Climate Dynamics (2026) 64:365
https://doi.org/10.1007/s00382-026-08239-4

ORIGINAL ARTICLE

Generating realistic global precipitation fields from modelled atmospheric circulation

Michael Aich[1,2] · Sebastian Bathiany[1,2] · Philipp Hess[1,2] · Yu Huang[1,2] · Niklas Boers[1,2,3]

Received: 9 July 2025 / Accepted: 14 May 2026 / Published online: 7 August 2026
© The Author(s) 2026

Abstract

Improving the representation of precipitation in Earth system models (ESMs) is critical for assessing the impacts of climate change and especially of extreme events like floods and droughts. In existing ESMs, precipitation is not resolved explicitly, but represented by parameterizations. These typically rely on resolving approximated but computationally expensive column-based physics, not accounting for interactions between locations. They struggle to capture fine-scale precipitation processes and introduce systematic biases. We present a novel approach, based on generative machine learning, which combines a conditional diffusion model with a regression model to generate accurate, high-resolution ($0.25°$) global daily precipitation fields from a small set of prognostic atmospheric variables. Unlike traditional parameterizations, our framework efficiently produces ensemble predictions, capturing uncertainties in precipitation, and does not require fine-tuning by hand. We train our model on the ERA5 reanalysis and present a method that allows us to apply it to unseen ESM data across different climate scenarios, enabling fast probabilistic generation of precipitation fields. By leveraging interactions between global prognostic variables, our approach provides an alternative parameterization approach that mitigates biases present in the ESM precipitation while maintaining consistency with its large-scale (annual) trends. This work demonstrates that complex precipitation patterns can be learned directly from large-scale atmospheric variables, offering a computationally efficient method to obtain high-resolution precipitation without the cost of running the dynamical model at such high resolution.

1 Introduction

Heavy precipitation events will increase in frequency and magnitude in response to global warming (Lee et al. 2023; Prein et al. 2017). Extreme precipitation amplifies flood risks and leads to significant infrastructure, economic, and ecological damages (Davenport et al. 2021). Earth system models (ESMs) are used to simulate precipitation in current and future scenarios to help us understand changes and their impacts on the Earth system and society. Accurate high-resolution precipitation simulations are important for managing the risks of extreme precipitation events by guiding adaptation and risk mitigation measures.

However, simulating precipitation accurately at global scales is difficult because of the wide range of physical processes involved, from microphysics at the cloud scale to large-scale dynamics of the atmosphere. Earth system model simulations are still extremely expensive and time consuming to run, as they need to numerically solve complex partial differential equations describing those processes. This greatly restricts the ability to create large ensemble projections for different climate change scenarios for the next century.

ESMs represent precipitation as a diagnostic variable that is computed from the dynamically resolved prognostic atmospheric variables. ESMs need to approximate the atmospheric circulation on discretized grids with often coarse spatial resolution. This leads to problems in correctly resolving small-scale dynamics that are important for inferring precipitation. Precipitation is a highly complex process

✉ Michael Aich
michael.aich@tum.de

[1] TUM School of Engineering and Design, Department of Aerospace and Geodesy, Earth System Modelling Group, Munich Climate Center, Technical University of Munich, Munich, Germany

[2] Potsdam Institute for Climate Impact Research, Potsdam, Germany

[3] Global Systems Institute and Department of Mathematics, University of Exeter, Exeter, UK

and can not be directly resolved by the ESM, but has to be parameterized. Parameterizations are simplified representations of the subgrid-scale processes that cannot be explicitly resolved due to computational constraints. At each time step, the ESM solves the atmospheric dynamics and applies parameterizations to account for subgrid-scale processes like convection and precipitation, updates the atmospheric state, and proceeds to the next step. Traditional parameterization schemes are often column-based, treating each ESM grid column independently. While this approach is well-justified for many sub-grid processes (Arakawa 2004), it is a simplification that can struggle with processes that depend on horizontal context. For example, neglecting feedback between neighboring columns makes it difficult to capture mesoscale convective organization (Rio et al. 2019). Additionally, the assumed timescales for convection closure can introduce further discrepancies, as convective processes in reality operate on a continuum of timescales that are not well represented in simple quasi-equilibrium closures (Rio et al. 2019). This can introduce systematic biases, leading to over- or underestimation of precipitation intensity and is especially problematic for the ESMs' ability to represent extreme events. The bias is often particularly pronounced in the tropics, a prominent example being the double Intertropical Convergence Zone (ITCZ) bias (Hwang and Frierson 2013), where models produce excessive precipitation in the southern tropics.

We must distinguish between different small-scale processes. Cloud microphysics (e.g. drop nucleation or coalescence), which occur at the micrometer scale, will effectively never be directly resolved in global models. Convection, however, could theoretically be resolved with sufficiently fine grid spacing (e.g. in cloud-resolving models), but doing so for large ensembles at the global scale remains computationally out of reach for the near future due to the prohibitive computational cost associated with such high resolution. Furthermore, many parameterizations are formulated as heuristic equations that may miss key processes and may be to some extent structurally inadequate (Schneider et al. 2017). Parameterizations rely on subjective parameter calibration choices, guided by observed climatologies (Hourdin et al. 2023). This limits their ability to generalize to changing climate conditions (Bordoni et al. 2025) and can create inter-model disagreements (Schneider et al. 2017).

The combination of coarse spatial discretization, simplifications like treating grid columns independently, inadequate functional forms in parameterizations, and subjective tuning all contribute to persistent biases in ESM-simulated precipitation. This makes impact assessments and water resource and flood management, which require precise spatial data at high resolution (Gutmann et al. 2014), extremely challenging. Also, ensembles of high-resolution precipitation projections are imperative for impact assessments but are currently not available due to computational limitations (Bordoni et al. 2025).

Recent studies have shown remarkable success utilizing deep learning (DL) models for weather-forecasting (Bi et al. 2023; Price et al. 2025; Lam et al. 2023; Bodnar et al. 2024; Chantry et al. 2025) and long-term climate simulations (Watt-Meyer et al. 2023). Generative machine learning (ML) methods have been applied for downscaling, bias correction and forecasting tasks, adding the ability to produce ensembles that reflect intrinsic variability (Price et al. 2025; Aich et al. 2025; Hess et al. 2025; Addison et al. 2022; Hess et al. 2022). Most prominent are generative adversarial networks (GANs) (Goodfellow et al. 2020) and more recently diffusion models (Ho et al. 2020). GANs were among the first generative machine learning methods that were successfully applied to weather and climate data (Harris et al. 2022; Leinonen et al. 2020). Diffusion models (DMs) have in recent years often been preferred over GANs because their training is more stable and converges more easily. Also, GANs are prone to suffer from mode collapse, an issue where they are not able to approximate the full distribution, another area where DMs are superior.

There have been efforts using deep learning to replace traditional column-based physical parameterizations in atmospheric models, while preserving the column structure (Rasp et al. 2018; Gentine et al. 2018; Yuval and O'Gorman 2020). In such approaches, more accurate but expensive simulations (e.g. cloud-resolving or large-eddy simulations) serve as training data, capturing small-scale processes that standard parameterizations approximate. Neural networks are then trained to emulate the complex subgrid-scale physics for much less computational cost. For instance, in ClimSim (Yu et al. 2023) each atmospheric column can be treated as an independent regression task, mapping one-dimensional state variables to one-dimensional subgrid-scale tendencies. Despite promising results, a central limitation is that as these methods are trained on simulations they inherit biases present in them. As the neural network training is performed under specific climate conditions, deep-learning-based parameterizations may struggle to generalize to changing climate states in decadal to centennial projections. As the simulations are very expensive, there is limited training data, which can hurt the generalization performance of data-hungry state-of-the-art DL methods. A recent approach (Kochkov et al. 2024) combined a fully differentiable dynamical core with a neural network to emulate small-scale processes in a hybrid model setup designed for weather prediction. In a follow up work the authors introduce an additional neural network that infers precipitation rate from the atmospheric column state for a 2.8° grid spacing of satellite-based precipitation (Yuval et al. 2024).

We propose an alternative to column-based parameterizations, a probabilistic ML-based model that directly predicts high-resolution precipitation from low-resolution global atmospheric fields. A notable difference to column-based methods is that ML-based models can act at global scales to capture spatial interactions between various locations and also between different variables. Our model is trained exclusively on the ERA5 reanalysis, which integrates a wide range of observational data through advanced data assimilation techniques within a high-resolution numerical weather prediction model. This eliminates the need for emulating expensive small-scale simulations, such as those from nested cloud resolving models. Additionally, our proposed approach can also operate as a post-processing tool to derive refined precipitation fields from the atmospheric state once an ESM run has concluded.

We present a two-stage ML framework that operates on a small set of global two-dimensional atmospheric variables, rather than on individual columns, to produce high-resolution daily precipitation at $0.25°$ (for a schematic overview see Fig. 1). Our goal is to show that such a framework can serve as a computationally efficient, probabilistic alternative for modeling precipitation, one that can reduce biases and capture realistic fine-scale spatial patterns. While our model is not intended as an operational replacement for fully coupled three-dimensional parameterizations, it serves as a proof-of-concept demonstrating the potential of spatially aware generative machine learning to perform the role of column-based precipitation parameterizations while addressing key structural limitations of traditional schemes.

