RAINATLAS: A MULTI-CONTINENTAL DATASET FOR PRECIPITATION DOWNSCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Extreme rainfall events are increasing in intensity and frequency as climate change accelerates. While kilometer-scale precipitation forecasts are critical for supporting local decision-making, the limited availability of high-resolution precipitation observations hinders their accuracy, especially in under-resourced regions. Machine learning models are widely used to downscale precipitation data to km-scale, but their application to unseen geographies presents challenges. First, processing raw high-resolution precipitation datasets across regions requires significant engineering and domain expertise. Second, generalization across regions remains difficult. To help overcome these barriers, we release `RainAtlas`, a large-scale, ML-ready and multi-continental dataset for precipitation downscaling. Covering three continents, `RainAtlas` harmonizes heterogeneous hourly km-scale observations to a common 2-km grid. Each regional partition contains around 210,000 aligned low- and high-resolution precipitation pairs, respectively from ERA5 reanalysis and direct observations. We benchmark state-of-the-art ML-based downscaling models across `RainAtlas` using a wide range of metrics. Our evaluation reveals substantial variance in out-of-domain generalization depending on the training regions. This underscores the need for cross-regional, multi-source km-scale evaluation, establishing `RainAtlas` as a well-positioned benchmark for precipitation downscaling research.

1 INTRODUCTION

Extreme precipitation is set to increase in both intensity and frequency as a consequence of climate change (Feng et al., 2023), causing severe socio-economic consequences (Špitalar et al., 2014; Liang, 2022). Accurately capturing extreme precipitation events requires kilometer-scale resolution to resolve fine-scale processes such as convection and local orographic interactions (Prein et al., 2015). In practice, many users leverage coarse reanalysis data products, such as ERA5, which reconstruct past weather by blending historical observations with physical models, but these severely underestimate extreme rainfall (Chen et al., 2024), and satellite-based data products still show substantial bias (Hooker et al., 2026). Reliable and accurate km-scale observations remain globally scarce and unequally distributed: less than 15 % of the world has access to accurate rain-gauge data Su et al. (2026), while ground-based radars suffer from continent-scale coverage gaps, as shown in Figure 1. Standard physics-based climate models could theoretically bridge this gap, but their prohibitive computational cost limits widespread application (Neumann et al., 2019).

Using machine learning (ML) to *downscale* precipitation data – that is, to enhance the resolution of coarse variables (spatially, temporally, or both) – has the potential to offer an efficient alternative, leveraging advances such as score-based models trained on historical observations to enhance spatial resolution (Rampal et al., 2026). Nevertheless, we identify two primary limitations hindering the practical application of ML for precipitation downscaling: (1) data processing for ML compatibility, and (2) limited geographical generalization. First, rain gauges and radars are subject to systemic bias that need to be mitigated with hydrological merging methods (McKee & Binns, 2016). Additionally, radar observations require expert knowledge to be processed, differ in quality, spatial and temporal resolution, projection, format, and quality-control procedures (Heistermann et al., 2013; Thorndahl et al., 2017). These differences create major interoperability barriers and require substantial preprocessing, including re-gridding, temporal aggregation, subsampling, and harmonization (Heistermann et al., 2013). When precipitation archives are publicly available, transforming raw multi-terabyte

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

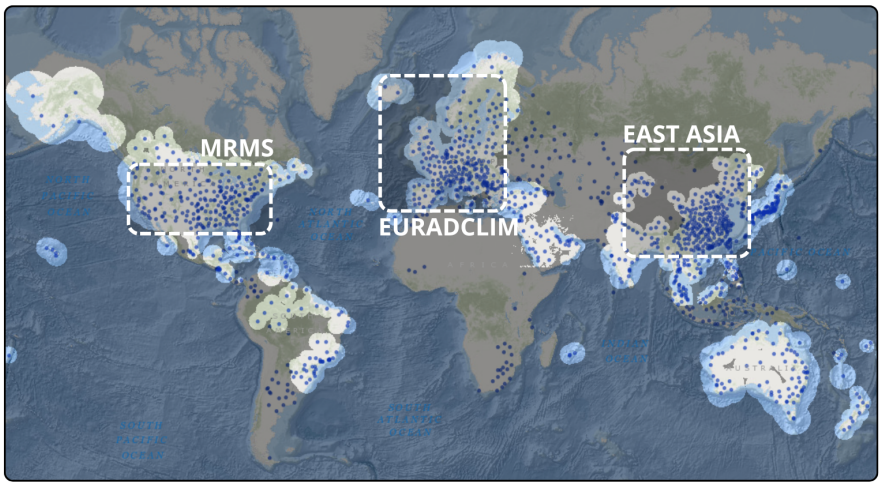


Figure 1: Global distribution of ground-based radar stations **(blue dots)** and their available coverage **(white transparent circles)**. Some regions have excellent coverage (USA, Europe, East-Asia), while other densely populated regions have little to no precipitation observations available (Africa, South America, Central Asia). `RainAtlas` consists of the three multi-source precipitation products highlighted in the figure, covering three Northern Hemisphere regions.

datasets into ML-ready datasets demands substantial computational resources and domain expertise. Second, the resulting scarcity of high-quality and ML-ready high-resolution data makes geographical generalization a challenge for ML-based downscaling models, which are rarely trained nor evaluated on multiple regions. Most approaches do not explicitly model underlying physics, implicitly assuming unchanged statistical relationships between low- and high-resolution precipitation across geographical regions. This assumption is frequently violated in practice, as precipitation patterns and extremes are governed by processes that vary substantially with local orography and regional climate phenomena. In fact, previous works report significant drops in performance when evaluating models on out-of-domain (Harder et al., 2026), greatly limiting their potential for downstream applications.

These challenges highlight the need for harmonized, ML-ready and multi-region datasets to support the training and evaluation of km-scale precipitation downscaling models, encourage systematic benchmarking, and drive research efforts on geographical generalization. In response, we introduce `RainAtlas` [1] [2]: a multi-source dataset and benchmark for km-scale ML-based precipitation downscaling and geographical generalization. Our dataset spans the contiguous United States, Europe, and East Asia, and include nine years of hourly observations in an ML-ready format, harmonizing three heterogeneous precipitation observation datasets to 2 km equal-area grids. Our contributions are the following:

- We release, to our knowledge, the first harmonized, multi-continental dataset for km-scale precipitation downscaling and geographical generalization, including over 1.2M samples.
- We introduce a modular framework for generating ML-ready precipitation datasets with flexible spatial and temporal subsampling configurations.
- We benchmark ML-based downscaling models for geographical generalization.

2 Related work

ML-based precipitation downscaling Generative models account for the majority of recent work on ML-based precipitation downscaling. Price & Rasp (2022) applied conditional GANs to radar observations to downscale coarse precipitation fields to km-scale. Score-based and diffusion models

[1] All partitions of `RainAtlas` are made publicly available on Hugging-Face.
[2] We release our code on an anonymized repository for review.

have also gained significant traction for km-scale precipitation downscaling (Liu et al., 2025; Dai & Ushijima-Mwesigwa, 2025; Yi et al., 2026). In particular, Mardani et al. (2025) proposed a two-step *corrective* diffusion model that demonstrated strong skill in reconstructing high spatial frequencies and precipitation extremes. More recently, Hess et al. (2025) trained a consistency model for unpaired km-scale downscaling, achieving similar performance as standard diffusion, at a fraction of its computational cost. Finally, Keisler et al. (2026) introduced SerpentFlow, an unpaired framework for climate downscaling using flow matching. Despite these advances, almost all exclusively focus on single geographical regions. Only a few studies evaluate models across multiple domains (Glawion et al., 2025; Vicens-Miquel et al., 2025), while none investigate multi-region training.

Benchmarks for precipitation downscaling Benchmarks with standardized datasets and evaluation protocols offer an avenue to address these limitations, by enabling systematic model comparison and supporting methodological progress in ML for Earth system sciences. RainBench Schroeder de Witt et al. (2021) proposed a global, multi-modal benchmark for precipitation forecasting at a resolution around 30 km. RainNet Chen et al. (2022) provided a fine-scale precipitation downscaling dataset, with a target resolution of 4 km, but is limited to the U.S. East Coast. RainShift Harder et al. (2026) introduced a global benchmark for evaluating geographical transferability in precipitation downscaling. However, it considers a comparatively coarse target resolution of around 10 km, and uses satellite observations as ground truth. Therefore, geographical generalization at km-scale resolution remains largely unexplored for ML-based precipitation downscaling.

3 RAINATLAS: KILOMETER-SCALE PRECIPITATION DOWNSCALING ACROSS CONTINENTS

`RainAtlas` is a multi-continental dataset and benchmark designed for km-scale precipitation downscaling and geographical generalization across diverse regions. It harmonizes high-resolution observations from heterogeneous sources with ERA5 reanalysis and static covariates over large continental domains, providing aligned low- and high-resolution pairs ready for ML-based downscaling. Train, validation and test datasets are subsampled from the harmonized domain-wide datasets along their spatial and temporal dimensions. Our framework is built with a modular structure to allow for the addition of new observational datasets and subsampling strategies.

3.1 HIGH-RESOLUTION DATASETS

United States We use the Multi-Radar/Multi-Sensor System (MRMS) precipitation monitoring system Smith et al. (2016). MRMS quantitative precipitation estimation products automatically integrate data streams from a network of over 200 ground-based radars, satellites, numerical weather prediction (NWP) models, and rain gauges. MRMS has a native resolution of $0.01^\circ \times 0.01^\circ$ on a regular latitude-longitude grid. Hourly observations from the most recent operational version are available from late 2020 onward; we include data from 2021 to 2025.

Europe We use the EURADCLIM dataset Overeem et al. (2023), also referred to as EUR. EURADCLIM provides precipitation observations on a 2 km grid, integrating observations from 138 ground-based radars and more than 7,700 rain gauges. Around 78 % of Europe is covered, with data missing only for Italy. EURADCLIM contains a decade of precipitation observations, from 2013 to 2023. We include a subset of this period (2017 to 2022) in `RainAtlas`.

East Asia We use a km-scale dataset that covers most of China and some neighboring regions, referred to as EA or EASTASIA (Xia & Wang, 2025a). EA integrates ground-based radar observations, IMERG satellite data, and around 2,700 rain gauge measurements using machine learning (Xia & Wang, 2025b). The resulting dataset has a native $0.01^\circ \times 0.01^\circ$ regular latitude-longitude grid, and covers 2017 to 2022.

Table 1: ERA5 and static input data variables.

Variable	Description	Unit	Level
tp	Total precipitation	mm	surface
cp	Convective precipitation	mm	surface
cape	Convective potential energy	J· kg^{-1}	surface
sp	Surface pressure	Pa	surface
tisr	Top-of-the-atmosphere incident solar radiation	J · m^{-2}	surface
tcw	Total column water	kg · m^{-2}	total column
tclw	Total column cloud liquid water	kg · m^{-2}	total column
u	Eastward wind velocity	m · s^{-1}	700 hPa
v	Northward wind velocity	m · s^{-1}	700 hPa
lsm	Land-sea mask	(0, 1)	surface
$elev_{mean}$	Elevation mean	m	surface
$elev_{std}$	Elevation standard deviation	m	surface

3.2 Reanalysis and static covariates

ML-based precipitation downscaling aims to provide high-resolution precipitation forecasts or projections globally, especially for data-sparse regions. This requires model inputs, such as coarse forecast or reanalysis fields and auxiliary covariates, to be globally available.

As coarse inputs, we use the ERA5 atmospheric reanalysis from the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al., 2020), which assimilates hourly numerical weather predictions (NWP) with observations at a global $0.25° \times 0.25°$ regular latitude-longitude resolution. Following previous work and domain knowledge (Harder et al., 2026; Hewson & Pillosu, 2021), we select nine variables known to have an impact on high-resolution precipitation, which we present in Table 1.

As static covariates, we use the land-sea mask derived from NASA's MODIS elevation product at its native 250 m resolution (Carroll et al., 2024). Coastal regions are known for having complex precipitation patterns, due to the frontier between oceanic and continental atmospheric processes. Elevation strongly influences local climates, so we also include the average, as well as the standard deviation of elevation from the 1 km resolution GMTED2010 dataset (Danielson & Gesch, 2011).

3.3 Re-projection to equal-area kilometer grids

Regular latitude-longitude grids introduce distortions in grid cell area. For example, a $0.01° \times 0.01°$ grid cell in Oslo, Norway would cover 0.56 km^2, compared to 0.94 km^2 for San Diego, USA. To reduce geographical bias and better preserve translation invariance, we reproject all low- and high-resolution datasets to region-specific equal-area grids.

MRMS, EA, and their corresponding ERA5 and static covariates are reprojected to Albers equal-area conic projections centered on their domains, which minimizes shape distortions for regions with large east-west extent. EURADCLIM already uses the Lambert azimuthal equal-area projection, commonly used for European domains. We use its projection to reproject ERA5 and the static variables over the European domain.

We resample the MRMS and EA target datasets to a common 2×2 km resolution, and all ERA5 fields to a 24×24 km resolution, using nearest-neighbour resampling. EURADCLIM is natively on a 2 km grid and is not resampled. This results in a downscaling factor of 12.

3.4 Preprocessing and format

Raw precipitation observations contain coverage heterogeneities, artifacts, and unphysical extremes that must be corrected. First, we mask out parts of the MRMS spatial domain to ensure regions containing only upsampled coarse forecasts are removed. For EURADCLIM, we identify phyiscally incoherent extremes and filter out any timestep containing values above a conservative threshold

aligned the data with historical European records. We additionally remove isolated artifacts in EURADCLIM and EASTASIA. We automatically ensure complete spatial alignment between high-resolution targets and ERA5 upon reprojection and enforce temporal consistency across hourly accumulations. Finally, all pre-processed high-resolution observational targets and their corresponding ERA5 input datasets are converted into cloud-optimized *Zarr* stores, chunked along the temporal dimension, and compressed without loss of information. Static covariates are stored separately in standardized NetCDF files. We provide additional details in Appendix A.

3.5 Stochastic spatio-temporal subsampling of precipitation events

Since the spatial domains of the source datasets included in `RainAtlas` are rather large (*e.g.,* 1,900 $\times$ 2,100 grid cells for EURADCLIM), we split them into spatial crops of dimensions 256×256, each covering around 512×512 km^2. Potential crops are delineated using a sliding window with 32 strides, and 32 margins, and assigned a sampling probability. All crops with more than $75\ \%$ of missing values are discarded, which removes crops outside the actual dataset's coverage, but still includes partial coverage gaps and coastal regions. When a crop is sampled, random horizontal and vertical offsets, ranging between -32 and 32 grid cells, are applied to ensure exhaustive coverage of the spatial domain (Ravuri et al., 2021).

Precipitation datasets are highly sparse, dominated by dry events where no measurable rain occurs. An analysis of randomly subsampled datasets from MRMS, EURADCLIM, and EA (see Table 2 in Appendix B) shows that only $0.88\ \%$, $8.89\ \%$ and $0.75\ \%$ of crops, respectively, contain more than $50\ \%$ non-zero precipitation values. Training on a purely random distribution of spatial crops may lead to unreliable estimates of comparatively rare, heavy precipitation events, which are of great importance for downstream applications.

