1 Supplementary Text

1.1 Unconditional Diffusion Models

Diffusion Models can be separated into two parts, a forward and a backward diffusion process. The forward diffusion process is a probabilistic model $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ that produces a noisy version of a given image $\mathbf{x}_t$ in t noising steps. The model is chosen to be a Gaussian model: $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mu(\mathbf{x}_{t-1}), \beta_t\mathbf{I})$, where β_t controls the amount of noise that is added in each step. In other studies, the model is often chosen to be of the form $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$ [32]. In practice, we use the reparametrization trick to sample from a Gaussian distribution by $\mathcal{N}(\mu, \sigma) = \mu + \sigma\epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$. Thus, a noisy version of $\mathbf{x}_0$ can be obtained as $\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \beta_t\epsilon$ after the t noising steps. The noise scheduler β_t is chosen to add small amounts of noise in the beginning and larger amounts later, to preserve a reasonable amount of information throughout the process.

The backward process $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ models how to restore the previous version $q(\mathbf{x}_{t-1})$ of a given image $\mathbf{x}_t$ at a certain noise step t. This process is also modelled by a Gaussian $q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu(\mathbf{x}_t), \sigma(\mathbf{x}_t))$. The problem is that $\mu(\mathbf{x}_t)$ is not known.

Using Bayes formula, the model can be rewritten as a product of Gaussians: $q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1})}{q(\mathbf{x}_t)}$. Each term is a Gaussian distribution, and their product is also a Gaussian distribution. Computing the product and taking the mean of that expression is a valid way to model the backward process. However, in practice, the distribution $q(\mathbf{x}_{t-1})$ is unknown, so we cannot explicitly compute $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$.

Predicting the state before a noising operation $\mathbf{x}_{t-1}$ can be done by conditioning on the noisy image $\mathbf{x}_t$ and the noise free image $\mathbf{x}_0$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \tag{1}$$

The terms on the right-hand side are Gaussian and can be explicitly computed. The resulting Gaussian has a mean term that depends on $\mathbf{x}_t$ and x_0, while the variance is a constant depending on the time step t.

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \sigma_t^2\mathbf{I}) \tag{2}$$

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \tag{3}$$

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \tag{4}$$

with $\alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$.

The following equation describes how $\mathbf{x}_0$ is connected to $\mathbf{x}_t$, when applying the forward diffusion model T times:

$$\begin{aligned}\mathbf{x}_t &= \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_{t-1} \\ &= \sqrt{\alpha_t}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t}\epsilon_{t-2} \\ &= \dots \\ &= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}\end{aligned} \tag{5}$$

where $\epsilon, \dots, \epsilon_{t-2}, \epsilon_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Solving for x_0 yields:

$$\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon\right). \tag{6}$$

Combining Eq. 6 with Eq. 4 leads to:

$$\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right).$$

The backward process (Eq. 2) is then modelled as:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{\bar{\alpha}_t}}\epsilon\right), \sigma_t^2\mathbf{I}\right), \tag{7}$$

so given a noisy image in step t, this model will tell us how the less noisy version of that image $\mathbf{x}_{t-1}$looks like. The only unknown in this equation is ϵ. The idea is to parameterize ϵ with a neural network ϵ_θ. The objective of the network is then basically to estimate the noise that was added to a (noisy) image $\mathbf{x}_{t-1}$ at each time step t:

$$\tilde{\boldsymbol{\mu}_\theta}\left(\mathbf{x}_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right) \tag{8}$$

Using the reparametrization trick and inserting equation (Eq. 8) into equation (Eq. 2), the backward diffusion process (also called de-noising process) denotes as:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right), \sigma_t^2\mathbf{I}\right). \tag{9}$$

Following the reparametrization trick, every iteration of the backward process takes the form:

$$\boldsymbol{x}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(\mathbf{x}_t, t\right)\right) + \sigma_t\boldsymbol{\epsilon}_t \tag{10}$$

with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The neural network in the backward diffusion process can be learned through the following algorithm proposed by [32]:

Algorithm 1: Training

```
repeat
    $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
    $t \sim \text{Uniform}(\{1, \ldots, T\})$
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Take gradient descent step on
    $\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left( \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, t \right) \right\|^2$
    until converged
```

Model inference can be achieved with the following algorithm [32]:

Algorithm 2: Sampling

```
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ do
    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1-\alpha t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$
```

The inference corresponds to generating a noise-free image, given a noisy input image.

1.2 Conditional Diffusion Models

The goal of a conditional diffusion model is to learn an approximation of the distribution $p(\mathbf{x}|c)$, where c is some conditional information. The method learns to model the reverse diffusion process as $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c)$. Starting with a pure noisy image, it gets denoised in T steps. The forward diffusion process is identical to the unconditional case. The difference to the unconditional case is that the model has knowledge about the condition during the backward process. Theoretically, the model could also be conditioned during the forward process. The backward process looks as follows:

$$\boldsymbol{x}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta \left(\mathbf{x}_t, c, t \right) \right) + \sigma_t \boldsymbol{\epsilon}_t. \tag{11}$$

The condition is integrated by concatenating the condition and the noisy image in the color channel of the images. The network takes a two-channel image as input and produces a one-channel image as output in each backward step.

2 Comparative Analysis of Diffusion Model Variants

2.1 No noise

In order to show the importance of noising the biased small scales of the ESM condition, we conducted an experiment where no noise was applied to the condition during both training and inference. The climatology (fig. S10D) of the noiseless model is very similar to our proposed model (fig. S10E), with a mean absolute bias of 0.29 mm/d. The difference between the two methods becomes evident when examining the spatial PSD (fig. S11A), where our method is far superior in correcting the small-scale biases. The histogram (fig. S11B) further highlights the superiority of our approach (noise level n=50), particularly in the range from 50-150 mm/d. Furthermore, the latitudinal and longitudinal profiles (fig. S11C and fig. S11D) reveal that the output of the noise-free model is overall less accurate and exhibits too much variability compared to our proposed approach. We conclude that the model's poorer performance stems from its overreliance on biased information from the GFDL conditions during inference. During training, the model was conditioned on low-resolution ERA5 fields, which are nearly unbiased compared to high-resolution ERA5. As a result, the model only had to make minimal corrections during training and thus will do the same at inference. The good performance of our method is not limited to daily precipitation statistics; we also find that it performs better in representing the ERA5 statistics of consecutive dry and consecutive wet days (fig. S12C and fig. S12F). For the precipitation extremes, the R95p for not noising the data looks slightly worse than with our model (fig. S13A and fig. S13C).

2.2 Different noise levels

We tested the effect of different noise levels (added to the conditions) by analyzing their impact on the downscaling results, maintaining the same noise level during both training and inference. The results for the noiseless case (n=0) were discussed in the previous paragraph, while our originally proposed approach uses a noise level of n=50. The dependence of climatology on the noise level appears small for the smaller noise levels n=0 (fig. S10D and fig. S14D) and n=50 (fig. S10E and fig. S14E), both with a mean absolute bias of 0.29 mm/d. However, at a higher noise level (n=80), the climatology deviates slightly (fig. S10F) and the bias becomes more pronounced (fig. S14F), with the mean absolute bias increasing to 0.36 mm/d. When looking at the PSD (fig. S15A) we find that our model (n=50) performs best. Also the histogram (fig. S15B) and latitudinal as well as longitudinal profiles (fig. S15B and fig. S15C) show that the choice of noise level n=50 is optimal. One could argue that, since high noise levels (n=80) remove more information, the model's reconstruction task is harder, because it needs to learn more dependencies and cannot rely on the condition as much. This shows that the conditional information is indeed valuable for the generation. Hence, as we train all models for the same amount of time, we would expect the diffusion model to initially perform worse when removing more information under the same amount of training time.

2.3 No quantile mapping

Fig. S11 (orange) shows that the diffusion model without quantile mapping (QM) struggles to fully correct the characteristic biases in the GFDL data. The climatology and mean average bias presented in fig. S10C and fig. S14C highlight the critical role of QM as part of the embedding transformation. Without QM, the model reduces the mean absolute bias from 0.69 mm/d to 0.58 mm/d, but falls short of the 0.29 achieved when QM is applied. The spatial Pearson correlation between the temporally averaged fields of the DM bias-corrected GFDL data without quantile mapping is 0.89, which is still an improvement over 0.83 of the raw GFDL data. The PSD (fig. S11A) shows that the diffusion model is still perfectly able to correct the spectrum of the embedded GFDL data without applying quantile mapping. The reason why the method without QM struggles with the histogram (fig. S11B) and latitudinal and longitudinal profiles (fig. S11C and fig. S11D) is that without QM, the training and inference distribution of the condition are very different from the distribution of the training condition. The transformation learned by the model during training does not generalize to the out-of-distribution condition when QM is excluded. The PSD is almost the same as before, as applying QM has almost no effect on the spatial variability of individual fields. The application of QM also helps to represent the ERA5 statistics of consecutive dry and consecutive wet days (fig. S12B and fig. S12E). Extreme precipitation, as represented by the R95p metric in fig. S13A with QM and fig. S13B without QM, also benefits from the application of QM.