2 Methods

Our framework consists of a two-stage deep learning pipeline: (1) a deterministic regression model to estimate $1°$ precipitation from $1°$ atmospheric variables and (2) a conditional diffusion model (DM) to perform bias correction and probabilistic downscaling from $1°$ to $0.25°$ (Fig. 1). Both models are trained exclusively on ERA5 reanalysis data. At

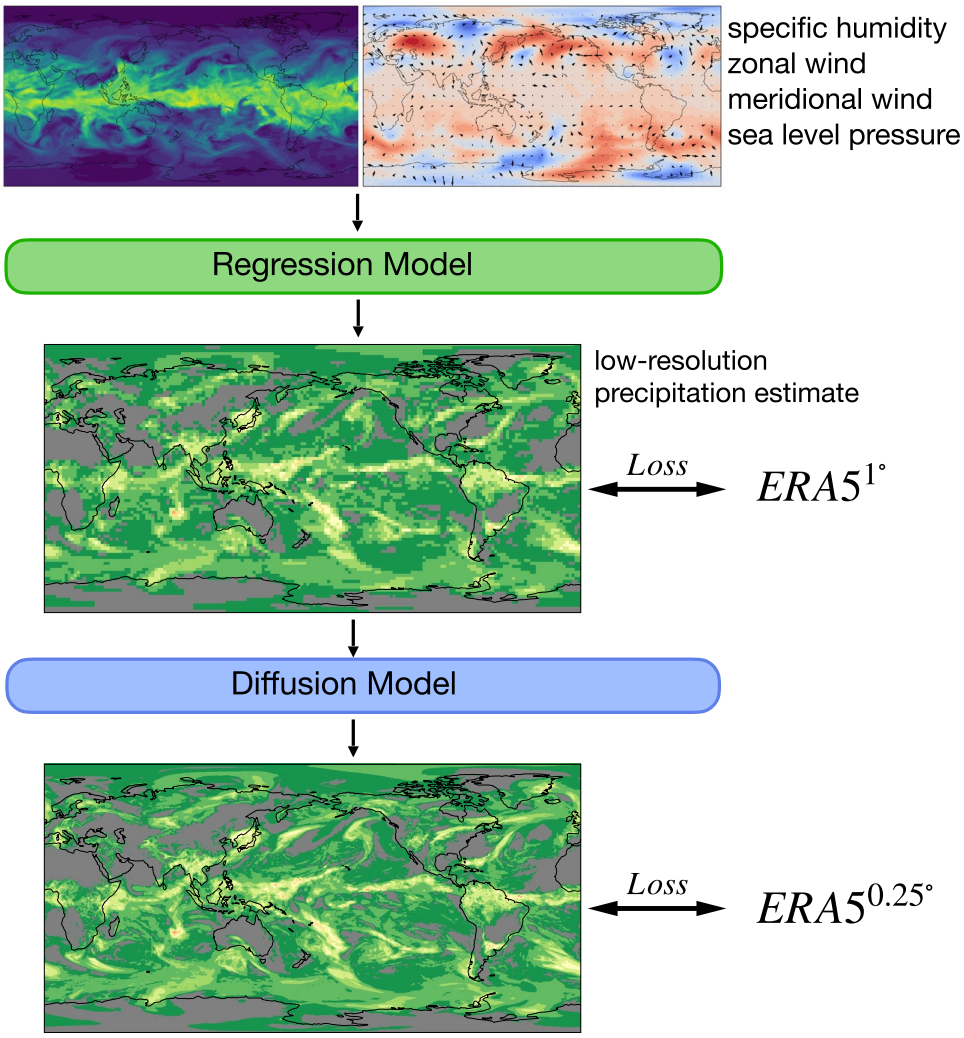


Fig. 1 Schematic overview of our two-stage approach for generating high-resolution global precipitation from atmospheric variables. 1) Deterministic UNet regression model. We first train a regression model to learn the mapping from four atmospheric variables at $1°$ resolution (specific humidity at 850 hPa, near-surface eastward and northward wind components at 10 m, and sea-level pressure) to $1°$ precipitation. During training, both the inputs and the precipitation targets come from ERA5 reanalysis. 2) Generative diffusion model. We train a conditional diffusion model to generate high-resolution ($0.25°$) precipitation fields, conditioned on coarse precipitation fields. To create this training condition, we interpolate ERA5 fields from $0.25°$ to $1°$ and inject noise. The diffusion model is trained to restore small-scale spatial patterns of the high-resolution ERA5 precipitation target. At inference we take atmospheric variables from the ESM and apply the trained UNet regression model to generate a $1°$ deterministic precipitation estimate. Before we use this estimate as a condition for our DM, we apply quantile delta mapping. The diffusion model is then conditioned on noisy quantile mapped precipitation estimates

inference we apply the trained models to ESM data, with specific domain adaptation steps (quantile delta mapping (QDM) and adding noise) to bridge the gap between the reanalysis and ESM data distributions.

2.1 Datasets and pre-processing

2.1.1 Datasets

ERA5: ERA5 (Hersbach et al. 2020) is a state-of-the-art atmospheric reanalysis dataset provided by the European Center for Medium-Range Weather Forecasting (ECMWF). We use ERA5 for training both of our models. The training period is 1980-01-01 to 2018-05-02, with 2018-05-03 to 2020-07-09 used for validation.

GFDL-ESM4: For inference, we use data from the GFDL-ESM4 model (Dunne et al. 2020), developed by the NOAA Geophysical Fluid Dynamics Laboratory (GFDL). GFDL-ESM4 is a member of Phase 6 of the Coupled Model Intercomparison Project (CMIP6). We use the historical period (1970-01-01 to 2014-12-31) and the SSP3−7.0 scenario (O'Neill et al. 2016) (2015-01-01 to 2100-12-31). For simplicity, we refer to the GFDL-ESM4 model output as GFDL from now on.

MPI-ESM: We also evaluate our framework on the MPI-ESM-HR model (Müller et al. 2018). We use daily data at a spatial resolution of $0.9375° \times 0.9375°$, spanning from 1970-01-01 to 2014-12-31. For simplicity, we will from now on refer to the MPI-ESM-HR data as MPI.

2.1.2 Data pre-processing

We use four prognostic variables on daily-resolution as model inputs: specific humidity at 850 hPa, eastward (U) and northward (V) wind components at 10 m, and mean sea level pressure (MSLP). The target variable is daily total precipitation in mm/day. We pre-process the data prior to the train/test split:

- GFDL data is bi-linearly interpolated from its native $1° \times 1.25°$ grid to a $1° \times 1°$ grid.
- MPI data is bi-linearly interpolated from its native $0.9375° \times 0.9375°$ grid to a $1° \times 1°$ grid.
- For precipitation we first add +1 mm/day to the data and then apply a log-transformation (following Hess et al. (2025)).
- All variables are projected to the range $[-1, 1]$ to stabilize training.
- High-resolution (0.25°) ERA5 data is bi-linearly interpolated to 1° (cdo command "remapbil") to create target data for training the regression model.

2.2 Stage 1: Deterministic regression model

Objective: The training objective is to learn a mapping from 1° daily atmospheric variables to 1° daily precipitation.

Architecture: We use the memory-efficient UNet architecture (Saharia et al. 2022). Specifically, the UNet has four input channels, one channel per atmospheric variable, and one output channel for precipitation. We increase the feature dimension across four downsampling stages using multipliers of (1,2,2,2), incorporating a ResNet (He et al. 2016) block at each stage, and include only a single bottleneck attention layer to keep computational demand minimal. This progressively reduces the spatial resolution to capture larger-scale patterns, while simultaneously increasing the number of feature channels. The model has approximately 24 million parameters. Training and inference were performed on a single Nvidia H100 GPU with 80 GB of RAM.

Training: The model is trained on ERA5 data for 2000 epochs with the Adam optimizer (Kingma et al. 2015), using a learning rate of $1e^{-4}$. Training for one epoch takes around 3 min, while inference of 1 field takes around 9 ms. The 1° input data (180×360 grid cells) is padded (zero-padding) to 192×368 grid cells to be compatible with the UNet's downsampling blocks. After doing inference with our model, we index the precipitation output over the original 180×360 grid cells ERA5 domain. The padding and cropping is a purely technical choice that does not introduce artifacts at the prime meridian.

2.3 Stage 2: Probabilistic diffusion model

Objective: The model learns to generate high-resolution (0.25°) precipitation fields, conditioned on a low-resolution (1°) precipitation field.

Architecture: We use a Denoising Diffusion Probabilistic Model (DDPM) (Ho et al. 2020) built upon the Efficient UNet architecture (Saharia et al. 2022), downscaling from a 180×360 grid cell condition to a 720×1440 grid cells output. The diffusion model architecture utilizes a cosine schedule for noising the target data and a linear schedule for the condition during noise condition augmentation. The model has approximately 17 million parameters. Training and inference were performed on a single Nvidia H100 GPU with 80 GB of RAM.

Training: The model is trained for 1100 epochs with the Adam optimizer, a batch size of 1, a learning rate of $1e^{-4}$, and 1000 noising steps. Training for one epoch takes around 44 min, while inference of 1 field takes around 2 s. The diffusion model is trained only on ERA5 data. The 1° precipitation fields that serve as the condition are now bi-linearly interpolated to 0.25° to match the grid cell count (720×1440 grid cells) of the 0.25° precipitation target. To

match the data distribution expected during inference, we add a fixed level of Gaussian noise to the 1° ERA5 condition during training (Sect. 2.4). We add Gaussian white noise that is independent for every grid cell following the forward process of the diffusion model. The added noise corrupts small-scale information in the condition. There is a direct relationship between the magnitude of the added noise and the spatial scales that are corrupted by the noise. Adding more noise (noise with larger amplitude) replaces the original information with noise up to larger spatial scales, which can be seen in a power spectral density (PSD) plot (Aich et al. 2025; Hess et al. 2025). The DM learns to extract large-scale information that is unaffected by the noise and use this information to constrain its conditional generation task. The DM learns to generate the small-scale patterns (which we replaced with noise) in accordance to the large-scale spatial patterns that are unaffected by the noise. The model learns to extract large-scale information from the condition, which is helpful for generating the high-resolution target (as it shares the large-scale patterns with the condition).