To mitigate this, we adopt an importance sampling strategy that increases the relative probability of selecting crops with heavier precipitation, based on the rain-rate saturation logic proposed by Ravuri et al. (2021). For each candidate crop x_n, we compute an acceptance score S_n by aggregating grid-cell intensities across the spatial dimensions:

$$S_n(x_n) = \min\left\{1, \frac{m}{C}\sum_{c=1}^{C} g(x_{n,c})\right\}, \tag{1}$$

where C is the total number of grid cells in the crop, and m is a scaling multiplier. The term $g(x_{n,c})$ represents the rain-rate saturation function:

$$g(x) = 1 - \exp(-x/s)\,. \tag{2}$$

Here, s is a saturation constant that controls the sensitivity to high-intensity rain. The exponential form ensures that the score S_n reflects the spatial extent of significant precipitation rather than being dominated by a single extreme pixel value. Once the scores of all candidate crops are computed, the final acceptance probability p_n is obtained by normalizing S_n by the total sum of scores across all potential crops. Throughout our experiments, we use $m = 0.1$ and $s = 1.0$ for training and validation, and $m = 0.2$ and $s = 30.0$ for test, as suggested in Ravuri et al. (2021) for MRMS. We find that this subsampling strategy significantly increases the representation of heavy precipitation events in the training data, as shown in Table 2, and use it for all our experiments. Future work may explore the impact of alternative sampling strategies.

4 Evaluation

4.1 Temporal and geographical generalization

Due to the limited availability and uneven temporal coverage of the high-resolution precipitation products, we must often work with observations datasets that do not fully overlap in time. The atmosphere is a chaotic system, with an effective predictability limit of approximately 14 days (Lorenz, 1969). Therefore, we assume that there is no data leakage when using recent observations for training while evaluating on past observations. We note that, to minimize overfitting to global

climate patterns, ML models must be trained and evaluated on distinct time periods when assessing geographical transferability. Given the temporal coverage of `RainAtlas`'s source datasets, we use observations from 2021 for validation, 2022 for test, 2017 to 2020 for subsampling training crops from EURADCLIM and EA, and 2023 to 2025 for MRMS. A visual representation of the splits is proposed in Figure 6 in Appendix C.

To evaluate geographical generalization across distinct climatic regimes, we additionally perform for following train-test splits: (1) train on each single region and test on the remaining two held out regions, and (2) train on each pair of regions and test on the remaining region.

4.2 Baseline models

We compare several state-of-the-art ML-based downscaling approaches: a deterministic UNet and four generative models, namely an EDM-style diffusion model, *corrective* diffusion, SerpentFlow and a consistency model. SerpentFlow is designed for unpaired data; we train the consistency model on paired data. All models share the same UNet backbone and are conditioned on the static covariates and the bilinearly upsampled ERA5 variables. We do not include any temporal context. As a non-ML reference, we also report the bilinearly upsampled ERA5 *total precipitation*.

UNet We adopt the UNet architecture introduced in ClimateDiffuse (Watt & Mansfield, 2024). The model is trained to predict the residual between the high-resolution target and the bilinearly upsampled ERA5 *total precipitation* variable with l_2 loss.

EDM-style Diffusion We adopt the conditional diffusion approach from ClimateDiffuse (Watt & Mansfield, 2024), based on EDM diffusion framework proposed by Karras et al. (2022) that optimizes noise scheduling, network preconditioning, and 2nd-order ODE sampling for faster, more stable training and generation.

***Corrective* Diffusion** Inspired by CorrDiff (Mardani et al., 2025), *corrective* diffusion decomposes the downscaling task into two stages. First, a deterministic UNet predicts a conditional mean. A diffusion model then learns to generate the conditional variance of the high-resolution target. For both stages, we reuse the UNet and EDM-style diffusion model described above.

SerpentFlow Following Keisler et al. (2026), we use SerpentFlow: a two-step generative framework for downscaling with unpaired data. SerpentFlow first identifies a scale below which the low- and high-resolution datasets can no longer be distinguished, using low-pass filtering and a classifier as a cut-off criterion. After replacing the domain specific high-resolution information with noise, a model is trained to recover it through a flow-matching objective. Unlike the other models, SerpentFlow is not conditioned on the ERA5 *total precipitation* variable, but ERA5 fields are used to define the shared coarse domain and as part of the input at inference.

Consistency model We further include a consistency model (Song et al., 2023; Song & Dhariwal, 2024). The model is trained with a consistency objective that encourages predictions from different noise levels along the same probability-flow ODE trajectory to agree. This enables high-fidelity generation in a single step, bypassing the costly iterative sampling required by standard diffusion models, while still permitting multi-step sampling to trade compute for fidelity.

Further details about architectures and training algorithms are provided in Appendix C.

4.3 Metrics

Precipitation downscaling is a multi-faceted task (Maraun et al., 2010), for which no single metric or unified score can adequately assess model performance. Depending on the downstream applications and end-user needs, different aspects of the evaluation may vary in importance. We therefore evaluate a diverse set of metrics capturing point-wise accuracy, scale-dependent skill, spectral and distributional fidelity, and probabilistic calibration. We detail mathematical formula in Appendix D.

Point-wise and probabilistic errors (MAE, CRPS and bias) We compute the bias, the mean absolute error (MAE) and its probabilistic analogue, the continuous ranked probability score (CRPS)

using an 8-member ensemble (Hersbach, 2000). Pointwise errors are affected by a double-penalty problem. To reduce sensitivity to the small-scale displacement errors, we additionally evaluate these metrics after spatially averaging predictions and observations by factors of 4, 8 and 12.

Spatial and scale-dependent skill (FSS and variogram) To circumvent the double-penalty problem and assess intense precipitation accuracy, we compute the fractions skill score (FSS) (Roberts & Lean, 2008) across multiple spatial neighbourhood radius and precipitation intensity thresholds. Additionally, we compute the variogram score (Scheuerer & Hamill, 2015), a proper scoring rule that penalizes errors in the spatial structure and pairwise dependencies of ensemble predictions.

Spectral and intensity fidelity (RAPSD and LHD) We compute the radially averaged power spectral density (RAPSD) to evaluate whether ML models are able to reproduce the observed distribution of spatial variability across scales (Ruzanski & Chandrasekar, 2011). As for the fidelity of precipitation intensities, we analyze intensity histograms and compute their logarithmic distance (LHD) relative to observations.

Calibration and uncertainty (SSR and rank histograms) Rank histograms are computed for generative models to measure if their ensemble spread capture the true variability of the observations (Hamill, 2001). We also compute the spread-skill ratio, which benchmarks internal model uncertainty against average error to further evaluate ensemble calibration.

5 RESULTS

5.1 IN-DOMAIN PERFORMANCE

Figure 2 presents a visualization of in-domain unit scores across regions on the importance-subsampled test datasets. The numerical results are provided in Table 4.

Generative models outperform deterministic baselines across the evaluation metrics. Trained on the conditional mean, the UNet baseline produces excessive smoothness and fails to capture high-intensity precipitation events, as shown by its poor performance on spectral and distributional fidelity (see Figure 8), and its low FSS on intensites above 5 mm (see Figure 9). Bilinear interpolation of ERA5 maintains near-zero bias, except in EA (which indicates a misalignment between ERA5 and the km-scale observational dataset), but results in a much larger MAE and similar failures as the UNet.

Regarding generative baselines, diffusion-based models show consistently low CRPS and high intensity fidelity across the three datasets. However, we note that the *corrective* diffusion model slightly under-perform on spectral fidelity compared to the diffusion-only model. This stems from the UNet's poor spectral fidelity, upon which the *corrective* diffusion model is conditioned on, which creates a larger RAPSD gap to recover compared to bilinear interpolation. Despite being trained on unpaired data, SerpentFlow exhibits a surprisingly low overall bias across datasets. Similarly, while we employ only one-step sampling, results from the consistency model yield relatively low CRPS, but high variance when predicting extreme precipitation across all spatial scales, as shown by its large relative RAPSD for each region (see Figure 8, bottom). Additionally, both SerpentFlow and the consistency models suffer from right-tail overestimation as reflected by their poor LHD scores, which hinders their overall performance. We find all baselines be well calibrated, the diffusion model outperforming others, and the consistency model being slightly over-dispersed as a result of its higher variance. Overall, the diffusion model outperform the other generative models in-domain, scoring best or second across all metrics.

5.2 GEOGRAPHICAL GENERALIZATION

In this section, we choose to focus our main analysis on pointwise errors and spatial structure, since it is more challenging to improve these metrics through post-hoc bias correction and quantile mapping techniques. Additional results for the rest of the metrics are provided in Appendix E.2.

As can be seen on Figure 3, the generalization performance on unseen regions relative to in-domain results is mostly determined by the training dataset. What stands out is the generally moderate out-of-domain performance degradation when models are trained on the MRMS dataset only, with a

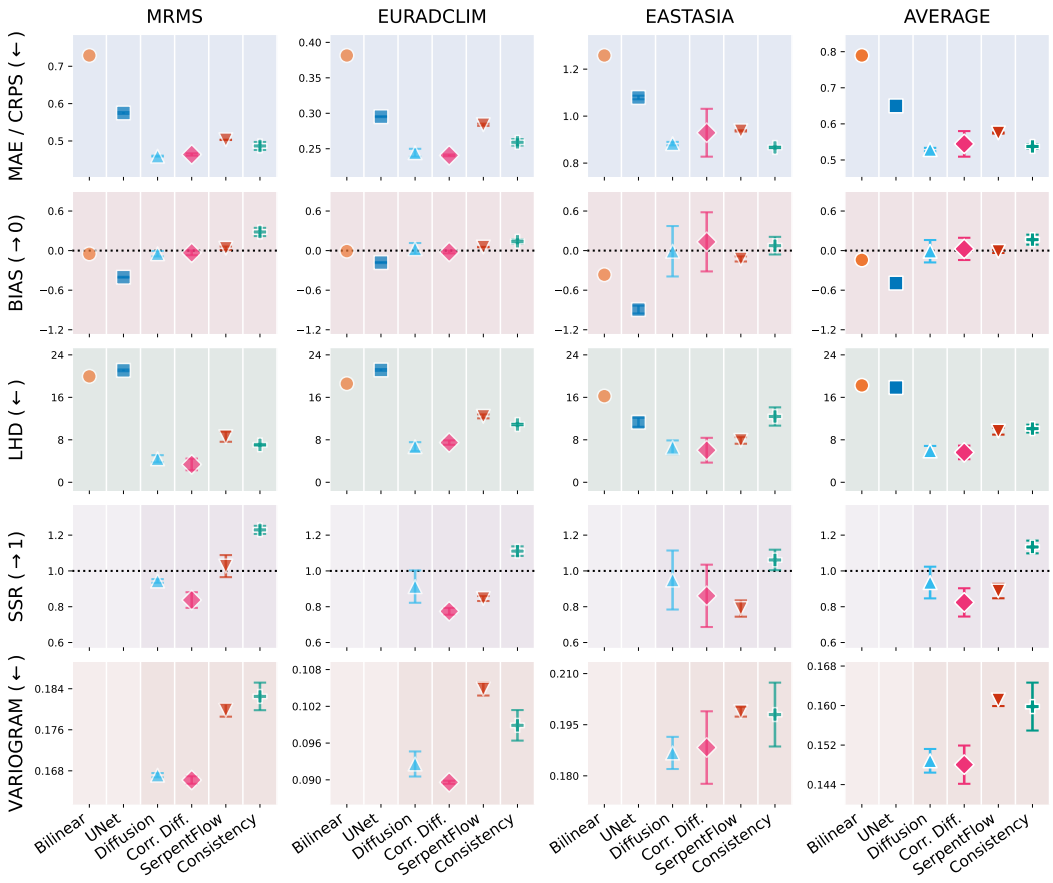


Figure 2: **Diffusion-based models outperform other baselines.** In-domain performance evaluation on importance-subsampled test sets. MAE/CRPS, and bias are computed without artificial coarsening. MAE/CRPS, and variogram rows have region-specific y-axis ranges to facilitate the visualization. Arrows indicate the desired direction for each metric. Numerical results are detailed in Table 4.

maximum relative increase in CRPS, among generative baselines and evaluation domains, of 11.24% for the *corrective* diffusion model when evaluated on EUR. On average, probabilistic models trained on MRMS yield a relative CRPS degradation of 4.97% on EUR, and 1.36% on EA. Contrastively, the UNet shows the most important CRPS degradations across baselines, with 9.35% and 3.53% for EUR and EA respectively. Surprisingly, with the exception of the UNet model, combining another region with MRMS during training did not enhance geographical performance, however it almost consistently reduced degradation compared to solely training on the additional region. CRPS degradation drops from 12.57% to 5.44% on average when evaluated on EUR, and from from 6.89% to 2.09% on EA. Results for the variogram-score show a similar agreement, with performance maintaining a similar region-wise ranking. The average out-of-domain relative degradation on EUR equals to 3.36% versus 13.64% when trained respectively on MRMS or EA, and similarly when evaluated on EA, the models trained on MRMS show a 13.28% relative degradation, compared to 17.09% when trained on EUR. There two likely factors explaining these results: (1) MRMS is the most homogeneous product in terms of observational sources, with the most ground-based radars from a single unified radar network, and (2) it is the most diverse domain in terms of climatic conditions (Beck et al., 2023). An interesting research avenue for future work would hence be investigating whether sampling from specific climate types, related to the target domains, improves geographical generalization.

We now turn to the geographical generalization specifically on the MRMS dataset. Interestingly, we observe the opposite trend in this case: higher degradation from single-region training than with multi-region training. While reaching high increase in relative CRPS when trained solely

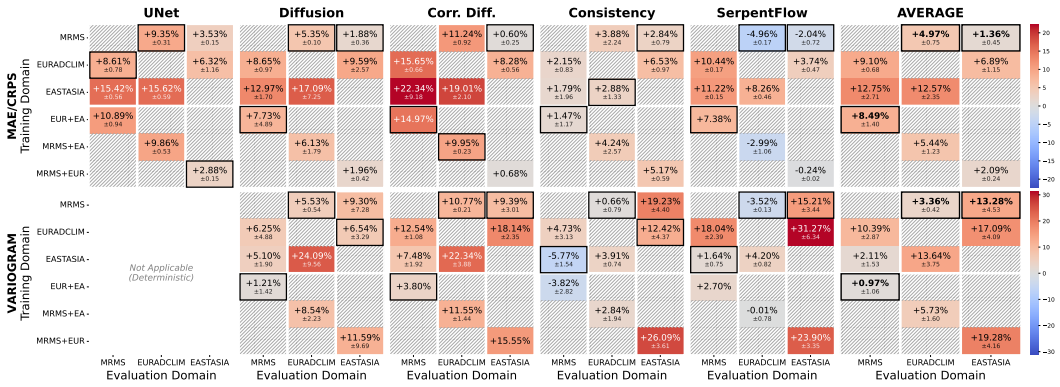


Figure 3: **Models trained on MRMS show more robust geographical generalization.** Relative MAE, CRPS and VARIOGRAM degradation (%) with respect to in-domain training. Additional results are detailed in Figures 15 and 16.

on EUR or EA, models are able to better generalize to the MRMS domain when trained on the two other datasets combined. This observation is specifically true for the diffusion-based models, which achieve substantial improvements in variogram-score: from 6.25% (EUR) and 5.10% (EA) to 1.21% degradation for the diffusion baseline, and from 12.54% (EUR) and 7.48% (EA) to 3.80% degradation for *corrective* diffusion. Again, we observe an important agreement on average results between CRPS and the variogram-score, showing that training on a combination of regions can significantly improve generalization performance when individual regions include limited climatic diversity, and suffer from more important observational bias: incoherent extreme values for EUR, and scarcer radar coverage for EA.