Overall, we find that QM is an essential part of our method, by ensuring that the embedded GFDL data is distributed like the embedded ERA5 data. In other words, it is important because of our unpaired data setting. Alone, however, it can not correct the spatial PSD (Fig. 5A) and is therefore not useful for downscaling. The histogram (Fig. 5B) is also corrected a lot better with our diffusion model. Applying our DM also improves the CDD (fig. S12A and fig. S12B) and CWD (fig. S12D and fig. S12E) statistics compared to our QM benchmark and.

3 Comparative Analysis to VQ-VAE model

We have included an additional comparison with an established generative model, the VQ-VAE. This model is primarily designed for representation learning and compression, with its training objective focused on mapping data to a latent space and reconstructing it. To adapt the VQ-VAE for our downscaling task, we employ a two-step training process inspired by the original VQ-VAE work [36].

Step 1: We train the VQ-VAE to compress and reconstruct high-resolution (HR) ERA5 images. Step 2: Using the encoder of the trained model, we construct a dataset of latent representations for each HR field. A conditional PixelCNN is then trained to autoregressively model the prior distribution $p(z \mid c)$ where c are samples from the embedded ERA5 distribution. This allows us to sample latents z conditioned on the embedded ERA5 fields, which are subsequently decoded by the VQ-VAE decoder to generate HR ERA5 fields.

A notable limitation of the VQ-VAE is its inability to generate high-frequency information, as shown by the power spectral density compared to the HR ERA5 ground truth (fig. S16A). Furthermore, the histogram reveals that the VQ-VAE performs significantly worse in capturing the target distribution for high precipitation values (fig. S16B). The model also struggles to consistently match the latitudinal and longitudinal means with the ground truth (fig. S16C, fig. S16D).

Overall, the VQ-VAE lacks the fine detail in small-scale variability and falls short in overall accuracy compared to our diffusion model approach.

4 Ensemble uncertainty evaluation

For the following evaluations, we generate a 50 member DM ensemble by conditioning our model 50 times on the same low-resolution ERA5 year. This results in a 50-member ensemble of one-year trajectories. The resulting ground truth will be the corresponding high-resolution ERA5 year.

4.1 Ensemble mean evaluation

To evaluate the accuracy of the DM ensemble in reproducing precipitation patterns, we compare the spatially averaged daily precipitation of a 50-member ensemble of downscaled high-resolution fields, obtained by conditioning our DM on low-resolution (i.e. upscaled) ERA5 fields, to the corresponding ERA5 high-resolution ground truth, at 0.25° resolution over one year. Figure S17 illustrates that the ensemble mean generated by the DM closely aligns with the ERA5 high-resolution ground truth throughout the annual cycle. This demonstrates the ability of the diffusion model ensemble to – on average – capture the temporal variability of precipitation while maintaining well-calibrated ensemble members.

4.2 Continuous Ranked Probability Score

We evaluate the probabilistic downscaling performance of our model using the Continuous Ranked Probability Score (CRPS), which extends the concept of the mean absolute error (MAE) to probabilistic forecasts (for details and definition, see [37]In our case, each downscaling realization corresponds to a forecast realization in tasks like weather forecasting. To compute the CRPS, we compare 50 downscaled outputs from our diffusion model to the corresponding high-resolution ERA5 ground truth. The 50-member ensemble is generated by running the diffusion model 50 times, always conditioned on the same low-resolution ERA5 year. As a baseline, we use a deterministic, bi-linearly upsampled version of the low-resolution ERA5 reference year and a 50-member VQ-VAE ensemble generated the same way as the DM. The corresponding high-resolution ERA5 year serves as the ground-truth reference. This results in a CRPS for every day of the year and every spatial location. Comparing the spatially and temporally averaged CRPS values (0.90 mm/day vs 0.76 mm/day) and the maximum CRPS values (14.31 mm/day vs 26.66 mm/day), our DM significantly outperforms the baseline of bilinear upsampling, while also slightly surpassing the VQ-VAE in terms of both the mean CRPS value (0.80 mm/day) and the maximum CRPS value (15.79 mm/day) (lower CRPS is better).

In fig. S18A we compare the CRPS time series over 1 year, where we averaged the spatial dimensions. Our model is consistently below the bi-linearly upsampled baseline and performs on par with the VQ-VAE. Next, the annual mean CRPS is computed, and then the difference is taken between the bilinearly upsampled ERA5 and the diffusion model ensemble (fig. S18B), as well as between the bi-linearly upsampled

ERA5 and the VQ-VAE ensemble (fig. S18C). The results show that both models outperform the bilinear baseline over the continent. Positive differences (blue) indicate a higher (worse) CRPS value for the upsampling baseline. The performance of the DM and VQ-VAE is comparable, except in the southern part of the Andes, where our DM consistently outperforms the baseline.

4.3 Spread-skill plot

We evaluated the statistical consistency of our model using a spread-skill plot, which evaluates the relationship between the predicted root mean squared spread (RMSS) and the root-mean-squared error (RMSE) of the ensemble mean. The spread-skill plot relates the predicted model spread to the actual model error. We follow the implementation of [38], for more details see their work. The x-axis represents the average standard deviation of the DM ensemble distribution, while the y-axis shows the RMSE of the model's mean prediction. Each point on the plot corresponds to a bin of predicted spread values. The model uncertainty estimate is biased if spread values are above the 1:1 line (under-dispersive, over-confident) and if they are under the 1:1 line (over-dispersive / under-confident). A perfect calibration is a spread–skill of 1:1, along the diagonal line.

In our comparison, we use 50 ensemble members generated by conditioning once the DM and once our VQ-VAE model on one year of low-resolution ERA5 data. Both our DM and also our VQ-VAE show very good calibration overall. However, the diffusion model is superior for very large spreads, where the VQ-VAE is overconfident (fig. S19). As desired, our DM matches the 1:1 spread-skill line, indicating that its uncertainty is well calibrated.

Fig. S1

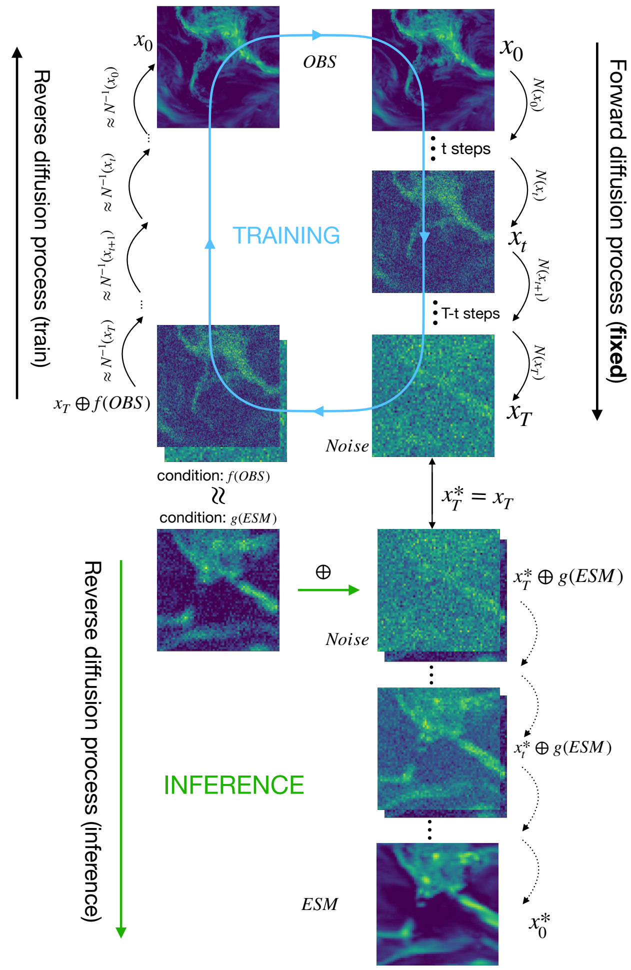

Fig. S1 Detailed visualisation of the training (blue) and inference process (green) of the conditional diffusion model introduced in our study. During the forward process (top to center), Gaussian noise $N(x_t)$ is added to an input image x_0 over t steps, following Eq.5, until x_t contains just noise. During the backward process (center to bottom) we concatenate a conditional image $f(OBS)$ with the noisy image x_t and train a neural network to remove the noise. The inference process uses the trained model. We also concatenate a purely noisy image x_t* with the condition $g(ESM)$ and remove the noise from x_t* with our trained model. The resulting image x_0* follows the same distribution as x_0, indicating that the bias correction and downscaling is achieved. The arrows between the training and inference part indicate that the image distribution at which the training ends is the same as the distribution the inference starts with.