2.4 Inference pipeline and domain adaptation

At inference, the goal is to predict precipitation from the atmospheric variables of an ESM. This is a domain adaptation task, as ESM data is biased compared to ERA5 and the train and inference data distributions are not identically distributed. It is crucial to the ML model's generalization performance that training data and inference data are approximately independent and identically distributed (i.i.d.). Our inference pipeline enables the trained ML model to predict precipitation using atmospheric variables from an ESM simulation, rather than the ERA5 data used during training.

2.4.1 Stage 1 inference and large-scale bias correction

First, the pre-processed 1° GFDL atmospheric variables are fed into the trained regression model, producing a deterministic 1° precipitation estimate. This precipitation estimate inherits systematic biases from the GFDL inputs. To correct these large-scale biases, we apply computationally efficient quantile delta mapping (Cannon et al. 2015) to the precipitation output. We fit QDM using the historical period (1980–2014) by mapping the UNet regression output (from historical GFDL atmospheric variables) to the 1° ERA5 precipitation. This learned correction is then applied to the UNet predictions for both the historical period and the future scenarios (2015–2100). The QDM-corrected 1° fields serve as the input for the next stage.

2.4.2 Stage 2 inference and small-scale distribution matching

The 1° QDM-corrected precipitation estimate is bi-linearly interpolated to 720×1440 grid cells to match the training setup. These fields, however, lack the realistic small-scale variability of reanalysis data. Because of that, the diffusion model's training and inference data distributions do not match.

To resolve this, we apply the same amount of noise as used during training (Sect. 2.3) to the regression models' interpolated precipitation output. Adding noise with the same magnitude as in training removes the smooth, biased small-scale patterns from the regression output and replaces them with noise. After applying quantile delta mapping, bilinear interpolation and adding the correct amount of noise to the regression model's precipitation estimate, the data distribution approximately matches the DM's training condition. The amount of noise that is added is chosen before training the diffusion model and represents a hyperparameter. To select this magnitude, we analyze the power spectra intersection of the regression output (bi-linearly interpolated to 0.25°) and the ERA5 fields (interpolated to 1° and then bi-linearly interpolated back to 0.25°). We look at the intersection for the bi-linearly interpolated 1° regression output and bi-linearly interpolated 1° ERA5 because the diffusion model internally interpolates the low-resolution conditioning input to the target resolution before generating high-frequency details. Even after quantile delta mapping, the regression model output and ERA5 fields will align only until a certain threshold scale in the spatial power spectrum, which we call s. We call the spatial scales that are above this point (around 900 km) large scales and the spatial scales below s small scales. The choice of s determines which spatial scales and therefore which size of precipitation events we want to preserve and which we want to regenerate according to the ERA5 distribution. Regenerating small-scale patterns that were overlaid by noise is effectively a bias correction, as long as the spatial patterns are preserved. Without preservation of some patterns in the input it would just be unconditional ERA5 emulation. For a detailed description of the choice of s and how it solves this domain adaptation problem we refer to Aich et al. (2025).

Finally, after quantile delta mapping, bilinear interpolation and noising, the regression model's precipitation is fed into the trained diffusion model as a condition, which iteratively denoises a completely noisy (Gaussian) field. The diffusion model regenerates ERA5-like, high-resolution (0.25°) small-scale patterns consistent with the bias-corrected large-scale structures of the condition. Our inference pipeline takes four atmospheric ESM inputs as condition to generate precipitation fields at a higher spatial resolution while

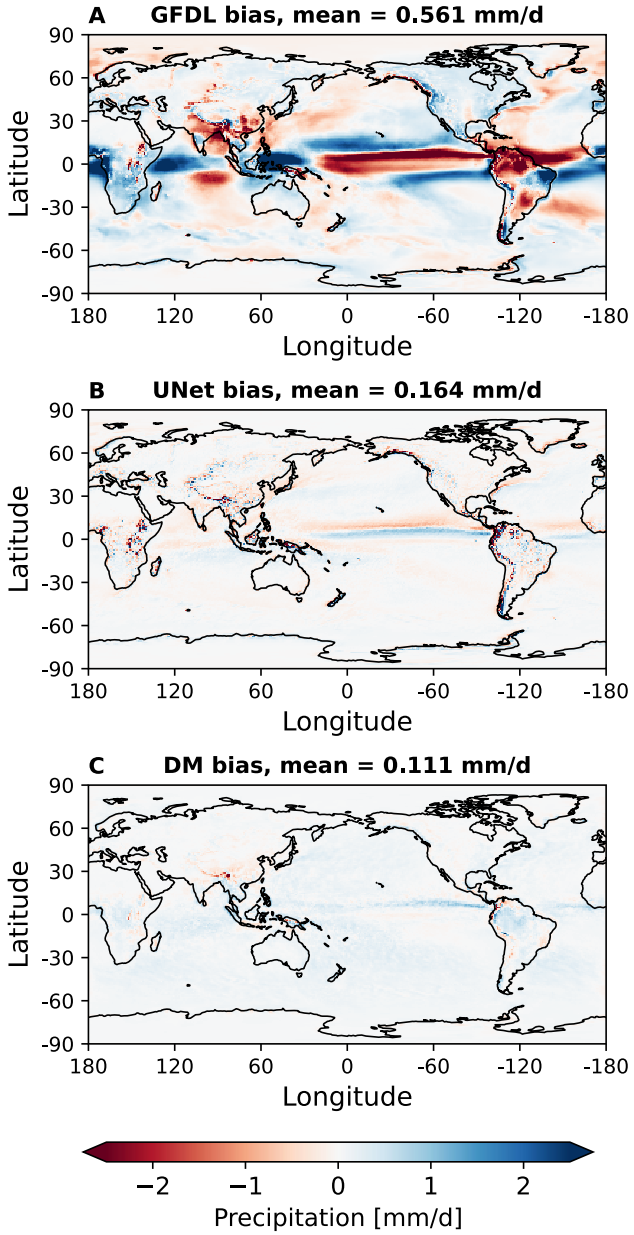


Fig. 2 Comparison of model biases. The maps show the precipitation biases of **A** GFDL, **B** the QDM UNet regression model, and **C** the diffusion model relative to ERA5. The difference map of GFDL shows pronounced bias in the tropical regions, including a double ITCZ. Our regression model already improves over GFDL, while the DM leads to further reduced deviations from ERA5 and the smallest mean absolute bias. Note that we interpolated ERA5 and the DM to 1° by applying average pooling for fair comparison to GFDL

simultaneously mitigating ESM biases. At inference, we use 100 noising steps (down from 1000 in training) which massively reduces inference time.

3 Results

To generate our results, we apply the two-stage inference framework detailed in Sect. 2. First, the UNet regression model estimates 1° precipitation from GFDL atmospheric variables. We then apply quantile delta mapping (QDM) to these fields to correct systematic large-scale biases. Unless otherwise stated, results labeled as 'UNet' or 'regression model' always refer to this QDM-corrected regression model output. The conditional diffusion model then uses this estimate to generate bias-corrected, 0.25° high-resolution ensembles. Visual inspection of representative snapshots (Fig. S1) confirms that the regression model matches the large-scale structure of the ERA5 validation ground truth well (Fig. S1A, B). The individual samples show that the DM preserves the large-scale structures (Figs. S1C, D, F & G) while generating plausible high-resolution details, confirming that the DM output is correctly conditioned on the atmospheric state.

3.1 Correction of spatial biases

The difference plots over 40 years (Fig. 2) reveal that GFDL (Fig. 2A) has by far the largest spatial bias, especially showcasing a strong double ITCZ bias, when compared to our regression model (Fig. 2B) and our DM (Fig. 2C). We aggregated both our DM and ERA5 data from 0.25° to 1° with average pooling for better comparability. The absolute climatologies (Fig. S2) further illustrate these discrepancies. The relative differences are shown in Fig. S3. GFDL has the highest mean absolute bias of 0.561 mm/day. The UNet regression model, even after applying QDM, yields a mean absolute bias of 0.164 mm/day. The diffusion model further reduces this bias to 0.111 mm/day. The GFDL precipitation is heavily biased in the tropics, with substantial regional over- and underestimation compared to ERA5. We also compare the climatologies of ERA5 (Fig. S4A) and our DM precipitation (Fig. S4B) at 0.25° high resolution (GFDL is only available at 1°). The climatologies over 40 years are very similar and only show small differences (Fig. S4C). This is in line with the small mean absolute bias of 0.136 mm/day. Our alternative precipitation estimates and in particular the DM exhibit substantially lower spatial biases compared to GFDL at 1°. Our DM has also only a very small spatial bias at the finer 0.25° ERA5-resolution, where GFDL is not available.

3.2 Improving precipitation statistics

The spatial power spectral density (PSD) shows that GFDL's amplitude of variability on small spatial scales is substantially reduced compared to ERA5 (Fig. 3A). The ESM produces blurry outputs lacking spatial variability at scales below 400 km. At scales between 100 km and 400 km, our regression model is aligned better with ERA5. Our DM can generate precipitation fields at 0.25° resolution that align well with the spatial spectrum of ERA5, demonstrating its ability to correctly model small-scale variability.