It is worth noting that among all baselines, the consistency model shows substantially better and more stable geographical generalization, with almost all relative CRPS degradation under 5% across training and evaluation configurations. We also find that SerpentFlow achieves low CRPS degradation, even improving in-domain performance in some cases. While this can be explained by its weaker in-domain CRPS, it still shows that unpaired downscaling can be a promising avenue for geographical generalization.

Another interesting finding is that the *corrective* diffusion model suffers on average from higher degradation than other generative baselines. A plausible explanation is that its input already diverges significantly, as shown by the UNet's high relative degradation on held-out domains. In fact, if we compare the results from these three baselines, we can infer that the *corrective* diffusion model's degradation follows an approximate addition of the UNet's and the diffusion's degradation.

6 Conclusion

We introduced `RainAtlas`, a harmonized, multi-continental dataset designed to address interoperability and generalization barriers in precipitation downscaling at km-scale. By standardizing heterogeneous radar, satellite and rain-gauges observations onto equal area grids and pairing them with ERA5 reanalysis and static covariates, we provide a modular framework tailored for machine learning applications. Benchmarking state-of-the-art architectures reveals that while generative models capture complex spatial patterns in-domain, generalizing to unseen geographies remains challenging. We note substantial variance in generalization performance, that we suggest depends heavily on the climatic conditions diversity of training datasets. Furthermore, we observe that models relying on sequential predictions, such as *corrective* diffusion, might suffer from compounded errors under geographical shift. These findings underscore the need for unified geographical benchmark in precipitation downscaling. By providing a multi-continental ML-ready benchmark, `RainAtlas` could help the Earth system science community more systematically investigate transferability dynamics, and support the development of models that produce reliable high-resolution precipitation estimates to regions lacking observational infrastructure.

AI USE STATEMENT

In this work, we used generative AI tools for improving the writing quality of our original drafts, as well as guidance for shortening some sections of the main text. We also used AI tools for assisting with parts of our code implementation, as well as for generating the code that produced the figures we present. We have not used generative AI tools for drafting entire sections of the paper nor for idea generation. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

`RainAtlas` is built from public observation products. MRMS is produced by NOAA and is in the public domain. EURADCLIM is released by KNMI under CC BY 4.0. The EA product is distributed by TPDC. The data contain no personal information. Downscaled precipitation from models trained on `RainAtlas` should not be used for operational warnings or risk decisions without local validation, in particular outside the three regions, since our results show that evaluation on unseen regions can degrade skill.

REPRODUCIBILITY STATEMENT

Dataset construction and preprocessing are described in Section 3 and Appendix A; the subsampling procedure in Section 3.5; architectures, training details and hyperparameters in Appendix C; and all metric definitions in Appendix D. The dataset is available on Hugging Face. We release the code for preprocessing, training and evaluation, and the seeds used for training (42, 84, 126).

REFERENCES

Hylke E. Beck, Tim R. McVicar, Noemi Vergopolan, Alexis Berg, Nicholas J. Lutsko, Ambroise Dufour, Zhenzhong Zeng, Xin Jiang, Albert I. J. M. van Dijk, and Diego G. Miralles. High-resolution (1 km) köppen-geiger maps for 1901–2099 based on constrained cmip6 projections. *Scientific Data*, 10(1), October 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02549-6. URL `http://dx.doi.org/10.1038/s41597-023-02549-6`.

Mark Carroll, Charlene DiMiceli, John Townshend, Robert Sohlberg, Alfred Hubbard, Margaret Wooten, Caleb Spradlin, Roger Gill, Savannah Strong, Amanda Burke, Mariana Blanco-Rojas, and Melanie Frost. MODIS/Terra land water mask derived from MODIS and SRTM L3 global 250m SIN grid V061, 2024.

Ting-Chen Chen, François Collet, and Alejandro Di Luca. Evaluation of ¡scp¿era5¡/scp¿ precipitation and 10-m wind speed associated with extratropical cyclones using station data over ¡scp¿north america¡/scp¿. *International Journal of Climatology*, 44(3):729–747, February 2024. ISSN 1097-0088. doi: 10.1002/joc.8339. URL `http://dx.doi.org/10.1002/joc.8339`.

Xuanhong Chen, Kairui Feng, Naiyuan Liu, Bingbing Ni, Yifan Lu, Zhengyan Tong, and Ziang Liu. Rainnet: A large-scale imagery dataset and benchmark for spatial precipitation downscaling. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. arXiv, 2022. doi: 10.48550/ARXIV.2012.09700. URL `https://arxiv.org/abs/2012.09700`.

Ting-Yu Dai and Hayato Ushijima-Mwesigwa. Precipdiff: Leveraging image diffusion models to enhance satellite-based precipitation observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):27932–27939, April 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i27.35010. URL `http://dx.doi.org/10.1609/aaai.v39i27.35010`.

Jeffrey J Danielson and Dean B Gesch. Global multi-resolution terrain elevation data 2010 (GMTED2010), 2011.

Taichen Feng, Xian Zhu, and Wenjie Dong. Historical assessment and future projection of extreme precipitation in ¡scp¿cmip6¡/scp¿ models: Global and continental. *International Journal of Climatology*, 43(9):4119–4135, April 2023. ISSN 1097-0088. doi: 10.1002/joc.8077. URL `http://dx.doi.org/10.1002/joc.8077`.

C. A. T. Ferro. Fair scores for ensemble forecasts: Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, December 2013. ISSN 0035-9009. doi: 10.1002/qj.2270. URL http://dx.doi.org/10.1002/qj.2270.

V. Fortin, M. Abaza, F. Anctil, and R. Turcotte. Why should ensemble spread match the rmse of the ensemble mean? *Journal of Hydrometeorology*, 15(4):1708–1713, 2014. ISSN 1525-7541. doi: 10.1175/jhm-d-14-0008.1. URL http://dx.doi.org/10.1175/JHM-D-14-0008.1.

Luca Glawion, Julius Polz, Harald Kunstmann, Benjamin Fersch, and Christian Chwala. Global spatio-temporal era5 precipitation downscaling to km and sub-hourly scale using generative ai. *npj Climate and Atmospheric Science*, 8(1), 2025. ISSN 2397-3722. doi: 10.1038/s41612-025-01103-y. URL http://dx.doi.org/10.1038/s41612-025-01103-y.

Thomas M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550 – 560, 2001. doi: 10.1175/1520-0493(2001)129⟨0550:IORHFV⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0550_iorhfv_2.0.co_2.xml.

Thomas M. Hamill and Stephen J. Colucci. Verification of eta–rsm short-range ensemble forecasts. *Monthly Weather Review*, 125(6):1312 – 1327, 1997. doi: 10.1175/1520-0493(1997)125⟨1312:VOERSR⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/125/6/1520-0493_1997_125_1312_voersr_2.0.co_2.xml.

Paula Harder, Luca Schmidt, Francis Pelletier, Nicole Ludwig, Matthew Chantry, Christian Lessig, Alex Hernandez-Garcia, and David Rolnick. Benchmarking the geographic generalization of deep learning models for precipitation downscaling. *Scientific Reports*, 16(1), January 2026. ISSN 2045-2322. doi: 10.1038/s41598-025-34557-4. URL http://dx.doi.org/10.1038/s41598-025-34557-4.

M. Heistermann, S. Jacobi, and T. Pfaff. Technical note: An open source library for processing weather radar data (¡i¿wradlib¡/i¿). *Hydrology and Earth System Sciences*, 17(2):863–871, February 2013. ISSN 1607-7938. doi: 10.5194/hess-17-863-2013. URL http://dx.doi.org/10.5194/hess-17-863-2013.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559 – 570, 2000. doi: 10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/15/5/1520-0434_2000_015_0559_dotcrp_2_0_co_2.xml.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Philipp Hess, Michael Aich, Baoxiang Pan, and Niklas Boers. Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning. *Nature Machine Intelligence*, 7(3):363–373, March 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-00980-5. URL http://dx.doi.org/10.1038/s42256-025-00980-5.

Timothy David Hewson and Fatima Maria Pillosu. A low-cost post-processing technique improves weather forecasts around the world. *Communications Earth amp; Environment*, 2(1), 2021. ISSN 2662-4435. doi: 10.1038/s43247-021-00185-9. URL http://dx.doi.org/10.1038/s43247-021-00185-9.

Helen Hooker, Jessica Steinkopf, Charles Langton Vanya, Genito Maure, Bernardino Nhantumbo, Francois Engelbrecht, Hannah Cloke, and Elisabeth Stephens. Extreme rainfall from tropical

cyclones is revealed by kilometer-scale downscaling in southeast africa. *Journal of Hydrometeorology*, 27(5):783–799, May 2026. ISSN 1525-7541. doi: 10.1175/jhm-d-25-0179.1. URL http://dx.doi.org/10.1175/JHM-D-25-0179.1.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022. URL https://arxiv.org/abs/2206.00364.

Julie Keisler, Anastase Alexandre Charantonis, Yannig Goude, Boutheina Oueslati, and Claire Monteleoni. Serpentflow: Generative unpaired domain alignment via shared-structure decomposition, 2026. URL https://arxiv.org/abs/2601.01979.

Xin-Zhong Liang. Extreme rainfall slows the global economy. *Nature*, 601(7892):193–194, January 2022. URL https://www.nature.com/articles/d41586-021-03783-x.

Yuhao Liu, James Doss-Gollin, Qiushi Dai, Ashok Veeraraghavan, and Guha Balakrishnan. Downscaling extreme precipitation with wasserstein regularized diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025. ISSN 1558-0644. doi: 10.1109/tgrs.2025.3611872. URL http://dx.doi.org/10.1109/TGRS.2025.3611872.

Edward N. Lorenz. The predictability of a flow which possesses many scales of motion. *Tellus*, 21 (3):289–307, 1969. doi: 10.3402/tellusa.v21i3.10086. URL https://doi.org/10.3402/tellusa.v21i3.10086.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth amp; Environment*, 6(1), February 2025. ISSN 2662-4435. doi: 10.1038/s43247-025-02042-5. URL http://dx.doi.org/10.1038/s43247-025-02042-5.

Jack L. McKee and Andrew D. Binns. A review of gauge–radar merging methods for quantitative precipitation estimation in hydrology. *Canadian Water Resources Journal / Revue canadienne des ressources hydriques*, 41(1-2):186–203, 2016. doi: 10.1080/07011784.2015.1064786. URL https://doi.org/10.1080/07011784.2015.1064786.

Philipp Neumann, Peter Düben, Panagiotis Adamidis, Peter Bauer, Matthias Brück, Luis Kornblueh, Daniel Klocke, Bjorn Stevens, Nils Wedi, and Joachim Biercamp. Assessing the scales in numerical weather and climate predictions: will exascale be the rescue? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 377(2142): 20180148, February 2019. ISSN 1364-503X. doi: 10.1098/rsta.2018.0148. URL https://doi.org/10.1098/rsta.2018.0148. _eprint: https://royalsocietypublishing.org/rsta/article-pdf/doi/10.1098/rsta.2018.0148/1412375/rsta.2018.0148.pdf.

A. Overeem, E. van den Besselaar, G. van der Schrier, J. F. Meirink, E. van der Plas, and H. Leijnse. Euradclim: the european climatological high-resolution gauge-adjusted radar precipitation dataset. *Earth System Science Data*, 15(3):1441–1464, 2023. doi: 10.5194/essd-15-1441-2023. URL https://essd.copernicus.org/articles/15/1441/2023/.

Andreas F. Prein, Wolfgang Langhans, Giorgia Fosser, Andrew Ferrone, Nikolina Ban, Klaus Goergen, Michael Keller, Merja Tölle, Oliver Gutjahr, Frauke Feser, Erwan Brisson, Stefan Kollet, Juerg

Schmidli, Nicole P. M. van Lipzig, and Ruby Leung. A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of Geophysics*, 53(2): 323–361, May 2015. ISSN 1944-9208. doi: 10.1002/2014rg000475. URL http://dx.doi.org/10.1002/2014RG000475.

Ilan Price and Stephan Rasp. Increasing the accuracy and resolution of precipitation forecasts using deep generative models. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10555–10571. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/price22a.html.

Neelesh Rampal, Bryn Ward-Leikis, Yun Sing Koh, Peter B. Gibson, Hong-Yang Liu, Vassili Kitsios, Tristan Meyers, Jeff Adie, Yang Juntao, and Steven C. Sherwood. An intercomparison of generative machine learning methods for downscaling precipitation at fine spatial scales, 2026. URL https://arxiv.org/abs/2512.13987.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03854-z. URL http://dx.doi.org/10.1038/s41586-021-03854-z.

Nigel M. Roberts and Humphrey W. Lean. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97, January 2008. ISSN 0027-0644. doi: 10.1175/2007mwr2123.1. URL http://dx.doi.org/10.1175/2007MWR2123.1.

Evan Ruzanski and V. Chandrasekar. Scale filtering for improved nowcasting performance in a high-resolution x-band radar network. *IEEE Transactions on Geoscience and Remote Sensing*, 49(6):2296–2307, 2011. ISSN 1558-0644. doi: 10.1109/tgrs.2010.2103946. URL http://dx.doi.org/10.1109/TGRS.2010.2103946.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321 – 1334, 2015. doi: 10.1175/MWR-D-14-00269.1. URL https://journals.ametsoc.org/view/journals/mwre/143/4/mwr-d-14-00269.1.xml.

Christian Schroeder de Witt, Catherine Tong, Valentina Zantedeschi, Daniele De Martini, Alfredo Kalaitzis, Matthew Chantry, Duncan Watson-Parris, and Piotr Bilinski. Rainbench: Towards data-driven global precipitation forecasting from satellite imagery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14902–14910, May 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i17.17749. URL http://dx.doi.org/10.1609/aaai.v35i17.17749.

Travis M. Smith, Valliappa Lakshmanan, Gregory J. Stumpf, Kiel L. Ortega, Kurt Hondl, Karen Cooper, Kristin M. Calhoun, Darrel M. Kingfield, Kevin L. Manross, Robert Toomey, and Jeff Brogden. Multi-radar multi-sensor (mrms) severe weather and aviation products: Initial operating capabilities. *Bulletin of the American Meteorological Society*, 97(9):1617 – 1630, 2016. doi: 10.1175/BAMS-D-14-00173.1. URL https://journals.ametsoc.org/view/journals/bams/97/9/bams-d-14-00173.1.xml.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097. arXiv, 2024. doi: 10.48550/ARXIV.2310.14189. URL https://arxiv.org/abs/2310.14189.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023. URL https://arxiv.org/abs/2303.01469.