208 # Fig. S2

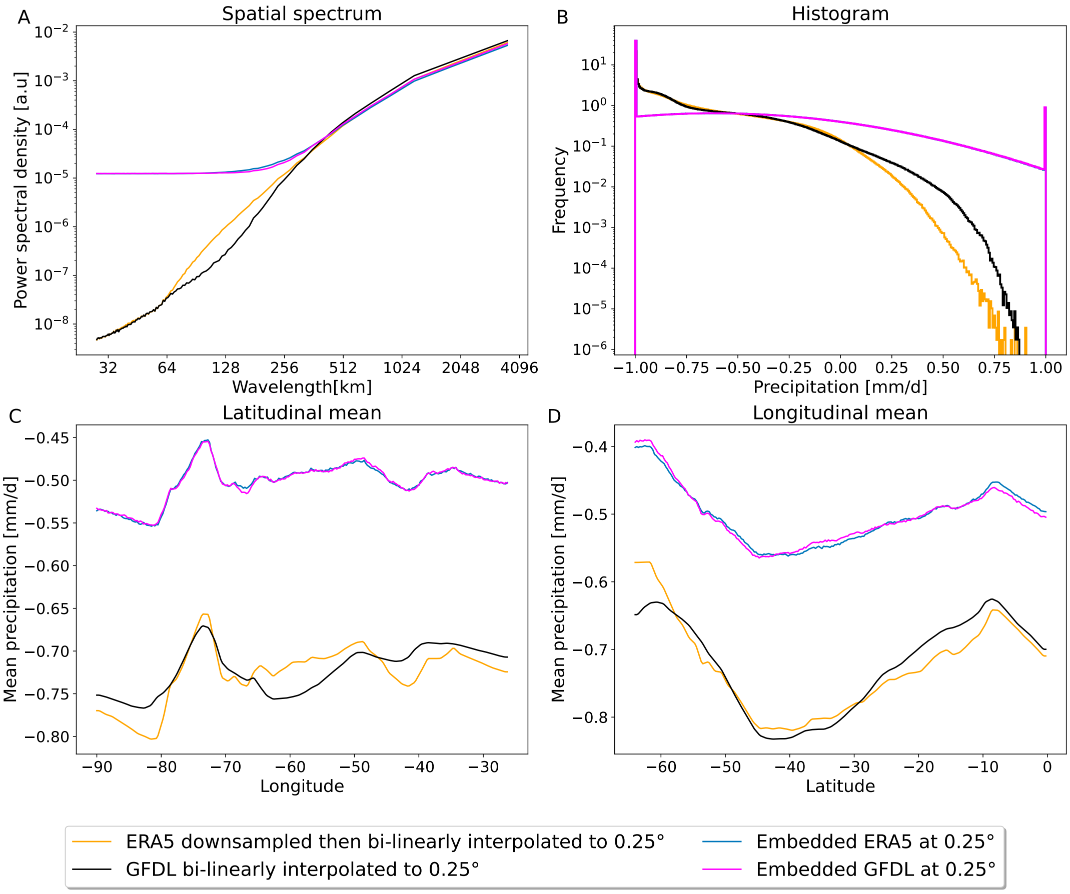

Fig. S2 We demonstrate the effect of the embedding transformations. We apply the transformations to the first downsampled, then bi-linearly upsampled 0.25° low-resolution ERA5 data and to the bi-linearly upsampled GFDL data at 0.25°. The data shown is after our pre-processing transformation, hence the different precipitation units. Both embedded ERA5 and embedded GFDL fields are also at 0.25°. The PSD **(A)** of the embedded datasets align very well. The horizontal lines in the PSD plot are a sign of the added noise that acts as a low-pass filter on small scales. The histogram **(B)** as well as latitudinal / longitudinal profiles **(C)** / **(D)** show a strong alignment between embedded ERA5 and embedded GFDL data. Note that we clip the data at -1 and 1 after noising, as this is the maximal range of our pre-processed data.

209 # Fig. S3

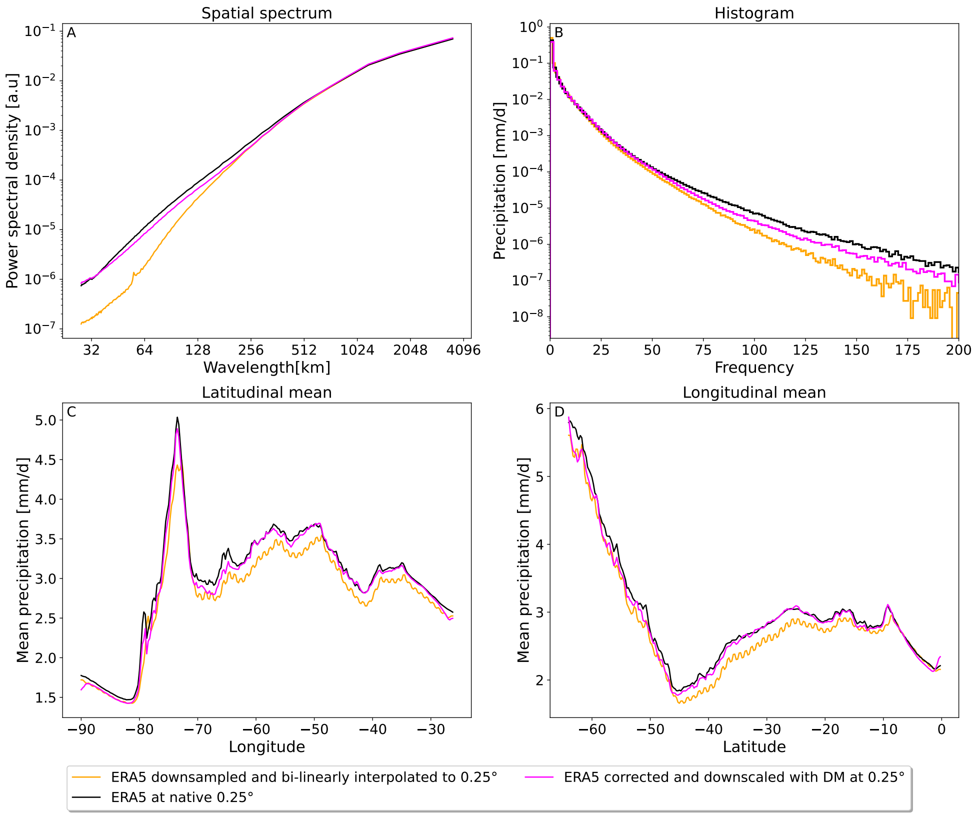

Fig. S3 Evaluation of the diffusion model's performance to reconstruct ERA5 at 0.25° resolution. The reconstruction starts from the embedded ERA5 data, obtained by downsampling to 1° via choosing only every fourth grid cell, and then bi-linearly interpolating back to 0.25°. In this case, the embedded ERA5 data mimics the ESM data in order to train the diffusion model (see Methods and main text). **(A)** Mean spatial power spectral densities (PSDs). **(B)** Histogram indicating the precipitation frequencies. **(C)** Latitude profile, given by the averaged longitudes over the validation period. **(D)** Longitude profile, given by the averaged latitudes over the validation period. Our diffusion model approximates the latitude and longitude profile of the original ERA5 reference data extremely well. The histogram shows also shows large improvements with slight deviations from the HR ERA5 reference data for extreme precipitation. The diffusion model manages to correct the small-scale spatial details and follows the target distribution closely.

