orography-dominated zones, which are not addressed in the present work. This limitation due to the scope of training data also bleeds in to our evaluation. We only evaluate the model on two years of data, close in time to the training data split which means that we do not probe the model's ability to generalise under different or non-stationary climate conditions. These aspects are important considerations for future applications, especially in the context of a changing climate.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

Although the modelling framework itself is agnostic and not architecturally tied to Denmark, it is not directly a plug-and-play model, and trained weights cannot be directly transferred to new climatic or geographic setting swithout retraining. Application to a new climatic or geographic region does require generation of a dedicated training dataset and a retraining of the model, which can both be done by following the preprocessing and training procedures from the code base. Thus our implementation should be understood as a domain-trained generative emulator and not a universal precipitation downscaler. Its portability lies in the framework, pipelines, and training protocol, and not in the trained weights themselves.

At the same time, the methodological choices in this study are motivated by a broader perspective than reanalysis downscaling alone. The model is designed to learn physically plausible mappings between large-scale atmospheric states and high-resolution locally realised fields, rather than to reproduce point-scale observations. As a consequence, we do not evaluate against point measurements and any systematic discrepancies between DANRA and in-situ observation will also be reflected in the generated fields, though they might fall within the spread of the generated ensembles.

Scale-transfer experiments across resolution gaps (e.g. CORDEX → DANRA) and bias-aware transfer strategies offer a promising route to testing generalisation across climates and forcing models. At present, however, high-resolution reanalysis products like DANRA constitute the only viable training target for this approach. Applying the methodology directly to climate projections introduces additional challenges related to distributional shifts between historical and future climate as well as to systematic biases in the driving models. Promising directions include scale-aware transfer strategies, bias-aware conditioning, and hybrid pipelines that bridge multiple intermediate resolutions (Hess et al., 2025). In this context, our work can be seen as a foundational step toward flexible, physically informed generative downscaling frameworks capable of supporting future climate impact studies - for example through multi-stage pipelines linking global climate models to regional reanalyses and kilometre-scale products.

Finally, the observed sensitivity to training seeds highlights the importance of explicitly accounting for stochasticity when evaluating generative models of any kind. We especially observed that seed variability is modest for bulk-distribution metrics (PSD, CRPS, ISS), but can materially effect tail and accumulation diagnostics, metrics highly relevant for decision-making. These results clearly highlight the importance of mullti-seed evaluation when assessing e.g. extremes. Nevertheless, training-seed dependence and architectural interactions are rarely discussed in the downscaling and generative modelling literature, where results are often reported for a single training instance without quantifying seed variability. EMA is commonly used to improve numerical stability and reduce sensitivity to training fluctuations and optimisation noise in generative models (Yazıcı et al., 2018). In our experiments, however, its effect was configuration- and seed-dependent: EMA modified seed-to-seed variability and, depending on the conditioning setup, could either modestly smooth solutions or improve agreement in tail and time-aggregated decision-critical statistics. Our results demonstrate that both seed selection and architectural choices can materially influence model behaviour, underscoring the need for standardised reporting practices, and particularly reporting mean performance together with seed spread, when comparing generative models - also for geophysical applications.

Despite these limitations, our results demonstrate that diffusion-based generative models can be constructed in a way that is both physically grounded and practically controllable. By anchoring stochastic generation to large-scale atmospheric forcing, combining soft physical constraints with probabilistic evaluation, and exposing scale-sensitivity, we show that generative

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

downscaling can move beyond purely aesthetic realism toward scientifically interpretable and application-relevant behaviour. While further work is required to improve climatological fidelity, extremes, and transferability, the framework introduced here provides a flexible foundation for future multi-stage climate downscaling pipelines, particularly in contexts where uncertainty, variability, and physical plausibility are as important as pointwise accuracy.

6 Conclusions

We presented a physically guided diffusion model for controllable downscaling of daily precipitation from ERA5 to DANRA resolution over Denmark, producing stochastic high-resolution ensembles conditioned on large-scale atmospheric forcing. The final model combines seasonal FiLM conditioning, geographic priors with optional SDF-weighted loss, and an auxiliary RainGate task to guide precipitation occurrence. To evaluate this final model, we propose an extensive multi-perspective evaluation framework, that balances stochastic performance, climatological fidelity, and spatial and scale-dependent realism.