Jiajia Su, Chiyuan Miao, Francis Zwiers, Hylke Beck, Phil Jones, Qiaohong Sun, Louise J. Slater, Wouter R. Berghuijs, Yoshihide Wada, Daniel Rosenfeld, Jiaojiao Gou, Yi Wu, Paolo Tarolli, Pasquale Borrelli, Panos Panagos, Lisa V. Alexander, Qi Zhang, Jinlong Hu, Seung-Ki Min, Luis Samaniego, Qingyun Duan, Georgia Destouni, Jose A. Marengo, Reza Modarres, and Soroosh

Sorooshian. Precipitation observing network gaps limit climate change impact assessment. *Nature*, 652(8108):119–125, March 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10300-5. URL http://dx.doi.org/10.1038/s41586-026-10300-5.

Søren Thorndahl, Thomas Einfalt, Patrick Willems, Jesper Ellerbæk Nielsen, Marie-Claire ten Veldhuis, Karsten Arnbjerg-Nielsen, Michael R. Rasmussen, and Peter Molnar. Weather radar rainfall data in urban hydrology. *Hydrology and Earth System Sciences*, 21(3):1359–1380, March 2017. ISSN 1607-7938. doi: 10.5194/hess-21-1359-2017. URL http://dx.doi.org/10.5194/hess-21-1359-2017.

Marina Vicens-Miquel, Amy McGovern, Aaron J. Hill, Efi Foufoula-Georgiou, Clement Guilloteau, and Samuel S. P. Shen. A diffusion-based framework for high-resolution precipitation forecasting over conus, 2025. URL https://arxiv.org/abs/2512.09059.

Robbie A. Watt and Laura A. Mansfield. Generative diffusion-based downscaling for climate, 2024. URL https://arxiv.org/abs/2404.17752.

Hanmeng Xia and Kaicun Wang. High resolution (0.01 °) hourly precipitation dataset in east asia based on satellite radar and rain gauge fusion (2017-2022), 10 2025a. URL https://dx.doi.org/10.11888/Atmos.tpdc.301339.

Hanmeng Xia and Kaicun Wang. Hourly, kilometer-scale precipitation merged from rain gauge, ground-based radar and satellites over east asia: methods, evaluation and applications. *Journal of Hydrology*, 662:134148, 2025b. ISSN 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2025.134148. URL https://www.sciencedirect.com/science/article/pii/S0022169425014866.

Chugang Yi, Minghan Yu, Weikang Qian, Yixin Wen, and Haizhao Yang. Efficient kilometer-scale precipitation downscaling with conditional wavelet diffusion. *Journal of Geophysical Research: Machine Learning and Computation*, 3(2):e2025JH000941, 2026. doi: https://doi.org/10.1029/2025JH000941. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2025JH000941.

Maruša Špitalar, Jonathan J. Gourley, Celine Lutoff, Pierre-Emmanuel Kirstetter, Mitja Brilly, and Nicholas Carr. Analysis of flash flood parameters and human impacts in the us from 2006 to 2012. *Journal of Hydrology*, 519:863–870, 2014. ISSN 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2014.07.004. URL https://www.sciencedirect.com/science/article/pii/S0022169414005216.

A PREPROCESSING OF LOW- AND HIGH-RESOLUTION SOURCE DATASETS

Masking MRMS for radar coverage Across the CONUS domain, the MRMS dataset shows regional heterogeneities in effective resolution (see Figure A). Further exploration revealed that some peripheral regions lack radar and rain gauge coverage. In these areas, the grid is instead filled with upsampled coarse satellite and NWP-derived data. To retain only genuine km-scale observations, we mask these regions following the spatial coverage of the variable `PrecipFlag`, which indicates the type of precipitation (*e.g.*, convective, stratiform, hail) from radar observations.

Incoherent extreme precipitation in EURADCLIM Authors of the EURADCLIM dataset report that some regions might exhibit physically incoherent values, resulting from radar artifacts and sparse rain-gauge networks. In particular, the raw data contains hourly precipitation intensities reaching up to 300 mm (Figure A, **left**), which is inconsistent with historical European extreme events records, which rarely exceed 150 mm/h. Setting a conservative threshold of 125 mm, we identified multiple clusters of affected grid-cells, especially over Norway, Estonia, Russia, Greece, Romania, Moldova, and some other coastal regions (Figure A **right**). Because only $\sim 5\%$ of the total timesteps contained at least one grid cell exceeding this threshold, we discarded these timesteps entirely from our dataset.

Individual artifacts in EURADCLIM and EA We removed the first 7 hours of 29 April 2019 from EURADCLIM due to a prominent spatial artifact over Spain. Similarly, for EA, we identified linear grid-line artifacts with unphysically large values occurring between 5 May and 7 May 2019, discarding the 26 affected timesteps from the final dataset.

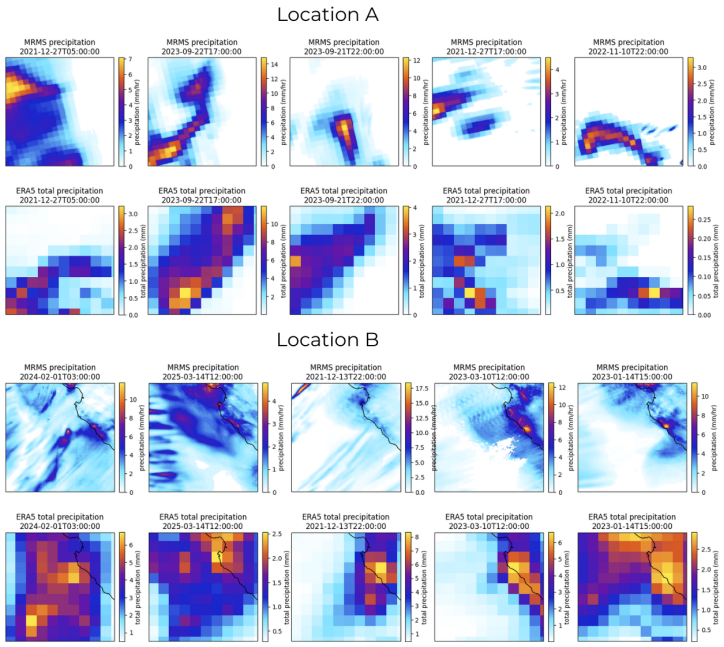


Figure 4: Comparison of MRMS and ERA5 precipitation samples. The high-resolution MRMS crops **(top)** are compared with the low-resolution ERA5 crops **(bottom)** for two different locations. In **Location A**, the MRMS data appears to be upsampled from a coarser grid. In contrast, **Location B** shows the typical fine-scale details expected from MRMS.

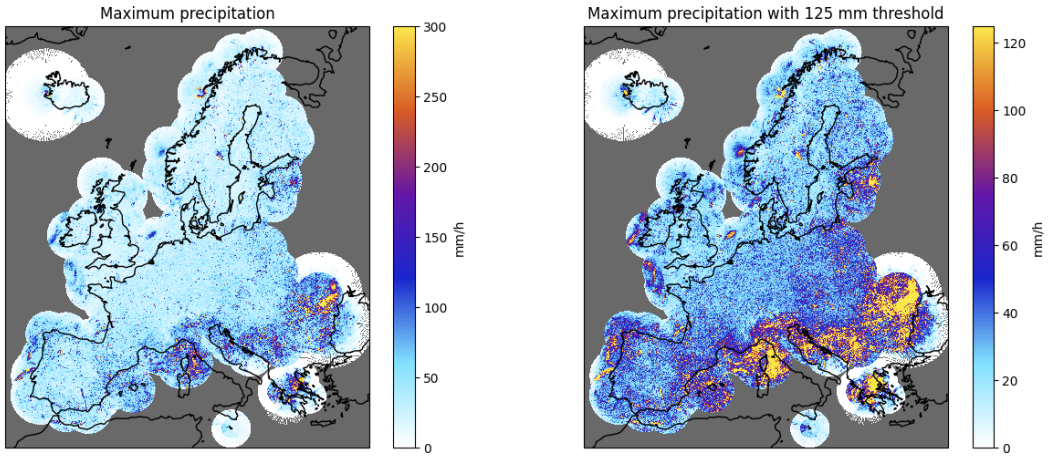


Figure 5: **(left)** Raw maximum hourly precipitation per grid-cell for EURADCLIM (2013 to 2023), and **(right)** same using a 125 mm threshold for easier identification of problematic grid-cells.

Spatio-temporal alignment of ERA5 and high-resolution datasets After processing the high-resolution precipitation source datasets, we employ an automated pipeline to retrieve and align the corresponding ERA5 reanalysis data. To maintain spatial alignment, we compute the geographic bounding box of the projected high-resolution domain, incorporating a coordinate buffer to ensure complete spatial coverage upon re-projection to a common equal-area projection. Regarding temporal alignment, we use hourly accumulated precipitation for low- and high-resolution precipitation data, and discard any timestamp that isn't include both in ERA5 and the high-resolution source datasets.

Storage and format Due to the large spatial extent and km-scale resolution of our domains, the raw data requires roughly a dozen terabytes of memory before re-projection and resampling. To handle this volume, we rely heavily on parallelization and lazy computation. We process the data incrementally and write it to disk in chunks along the temporal dimension using the *Zarr* format, which provides efficient compression. While the three source observational datasets would require 7.32 terabytes after preprocessing, each final *Zarr* store needs only 150 GB.

B STOCHASTIC SPATIO-TEMPORAL SUBSAMPLING OF PRECIPITATION CROPS

Precipitation observations are largely dominated by dry conditions. To illustrate the effect of our subsampling methodology, Table 2 compares the distributional statistics of 150k training crops extracted via uniform random sampling against rain-rate saturation importance sampling strategy proposed by Ravuri et al. (2021), across MRMS, EURADCLIM, and EA. Under uniform random sampling, dry grid cells (≤ 0 mm) account for 82.8% to 96.7% of observations, while fewer than 1% of spatial crops in MRMS and EA contain more than 50% wet grid cells. In contrast, importance sampling significantly enhances the representation of heavy rainfall regimes: it increases the average crop intensity by up to $7\times$, elevates the average maximum intensity up to 19.84 mm, and boosts the proportion of crops with more than 50% wet grid cells by more than an order of magnitude.

Table 2: Statistics of `RainAtlas`'s training datasets with 150k samples, generated with random or importance subsampling. The first section contains grid cell occurrences of multiple precipitation intensities, while the second section gives information on crops: average precipitation, average maximum intensity, and proportion of crops with more than $10\%, 25\%$ and 50% non-zero precipitation grid cells.

	MRMS		EURADCLIM		EA	
	Random	Importance	Random	Importance	Random	Importance
≤ 0 mm (%)	95.80	72.76	82.81	55.77	96.68	80.33
$0-1$ mm (%)	2.38	12.63	15.34	37.18	1.01	5.05
$1-4$ mm (%)	1.47	11.72	1.65	7.13	1.77	10.58
$4-5$ mm (%)	0.10	0.91	0.08	0.36	0.15	1.03
$5-8$ mm (%)	0.13	1.14	0.08	0.39	0.19	1.37
$8-10$ mm (%)	0.04	0.29	0.02	0.08	0.06	0.44
>10 mm (%)	0.08	0.55	0.02	0.09	0.15	1.19
average (mm)	0.07	0.55	0.07	0.29	0.09	0.66
maximum (mm)	6.53	19.84	5.83	13.10	4.61	16.60
nz >10 (%)	13.30	78.10	48.47	95.25	10.33	57.64
nz >25 (%)	4.58	45.99	27.27	78.55	3.66	31.39
nz >50 (%)	0.88	14.47	8.89	41.10	0.75	9.55

C ARCHITECTURES AND TRAINING

Transformation and normalization With the exception of the static land-sea mask, all variables undergo a fourth-root transformation ($y = x^{0.25}$) to mitigate skewness, and are then standardized independently for each dataset using z-score normalization.

C.1 TRAINING DETAILS

All models are trained using 187,000 steps with a batch-size of 32 samples and the AdamW Loshchilov & Hutter (2019) optimization algorithm. Learning rates follow a linear warm-up from $1\ \%$ to $100\ \%$ of the base learning rate over the first $2\ \%$ of steps, followed by cosine annealing to $\eta_{\min} = 10^{-6}$ over the remaining $98\ \%$. Mixed precision is used for all models. Gradients are clipped so that their L_2 norm doesn't exceed 1.0. Following their original implementation, SerpentFlow and Consistency are trained using Exponential Moving Average (EMA). Each configuration is trained with three random seeds (42, 84, 126). Due to time constraints, a few configurations are evaluated over only one or two

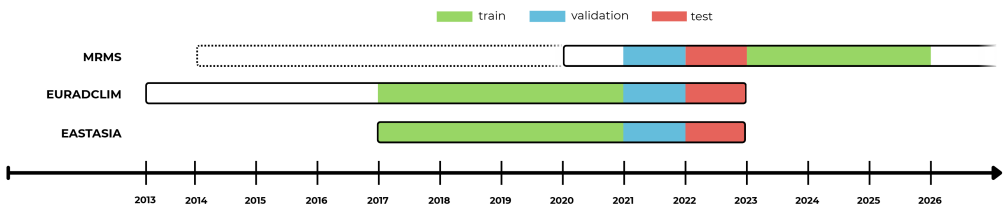


Figure 6: **Train, validation, and test temporal splits.** Because MRMS only partially overlap with EURADLCIM and EASTASIA, we reserve 2021 and 2022 for subsampling validation and test datasets. Since atmospheric predictability is limited to approximately 14 days, training on data from later periods does not induce data leakage.

random seeds. Results will be updated during the rebuttal period. Individual hyperparameters are defined in Table 3.

C.2 Deterministic UNet

The UNet backbone shared by all models is the EDM UNet of Karras et al. (2022) adapted to climate downscaling by (Watt & Mansfield, 2024). We use 64 base channels, a per-resolution multiplier of $[1, 2, 3, 4]$ over four spatial stages ($256 \to 128 \to 64 \to 32$), two encoder blocks plus three decoder blocks per stage, and decoder-only self-attention at the three deepest stages (32×32, 16×16, 8×8). The deterministic UNet baseline is trained on the standardized residual so that $\hat{y} = \mathrm{UNet}(x) + \bar{x}$, where $\bar{x}$ is the bilinearly upsampled ERA5 total-precipitation.

C.3 Diffusion

The diffusion baseline trains the EDM denoiser of Karras et al. (2022) directly on the precipitation target conditioned on the bilinearly upsampled ERA5 predictors and the static covariates. The denoiser shares the UNet backbone wrapped by the EDM preconditioner. We use the EDM training noise schedule with $P_{\mathrm{mean}} = -1.2$, $P_{\mathrm{std}} = 1.2$, and $\sigma_{\mathrm{data}} = 0.5$, as well as the EDM denoising loss (Karras et al., 2022).