210 # Fig. S4

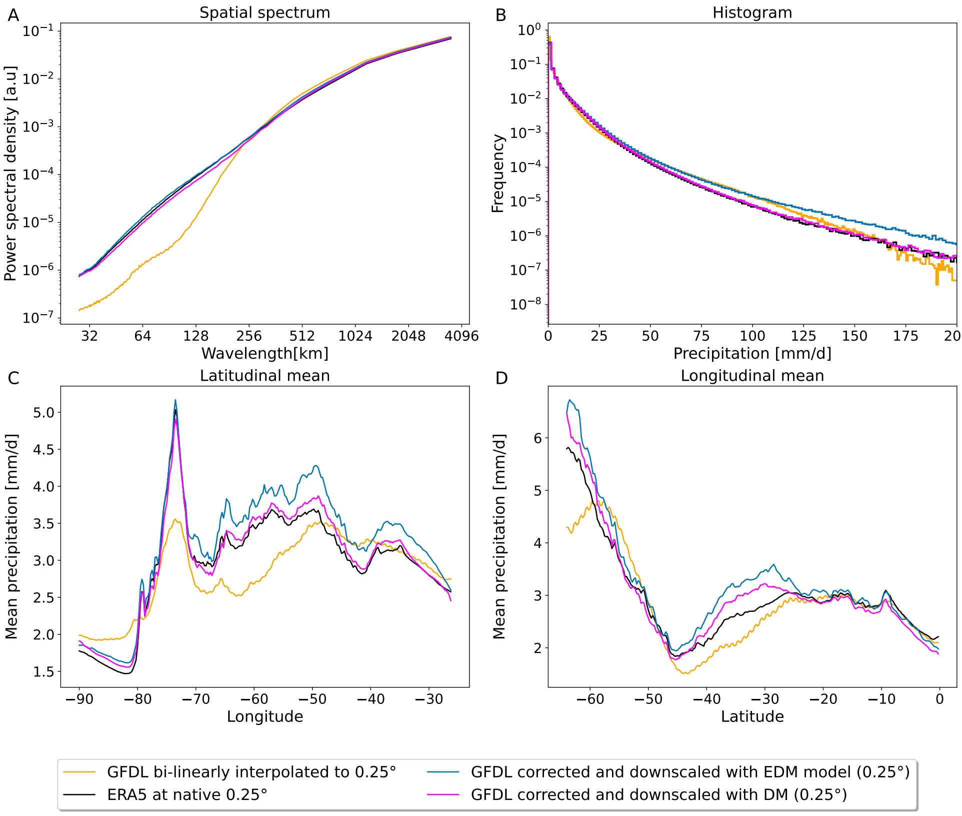

Fig. S4 Comparing the downscaling and bias correction performance of our DM to an EDM-diffusion model. Comparison of GFDL (bi-linearly upsampled to 0.25°) (orange) and ERA5 at 0.25° (black) to diffusion model-corrected GFDL fields with our proposed method at 0.25° (magenta) and diffusion model-corrected GFDL fields with the EDM diffusion model at 0.25° (blue). The Power spectral density (PSD) plot **(A)** shows that both diffusion models correct the small-scale spatial details very well. The spectrum aligns well with the high-resolution ERA5 target data. The EDM model is inferior in correcting the histogram **(B)** as well as the latitude **(C)** and longitude **(D)** profiles compared to our proposed method.

211 # Fig. S5

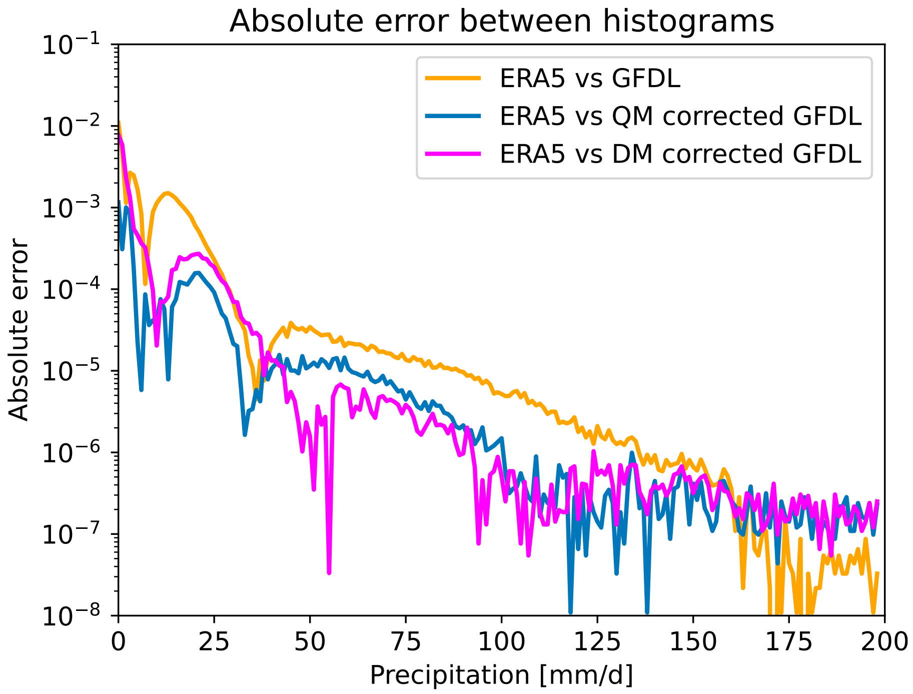

Fig. S5 Comparison of the absolute histogram errors. The histogram shows the absolute errors between high-resolution ERA5 data (0.25°) and GFDL (bi-linearly upsampled to 0.25°) (orange), ERA5 and DM-corrected GFDL (magenta) and ERA5 and QM-corrected GFDL data (blue). The dips around 10 mm/d and 40 mm/d correspond to points where the histograms of ERA5 and its comparisons intersect. For very large precipitation values, our diffusion model outperforms the QM baseline, while it has slightly larger absolute error for smaller precipitation values.

212 # Fig. S6

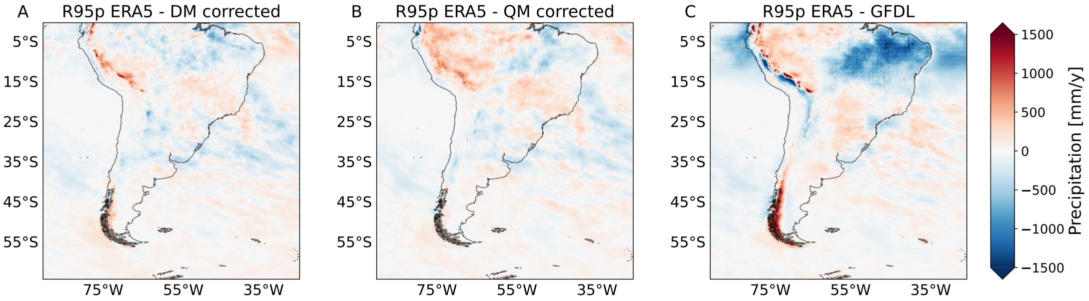

Fig. S6 Performance evaluation of the diffusion model in representing extreme rainfall events. The R95p metric represents the total annual precipitation from heavy rain days, calculated as the sum of daily precipitation on wet days (precipitation $>$ 1 mm/day) that exceed the 95th percentile of our reference period. We calculate the metric on the validation periods, giving the ERA5 training period as reference. We computed R95p for ERA5, GFDL, DM-corrected GFDL, and QM-corrected GFDL. For clarity, we plotted the difference between R95p values for ERA5 and DM-corrected GFDL **(A)**, ERA5 and QM-corrected GFDL **(B)**, and ERA5 and GFDL **(C)**. Our diffusion model effectively corrects the bias in extreme precipitation events, performing similarly to the quantile mapping correction.

213 # Fig. S7

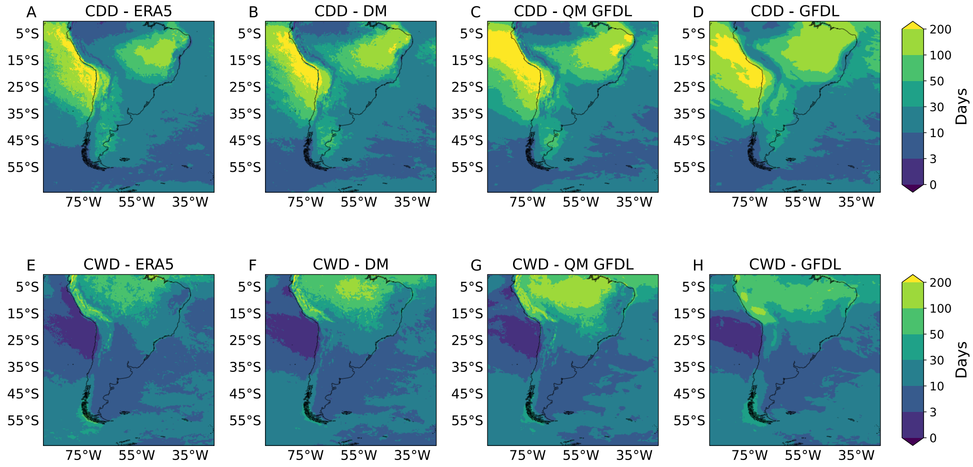

Fig. S7 Duration of consecutive dry and wet periods in the different datasets. The first row shows a comparison between consecutive dry days (CDD) between ERA5 at 0.25° **(A)**, diffusion model (DM) correction GFDL fields at 0.25° **(B)**, GFDL first quantile mapped and then bi-linearly upsampled to 0.25° **(C)** and the GFDL bi-linearly upsampled to 0.25° **(D)**. The second row shows a comparison between consecutive wet days (CWD) between ERA5 at 0.25° **(E)**, diffusion model correction GFDL fields at 0.25° **(F)**, GFDL first quantile mapped and then bi-linearly upsampled to 0.25° **(G)** and the GFDL bi-linearly upsampled to 0.25° **(H)**. Overall our diffusion model is the most similar to ERA5 in both, the case of CDD (**(A)** and **(B)**) and the case of CWD (**(E)** and **(F)**).