Across this evaluation framework on unseen years, the model produced somewhat-calibrated ensembles (CRPS, PIT), maintained multi-scale structure (ISS, PSD), but struggles with reproducing climatology (dry-bias) and tail statistics, otherwise key for impact applications. Along with the ability to generate high-resolution precipitation ensembles, our framework explicitly demonstrates the scale sensitivity of diffusion-based downscaling through σ^*, which reveals a predictable but limited area of physically supported stochastic variability.

Together, these insights show that diffusion-based downscaling is neither the final solution to climate downscaling nor only a visually appealing alternative to traditional methods. Instead, it is a powerful but delicate framework, where performance, metrics, strengths and challenges are intricately and non-linearly intertwined. Strengths like stochastic realism, multiscale structure, ensemble generation, and controllability are tightly coupled to persistent challenges in climatological fidelity, extremes and robustness. Through our systematic evaluation and controlled inference, we show that it is possible to expose these trade-offs, and our work reframes generative downscaling not as a purely performance-driven task, but as a model class that must be carefully investigated, constrained and interpreted.

...

. Code and data availability

DANRA data (v0.5) are publicly available via Zenodo (Yang et al., 2025a), https://doi.org/10.5281/zenodo.17294180, and described in detail in Yang et al. (2025b). ERA5 data are available from the Copernicus Climate Data Store (CDS) (Hersbach et al., 2020). The ERA5 download and preprocessing pipeline used in this work is included in the CEDDAR repository (Quistgaard, 2026b).

The CEDDAR v1.0.2 source code, is publicly available at GitHub and permanently archived on Zenodo (Quistgaard, 2026b) at https://doi.org/10.5281/zenodo.18643186.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

A small pre-processed example dataset (Data_DiffMod_small) for reproducibility testing as well as the trained baseline model (B0) and final model (B1_GSDF_RGBCE) are archived Zenodo (Quistgaard, 2026a), https://doi.org/10.5281/zenodo.18643307.

All scripts required to reproduce the reduced and small test experiments are included in the archived repository. Reproducing results for new regions requires generation of region-specific training datasets following the preprocessing procedure described in Section 3.1 and in Supplement Section S2

. *Author contribution*.

T.Q. conceived the study, developed the methodology, implemented the model code, performed the experiments and analyses, and prepared the original manuscript draft. All authors contributed to conceptual development, scientific discussion, and manuscript review and editing. P.L.L. contributed to conceptual development and methodological design. P.L.L. and S.S. supervised the research. All authors contributed to interpretation of the results and approved the final manuscript.

. *Competing interests.*

The authors declare no competing interests.

. *Acknowledgements.*

This research is part of the PEACE project, and was funded by the Independent Research Fund Denmark (DFF), grant no. 1127-00243B, and the Interdisciplinary Centre for Climate Change, iClimate, at Aarhus University. We gratefully acknowledge the Danish e-Infrastructure Centre (DeIC), Denmark, for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through DeIC, Denmark, which was instrumental for development and conducting experiments for this study. The authors would like to thank and acknowledge the collaboration of the Danish Meteorological Institute (DMI), and specifically Xiaohua Yang, who shared DANRA data before publicly available.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

References

Accarino, G., Chiarelli, M., Immorlano, F., Aloisi, V., Gatto, A., and Aloisio, G.: MSG-GAN-SD: A Multi-Scale Gradients GAN for Statistical Downscaling of 2-Meter Temperature over the EURO-CORDEX Domain, AI (Switzerland), 2, 600–620, https://doi.org/10.3390/ai2040036, 2021.

Baño-Medina, J., Manzanas, R., and Gutiérrez, J. M.: Configuration and Intercomparison of Deep Learning Neural Models for Statistical Downscaling, Geoscientific Model Development, 13, 2109–2124, https://doi.org/10.5194/gmd-13-2109-2020, 2020.

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., and Gentine, P.: Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems, Physical Review Letters, 126, 098 302, https://doi.org/10.1103/PhysRevLett.126.098302, 2021.

Denager, T., Christiansen, J. R., Schneider, R. J. M., Langen, P., Quistgaard, T., and Stisen, S.: Combined Water Table and Temperature Dynamics Control CO_2 Emission Estimates from Drained Peatlands under Rewetting and Climate Change Scenarios, Biogeosciences, 23, 441–462, https://doi.org/10.5194/bg-23-441-2026, 2026.

Dhariwal, P. and Nichol, A.: Diffusion Models Beat Gans on Image Synthesis, in: Advances in Neural Information Processing Systems, edited by Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., vol. 34, pp. 8780–8794, Curran Associates, Inc., 2021.