The histogram (Fig. 3B) shows that the 1° GFDL model overestimates precipitation between 25 mm/day and 100 mm/day and underestimates precipitation greater than 150 mm/day compared to 0.25° ERA5. Our 1° UNet regression model represents the precipitation more accurately, but also slightly underestimates the small but more frequent precipitation events with less than 25 mm/day. Our 0.25° DM performs best with only a very slight underestimation of extreme events with more than 150 mm/day, similar to the regression model. We confirm the statistical robustness using a bootstrap analysis (Fig. S5). We performed the bootstrap analysis by resampling the 40 years of daily data with replacement 1000 times. For each resampled dataset, we computed the precipitation histogram. The resulting aggregate histograms were used to calculate the 2.5th, 50th, and 97.5th percentiles for each bin. The 95% confidence intervals are generally narrow due to the large sample

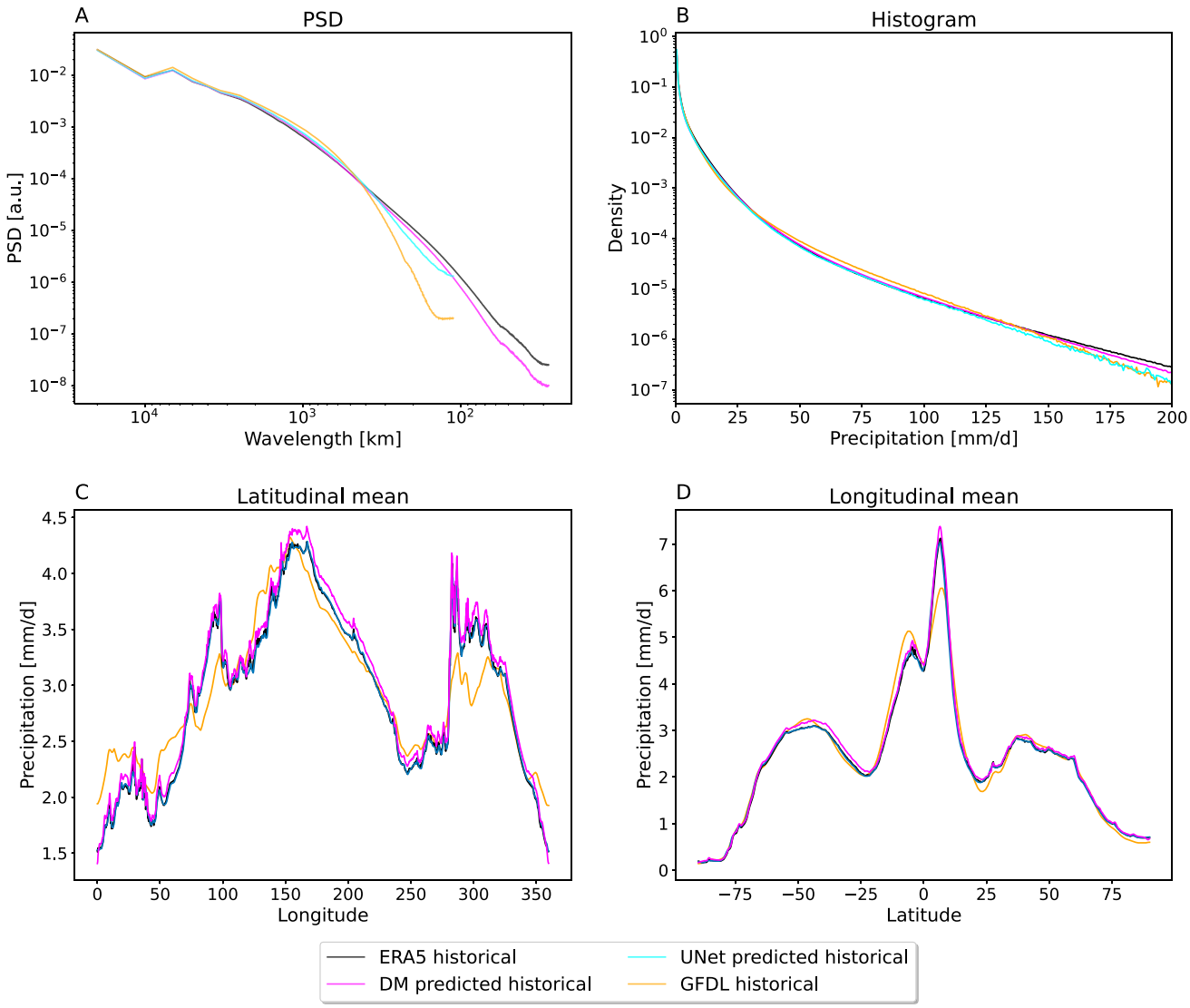


Fig. 3 Reproduction of the statistics of ERA5 at 0.25° resolution over 40 years for the GFDL-ESM. **A** Mean spatial power spectral density (PSD). **B** Histogram indicating the precipitation frequencies. **C** Longitude profile, given by the data averaged over all latitudes, weighted by the cosine of latitude to account for the varying grid-cell area. **D** Latitude profile, given by the data averaged over all longitudes. For panels (**B**–**C**), we bi-linearly interpolate the GFDL and our regression model predictions of precipitation (orange/cyan) from 1° to 0.25° to compare them against the 0.25° ERA5 and DM fields

size, becoming visible only at the extreme tails (>170 mm/day) where event frequencies are low. Notably, the GFDL uncertainty envelope is wider than that of the DM or ERA5, because the coarser 1° resolution of GFDL yields fewer spatial data points compared to the 0.25° high-resolution references.

For the latitudinal and longitudinal mean comparisons (Fig. 3C and D), we bi-linearly interpolated the regression model and GFDL precipitation to 0.25°. The plots show that both of our models align well with the ERA5 statistics, while GFDL sometimes strongly deviates, depending on the region. Because the regression model outputs are post-processed with QDM, they match the ERA5 means almost perfectly. In contrast, the DM shows slight deviations as a result of its stochastic generative process, which prioritizes realistic spatial variability over perfect mean preservation. The differences between the DM, UNet and GFDL compared to ERA5 are further highlighted in Fig. S6.

Furthermore, we demonstrate that our DM framework clearly outperforms a traditional statistical bias-correction and downscaling baseline (Fig. S7), especially in recovering high-frequency spatial details. We further substantiated this by applying our pre-trained model to the MPI-ESM for inference without retraining. The results (Fig. S8) confirm that our approach is not limited to a specific ESM, recovering missing small-scale details and improving the distribution of intermediate precipitation intensities compared to the raw MPI-ESM output (instead of GFDL).

3.3 Reproducing extreme events

To assess whether our diffusion model can adequately reproduce present-day climate extremes and adapt to unobserved future conditions, we compare the R95p extreme precipitation index across historical and projected climates under SSP3−7.0 (Fig. 4). The R95p index is computed as the annual total precipitation from days exceeding the 95th percentile. We use this index to compare historical extreme precipitation of ERA5, our UNet regression model, GFDL and our DM (Fig. S9A–D).

During the historical period (1980–2005), GFDL exhibits pronounced biases in the mid-latitudes and tropics, including a double ITCZ bias, leading to an overestimation of extreme rainfall around the equator (Fig. 4A). Our ML models follow ERA5 closely, exhibiting smaller deviations in both spatial extent and magnitude. The DM in particular reduces the wet bias of GFDL, and is more closely aligned with ERA5. The reference period for the index computation is 1980–2020. To verify the robustness of these findings for rarer extremes, we also evaluated the R99.9p index (Fig. S10). The results confirm that the DM maintains its skill for higher percentiles, also correcting the tropical wet biases of the input while capturing the signal of future intensification.

We further evaluate extreme-event performance by comparing consecutive wet days (CWD) and consecutive dry days (CDD) during the historical period (1980-2005) against ERA5 (Fig. S11). The largest differences appear over the tropics, where pronounced convective precipitation leads to larger deviations in CWD, and over subtropical regions where prolonged dry periods cause sizable shifts in CDD. Both our ML-based approaches, but especially the diffusion model, capture ERA5 patterns of CWD and CDD more accurately, clearly outperforming GFDL in representing CWD and CDD patterns.

3.4 Ensemble evaluation

A key strength of our generative DM approach compared to traditional deterministic column-based parameterizations is its ability to generate a diverse ensemble of possible, spatially consistent high-resolution precipitation realizations for a given low-resolution condition. This accounts for the aleatoric uncertainty inherent in inferring fine-scale details from coarse variables.

It is critical to evaluate the probabilistic skill of our diffusion model in the downscaling task. We investigate whether the uncertainty in a DM ensemble aligns with the inherent uncertainty of downscaling the data from 1° to 0.25° . Note that the diffusion model also performs bias correction by regenerating small-scale variability. We condition our DM on a 1° bi-linearly interpolated noisy ERA5 validation year and evaluate the DM 50 times for each day in the year, resulting in a 50-member DM ensemble of one-year trajectories. The downscaling ground truth will be the respective high-resolution ERA5 year. We compare the continuous ranked probability score (CRPS) of our 50-member DM downscaling with 1° ERA5 and the QDM-corrected 1° regression model's output, both bi-linearly interpolated to 0.25°. For this specific experiment, the QDM correction was fitted using the UNet output derived from coarse (1°) ERA5 inputs to ensure a consistent evaluation.