214 # Fig. S8

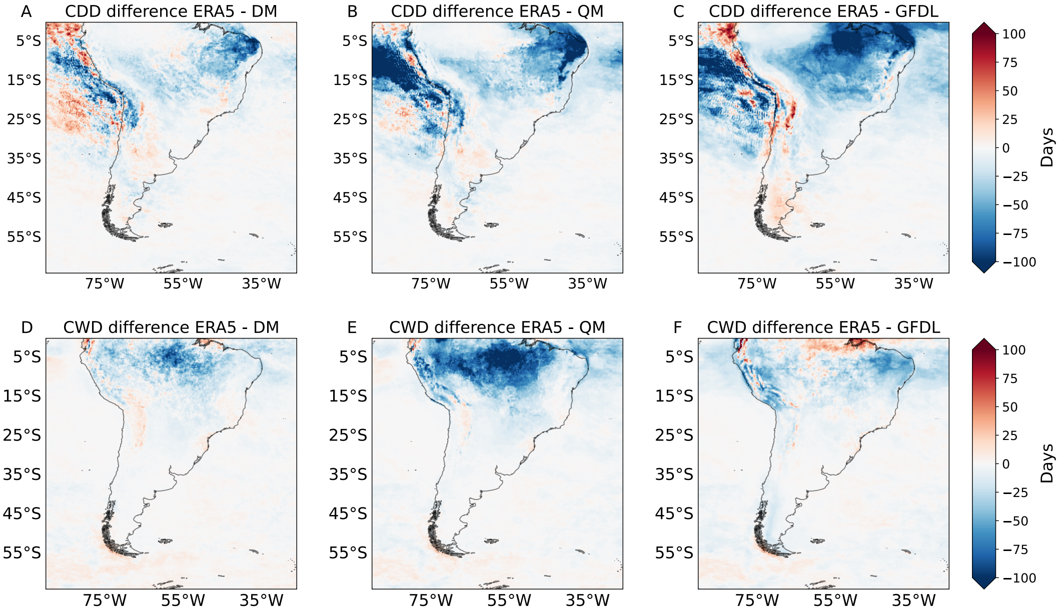
A CDD difference ERA5 - DM
B CDD difference ERA5 - QM
C CDD difference ERA5 - GFDL
D CWD difference ERA5 - DM
E CWD difference ERA5 - QM
F CWD difference ERA5 - GFDL
5°S
15°S
25°S
35°S
45°S
55°S
75°W
55°W
35°W
100
75
50
25
0
−25
−50
−75
−100
Days

Fig. S8 Performance of different methods regarding consecutive dry and wet period statistics. We compare the consecutive dry day (CDD) differences between 0.25° ERA5 and diffusion model (DM) correction GFDL fields at 0.25° **(A)**, as well as GFDL first quantile mapped and then bi-linearly upsampled to 0.25° **(B)** and the GFDL bi-linearly upsampled to 0.25° **(C)**. The second row shows a comparison between consecutive wet day (CWD) differences between ERA5 at 0.25° and diffusion model corrected GFDL fields at 0.25° **(D)**, as well as GFDL first quantile mapped and then bi-linearly upsampled to 0.25° **(E)** and the GFDL bi-linearly upsampled to 0.25° **(F)**. Our diffusion model exhibits substantially smaller errors than the benchmarks.

215 # Fig. S9

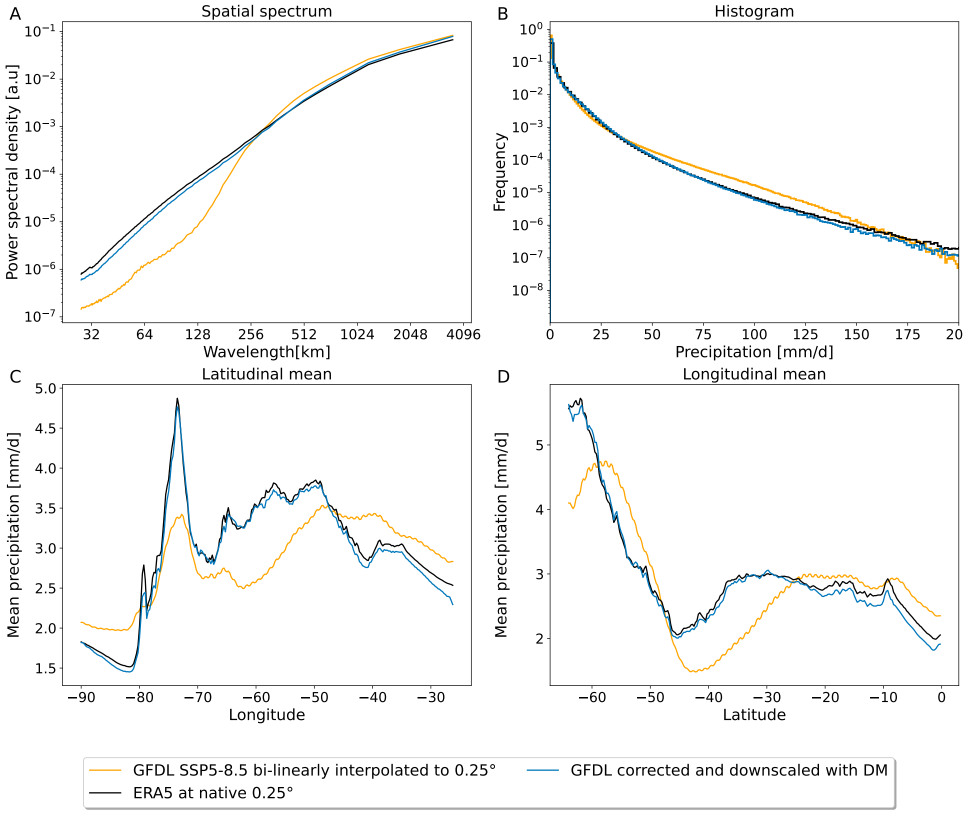

Fig. S9 Evaluation of the diffusion model's downscaling and bias correction performance of GFDL SSP5-8.5. Comparison of GFDL SSP5-8.5 (bi-linearly upsampled to 0.25°) (orange) and ERA5 at 0.25° (black) to DM-corrected GFDL SSP5-8.5 fields at 0.25° (blue). The PSD plot **(A)** shows that the diffusion model applied to the SSP5-8.5 GFDL data still results in a strong correction of the small-scales. The PSD aligns very well with the high-resolution ERA5 target data. The histograms **(B)** as well as the latitude **(C)** and longitude **(D)** profiles also shows the same improvements compared to the uncorrected SSP5-8.5 GFDL data.