Donat, M. G., Alexander, L. V., Yang, H., Durre, I., Vose, R., and Caesar, J.: Global Land-Based Datasets for Monitoring Climatic Extremes, Bulletin of the American Meteorological Society (BAMS), 94, 997–1006, https://doi.org/10.1175/BAMS-D-12-00109.1, 2013.

Durall, R., Keuper, M., and Keuper, J.: Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7887–7896, IEEE, Seattle, WA, USA, ISBN 978-1-7281-7168-5, https://doi.org/10.1109/CVPR42600.2020.00791, 2020.

Gires, A., Tchiguirinskaia, I., Schertzer, D., and Lovejoy, S.: Influence of the Zero-Rainfall on the Assessment of the Multifractal Parameters, Advances in Water Resources, 45, 13–25, https://doi.org/10.1016/j.advwatres.2012.03.026, 2012.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Monthly Weather Review, 133, 1098–1118, https://doi.org/10.1175/MWR2904.1, 2005.

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., and Palmer, T. N.: A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts, Journal of Advances in Modeling Earth Systems, 14, e2022MS003 120, https://doi.org/10.1029/2022MS003120, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 Global Reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hess, P., Aich, M., Pan, B., and Boers, N.: Fast, Scale-Adaptive and Uncertainty-Aware Downscaling of Earth System Model Fields with Generative Machine Learning, Nature Machine Intelligence, 7, 363–373, https://doi.org/10.1038/s42256-025-00980-5, 2025.

Ho, J., Jain, A., and Abbeel, P.: Denoising Diffusion Probabilistic Models, in: Advances in Neural Information Processing Systems, edited by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

IPCC: Climate Change 2022: Impacts, Adaptation and Vulnerability. Summary for Policymakers, Intergovernmental Panel on Climate Change (AR6 WGII), 2022.

Karpatne, A., Atluri, G., Faghmous, J. H., et al.: Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data, IEEE Transactions on Knowledge and Data Engineering, 29, 2318–2331, https://doi.org/10.1109/TKDE.2017.2720168, 2017.

Karras, T., Aittala, M., Laine, S., and Aila, T.: Elucidating the Design Space of Diffusion-Based Generative Models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, Nips '22, Curran Associates Inc., Red Hook, NY, USA, ISBN 978-1-7138-7108-8, 2022.

Kendon, E. J., Roberts, N. M., Fowler, H. J., Roberts, M. J., Chan, S. C., and Senior, C. A.: Heavier Summer Downpours with Climate Change Revealed by Weather Forecast Resolution Model, Nature Climate Change, 4, 570–576, https://doi.org/10.1038/nclimate2258, 2014.

Li, W., Chen, J., Li, L., Chen, H., Liu, B., Xu, C.-Y., and Li, X.: Evaluation and Bias Correction of S2S Precipitation for Hydrological Extremes, Journal of Hydrometeorology, 20, 1887–1906, https://doi.org/10.1175/JHM-D-19-0042.1, 2019.

Lovejoy, S. and Schertzer, D.: The Weather and Climate: Emergent Laws and Multifractal Cascades, Cambridge University Press, https://doi.org/10.1017/CBO9781139093811, 2013.

Maraun, D.: Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue, Journal of Climate, 26, 2137–2143, https://doi.org/10.1175/JCLI-D-12-00821.1, 2013.

Maraun, D., Wetterhall, F., Ireson, A. M., et al.: Precipitation Downscaling under Climate Change: Recent Developments to Bridge the Gap between Dynamical Models and the End User, Reviews of Geophysics, 48, https://doi.org/10.1029/2009RG000314, 2010.

Meresa, H., Tischbein, B., and Mekonnen, T.: Climate Change Impact on Extreme Precipitation and Peak Flood Magnitude and Frequency: Observations from CMIP6 and Hydrological Models, Natural Hazards, 111, 2649–2679, https://doi.org/10.1007/s11069-021-05152-3, 2022.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C.: FiLM: Visual Reasoning with a General Conditioning Layer, in: AAAI, 2018.

Quistgaard, T.: CEDDAR Reproducibility Example Dataset (Data_DiffMod_small) and Trained Models (B0 and B1_GSDF_RGBCE), https://doi.org/10.5281/ZENODO.18643307, 2026a.

Quistgaard, T.: CEDDAR v1.0.2 - Controllable Ensemble Diffusion Downscaling for Atmospheric Rainfall (Code), Zenodo, https://doi.org/10.5281/ZENODO.18643186, 2026b.