This evaluation is conducted over a one-year period from the ERA5 validation dataset, starting on 2018-05-03. This comparison assesses how well the DM's ensemble spread matches the inherent uncertainty. Our DM outperforms both deterministic baselines in terms of temporally averaged CRPS (Fig. 5). Note that for deterministic baselines, the CRPS reduces to the mean absolute error. The mean CRPS of our DM is the lowest at 0.56 mm/day, better than that of the ERA5 interpolated from 1° to 0.25° (0.73 mm/day) and the interpolated QDM-corrected prediction (1.42 mm/day). The baselines filter out the spatial variability between 0.25° and 1°, while our DM is able to reconstruct it. The spatially

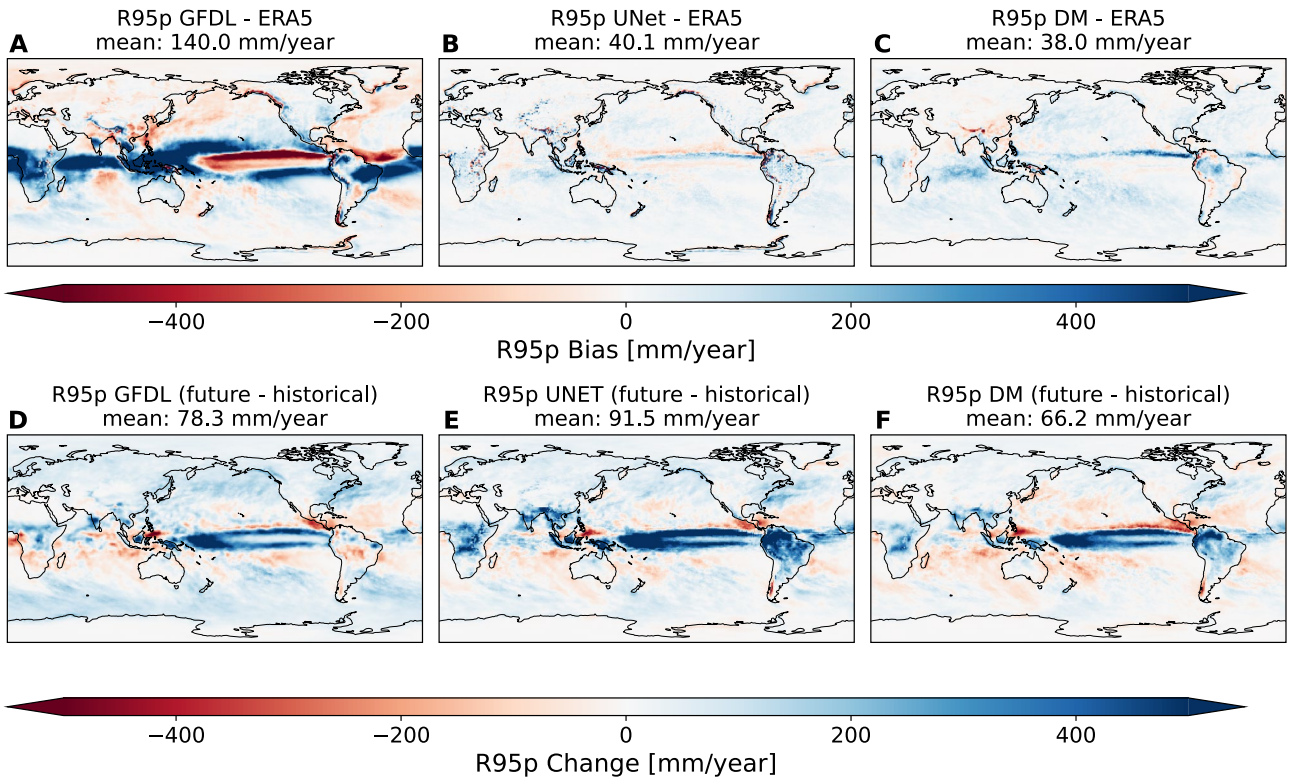


Fig. 4 Evaluating extreme event coverage for historical and future scenarios. We compare the total precipitation from days exceeding the 95th percentile of daily rainfall (R95p). **A**–**C**: Bias in historical data (1980–2005). **A** GFDL bias (GFDL - ERA5), **B** UNet bias (UNet - ERA5), and **C** DM bias (DM - ERA5). ERA5 and the DM were interpolated to 1° to match the resolution of GFDL. Positive values (blue) indicate overestimation of extreme precipitation by the respective model, while negative values (red) indicate underestimation. The GFDL model exhibits strong wet biases in the tropics during the historical period (**A**), whereas the DM and UNet models show reduced biases (**B**, **C**). Bottom row (**D**–**F**): Projected changes in R95p between the late 21st century (2075–2100, SSP3−7.0) and the historical period. **D** GFDL future change (future – historical), **E** UNet future change, and **F** DM future change. Positive values (blue) indicate an intensification of extreme precipitation, while negative values (red) indicate reductions. Under future warming, all models predict an increase in R95p over the tropics. While the UNet predicts the largest mean increase, the DM moderates this signal, resulting in a more conservative intensification compared to GFDL. The area-weighted mean absolute error (**A**–**C**) and mean change (**D**–**F**) are indicated in the subplot titles

averaged CRPS (Fig. 5) also shows that our model has the best scores overall. The spread-skill relationship (calculated following Haynes et al. 2023) roughly follows the 1:1 line, indicating an accurate, well-calibrated representation of uncertainty (Fig. 5). For larger spreads the RMSE is slightly too low (underconfident). To verify that the diffusion model accurately captures large-scale precipitation patterns, we also evaluated performance at the 1° scale by interpolating the predictions (Fig. S12). The relative performance ranking remains consistent with the high-resolution results. This confirms that the DM accurately captures large-scale structures while generating realistic small-scale variability.

3.5 Precipitation generation for climate projections

In the following, we apply our method to future climate projections. This is a challenging task, given that the underlying data distribution shifts over time in response to changing climate conditions. A comparison between our regression model (bi-linearly interpolated to 0.25°) and the diffusion model (Fig. S13) for SSP3−7.0 data between 2015–2100 shows that our DM is still able to increase the spatial resolution of the precipitation predictions from 1° to 0.25° with ERA5-like large-scale patterns and small-scale deviations consistent with those observed in the historical period (Fig. S13A).

3.5.1 Trend preservation

We specifically tested the ability of our diffusion model to extrapolate to an unseen future climate scenario, namely SSP3−7.0, which projects relatively high greenhouse gas emissions and therefore increasing specific humidity. First, we initialize our regression model with the SSP3−7.0 atmospheric variables and then apply quantile delta mapping. We find that the regression model predicts an increasing precipitation trend, probably driven largely by the higher specific humidity (Fig. 6 and Fig. S14). The spatial trend at

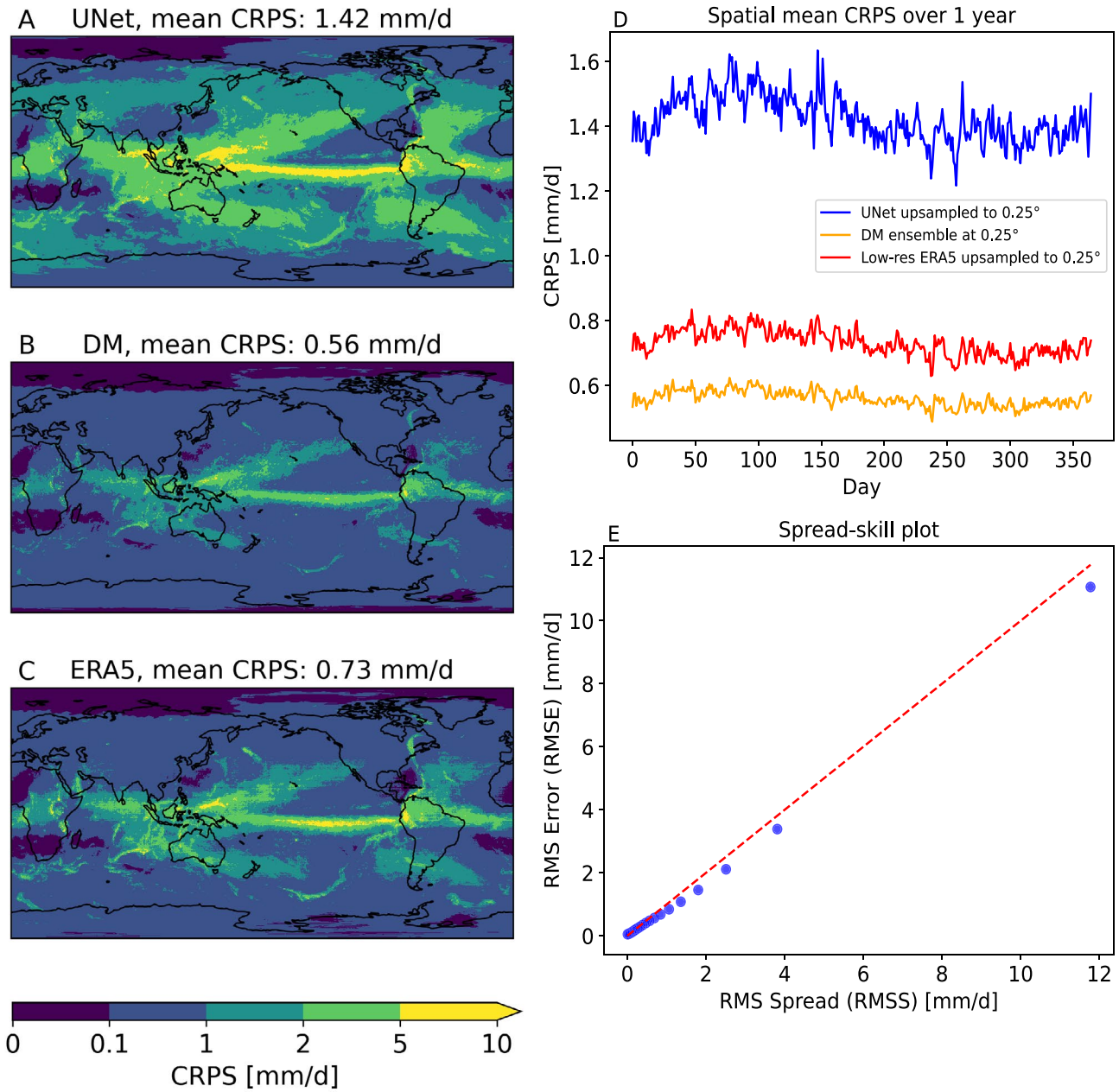


Fig. 5 Evaluation of the 50 DM ensemble members over one ERA5 reference year. Temporally averaged continuous ranked probability score (CRPS) (lower is better) for (**A**) applying our deterministic UNet, followed by QDM and then bi-linearly interpolating from 1° to 0.25°, **B** our DM and **C** the bi-linearly interpolated deterministic ERA5 baseline (from 1° to 0.25°). **D** Spatially averaged CRPS. **E** Spread-skill plot. Following the 1:1 line indicates well-calibrated spread of the ensemble