Fig. S10

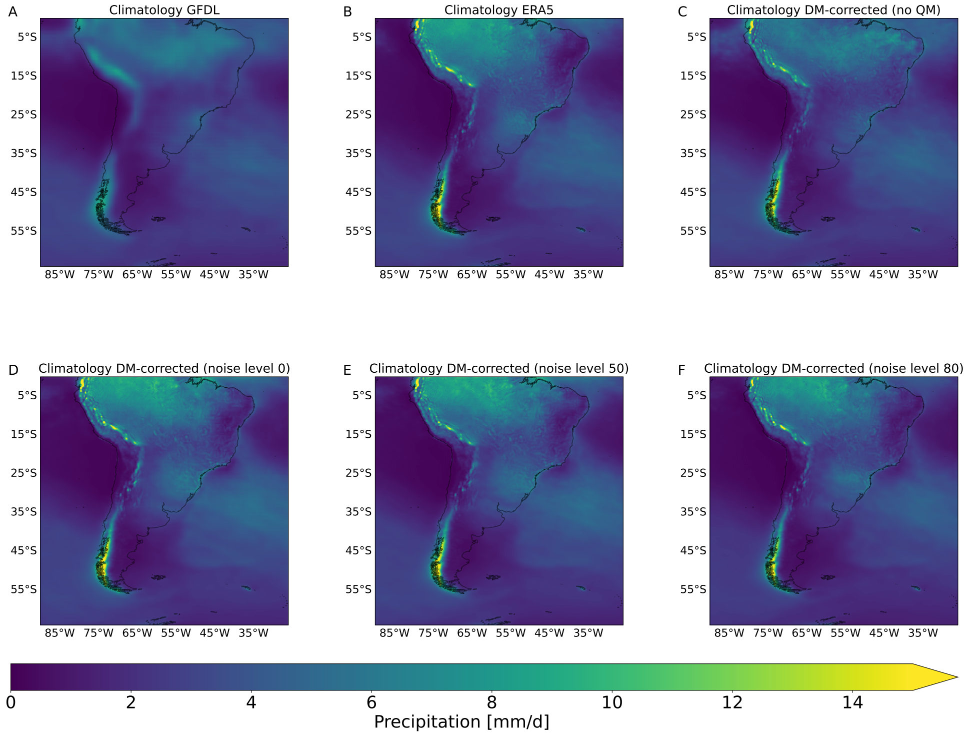
A
Climatology GFDL
B
Climatology ERA5
C
Climatology DM-corrected (no QM)
D
Climatology DM-corrected (noise level 0)
E
Climatology DM-corrected (noise level 50)
F
Climatology DM-corrected (noise level 80)
5°S
15°S
25°S
35°S
45°S
55°S
85°W 75°W 65°W 55°W 45°W 35°W
0
2
4
6
8
10
12
14
Precipitation [mm/d]

Fig. S10 Time-averaged precipitation for different diffusion model training and inference variants. We show GFDL upsampled to 0.25° **(A)** and 0.25° ERA5 **(B)** for reference. The climatology of the diffusion model corrected GFDL data without applying QM **(C)** looks noticeably different from ERA5. figures **(D)**-**(F)** all show the DM-corrected GFDL fields at 0.25° with different noise variations (n=0, n=50, n=80). Adding no noise during training and inference **(D)** and our proposed approach **(E)** are most similar to ERA5. Choosing a very large noise level (n=80) **(F)** still results in a similar climatology.

217 # Fig. S11

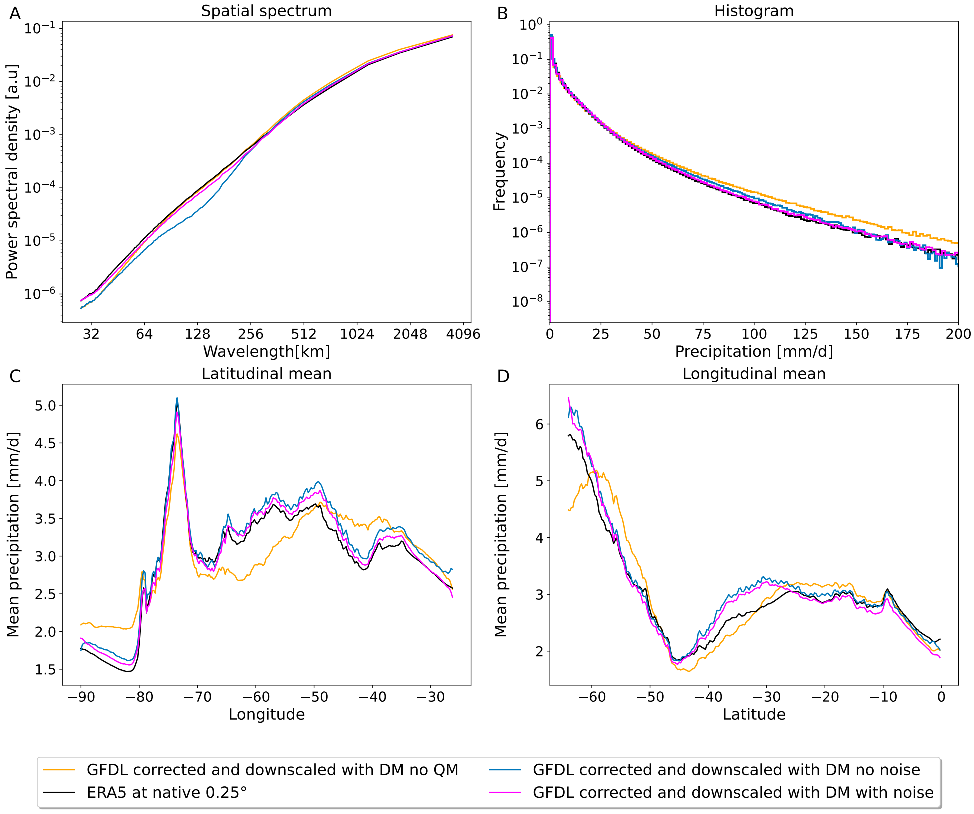

Fig. S11 Evaluating the role of QM and noising in the embedding transformation. We compare the diffusion model-corrected GFDL fields at 0.25° without applying quantile mapping in the embedding transformation (orange) and 0.25° ERA5 (black) to the regularly noised diffusion model-corrected GFDL at 0.25° (magenta) and diffusion model-corrected GFDL fields without adding noise at 0.25° (blue). The spatial Power spectral density (PSD) plot **(A)** shows that the regularly noised diffusion models correct the small-scale spatial details far better then the noiseless variant. The histogram **(B)** for the version without noise is also worse for the range 50-150 mm/d. The model without quantile mapping has a completely shifted histogram compared to the ground truth. The latitude **(C)** and longitude **(D)** profiles show that the noisy version is less smooth and the version without quantile mapping is completely shifted compared to ERA5.

Fig. S12

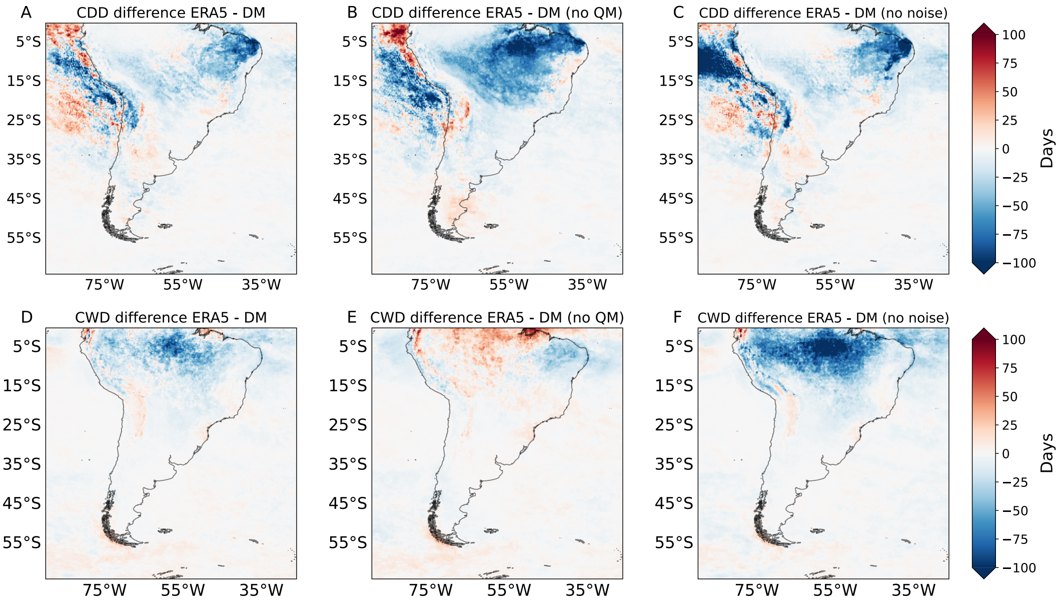
A
CDD difference ERA5 - DM
B
CDD difference ERA5 - DM (no QM)
C
CDD difference ERA5 - DM (no noise)
D
CWD difference ERA5 - DM
E
CWD difference ERA5 - DM (no QM)
F
CWD difference ERA5 - DM (no noise)
5°S
15°S
25°S
35°S
45°S
55°S
75°W
55°W
35°W
100
75
50
25
0
−25
−50
−75
−100
Days

Fig. S12 Comparing consecutive dry day (CDD) and consecutive wet days (CWD) differences. We compare the CDD/CWD between 0.25° ERA5 and diffusion model correction GFDL fields with our method at 0.25° **(A)**/**(D)**, as well as ERA5 and DM-corrected GFDL fields without using qunatile mapping at 0.25° **(B)**/**(E)**, ERA5, and DM-corrected GFDL fields at 0.25° without adding noise to the condition neither during training nor inference **(C)**/**(F)**. Our diffusion model exhibits substantially smaller differences than the versions without QM and without adding noise.