Rampal, N., Hobeichi, S., Gibson, P. B., Baño-Medina, J., Abramowitz, G., Beucler, T., González-Abad, J., Chapman, W., Harder, P., and Gutiérrez, J. M.: Enhancing Regional Climate Downscaling through Advances in Machine Learning, Artificial Intelligence for the Earth Systems, 3, 230 066, https://doi.org/10.1175/AIES-D-23-0066.1, 2024.

Rasp, S. and Thuerey, N.: Data-Driven Medium-Range Weather Prediction with a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench., Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2020MS002405, 2021.

Roe, G. H.: Orographic Precipitation, Annual Review of Earth and Planetary Sciences, 33, 645–671, https://doi.org/10.1146/annurev.earth.33.092203.122541, 2005.

Saha, A. and Ravela, S.: Statistical-Physical Adversarial Learning From Data and Models for Downscaling Rainfall Extremes, Journal of Advances in Modeling Earth Systems, 16, e2023MS003 860, https://doi.org/10.1029/2023MS003860, 2024.

Seneviratne, S. I., Nicholls, N., et al.: Changes in Climate Extremes and Their Impacts on the Natural Physical Environment, Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX), IPCC, pp. 109–230, 2012.

https://doi.org/10.5194/egusphere-2026-1139
Preprint. Discussion started: 9 March 2026

Song, Y., Sohl-Dickstein, J. N., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations, in: International Conference on Learning Representations (ICLR), 2021.

Stevens, B. and Bony, S.: What Are Climate Models Missing?, Science, 340, 1053–1054, https://doi.org/10.1126/science.1237554, 2013.

Themeßl, M. J., Gobiet, A., and Leuprecht, A.: Empirical-Statistical Downscaling and Error Correction of Daily Precipitation from Regional Climate Models, International Journal of Climatology, 31, 1530–1544, https://doi.org/10.1002/joc.2168, 2011.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., and Ganguly, A. R.: DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Kdd '17, pp. 1663–1672, Association for Computing Machinery, New York, NY, USA, ISBN 978-1-4503-4887-4, https://doi.org/10.1145/3097983.3098004, 2017.

Vogel, R., Bony, S., and Stevens, B.: Estimating the Shallow Convective Mass Flux from the Subcloud-Layer Mass Budget, Journal of the Atmospheric Sciences, 77, 1559–1574, https://doi.org/10.1175/JAS-D-19-0135.1, 2020.

Watt, R. A. and Mansfield, L. A.: Generative Diffusion-based Downscaling for Climate, https://doi.org/10.48550/ARXIV.2404.17752, 2024.

Wilby, R. L., Charles, S. P., Zorita, E., Timbal, B., Whetton, P., and Mearns, L. O.: Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods, IPCC Task Group on Scenarios for Climate Impact Assessment, 2004.

Yang, X., Peralta, C., Amstrup, B., Hintz, K. S., Thorsen, S. B., Denby, L., Christiansen, S. K., Schulz, H., Pelt, S., and Schreiner, M.: DANRA, https://doi.org/10.5281/ZENODO.17294180, 2025a.

Yang, X., Peralta, C., Amstrup, B., Hintz, K. S., Thorsen, S. B., Denby, L., Christiansen, S. K., Schulz, H., Pelt, S., and Schreiner, M.: DANRA: The Kilometer-Scale Danish Regional Atmospheric Reanalysis, https://doi.org/10.5194/essd-2025-610, 2025b.

Yazıcı, Y., Foo, C.-S., Winkler, S., Yap, K.-H., Piliouras, G., and Chandrasekhar, V.: The Unusual Effectiveness of Averaging in GAN Training, https://doi.org/10.48550/ARXIV.1806.04498, 2018.

Zarr Developers, Z. D.: Zarr: Storage of Large N-dimensional Typed Arrays, https://zarr.dev/, 2026.

Zhou, F., Zhao, T., Nguyen, L. V., and Yao, Z.: A Parallel Gumbel-Softmax VAE Framework with Performance-Based Tuning, in: Frontiers in Artificial Intelligence and Applications, edited by Endriss, U., Melo, F. S., Bach, K., Bugarín-Diz, A., Alonso-Moral, J. M., Barro, S., and Heintz, F., IOS Press, ISBN 978-1-64368-548-9, https://doi.org/10.3233/FAIA240689, 2024.