each location (Fig. S15A–C) reveals that applying the DM on the QDM UNet precipitation preserves the spatial patterns of the trend. While the spatial patterns are similar, the DM corrects the local structures and intensities. The spatial patterns of trends predicted by the regression model differ in sign from the GFDL projections in some regions. The spatially averaged annual mean trend reveals that the UNet regression model has a different annual mean precipitation and exhibits an overly large variance, resulting in too large trend. This high inter-annual variance is the direct result of a systematic spatial error in the daily fields, which is further amplified when QDM is applied to these overly smooth outputs. The regression model has to make predictions in an out-of-distribution scenario for future scenario GFDL fields and therefore produces fields with flawed spatial structures and unrealistic intensities. Quantile delta mapping cannot fix this problem entirely. While QDM can correct the long-term climatological average, by adjusting the mean precipitation,

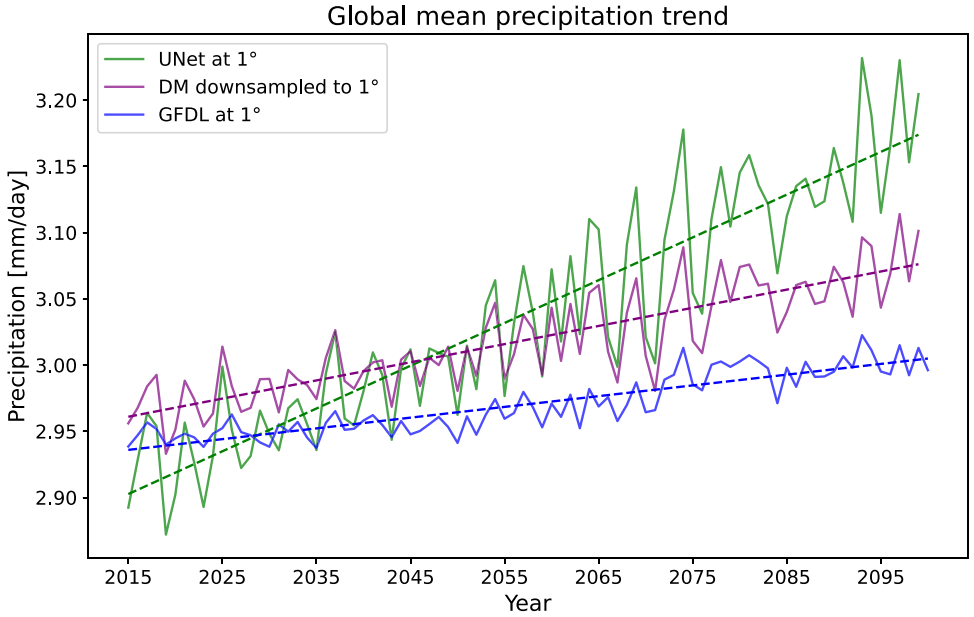


Fig. 6 Precipitation trends under the SSP3−7.0 scenario. Comparison of globally and annually averaged precipitation trends for the SSP3−7.0 scenario for GFDL (blue), our UNet regression model (green) and our DM (purple). For comparison, the DM was interpolated to 1°. The dashed lines represent the linear regression trends for each model. The spatial and globally averaged trend of the UNet (even after quantile delta mapping) differs from GFDL. The DM corrects the global average trend of the UNet, aligning it more with GFDL. The DM trend is slightly larger than that of GFDL, while the mean values are slightly lower. Note that the GFDL trend is not a strict ground truth but rather serves as a reference for comparison, as the DM performs a correction of the data towards the ERA5 distribution

it cannot correct the flawed small scales. The DM, however, will generate correct small-scale information, reducing the variance in the quantile delta mapped regression output and thereby correcting the overly steep annual mean trend. This is especially relevant for the far future (post 2055), where the UNet predictions have increasingly larger variability, resulting in a steep trend. When we correct the regression model's precipitation with our DM, the DM dampens the variance such that the annual precipitation trend matches the trend of the GFDL SSP3−7.0 projection more closely (Fig. 6). After applying the DM, there is only a small deviation in the globally averaged mean values and trend compared to the GFDL projection. It is important to note that the GFDL simulation represents just one possible realization of a future climate scenario and acts as a reference rather than a ground truth target. Our method relies on the assumption that the statistical relationship between large-scale circulation and small-scale precipitation variability, as learned from ERA5, remains valid under future warming. Consequently, our DM robustly adapts to these future scenarios by preserving the large-scale structures projected by GFDL while generating consistent, ERA5-like small-scale variability.

To verify that trend preservation is a robust feature of our framework and not limited to a specific model or time period, we analyzed the combined historical and future period from 1970–2100 for both GFDL-ESM and the independent MPI-ESM (Fig. S14). The evaluation confirms that trend preservation is maintained across the full 1970–2100 period for both ESMs. In both cases, the quantile delta mapped regression model produces unrealistically high variance (and thus steeper trends) which the DM corrects. Crucially, the experiment reveals that our framework does not simply reproduce the ERA5 trend during the historical period. Instead, the DM-corrected trend aligns with the respective ESM rather than forcing the data to match the steeper historical trend of ERA5, confirming that our framework adapts to the specific input climate signal rather than overfitting to the training distribution. Consequently, our framework allows us to correct spatial patterns while preserving the overarching global trend information of the driving atmospheric variables. We note, however, that the output of our generation routine does not strictly replicate the native precipitation trend of the driving ESM. The regression model infers precipitation by mapping the climate change signal of the ESM's atmospheric variables through the observational relationships learned from ERA5. The generative diffusion step is crucial because the intermediate quantile delta mapping applied to the initial regression predictions introduces an unrealistically high inter-annual variability (see Fig. S14). This variance inflation is an artifact that occurs when grid-wise quantile mapping is applied to overly smooth deterministic model outputs (Maraun 2013). The diffusion process naturally dampens this QDM artifact. The native ESM projections in Fig. 6 and Fig. S14

are therefore shown for reference rather than as an absolute ground truth. Overall, Fig. S14 demonstrates that our framework yields consistent predictions for two different ESMs with the slopes of the DM-corrected lines remaining largely consistent with the respective reference projections.

We note that although SSP3−7.0 is an out-of-distribution test in terms of temperature increases, in practice it represents a moderate distribution shift for precipitation, characterized primarily by an increased frequency of high values rather than entirely novel conditions outside the historical range.

3.5.2 Spatial climate change signal

Figure S16 shows the relative signal of climate change in precipitation (2075–2100 vs. 1980–2005) relative to the historical period for GFDL (Fig. S16A), the UNet regression model (Fig. S16B), and our diffusion model (Fig. S16C). Overall, both UNet and DM capture large-scale changes while offering more moderate local responses than GFDL. In the tropics and especially the ITCZ region, GFDL exhibits a stronger positive precipitation response compared to both data-driven methods, which generally temper the climate change signal. In particular, in parts of South America, the sign of the projected change switches from positive in GFDL to almost zero or slightly negative for the UNet and DM. As we show that the overall trend of precipitation is well approximated in statistical metrics and time series trends, this could suggest that the ML-based models correct for some of the potential biases of GFDL, resulting in different and, in some cases, opposite regional precipitation trends.

Under the future warming scenario SSP3−7.0 (between 2075–2100), all models predict an increase in R95p in the tropics relative to the historical period (Fig. 4D–F). Because model biases can alter the magnitude of these changes, we compare the sign of the change across the plots, which is largely consistent among GFDL, UNet regression, and DM. Overall, the UNet projects the largest mean intensification of precipitation extremes. However, the DM produces a more moderate increase than both UNet and GFDL, with some regional disagreements, particularly over South America. The double ITCZ bias remains visible in the future projections of GFDL. Visually, both GFDL and DM show stronger extremes within the tropical belt, although GFDL changes are larger in magnitude overall, possibly reflecting its known wet bias. Our DM reproduces similar spatial patterns (signs of change) but with a more modest amplitude, suggesting that it partially corrects the climatological biases of GFDL, while retaining the overarching signal of climate change.

4 Discussion and conclusions

We introduced a two-stage framework for inferring high-resolution global precipitation fields from a small set of atmospheric variables, composed of a deterministic UNet regression model together with a probabilistic conditional diffusion model. Our generative machine learning approach hence serves as a proof-of-concept for spatially aware generative machine learning as an alternative to deterministic column-based parameterization of precipitation for ESMs.

We chose this two-stage framework to map between the relatively smooth large-scale atmospheric inputs and the spatially intermittent, high-resolution precipitation target. We found that learning this complex mapping in a single generative step is challenging. This difficulty was also identified by Mardani et al. (2025), who also employ a two-stage approach to address this. A key advantage of our approach is that both our models are trained exclusively on a single dataset, in our case ERA5 reanalysis data. This training strategy ensures that the framework is not tied to the specific biases or parameterizations of any specific ESM. The only hyperparameter that must be adapted to the specific ESM is the noise level added to the DM's condition. While applying the framework to ESMs with significantly different resolution or biases requires retraining with a noise level calibrated to that specific ESM, we empirically confirmed that for ESMs with similar spectral characteristics (GFDL and MPI), our model works well without retraining. For ESMs with coarser resolution or stronger small-scale biases than those used for calibration (GFDL), not adapting the noise level could result in incomplete removal of small-scale biases.