At inference we run the EDM Heun sampler Karras et al. (2022) with 100 second-order steps for 8 ensemble members, 0.002, $\sigma_{\max} = 80$, $\rho = 7$, and stochastic churn parameters $S_{\mathrm{churn}} = 40$, $S_{\min} = 0$, $S_{\max} = \infty$, $S_{\mathrm{noise}} = 1$.

C.4 *Corrective* Diffusion

The *corrective* diffusion baseline is adapted from CorrDiff Mardani et al. (2025), decomposing the downscaling task into a deterministic mean prediction followed by residual generative modeling. First, a deterministic UNet predicts the conditional mean $\mu(x) = \mathbb{E}[y \mid x]$, where y denotes the target km-scale precipitation field and x denotes the ERA5 and static conditioning variables. Second, a conditional diffusion model is trained to model the residual distribution $r = y - \mu(x)$, capturing high-frequency spatial details and unresolved sub-grid variability.

The UNet conditional mean regression stage reuses the identical setting as the UNet deterministic baseline, while the generative stage shares the diffusion baseline setting.

C.5 SerpentFlow

SerpentFlow is a two-stage downscaling generative model designed for unpaired domain alignment (Keisler et al., 2026). Given low- and high-resolution domains, respectively $\mathcal{D}_A$ and $\mathcal{D}_B$, SerpentFlow assumes that there exists a shared latent space $\mathcal{Z}$ and a bijective mapping μ such that: $\mu(\mathcal{D}_A) = \mathcal{B}_A \subset \mathcal{Z}$ and $\mu(\mathcal{D}_B) = \mathcal{B}_B \subset \mathcal{Z}$. Further, both the target and input latent domains can be decomposed into shared and specific components: $\mathcal{B}_A = \mathcal{B}^S \oplus \mathcal{B}_A^D$ and $\mathcal{B}_B = \mathcal{B}^S \oplus \mathcal{B}_B^D$.

Instead of training on paired low- and high-resolution samples, SerpentFlow constructs pseudo-pairs from the target samples by replacing their latent specific components with white noise. Given a target sample y^i, $\epsilon \sim \mathcal{N}(0,1)$, and its transformation into the shared latent space $\mu(\epsilon) = z^S + z^D_\epsilon \in \mathcal{Z}^S \oplus \mathcal{Z}^D_\epsilon \subset \mathcal{Z}$, we construct the pseudo-view $\tilde{y}^i$ such that:

$$(\tilde{y}^i, y^i) \in (\mu^{-1}(\mathcal{B}^S \oplus \mathcal{B}^D_\epsilon) \times \mathcal{D}_B), \quad \text{with } \tilde{y}^i = \mu^{-1}(z^{i,S}_B + z^D_\epsilon), \tag{3}$$

where $z^{i,S}_B = \mu(y^i)^S$ represents the shared latent component of the target sample.

An informed choice for $\mathcal{Z}$ is the Fourier domain, where low-frequency signals correspond to large-scale synoptic processes shared by low- and high-resolution precipitation datasets. Using a cut-off frequency ω_c, SerpentFlow constructs the shared and specific elements of the latent space $\mathcal{Z}$ following:

$$\mathcal{B}^S = \{z(\xi) : \|\xi\| < \omega_c\}, \quad \mathcal{B}^D = \{z(\xi) : \|\xi\| \geq \omega_c\}, \tag{4}$$

where ξ represents the frequency components in the Fourier domain.

SerpentFlow proposes an automatic method for selecting the cut-off frequency. Given a candidate ω_c, a low-pass filter is applied to each sample from the input and target domains:

$$x^S = \mathcal{F}^{-1}[\mathbb{1}_{\{\|\xi\| < \omega_c\}} \cdot \mathcal{F}(x)] \tag{5}$$

where $\mathcal{F}$ is the Fourier transform. Starting from a high candidate cut-off frequency, SerpentFlow trains a classifier D_ψ to predict whether x^S originates from $\mathcal{D}_A$ or $\mathcal{D}_B$ using the standard binary cross-entropy loss. ω_c is decreased iteratively until the validation accuracy of D_ψ reaches an indiscriminable level:

$$Acc(D_\psi, \omega^*_c) \approx 0.5, \tag{6}$$

which indicates that there is no more specific information contained in the samples given the optimal cut-off frequency ω^*_c.

We use the same 3-layer convolutional neural network as Keisler et al. (2026) for the classifier D_ψ. Once ω^*_c is reached, we train a flow-matching model f_θ for reconstructing y^i from $\tilde{y}^i_{\omega^*_c}$ following the same setting as the SerpentFlow authors. At inference, we sample the low-resolution precipitation field $x^i \in \mathcal{D}_A$ and predict the corresponding high-resolution target $\hat{y}^i = f_\theta(\tilde{x}^i_{\omega^*_c})$ using the flow-matching model.

C.6 Consistency

The consistency baseline reproduces the continuous-time framework of Song et al. (2023), which is designed to overcome the slow, iterative sampling process of traditional diffusion models. We utilize the identical EDM UNet backbone and conditioning strategy on the bilinearly upsampled ERA5 predictors and static covariates as our diffusion baseline.

Building upon this continuous-time formulation, consistency models learn a mapping $f_\theta(y_t, t, x)$ that points from any intermediate noisy state y_t directly back to its clean origin y. This requires the model to satisfy the consistency property, meaning that evaluations at any two arbitrary time steps t and t' along the same diffusion trajectory must yield the same prediction:

$$f_\theta(y_t, t, x) = f_\theta(y_{t'}, t', x), \quad \forall t, t' \in [\epsilon, T]. \tag{7}$$

A critical requirement is the boundary condition $f_\theta(y_\epsilon, \epsilon, x) = y_\epsilon$. To strictly enforce this, the model prediction is parameterized using modified time-dependent skip connections:

$$f_\theta(y_t, t, x) = c_{\text{skip}}(t) y_t + c_{\text{out}}(t) F_\theta(y_t, t, x), \tag{8}$$

where F_θ is the raw neural network output. Unlike the standard EDM preconditioner, the consistency weighting functions are explicitly constrained such that $c_{\text{skip}}(\epsilon) = 1$ and $c_{\text{out}}(\epsilon) = 0$, ensuring the model perfectly returns the input at the minimum noise level ϵ.

To enforce self-consistency across the trajectory, the model minimizes the discrepancy between predictions made at adjacent time steps (t_{n+1}, t_n) derived from the discretized EDM noise schedule:

$$\mathcal{L}(\theta, \theta^-) = \mathbb{E}_{y,x,n,z}\left[\lambda(\sigma_i) d\left(f_\theta(y_{t_{n+1}}, t_{n+1}, x), f_{\theta^-}(y_{t_n}, t_n, x)\right)\right], \tag{9}$$

where θ^- represents the exponential moving average (EMA) of the online network parameters θ, $\lambda(\sigma_i)$ weights the relative contribution of different noise levels, $d(\cdot,\cdot)$ is a distance metric, and $z \sim \mathcal{N}(0, I)$ is the noise applied to sample $y_{t_{n+1}}$ and y_{t_n}. This EMA target network stabilizes training and ensures a continuous learning trajectory relative to the online parameters. Following recommendations from Song & Dhariwal (2024), we set the EMA decay parameter to 0, formally fixing $\theta = \theta^-$. This provably eliminates the inherent bias and approximation error introduced by a lagging target network.

We adopt several additional techniques from Song & Dhariwal (2024) to improve performance. First, we replace the L_2 distance in the consistency loss $\mathcal{L}$ with the Pseudo-Huber metric: $d(x, y) = \sqrt{\|x - y\|_2^2 + c^2} - c$. Following the authors, we fix $c = 0.00054\sqrt{d}$ for samples of dimension d (we note this hyperparameter was originally optimized for natural images and may be suboptimal for precipitation fields). Second, we replace the uniform weighting function $\forall\, i, \lambda(\sigma_i) = 1$ with $\lambda(\sigma_i) = \frac{1}{\sigma_{i+1} - \sigma_i}$, and we implement the proposed discretized lognormal noise schedule. Finally, we set the minimum number of discretization bins to 10, and for reasons of training stability, we lower the number of maximal bins to 50.

At inference, starting from pure Gaussian noise $y_T \sim \mathcal{N}(0, T^2 I)$ and the given conditioning variables x, the model predicts the corresponding high-resolution precipitation field $\hat{y}$ in a single forward pass:

$$\hat{y} = f_\theta(y_T, T, x). \tag{10}$$

We initially sought to reproduce the unpaired approach proposed by (Hess et al., 2025). This method relies on finding the wavenumber k^* where the Power Spectral Densities (PSDs) of the low- and high-resolution datasets intersect to determine the corresponding noise timestep t^*. Inference is then performed from the perturbed state $\tilde{x} = x + \epsilon_{t^*}$, with $\epsilon_{t^*} \sim \mathcal{N}(0, \sigma^2(t^*)\mathbf{I})$ and $\sigma(t) = N^2\text{PSD}(k)$ for grid size N. However, because the PSDs of our low- and high-resolution datasets do not intersect at any wavenumber, applying this specific unpaired strategy proved impossible.

Table 3: Training and architecture hyperparemeters for every model. Additional hyperparameters are specified under each model specification when they are not shared with others. We note that most of these hyperparameters are certainly sub-optimals as we didn't perform any extensive hyperparameters seach. Most values are taken from previous work (Watt & Mansfield, 2024; Karras et al., 2022; Song & Dhariwal, 2024; Keisler et al., 2026).

	UNet	Diffusion	*Corrective* Diffusion	SerpentFlow	Consistency
Loss	L_2	EDM denoising	EDM denoising	L_2	Pseudo-Huber
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Learning rate	2×10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}
Weight decay	10^{-4}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
Residual	✓	—	✓(UNet)	—	—
Dropout	0.1	0.1	0.1	0.1	0.1
EDM $(P_{\text{mean}}, P_{\text{std}}, \sigma_{\text{data}})$	—	$(-1.2,\ 1.2,\ 0.5)$	$(-1.2,\ 1.2,\ 0.5)$	—	$(-1.1,\ 2.0,\ 0.5)$
EDM $(\sigma_{\min}, \sigma_{\max}, \rho)$	—	$(0.002,\ 80,\ 7)$	$(0.002,\ 80,\ 7)$	—	$(0.002,\ 80,\ 7)$
EMA decay	—	—	—	0.999	0
ODE solver	—	Heun	Heun	dopri5	—
Sampling steps or (r_{tol}/a_{tol})	—	100	100	$(10^{-3}/10^{-4})$	1

D METRICS

Let $y \in \mathbb{R}^{H\times W}$ denote the ground-truth high-resolution precipitation field ($H = W = 256$), and let $\hat{y} \in \mathbb{R}^{H\times W}$ denote the prediction from a deterministic baseline. For generative baselines, models generate an ensemble of M realizations, denoted by $\{\hat{y}^{(m)}\}_{m=1}^{M}$ ($M = 8$ across all experiments), with empirical ensemble mean $\bar{y} = \frac{1}{M}\sum_{m=1}^{M} \hat{y}^{(m)}$.

D.1 POINT-WISE AND PROBABILISTIC ERRORS

To determine whether predictions systematically under- or over-estimate precipitation, we compute the mean BIAS. For deterministic baselines, point-wise error is measured via the Mean Absolute

Error (MAE):

$$\text{MAE}(y, \hat{y}) = \frac{1}{|\Omega|} \sum_{i \in \Omega} |y_i - \hat{y}_i|, \qquad \text{BIAS}(y, \hat{y}) = \frac{1}{|\Omega|} \sum_{i \in \Omega} (\hat{y}_i - y_i), \tag{11}$$

where Ω denotes the set of grid cells with non-missing values within the spatial domain. For generative models, the BIAS is computed by substituting $\hat{y}$ for the ensemble mean $\bar{y}$.

For generative models, point-wise errors are evaluated using the MAE's probabilistic generalization: the Continuous Ranked Probability Score (CRPS) (Hersbach, 2000). Since computing the CRPS in its integral form is intractable with a limited number of ensemble members, it is evaluated pointwise at each grid cell i via its energy form representation, given by:

$$\text{CRPS}(y, \{\hat{y}^{(m)}\}_{m=1}^{M}) = \frac{1}{|\Omega|} \sum_{i \in \Omega} \left(\frac{1}{M} \sum_{m=1}^{M} |\hat{y}_i^{(m)} - y_i| - \frac{1}{2M(M-1)} \sum_{m=1}^{M} \sum_{n=1}^{M} |\hat{y}_i^{(m)} - \hat{y}_i^{(n)}| \right). \tag{12}$$

We use this fair estimator, which is unbiased for the CRPS of the distribution the ensemble is drawn from (Ferro, 2013). For a deterministic prediction ($M = 1$), the second term of the CRPS vanishes and the score reduces to the MAE. Therefore, we compare the MAE of deterministic models with the CRPS of generative models. To evaluate sensitivity to the double-penalty problem, BIAS, MAE and CRPS are also computed after spatially pooling fields with average pooling filters of stride and kernel sizes $k \in \{4, 8, 12\}$.

D.2 Scale-dependent skill

The fractions skill score (FSS) Roberts & Lean (2008) compares the fraction of grid cells above a threshold q within a square neighbourhood of size n in the prediction, f_i, and in the observation, o_i:

$$\text{FSS}_{q,n} = 1 - \frac{\sum_{i \in \Omega} (f_i - o_i)^2}{\sum_{i \in \Omega} f_i^2 + \sum_{i \in \Omega} o_i^2}. \tag{13}$$

We use $q \in \{1, 5, 10, 20\}$ mm h^{-1} and $n \in \{2, 4, 8, 16, 32\}$ grid cells. For generative models, f_i is computed from the ensemble mean. The numerator and denominator are summed over all test samples before taking the ratio, so samples with no exceedance of q in either field do not contribute. FSS ranges from 0 to 1, and higher is better.

D.3 Spatial structure

We evaluate the variogram score of order p to assess whether generated high-resolution fields reproduce observed spatial variability, local gradients, and sub-grid structure across scales (Scheuerer & Hamill, 2015). Unlike point-wise metrics, the variogram score is computed through pair-wise differences, making it a more robust scoring rule against the double-penalty problem:

$$\text{VS}_p(y, \{\hat{y}^{(m)}\}_{m=1}^{M}) = \frac{1}{\sum_{i \in \Omega} \sum_{j \in \Omega_i} w_{i,j}} \sum_{i \in \Omega} \sum_{j \in \Omega_i} w_{i,j} \left(|y_i - y_j|^p - \frac{1}{M} \sum_{m=1}^{M} |\hat{y}_i^{(m)} - \hat{y}_j^{(m)}|^p \right)^2, \tag{14}$$

where $\Omega_i = \{j \in \Omega : 0 < \|r_i - r_j\|_2 \leq r_{\max}\}$ defines the neighbourhood of grid cell i up to a maximum separation radius $r_{\max}$. Following standard meteorological practice, we set $p = 0.5$ and use an inverse Euclidean distance weighting scheme, $w_{i,j} = \frac{1}{\|r_i - r_j\|_2}$, evaluated over separation distances up to $r_{\max} = 10$ km (5 grid units on the 2 km target grid).