Fig. S13

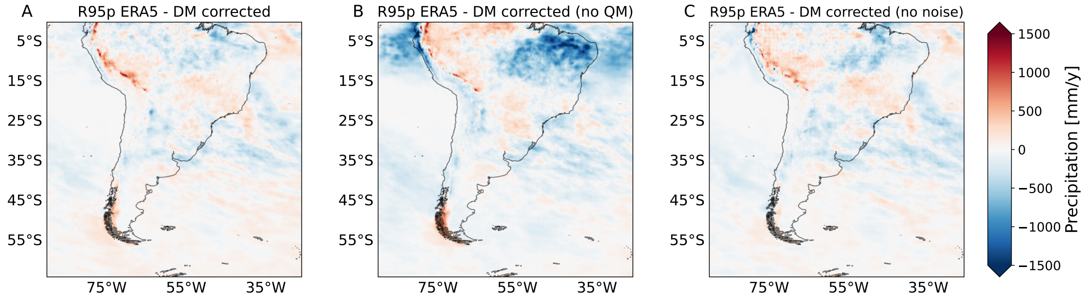

Fig. S13 We use the R95p metric to investigate the performance of different diffusion model variants on extreme events. For clarity, we plotted the difference between R95p values for ERA5 and DM-corrected GFDL with our proposed model **(A)**, ERA5 and DM-corrected GFDL fields without applying quantile mapping **(B)**, and ERA5 and DM-corrected GFDL fields without adding noise to the condition neither during training nor inference **(C)**. Our proposed model is most effective in correcting extreme precipitation events.

Fig. S14

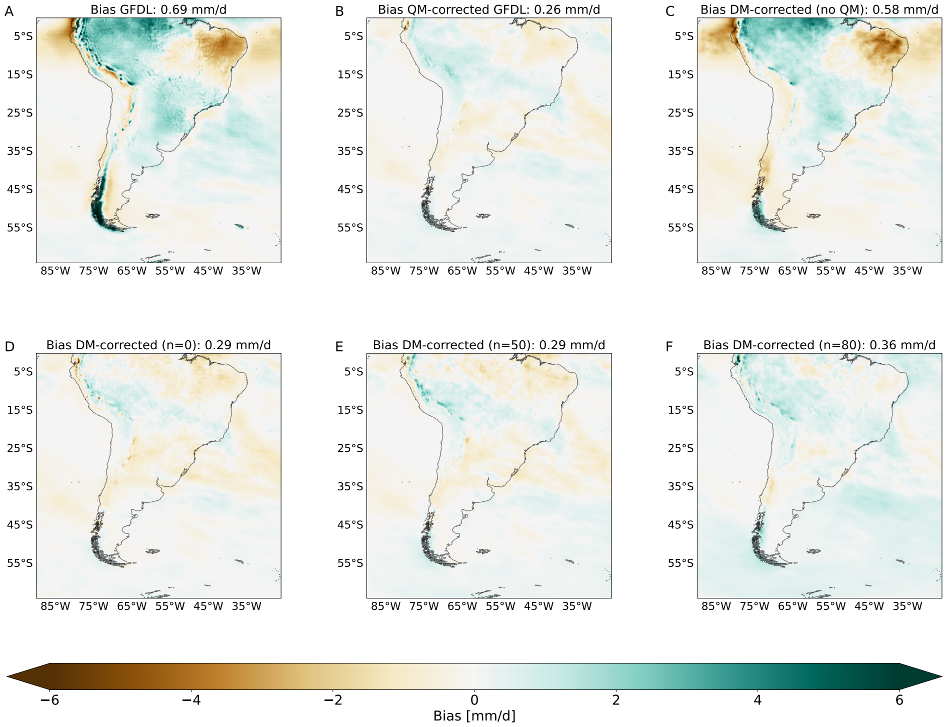
A
Bias GFDL: 0.69 mm/d
B
Bias QM-corrected GFDL: 0.26 mm/d
C
Bias DM-corrected (no QM): 0.58 mm/d
D
Bias DM-corrected (n=0): 0.29 mm/d
E
Bias DM-corrected (n=50): 0.29 mm/d
F
Bias DM-corrected (n=80): 0.36 mm/d
5°S
15°S
25°S
35°S
45°S
55°S
85°W
75°W
65°W
55°W
45°W
35°W
−6
−4
−2
0
2
4
6
Bias [mm/d]

Fig. S14 Time-averaged precipitation differences for various diffusion model training and inference variants. We show GFDL upsampled to 0.25° **(A)** and 0.25° QM-corrected GFDL **(B)** for comparison. The bias of the diffusion model corrected GFDL data without applying quantile mapping **(C)** looks noticeably different from ERA5. figures **(D)**-**(F)** all show the DM-corrected GFDL fields at 0.25° with different noise variations (n=0, n=50, n=80). Adding no noise during training and inference **(D)** and our proposed approach **(E)** are most similar to ERA5. Choosing a very large noise level **(F)** leads still results in a similar climatology.

221 # Fig. S15

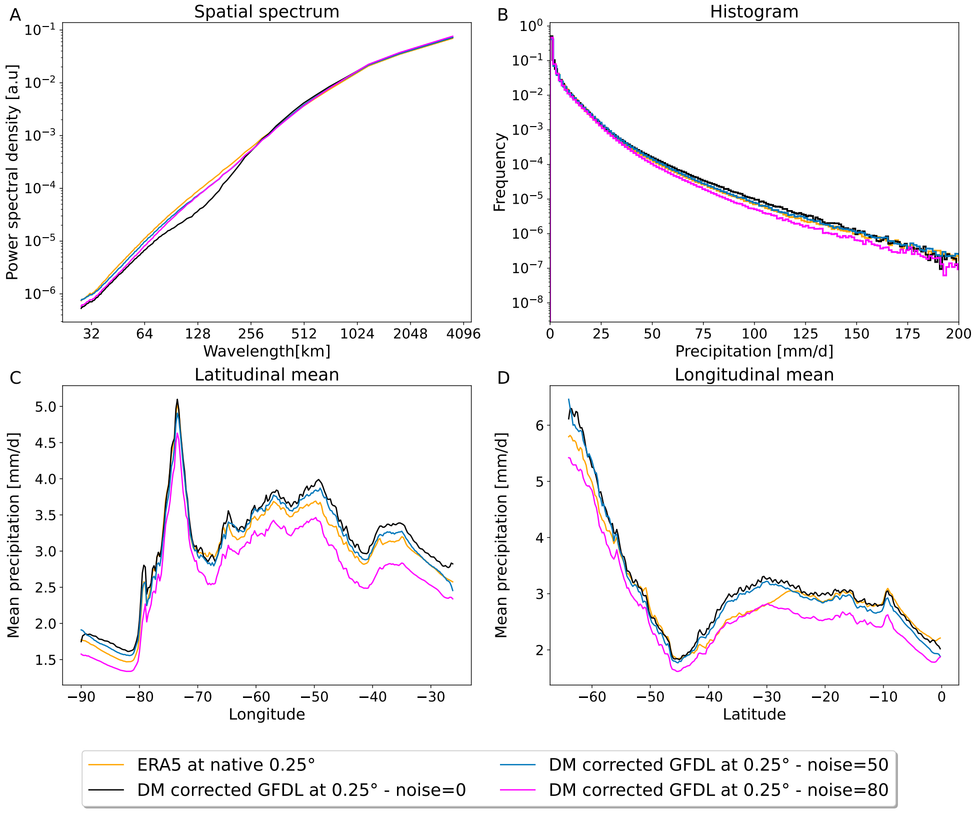

Fig. S15 Evaluating the role of different noising strengths in the embedding transformation. Comparison of 0.25° ERA5 (orange), the diffusion model-corrected GFDL fields without adding noise (n=0) at 0.25° (black), the regularly noised (n=50) diffusion model-corrected GFDL at 0.25° (blue), and the diffusion model-corrected GFDL with a higher noise level (n=80) at 0.25° (magenta). The power spectral density plot **(A)** shows that the model with n=50 corrects the small-scale spatial details better than the n=80 version and a lot better than the n=0 version. The histogram **(B)** for both n=0 and n=50 are superior to n=80. The n=0 model falls behind the n=50 model in the range of 50-150 mm/d. The latitude **(C)** and longitude **(D)** profiles also show the superiority of the n=50 model over both alternatives, especially the n=80 model.