Our setup aligns with the so-called perfect model approach (van der Meer et al. 2023), i.e. learning to reconstruct high-resolution target fields from artificially coarsened input fields. Recent studies (van der Meer et al. 2023; Rampal et al. 2024) suggest that the imperfect model approach, mapping low-resolution ESM output directly to high-resolution observations, can be advantageous for implicitly correcting ESM biases. However, it requires temporal correspondence between input and target data. Such paired data is available when using regional climate models (RCMs) that are driven by ESM boundary conditions. Since our problem setup involves mapping a free-running ESM to ERA5 reanalysis, which represent two independent realizations of internal variability without alignment, the imperfect strategy is not applicable.

Another advantage of our framework is its computational efficiency. With inference times of roughly two seconds per global 0.25° precipitation prediction (on a single Nvidia H100 GPU), our framework is fast compared to running an ESM at 0.25° resolution.

This efficiency is partly due to our deliberate choice of a minimal set of two-dimensional predictor variables. While traditional parameterizations rely on the full vertical profiles, our results suggest that modern deep learning models can extract implicit information about the three-dimensional atmospheric state from surface fields. For instance, MSLP captures the integrated mass of the atmospheric column, while our surface winds are intrinsically coupled to the larger-scale three-dimensional circulation. Our evaluation confirms that these two-dimensional fields contain sufficient information to reconstruct high-resolution precipitation patterns without explicitly providing the vertical dimension.

Our approach efficiently produces ensembles of high-resolution precipitation fields from low-resolution (1°) atmospheric variables, which faithfully represent the inherent uncertainty in the downscaling process. Our method can also be used in post-processing for generating large ensembles and long-term precipitation projections, which are essential for impact and risk assessments.

Our evaluation demonstrates that our method substantially reduces the bias compared to our ESM reference. Our approach successfully avoids the prominent double ITCZ bias which is strongly pronounced in the GFDL-ESM. Our method also reproduces accurate small-scale spatial patterns with which ESMs often struggle. The power spectral density we confirm that the small-scale spatial variability of the downscaled precipitation aligns closely with the ERA5 reanalysis data, which is an improvement over the comparatively blurry GFDL precipitation fields. The spatial variability of small scales is essential for accurate simulation of regional precipitation extremes and associated impact assessments.

While ERA5 is used as the ground truth target in this study, we acknowledge that reanalysis precipitation data contains inherent biases. Specifically, ERA5 has a wet bias and can underestimate the intensity of extreme precipitation events (Lavers et al. 2022). However, given its global coverage, high resolution and physical consistency, ERA5 remains the standard dataset for training machine learning models. Our framework is data-agnostic and can be retrained on observational datasets for specific operational applications.

Importantly, our framework largely preserves the climate change signal inherent in the specific humidity under the SSP3−7.0 scenario, while correcting small-scale spatial patterns in accordance with reanalysis data. Compared to the quantile delta mapped UNet regression model, our diffusion model adjusts the spatial variability of precipitation more effectively, aligning it with the native GFDL projections, with only slight deviations in the global mean. However, GFDL should not be considered as ground truth in our setup, as the diffusion model performs a correction toward ERA5, balancing consistency with climate model trends and alignment with observational data. In contrast, our quantile delta mapped UNet regression model alone fails to reproduce the mean and predicts too high variance. An essential part of our method is to add noise to the quantile delta mapped regression model's 1° output before conditioning the diffusion model on it. This removes the flawed small-scale features while preserving large-scale dynamics, effectively acting as a bias correction for the small-scales. We show that the diffusion model preserves the spatial patterns of the trend (i.e. regions of increase and decrease) predicted by the UNet, while refining the magnitude of these changes by regenerating realistic small-scale variability. By regenerating improved, reanalysis-like, small-scale structures, the DM aligns the global, spatially averaged, trend with the trend of the GFDL projection. Our model demonstrates robustness to out-of-distribution scenarios, as shown by its ability to preserve trends and to capture extreme events under the SSP3−7.0 scenario. Overall, evaluating the model's performance on future scenarios is challenging due to the lack of high-resolution ground truth. Future research could compare against dynamically downscaled datasets like Regional Climate Models (RCMs), although such models introduce their own biases.

Generally, we acknowledge that our model generates reanalysis-like fine-scale details based on the historical training data. This implicitly assumes that the relationship between large-scale states and small-scale structures remains largely constant under climate change. Investigating this relationship could be an avenue of research in future work.

Our method is trained on a frame-by-frame basis, having no information from neighboring time steps. The temporal consistency of the large-scale precipitation patterns is inherited directly from the evolution of the atmospheric variables, which give us temporal consistency for large-scale features. However, we acknowledge that the fine-scale features generated by the diffusion model are stochastic and not explicitly constrained to be temporally consistent from one time step to the next. It is worth noting, however, that such consistency is, in the case of precipitation, primarily relevant for high-frequency data (e.g. sub-daily) where substantial autocorrelation exists. For our daily precipitation fields, the temporal autocorrelation of these fine-scale features is less pronounced compared to sub-daily scales. While our approach captures spatial statistics and large-scale dynamics, extending the framework to a video diffusion model to enforce temporal consistency for fine-scale structures remains a valuable direction for future development, for applications involving sub-daily data.

An important practical benefit of our framework producing 0.25° resolution is that it enables coupling the

precipitation outputs to impact or land models without requiring a higher resolution for the 1° atmospheric variables. This decoupling of scales can substantially lower computational costs while still providing the detailed precipitation fields needed for impact studies.

Several avenues for future research remain. An important one is to integrate our DM-based parameterization directly into a climate model and evaluate its stability in a fully coupled system. Such experiments can help determine whether further constraints on the UNet outputs are necessary when deployed in operational settings. This includes incorporating physical constraints to ensure that conservation laws are fulfilled (Harder et al. 2023), or adopting alternative loss functions to better capture specific features such as precipitation extremes (Doury et al. 2024).

Bridging the gap from our proof-of-concept to a fully operational parameterization would also require extending the framework to predict three-dimensional tendencies, potentially incorporating vertical profiles and ensuring its coupling with other physical processes. This is non-trivial given the known challenges regarding stability and drift for this task (Brenowitz and Bretherton 2018; Yuval and O'Gorman 2020). Optimizing the inference strategy with techniques such as model distillation (Luhman and Luhman 2021) could be explored to accelerate the sampling process and further reduce computational cost.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00382-026-08239-4.

Funding Open Access funding enabled and organized by Projekt DEAL. Funded under the Excellence Strategy of the Federal Government and the Länder through the TUM Innovation Network EarthCare. This is ClimTip contribution #140; the ClimTip project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101137601: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. The Past to Future (P2F) project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101184070: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. PH, SB, and NB acknowledge funding by the Volkswagen Foundation. YH acknowledges funding by the Alexander von Humboldt Foundation.

Data availability The ERA5 reanalysis was accessed from https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_DAILY. The GFDL-ESM4 and MPI-ESM-HR data is available at https://esgf-data.dkrz.de/search/cmip6-dkrz/.

Code availability The code is available on GitHub at https://github.com/aim56009/precipitation_from_atmosphere.git.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Addison H, Kendon E, Ravuri S, Aitchison L, Watson PA (2022) Machine learning emulation of a local-scale uk climate model, arXiv:2211.16116,

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T, Ho J, Fleet D, Norouzi M (2022) Photorealistic text-to-image diffusion models with deep language understanding. Adv Neural Inf Process Syst 35:36479–36494

Aich M, Hess P, Pan B, Bathiany S, Huang Y, Boers N (2025) Conditional diffusion models for downscaling & bias correction of earth system model precipitation. EGUsphere 2025:1–21

Arakawa A (2004) The cumulus parameterization problem: past, present, and future. J Clim 17(13):2493–2525

Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q (2023) Accurate medium-range global weather forecasting with 3d neural networks. Nature 619(7970):533–538

Bodnar C, Bruinsma WP, Lucic A, Stanley M, Brandstetter J, Garvan P, Riechert M, Weyn J, Dong H, Vaughan A, et al (2024) Aurora: a foundation model of the atmosphere, arXiv:2405.13063

Bordoni S, Kang SM, Shaw TA et al (2025) The futures of climate modeling. npj Clim Atmos Sci 8:99

Brenowitz ND, Bretherton CS (2018) Prognostic validation of a neural network unified physics parameterization. Geophys Res Lett 45(12):6289–6298

Cannon AJ, Sobie SR, Murdock TQ (2015) Bias correction of gcm precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? J Clim 28(17):6938–6959

Chantry M, Lang S, Alexe M, Dramsch J, Raoult B, Clare M, Santa Cruz M, Hahner S, Adewoyin R, Pinault F, et al. (2025) Aifs–ecmwf's data-driven forecasting system, In: 105th AMS Annual Meeting, AMS

Davenport FV, Burke M, Diffenbaugh NS (2021) Contribution of historical precipitation change to us flood damages. Proc Natl Acad Sci 118(4):e2017524118

Doury A, Somot S, Gadat S (2024) On the suitability of a convolutional neural network based rcm-emulator for fine spatio-temporal precipitation. Clim Dyn 62(9):8587–8613

Dunne JP, Horowitz LW, Adcroft AJ, Ginoux P, Held IM, John JG, Krasting JP, Malyshev S, Naik V, Paulot F, Shevliakova E, Stock CA, Zadeh N, Balaji V, Blanton C, Dunne KA, Dupuis C,