D.4 Spectral and intensity fidelity

Since high-intensity precipitation events are not well represented in the input *total precipitation* variable from the coarse ERA5 reanalysis, we assess how baselines reconstruct the ground-truth intensity distribution by computing intensity histograms, and the resulting Logarithmic Histogram Distance (LHD). Additionally, km-scale observed precipitation has much finer spatial details than ERA5, often combined with intense precipitation, which results in higher spatial frequencies. To evaluate whether models reproduce these frequencies without introducing spectral artifacts, we compute the Radially Averaged Power Spectral Density (RAPSD).

Intensity histograms and Logarithmic Histogram Distance (LHD). Intensity histograms are computed over B uniformly spaced bins between 0 and 300 mm, with a length of 1 mm per bin (*i.e.,* $B = 300$). Because precipitation distributions are heavy-tailed and dominated by dry grid cells, standard linear probability distances are heavily skewed toward low intensities and obscure errors in heavy rainfall regimes. To benchmark distributional fidelity across both light rain and rare extreme events, we compute the Logarithmic Histogram Distance (LHD). We define $\mathcal{B}_V = \{b : c_b + \hat{c}_b) > 10\}$ as the set of valid bins for computing the LHD, where c_b and $\hat{c}_b$ are the respective counts in bin b for the ground-truth and predictions. Let $p_b(y)$ and $p_b(\hat{y})$ denote the empirical probability mass of observation y and prediction $\hat{y}$ falling into valid bin $b \in \mathcal{B}_V$, normalized over grid cells with non-missing values Ω. The LHD is defined as:

$$\mathrm{LHD}(y, \hat{y}) = \sqrt{\frac{1}{|\mathcal{B}_V|} \sum_{b \in \mathcal{B}_V} (\log_{10} p_b(\hat{y} + \epsilon) - \log_{10}(p_b(y) + \epsilon))^2}, \tag{15}$$

where $\epsilon = 10^{-16}$ is a small regularization constant to prevent undefined logarithms for empty bins. For generative models, $p_b(\hat{y})$ is computed by summing counts across all M ensemble members prior to calculating the bin frequencies.

Radially Averaged Power Spectral Density (RAPSD). We compute the RAPSD Ruzanski & Chandrasekar (2011) to evaluate whether models reproduce the spatial frequencies observed in high-resolution observations, thereby producing realistic km-scale precipitation fields. To avoid introducing high-frequency spectral artifacts from zero-padding or imputation masks, spatial crops containing any missing values are excluded entirely from the spectral evaluation. Given a spatial field $x \in \mathbb{R}^{H \times W}$, we compute its two-dimensional discrete Fourier transform:

$$\hat{X}(k_u, k_v) = \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} x(u, v) \exp\left(-2\pi i \left(\frac{k_u u}{H} + \frac{k_v v}{W}\right)\right), \tag{16}$$

yielding the two-dimensional power spectrum $P(k_u, k_v) = |\hat{X}(k_u, k_v)|^2$. The radially averaged power spectrum $E(k)$ is obtained by integrating $P(k_u, k_v)$ over concentric annuli in wavenumber space:

$$E(k) = \frac{1}{|\mathcal{A}_k|} \sum_{(k_u, k_v) \in \mathcal{A}_k} P(k_u, k_v), \quad \mathcal{A}_k = \left\{(k_u, k_v) : k - \frac{\Delta k}{2} \leq \sqrt{k_u^2 + k_v^2} < k + \frac{\Delta k}{2}\right\}, \tag{17}$$

where k represents the isotropic radial wavenumber and $\Delta k = 1$. The radial wavenumbers correspond to physical spatial wavelengths $\lambda = \frac{L}{k}$, ranging from the domain extent ($L = 512$ km at $k = 1$) down to the Nyquist limit ($\lambda = 4$ km at $k = 128$). For generative baselines, $E(k)$ is computed independently for each ensemble member and averaged across members and evaluation crops.

D.5 Calibration and uncertainty

Evaluating model calibration is critical to ensure that the true uncertainty of the physical process is accurately captured by the spread of the ensemble predictions. To quantify dispersion and compare generative baselines using a single scalar metric, we compute the Spread-Skill Ratio (SSR). However, because the SSR does not reveal the underlying structure of this dispersion, we additionally construct Talagrand rank histograms (Hamill, 2001). By evaluating where observations fall within the ensemble's uncertainty distribution, these histograms allow us to diagnose exactly how and where the model's confidence deviates from the true probabilities.

Spread-Skill Ratio (SSR). The SSR compares the internal model uncertainty against the prediction error. For a perfectly calibrated ensemble, the ensemble spread should equal the Root Mean Square Error (RMSE) of the ensemble mean. The SSR is defined as:

$$\mathrm{SSR} = \sqrt{\frac{M+1}{M}} \frac{\sqrt{\frac{1}{|\Omega|} \sum_{i \in \Omega} \frac{1}{M-1} \sum_{m=1}^{M} (\hat{y}_i^{(m)} - \bar{y}_i)^2}}{\sqrt{\frac{1}{|\Omega|} \sum_{i \in \Omega} (\bar{y}_i - y_i)^2}}, \tag{18}$$

where the numerator is the root mean ensemble variance, the factor $\sqrt{(M+1)/M}$ corrects for the finite ensemble size Fortin et al. (2014), and the denominator is the RMSE of the empirical ensemble mean $\bar{y}$. Variances and squared errors are pooled over all grid cells and test crops before taking the square roots. An SSR ≈ 1 indicates a calibrated ensemble, whereas values strictly less than 1 indicate under-dispersion (over-confidence) and values greater than 1 indicate over-dispersion (under-confidence).

Rank Histograms. Talagrand rank histograms Hamill (2001) evaluate whether ground-truth observations could originate from the predictive ensemble distribution. For each valid grid cell $i \in \Omega$, the observation y_i is compared against the ensemble predictions $\{\hat{y}_i^{(m)}\}_{m=1}^{M}$. The rank $R_i \in \{1, \ldots, M+1\}$ is determined by counting the number of ensemble members strictly less than the observation. However, since precipitation fields are heavily zero-inflated (see Table 2), dry observations would often tie with several members but get classified under the first rank, producing artificial L-shape histograms. We set values below a threshold of 0.1 mm to zero in both predictions and observations. We therefore break ties at random Hamill & Colucci (1997):

$$R_i = 1 + \sum_{m=1}^{M} \mathbb{1}_{\{\hat{y}_i^{(m)} < y_i\}} + U_i, \quad U_i \sim \mathcal{U}\{0, \ldots, T_i\}, \quad T_i = \sum_{m=1}^{M} \mathbb{1}_{\{\hat{y}_i^{(m)} = y_i\}}, \tag{19}$$

where $\mathbb{1}$ is the indicator function. By accumulating R_i across all evaluation samples, we obtain a histogram of ranks. A perfectly calibrated model produces a uniform (flat) histogram. An over-confident model yields a $\cup$-shaped histogram (indicating the ensemble spread is too narrow to cover the true variance), while an over-dispersed model results in a $\cap$-shaped histogram. An L-shape histogram is the result of over-estimation of precipitation values across the entire ensemble spread.

E Additional results

E.1 In-domain performance

E.1.1 Importance-subsampled test datasets

We present comprehensive in-domain evaluation results on the importance-subsampled test datasets across all three continental domains (MRMS, EURADCLIM, and EASTASIA). Table 4 provides numerical values across all evaluated metrics. Figure 7 examines the impact of spatial coarsening (from 2 km down to 24 km effective resolution) on pointwise errors and bias. Figure 8 displays the intensity distribution and spectral fidelity via radially averaged power spectral density (RAPSD). Figure 9 shows scale-dependent Fractions Skill Scores (FSS) across four precipitation thresholds. Finally, Figure 10 presents rank histograms evaluating ensemble calibration for the generative baselines.

Table 4: **In-domain performance evaluation on importance-subsampled test datasets.** MAE, CRPS, and BIAS are computed on the original resolution without artificial coarsening. Probabilistic metrics (SSR and VARIOGRAM) are evaluated exclusively for generative models, omitting the deterministic Bilinear (ERA5) and UNet baselines. Arrows in the header indicate the optimal direction for each metric. The best and second-best results are **bolded** and underlined, respectively. Subscripts denote standard deviations.

		MAE / CRPS (↓)	BIAS (→ 0)	LHD (↓)	SSR (→ 1)	VARIOGRAM (↓)
MRMS	Bilinear	0.7288	-0.0490	19.9497	-	-
	UNet	0.5750 ± 0.0026	-0.4033 ± 0.0069	21.0799 ± 0.1626	-	-
	Diffusion	**0.4593** ± 0.0014	-0.0472 ± 0.0413	4.4261 ± 0.6763	0.9437 ± 0.0103	0.1671 ± 0.0004
	Corrective Diffusion	0.4639 ± 0.0031	**-0.0341** ± 0.0358	**3.3621** ± 1.0940	0.8373 ± 0.0436	**0.1662** ± 0.0007
	SerpentFlow	0.5031 ± 0.0005	0.0353 ± 0.0271	8.5822 ± 0.9519	**1.0268** ± 0.0616	0.1798 ± 0.0012
	Consistency	0.4867 ± 0.0107	0.2825 ± 0.0623	7.0539 ± 0.2034	1.2294 ± 0.0225	0.1825 ± 0.0027
EURADCLIM	Bilinear	0.3815	**-0.0077**	18.5740	-	-
	UNet	0.2953 ± 0.0007	-0.1813 ± 0.0085	21.1591 ± 0.1674	-	-
	Diffusion	0.2444 ± 0.0056	0.0290 ± 0.0861	**6.7102** ± 0.8443	**0.9124** ± 0.0904	0.0926 ± 0.0020
	Corrective Diffusion	**0.2407** ± 0.0012	-0.0208 ± 0.0214	7.4758 ± 0.4153	0.7743 ± 0.0192	**0.0896** ± 0.0002
	SerpentFlow	0.2839 ± 0.0016	0.0537 ± 0.0034	12.4180 ± 0.3756	0.8458 ± 0.0139	0.1048 ± 0.0011
	Consistency	0.2591 ± 0.0045	0.1414 ± 0.0306	10.8817 ± 0.2944	1.1105 ± 0.0268	0.0989 ± 0.0025
EA	Bilinear	1.2581	-0.3655	16.2220	-	-
	UNet	1.0794 ± 0.0077	-0.8937 ± 0.0591	11.2942 ± 0.8373	-	-
	Diffusion	0.8829 ± 0.0074	**-0.0099** ± 0.3823	6.5990 ± 1.2842	**0.9492** ± 0.1648	**0.1867** ± 0.0047
	Corrective Diffusion	0.9292 ± 0.1019	0.1327 ± 0.4484	**6.0387** ± 2.3348	0.8608 ± 0.1742	0.1883 ± 0.0106
	SerpentFlow	0.9375 ± 0.0029	-0.1257 ± 0.0391	7.8810 ± 0.6203	0.7899 ± 0.0456	0.1988 ± 0.0014
	Consistency	**0.8667** ± 0.0051	0.0749 ± 0.1336	12.3931 ± 1.7233	1.0613 ± 0.0574	0.1980 ± 0.0094
AVERAGE	Bilinear	0.7895	-0.1408	18.2486	-	-
	UNet	0.6499 ± 0.0037	-0.4928 ± 0.0248	17.8444 ± 0.3891	-	-
	Diffusion	**0.5289** ± 0.0048	**-0.0093** ± 0.1699	5.9118 ± 0.9349	**0.9351** ± 0.0885	0.1488 ± 0.0024
	Corrective Diffusion	0.5446 ± 0.0354	0.0260 ± 0.1685	**5.6255** ± 1.2814	0.8242 ± 0.0790	**0.1480** ± 0.0039
	SerpentFlow	0.5748 ± 0.0016	-0.0122 ± 0.0232	9.6271 ± 0.6492	0.8875 ± 0.0403	0.1611 ± 0.0012
	Consistency	0.5375 ± 0.0067	0.1663 ± 0.0755	10.1096 ± 0.7404	1.1337 ± 0.0356	0.1598 ± 0.0049

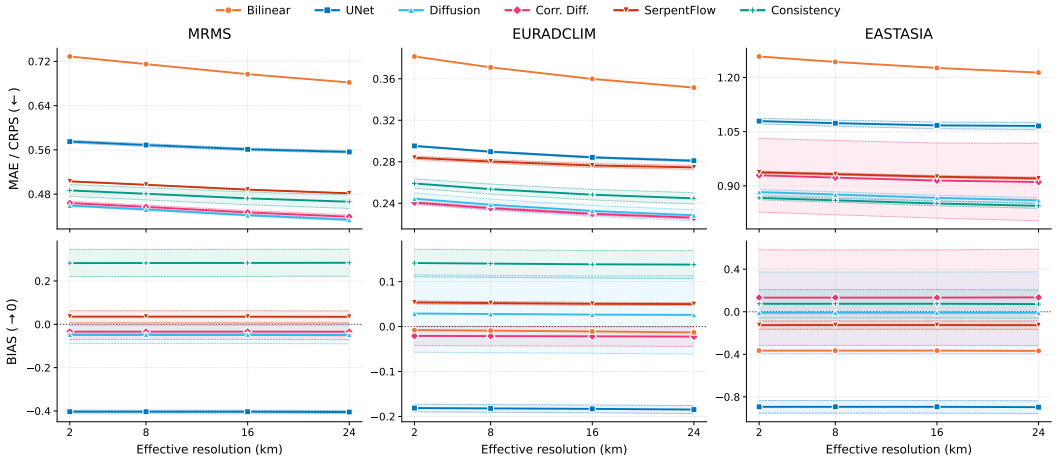


Figure 7: **Effect of spatial coarsening on pointwise error and bias (importance-subsampled test sets).** MAE / CRPS (**top**, ↓) and bias (**bottom**, → 0) evaluated across four effective spatial resolutions (native 2 km, 8 km, 16 km, and 24 km) obtained by average pooling. Shaded bands denote ± standard deviation across three random seeds. Pointwise errors systematically decrease with spatial aggregation due to reduced sensitivity to small-scale displacement errors (double-penalty effect), while bias remains scale-invariant. Relative performance rankings remain unchanged across resolutions.

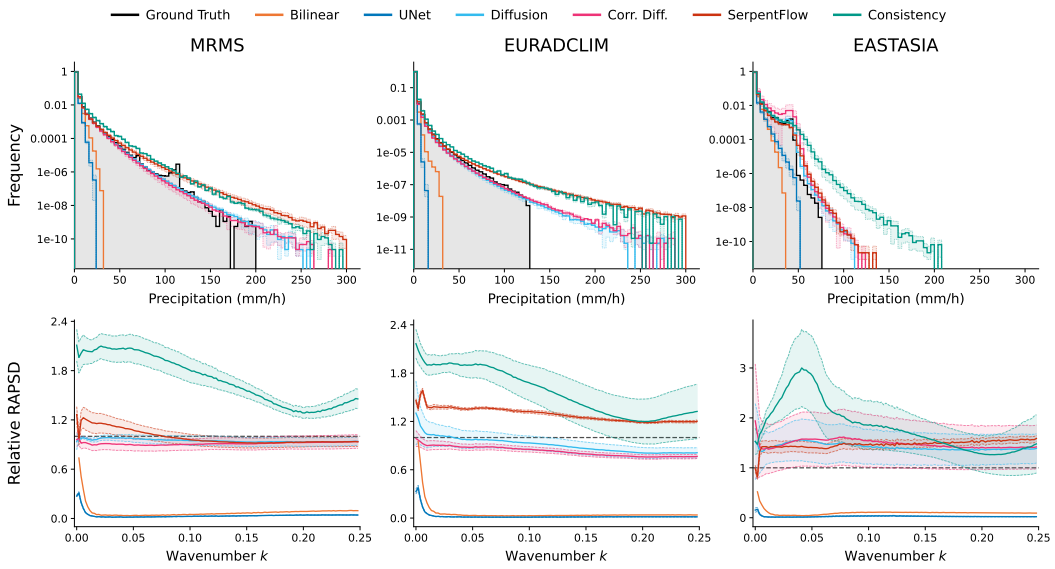


Figure 8: **In-domain intensity distribution and spectral fidelity (importance-subsampled test sets). (Top)** Precipitation intensity histograms comparing model predictions against ground truth up to 300 mm h^{-1} on a logarithmic scale. **(Bottom)** Radially Averaged Power Spectral Density (RAPSD) normalized by observed power spectrum across radial wavenumbers k. The horizontal dashed line at $y = 1.0$ indicates perfect spectral agreement with observations. Shaded bands represent $\pm$ standard deviation across evaluation crops and ensemble members.

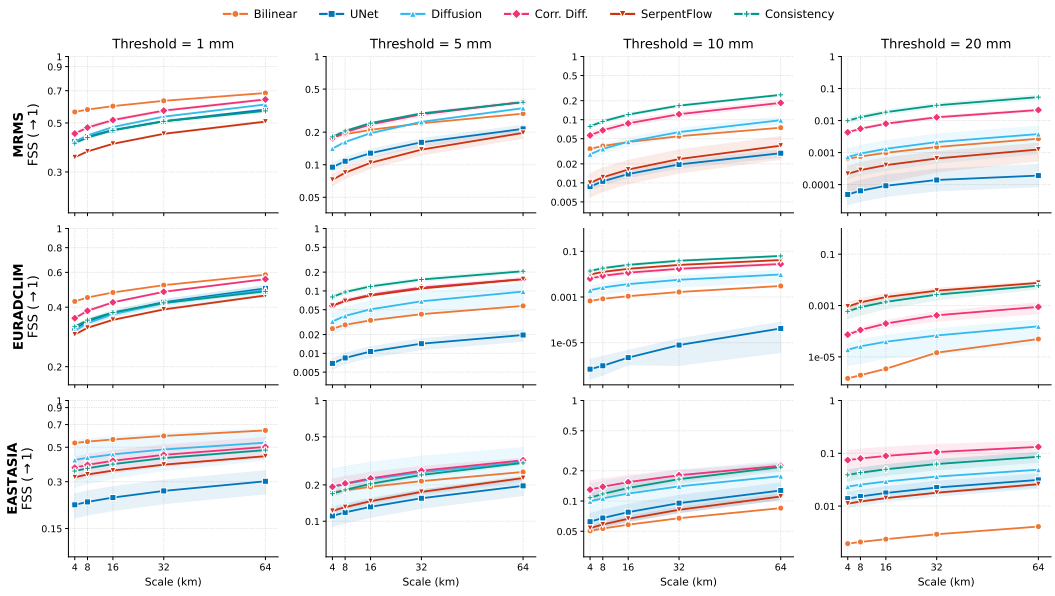


Figure 9: **Fractions Skill Score across precipitation thresholds (importance-subsampled test sets).** Fractions Skill Score (FSS, $\uparrow$ 1) evaluated across neighborhood scales $(4, 8, 16, 32, 64$ km) for thresholds $q \in \{1, 5, 10, 20\}$ mm h^{-1} **(left to right)**. Shaded bands denote $\pm$ standard deviation across three random seeds. Generative baselines, especially diffusion and *corrective* Diffusion, retain high skill for extreme precipitation ($q \geq 5$ mm h^{-1}), whereas deterministic baselines degrade sharply.

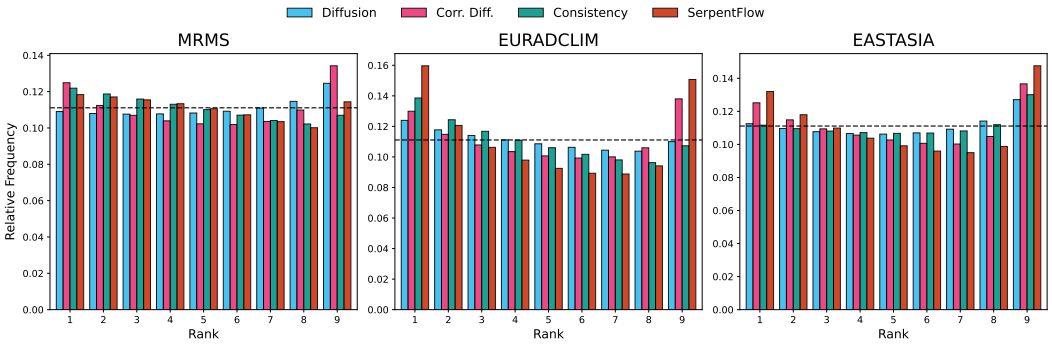


Figure 10: **Ensemble calibration via rank histograms (importance-subsampled test sets).** Talagrand rank histograms for generative models evaluated using an 8-member ensemble (9 rank bins). The horizontal dashed line at $y \approx 0.111$ marks the ideal uniform distribution corresponding to perfect ensemble calibration. Random tie-breaking is applied to values below 0.1 mm h^{-1}.

E.1.2 RANDOMLY-SUBSAMPLED TEST DATASETS

To evaluate model performance under unweighted climatological conditions dominated by dry and light precipitation events, we report in-domain results on the randomly-subsampled test sets. Table 5 lists numerical scores across all baselines. Figure 11 illustrates the sensitivity of MAE/CRPS and bias across spatial aggregation scales. Figure 12 displays intensity histograms and RAPSD curves. Figure 13 details the scale-dependent Fractions Skill Score across intensity thresholds, and Figure 14 presents rank histograms evaluating ensemble calibration under random sampling.

Table 5: **In-domain performance evaluation on randomly-subsampled test datasets.** Details regarding metrics and formatting are identical to Table 4. Probabilistic metrics are evaluated exclusively for generative models. The best and second-best results are **bolded** and underlined, respectively. Subscripts denote standard deviations.

		MAE / CRPS (↓)	BIAS (→ 0)	LHD (↓)	SSR (→ 1)	VARIOGRAM (↓)
MRMS	Bilinear	0.1115	**0.0144**	18.7129	-	-
	UNet	0.0719 $\pm$0.0001	-0.0513 $\pm$0.0011	19.8704 $\pm$0.1937	-	-
	Diffusion	**0.0609** $\pm$0.0003	0.0323 $\pm$0.0063	5.0397 $\pm$0.6527	1.2478 $\pm$0.0158	0.0314 $\pm$0.0004
	Corrective Diffusion	0.0616 $\pm$0.0008	0.0224 $\pm$0.0075	**3.6442** $\pm$1.6567	**1.0549** $\pm$0.0788	**0.0304** $\pm$0.0006
	SerpentFlow	0.0742 $\pm$0.0002	0.0349 $\pm$0.0028	9.2778 $\pm$0.5459	1.2290 $\pm$0.0520	0.0340 $\pm$0.0005
	Consistency	0.0659 $\pm$0.0017	0.0716 $\pm$0.0109	7.3231 $\pm$0.2379	1.4877 $\pm$0.0318	0.0347 $\pm$0.0008
EURADCLIM	Bilinear	0.1091	**0.0182**	18.7759	-	-
	UNet	0.0741 $\pm$0.0020	-0.0379 $\pm$0.0030	19.9062 $\pm$0.4600	-	-
	Diffusion	0.0650 $\pm$0.0053	0.0535 $\pm$0.0360	5.3655 $\pm$1.4171	1.2050 $\pm$0.1036	0.0308 $\pm$0.0031
	Corrective Diffusion	**0.0616** $\pm$0.0009	0.0198 $\pm$0.0072	**5.3510** $\pm$0.1151	**0.9797** $\pm$0.0232	**0.0275** $\pm$0.0004
	SerpentFlow	0.0774 $\pm$0.0008	0.0403 $\pm$0.0019	9.4164 $\pm$0.8482	0.9559 $\pm$0.0145	0.0358 $\pm$0.0008
	Consistency	0.0687 $\pm$0.0017	0.0699 $\pm$0.0107	9.9300 $\pm$0.6042	1.3335 $\pm$0.0390	0.0325 $\pm$0.0015
EASTASIA	Bilinear	0.1609	**0.0291**	14.2227	-	-
	UNet	0.0921 $\pm$0.0002	-0.0793 $\pm$0.0037	16.0623 $\pm$1.2856	-	-
	Diffusion	**0.0845** $\pm$0.0042	0.0758 $\pm$0.0291	7.7125 $\pm$0.4733	1.3349 $\pm$0.0504	**0.0284** $\pm$0.0013
	Corrective Diffusion	0.0889 $\pm$0.0100	0.0970 $\pm$0.1427	**6.4958** $\pm$3.0891	1.1446 $\pm$0.4056	0.0306 $\pm$0.0094
	SerpentFlow	0.1124 $\pm$0.0013	0.0843 $\pm$0.0019	8.9906 $\pm$0.2691	**1.0521** $\pm$0.0231	0.0380 $\pm$0.0011
	Consistency	0.0861 $\pm$0.0028	0.0846 $\pm$0.0184	12.7561 $\pm$1.3035	1.3858 $\pm$0.0477	0.0299 $\pm$0.0017
AVERAGE	Bilinear	0.1272	**0.0206**	17.2372	-	-
	UNet	0.0794 $\pm$0.0007	-0.0562 $\pm$0.0026	18.6130 $\pm$0.6464	-	-
	Diffusion	**0.0701** $\pm$0.0033	0.0539 $\pm$0.0238	6.0392 $\pm$0.8477	1.2626 $\pm$0.0566	0.0302 $\pm$0.0016
	Corrective Diffusion	0.0707 $\pm$0.0039	0.0464 $\pm$0.0524	**5.1637** $\pm$1.6203	**1.0598** $\pm$0.1692	**0.0295** $\pm$0.0035
	SerpentFlow	0.0880 $\pm$0.0008	0.0531 $\pm$0.0022	9.2283 $\pm$0.5544	1.0790 $\pm$0.0299	0.0360 $\pm$0.0008
	Consistency	0.0736 $\pm$0.0020	0.0754 $\pm$0.0133	10.0031 $\pm$0.7152	1.4023 $\pm$0.0395	0.0324 $\pm$0.0013

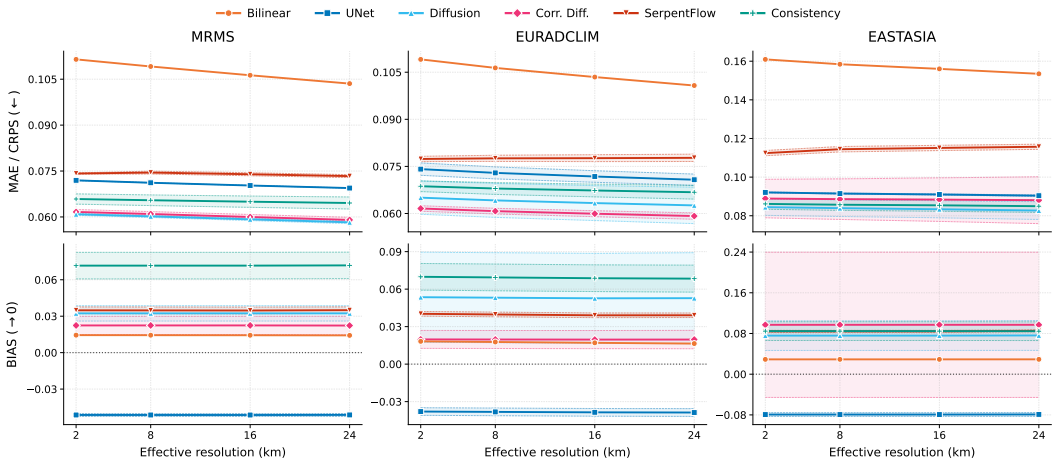


Figure 11: **Effect of spatial coarsening on pointwise error and bias (randomly-subsampled test sets).** Settings and results are similar to Figure 7.

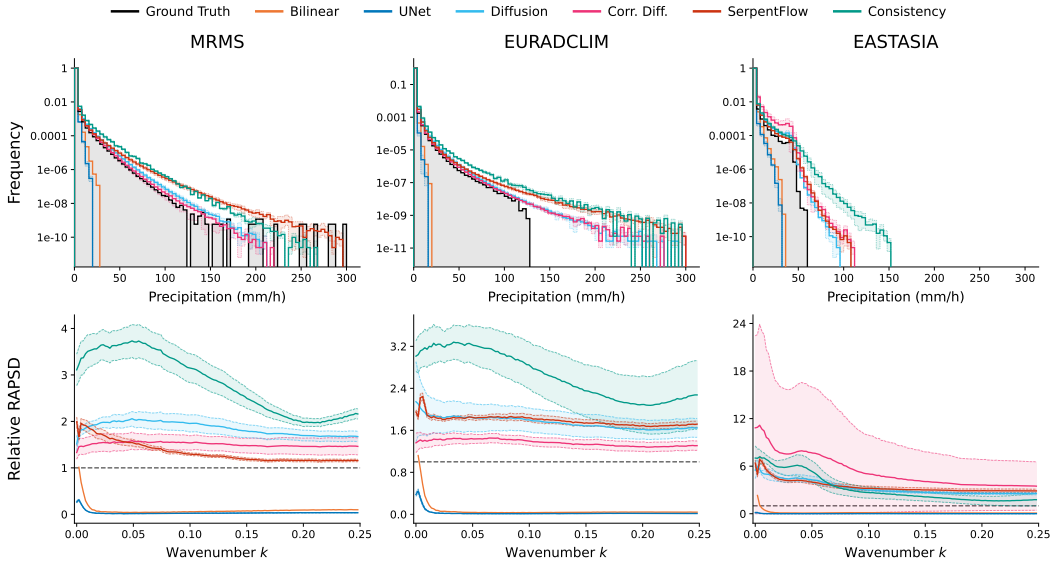


Figure 12: **In-domain intensity distribution and spectral fidelity (randomly-subsampled test sets).** Settings are similar to Figure 8. Generative baselines show much larger RAPSD, with high variance on EA, compared to evaluation on importance-subsampled test datasets.