Durachta J, Dussin R, Gauthier PPG, Griffies SM, Guo H, Hallberg RW, Harrison M, He J, Hurlin W, McHugh C, Menzel R, Milly PCD, Nikonov S, Paynter DJ, Ploshay J, Radhakrishnan A, Rand K, Reichl BG, Robinson T, Schwarzkopf DM, Sentman LT, Underwood S, Vahlenkamp H, Winton M, Wittenberg AT, Wyman B, Zeng Y, Zhao M (2020) The GFDL earth system model version 4.1 (GFDL-ESM 4.1): overall coupled model description and simulation characteristics. J Adv Model Earth Syst, 12(11): e2019MS002015

Gentine P, Pritchard M, Rasp S, Reinaudi G, Yacalis G (2018) Could machine learning break the convection parameterization deadlock? Geophys Res Lett 45(11):5742–5751

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144

Gutmann E, Pruitt T, Clark MP, Brekke L, Arnold JR, Raff DA, Rasmussen RM (2014) An intercomparison of statistical downscaling methods used for water resource assessments in the united states. Water Resour Res 50(9):7167–7186

Harder P, Hernandez-Garcia A, Ramesh V, Yang Q, Sattegeri P, Szwarcman D, Watson C, Rolnick D (2023) Hard-constrained deep learning for climate downscaling. J Mach Learn Res 24(365):1–40

Harris L, McRae AT, Chantry M, Dueben PD, Palmer TN (2022) A generative deep learning approach to stochastic downscaling of precipitation forecasts. J Adv Model Earth Syst, 14(10): e2022MS003120

Haynes K, Lagerquist R, McGraw M, Musgrave K, Ebert-Uphoff I (2023) Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. Artif Intell Earth Syst 2(2):220061

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15(5):559–570

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M, De Chiara G, Dahlgren P, Dee D, Diamantakis M, Dragani R, Flemming J, Forbes R, Fuentes M, Geer A, Haimberger L, Healy S, Hogan RJ, Hólm E, Janisková M, Keeley S, Laloyaux P, Lopez P, Lupu C, Radnoti G, de Rosnay P, Rozum I, Vamborg F, Villaume S, Thépaut J-N (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146(730):1999–2049 (https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803)

Hess P, Drüke M, Petri S, Strnad FM, Boers N (2022) Physically constrained generative adversarial networks for improving precipitation fields from earth system models. Nat Mach Intell 4(10):828–839

Hess P, Aich M, Pan B, Boers N (2025) Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning, Nature Machine Intelligence, pp. 1–11

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Process Syst 33:6840–6851

Hourdin F, Ferster B, Deshayes J, Mignot J, Musat I, Williamson D (2023) Toward machine-assisted tuning avoiding the underestimation of uncertainty in climate change projections. Sci Adv 9(29):2758

Hwang Y-T, Frierson DM (2013) Link between the double-intertropical convergence zone problem and cloud biases over the southern ocean. Proc Natl Acad Sci 110(13):4935–4940

Kingma D, Ba J (2015) Adam: a method for stochastic optimization, In: International Conference on Learning Representations (ICLR), (San Diega, CA, USA)

Kochkov D, Yuval J, Langmore I, Norgaard P, Smith J, Mooers G, Klöwer M, Lottes J, Rasp S, Düben P et al (2024) Neural general circulation models for weather and climate. Nature 632(8027):1060–1066

Lam R, Sanchez-Gonzalez A, Willson M, Wirnsberger P, Fortunato M, Alet F, Ravuri S, Ewalds T, Eaton-Rosen Z, Hu W et al (2023) Learning skillful medium-range global weather forecasting. Science 382(6677):1416–1421

Lavers DA, Simmons A, Vamborg F, Rodwell MJ (2022) An evaluation of era5 precipitation for climate monitoring. Q J R Meteorol Soc 148(748):3152–3165

Lee H, Calvin K, Dasgupta D, Krinner G, Mukherji A, Thorne P, Trisos C, Romero J, Aldunce P, Barret K, Blanco G, Cheung WW, Connors SL, Denton F, Diongue-Niang A, Dodman D, Garschagen M, Geden O, Hayward B, Jones C, Jotzo F, Krug T, Lasco R, Lee Y-Y, Masson-Delmotte V, Meinshausen M, Mintenbeck K, Mokssit A, Otto FE, Pathak M, Pirani A, Poloczanska E, Pörtner H-O, Revi A, Roberts DC, Roy J, Ruane AC, Skea J, Shukla PR, Slade R, Slangen A, Sokona Y, Sörensson AA, Tignor M, van Vuuren D, Wei Y-M, Winkler H, Zhai P, Zommers Z, Hourcade J-C, Johnson FX, Pachauri S, Simpson NP, Singh C, Thomas A, Totin E, Arias P, Bustamante M, Elgizouli I, Flato G, Howden M, Méndez-Vallejo C, Pereira JJ, Pichs-Madruga R, Rose SK, Saheb Y, Rodríguez RS, Ürge-Vorsatz D, Xiao C, Yassaa N, Alegría A, Armour K, Bednar-Friedl B, Blok K, Cissé G, Dentener F, Eriksen S, Fischer E, Garner G, Guivarch C, Haasnoot M, Hansen G, Hauser M, Hawkins E, Hermans T, Kopp R, Leprince-Ringuet N, Lewis J, Ley D, Ludden C, Niamir L, Nicholls Z, Some S, Szopa S, Trewin B, van der Wijst K-I, Winter G, Witting M, Birt A, Ha M, Romero J, Kim J, Haites EF, Jung Y, Stavins R, Birt A, Ha M, Orendain DJA, Ignon L, Park S, Park Y (2023) Ipcc, 2023: Climate change 2023: synthesis report, summary for policymakers. contribution of working groups i, ii and iii to the sixth assessment reportof the intergovernmental panel on climate change [core writing team, H. Lee and J. Romero (eds.)]. Ipcc, Geneva, Switzerland., Technical Report https://doi.org/10.59327/IPCC/AR6-9789291691647.001, Intergovernmental Panel on Climate Change (IPCC), Geneva, Switzerland, 2023. Please access report for extended list of writing team, editors, scientific steering committee and other contributing authors

Leinonen J, Nerini D, Berne A (2020) Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. IEEE Trans Geosci Remote Sens 59(9):7211–7223

Luhman E, Luhman T (2021) Knowledge distillation in iterative generative models for improved sampling speed. arXiv:2101.02388

Maraun D (2013) Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. J Clim 26(6):2137–2143

Mardani M, Brenowitz N, Cohen Y, Pathak J, Chen C-Y, Liu C-C, Vahdat A, Nabian MA, Ge T, Subramaniam A et al (2025) Residual corrective diffusion modeling for km-scale atmospheric downscaling. Communications Earth & Environment 6(1):124

van der Meer M, de Roda Husman S, Lhermitte S (2023) Deep learning regional climate model emulators: a comparison of two downscaling training frameworks. J Adv Model Earth Syst, 15(6): e2022MS003593

Müller WA, Jungclaus JH, Mauritsen T, Baehr J, Bittner M, Budich R, Bunzel F, Esch M, Ghosh R, Haak H et al (2018) A higher-resolution version of the max planck institute earth system model (mpi-esm1. 2-hr). Journal of Advances in Modeling Earth Systems 10(7):1383–1413

O'Neill BC, Tebaldi C, Van Vuuren DP, Eyring V, Friedlingstein P, Hurtt G, Knutti R, Kriegler E, Lamarque J-F, Lowe J et al (2016) The scenario model intercomparison project (scenariomip) for cmip6: choices and design of the climate change scenario simulations. Geosci Model Develop 9(9):3461–3482

Prein AF, Rasmussen RM, Ikeda K, Liu C, Clark MP, Holland GJ (2017) The future intensification of hourly precipitation extremes. Nat Clim Chang 7(1):48–52

Price I, Sanchez-Gonzalez A, Alet F, Andersson TR, El-Kadi A, Masters D, Ewalds T, Stott J, Mohamed S, Battaglia P et al (2025) Probabilistic weather forecasting with machine learning. Nature 637(8044):84–90

Rampal N, Hobeichi S, Gibson PB, Baño-Medina J, Abramowitz G, Beucler T, González-Abad J, Chapman W, Harder P, Gutiérrez JM (2024) Enhancing regional climate downscaling through advances in machine learning. Artif Intell Earth Syst 3(2):230066

Rasp S, Pritchard MS, Gentine P (2018) Deep learning to represent subgrid processes in climate models. Proc Natl Acad Sci 115(39):9684–9689

Rio C, Del Genio AD, Hourdin F (2019) Ongoing breakthroughs in convective parameterization. Curr Clim Change Rep 5(2):95–111

Schneider T, Lan S, Stuart A, Teixeira J (2017) Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations. Geophys Res Lett 44(24):12–396

Watt-Meyer O, Dresdner G, McGibbon J, Clark SK, Duncan J, Henn B, Peters M, Brenowitz ND, Kashinath K, Pritchard M, Bonev B, Bretherton C (2023) Ace: a fast, skillful learned global atmospheric model for climate prediction, In: NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning

Yu S, Hannah W, Peng L, Lin J, Bhouri MA, Gupta R, Lütjens B, Will JC, Behrens G, Busecke J et al (2023) Climsim: a large multi-scale dataset for hybrid physics-ml climate emulation. Adv Neural Inf Process Syst 36:22070–22084

Yuval J, O'Gorman PA (2020) Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. Nat Commun 11(1):3295

Yuval J, Langmore I, Kochkov D, Hoyer S (2024) Neural general circulation models optimized to predict satellite-based precipitation observations, arXiv:2412.11973

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.