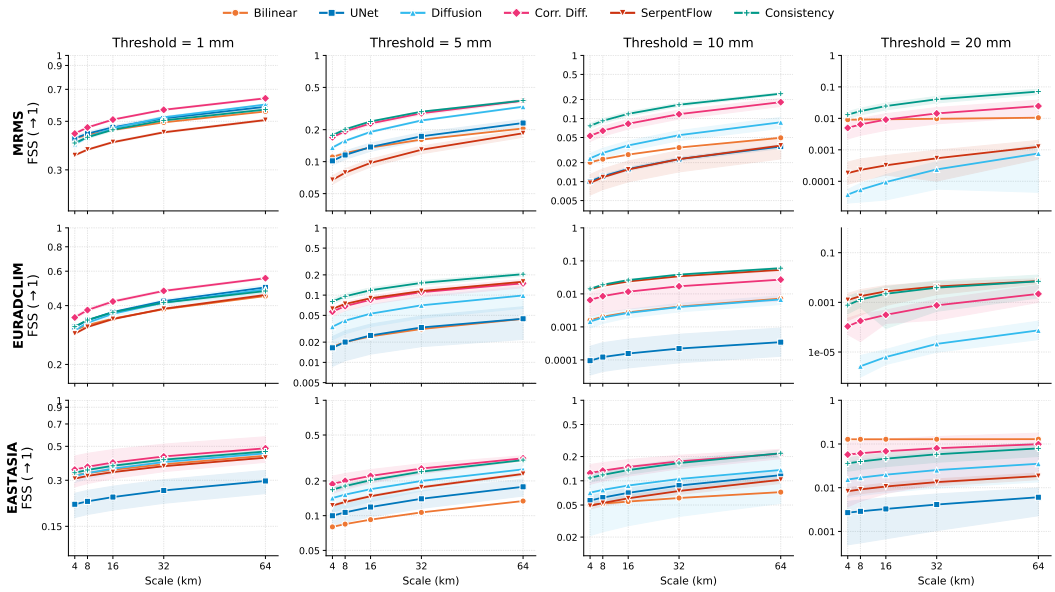


Figure 13: **Fractions Skill Score across precipitation thresholds (randomly-subsampled test sets).** Settings are similar to Figure 9. Generative models maintain notable skill for higher intensity thresholds even under climatologically unweighted random sampling.

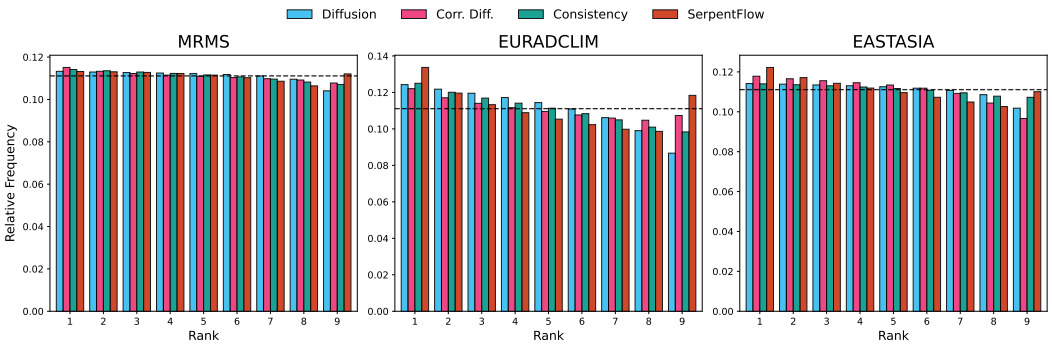


Figure 14: **Ensemble calibration via rank histograms (randomly-subsampled test sets).** Settings are similar to Figure 10. Ensemble calibration patterns remain consistent with the importance-subsampled results.

E.2 GEOGRAPHICAL GENERALIZATION

This section provides extended results on geographical generalization. Figures 15 and 16 present relative degradation heatmaps across all five evaluation metrics for importance-subsampled and randomly-subsampled test sets, respectively. Figure 17 details the resulting precipitation intensity histograms and relative power spectral density curves under geographical shift for importance-subsampled test sets.

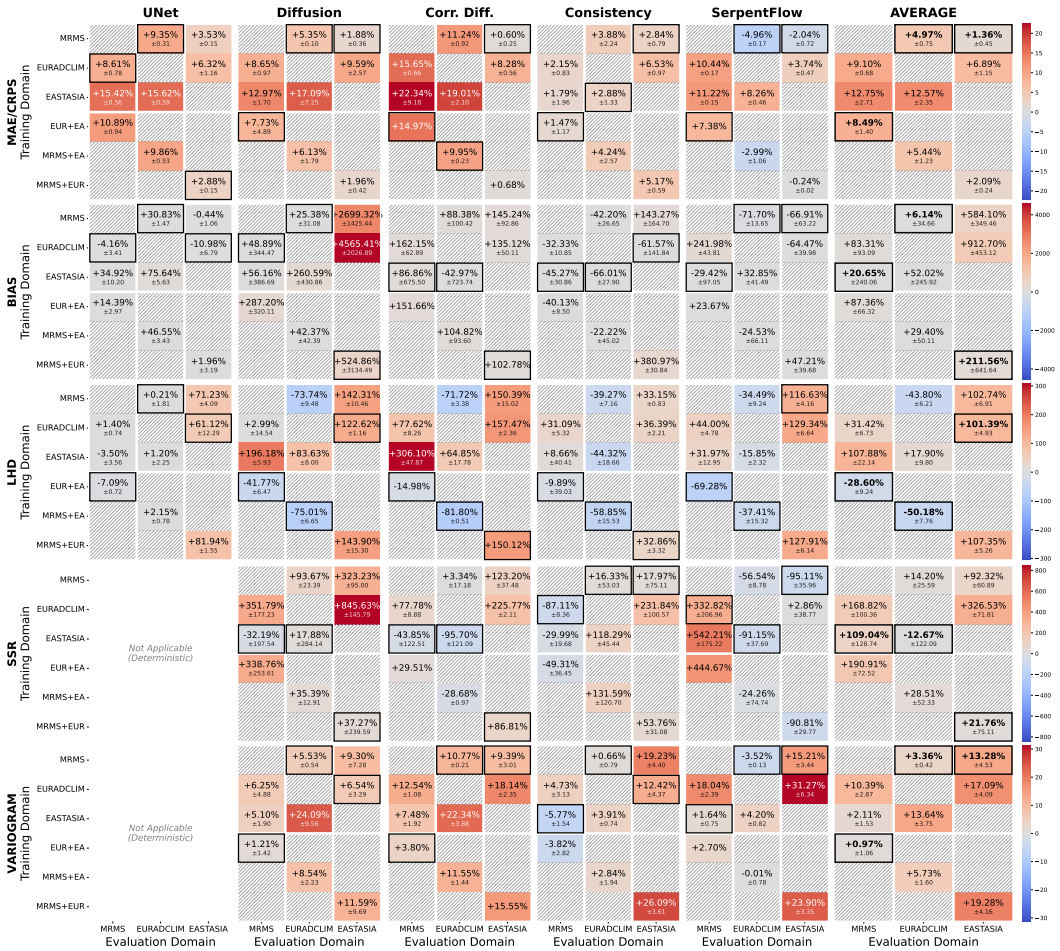


Figure 15: **Out-of-domain generalization relative degradation on importance-subsampled test sets.** Relative degradation (%) with respect to in-domain training across five evaluation metrics (MAE/CRPS, BIAS, LHD, SSR, and VARIOGRAM). Best-performing out-of-domain training domains per column are boxed in black. Scores are averaged over three random seeds, except for configurations with runs still in progress at submission, which are averaged over the seeds available. Full results will be updated during the rebuttal period.

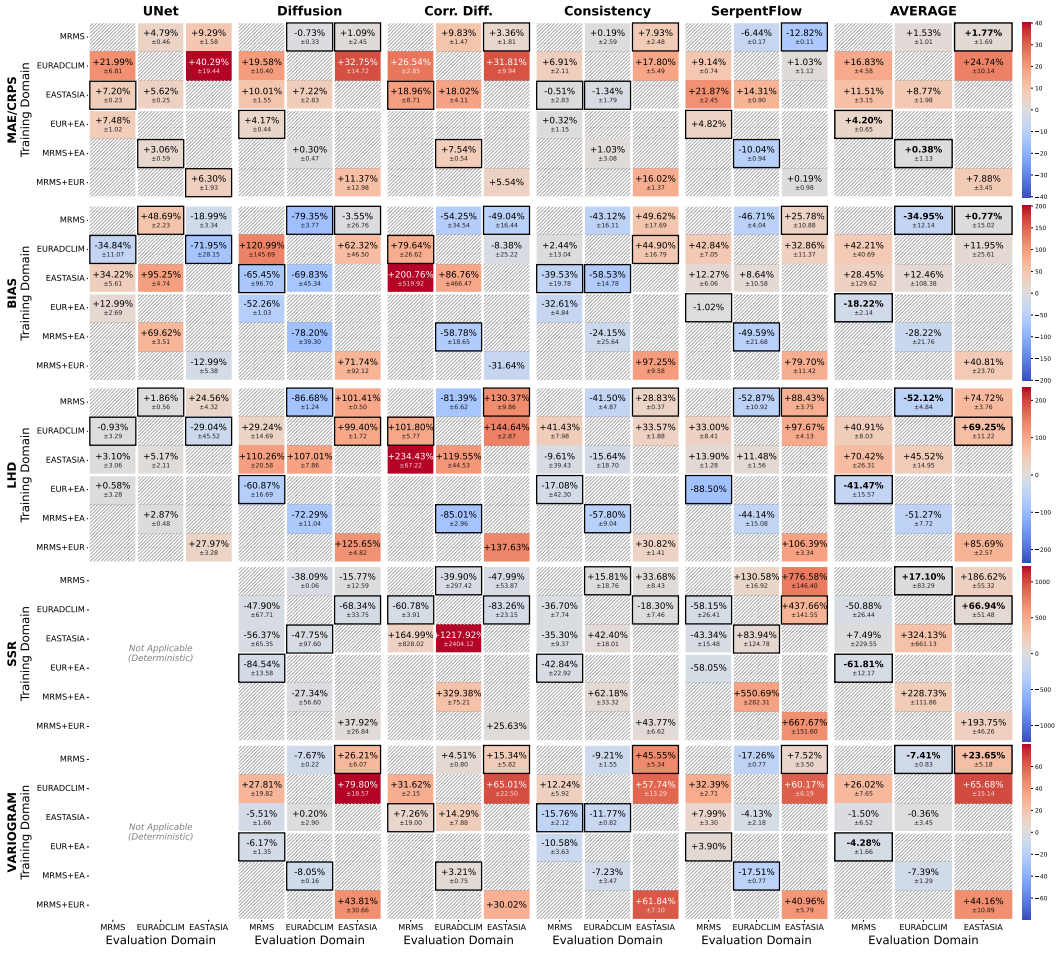


Figure 16: **Out-of-domain generalization relative degradation on randomly-subsampled test sets.** Relative degradation (%) with respect to in-domain training across five evaluation metrics (MAE/CRPS, BIAS, LHD, SSR, and VARIOGRAM). Scores are averaged over three random seeds, except for configurations with runs still in progress at submission, which are averaged over the seeds available. Full results will be updated during the rebuttal period.

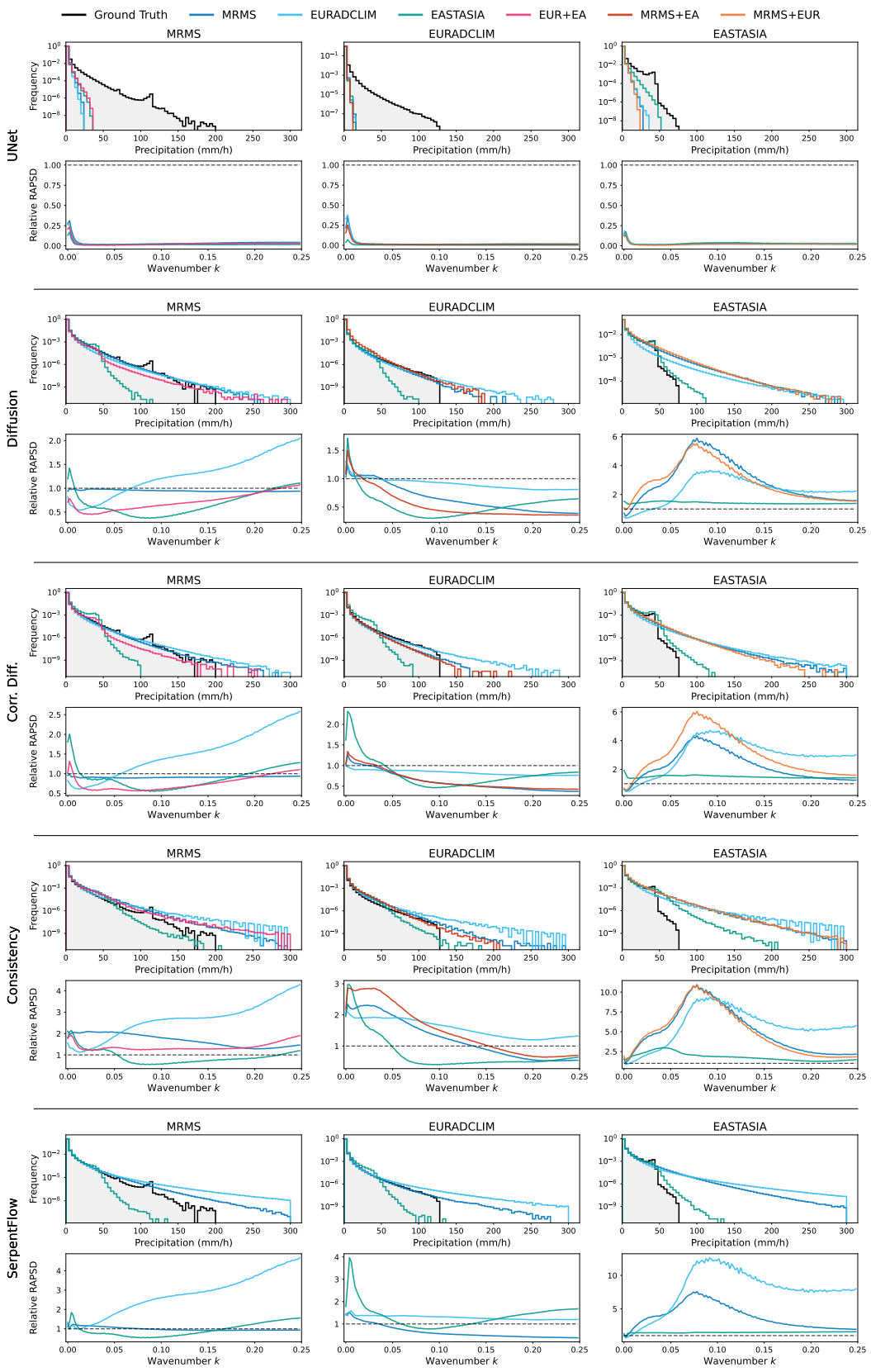


Figure 17: **Intensity histograms and relative RAPSD under geographical generalization (importance-subsampled test sets).** Line colors distinguish the different training domains (MRMS, EURADCLIM, EASTASIA, EUR+EA, MRMS+EA, MRMS+EUR) against ground truth (black).