Fig. S16

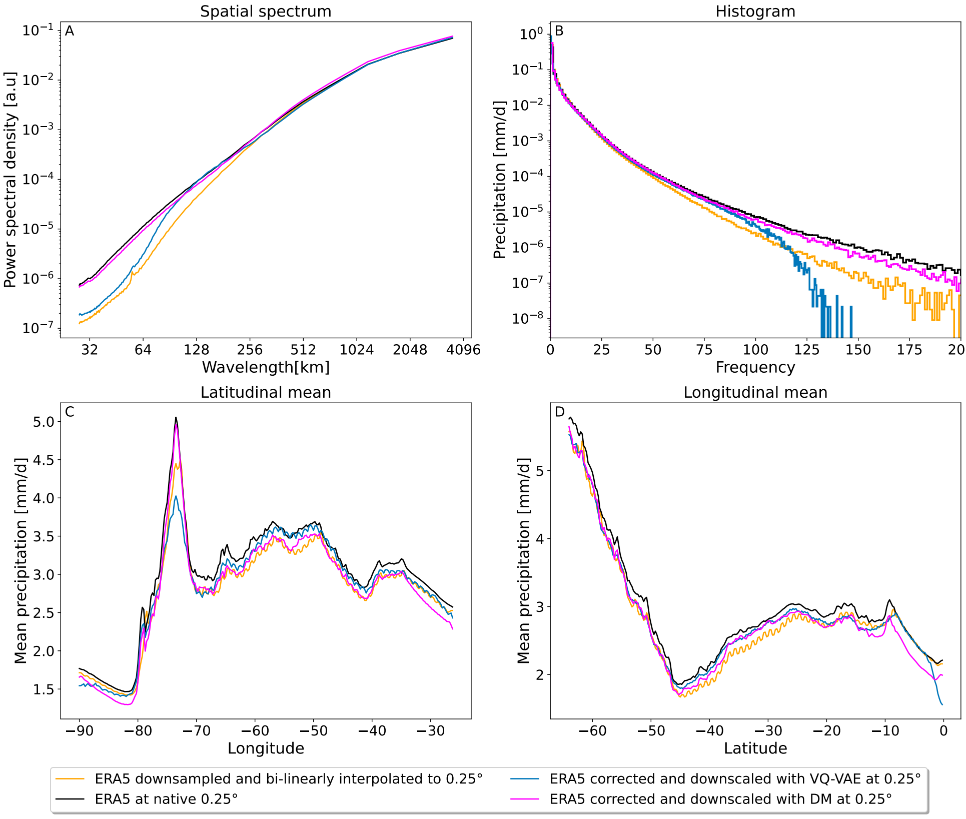

Fig. S16 Comparing the diffusion model's downscaling and bias correction performance to a VQ-VAE. We compare the VQ-VAE model (blue) to 0.25° ERA5 (black), ERA5 bi-linearly upsampled to 0.25° (orange) and the diffusion model-corrected ERA5 fields with our proposed method at 0.25° (magenta). **(A)** Power spectral density (PSD) illustrates the inability of the VQ-VAE to generate high-frequency information, diverging from the HR ERA5 ground truth. **(B)** The histogram highlights the VQ-VAE's poor performance in capturing the target distribution for high precipitation values. The VQ-VAE also struggles to consistently match the latitudinal **(C)** and longitudinal **(D)** means with the ERA5 ground truth.

223 # Fig. S17

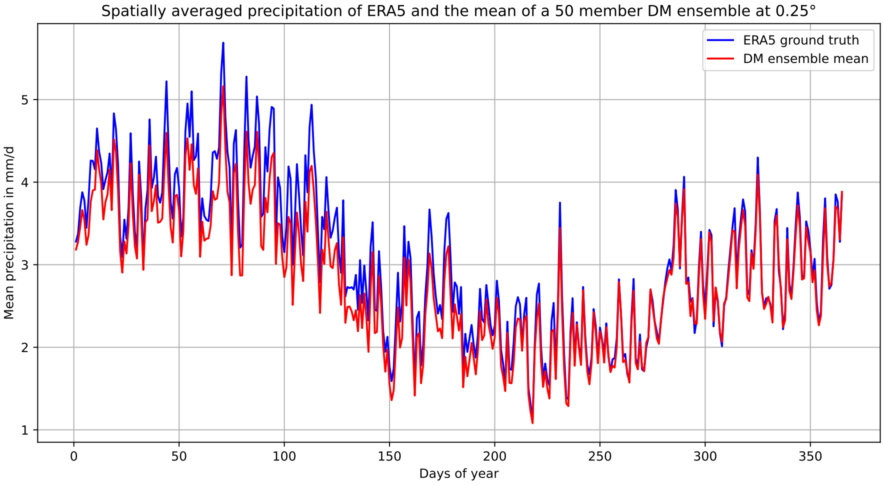

Fig. S17 Spatially averaged daily precipitation at 0.25° resolution over one year. We compare the ERA5 ground truth (blue) with the mean of a 50-member ensemble of DM-downscaled fields (red). The close alignment demonstrates the DM ensemble's ability to capture temporal variability and maintain accurate ensemble calibration on average.

224 # Fig. S18

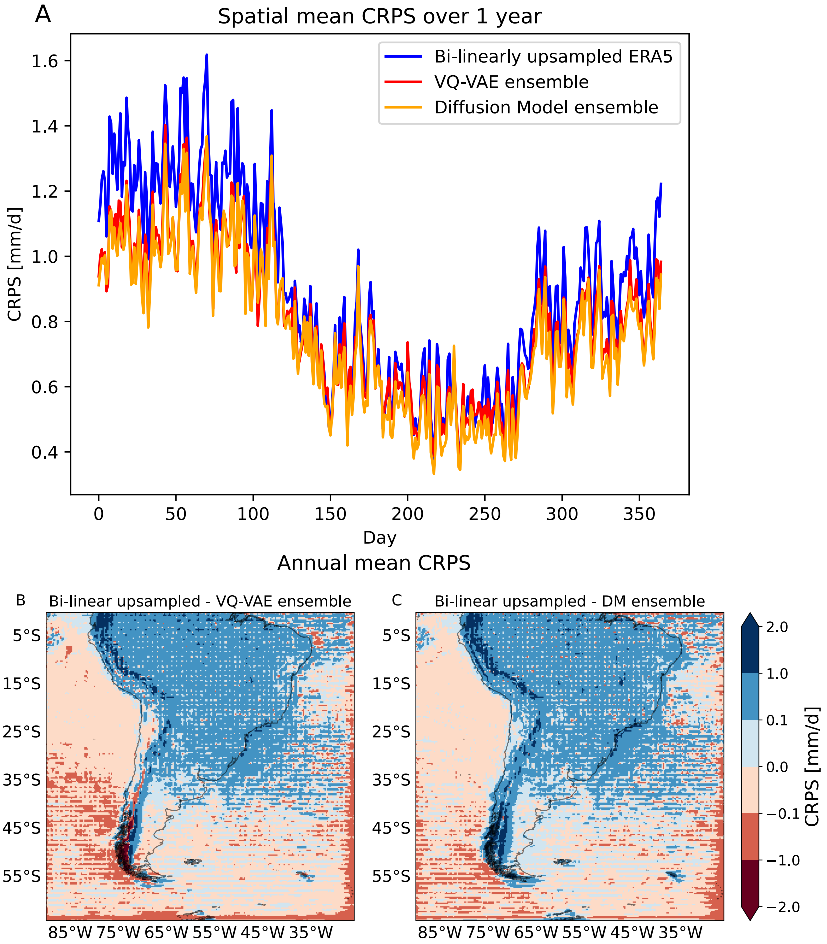

Fig. S18 Evaluating our DM using the Continuous Ranked Probability Score (CRPS). We compare a 50-member DM downscaling ensemble (orange), a 50-member VQ-VAE downscaling ensemble (red) and a deterministic baseline generated by bi-linearly upsampling 1 year of low-resolution ERA5 fields (blue). **(A)** shows a time series of CRPS over 1 year averaged over the spatial domain. Our DM ensemble as well as the VQ-VAE ensemble are consistently under the baseline (lower is better). **(B)** and **(C)** display the annual mean CRPS difference between the bi-linearly upsampled ERA5 and our DM ensemble, as well as the difference between the bi-linearly upsampled ERA5 and the VQ-VAE ensemble. Both models consistently outperform the bilinear baseline, achieving lower mean CRPS values across the continent. The blue regions indicate that the baseline has higher (worse) CRPS than the ML models. While our DM and VQ-VAE perform similarly overall, in the southern part of the Andes, our DM achieves lower CRPS.

Fig. S19

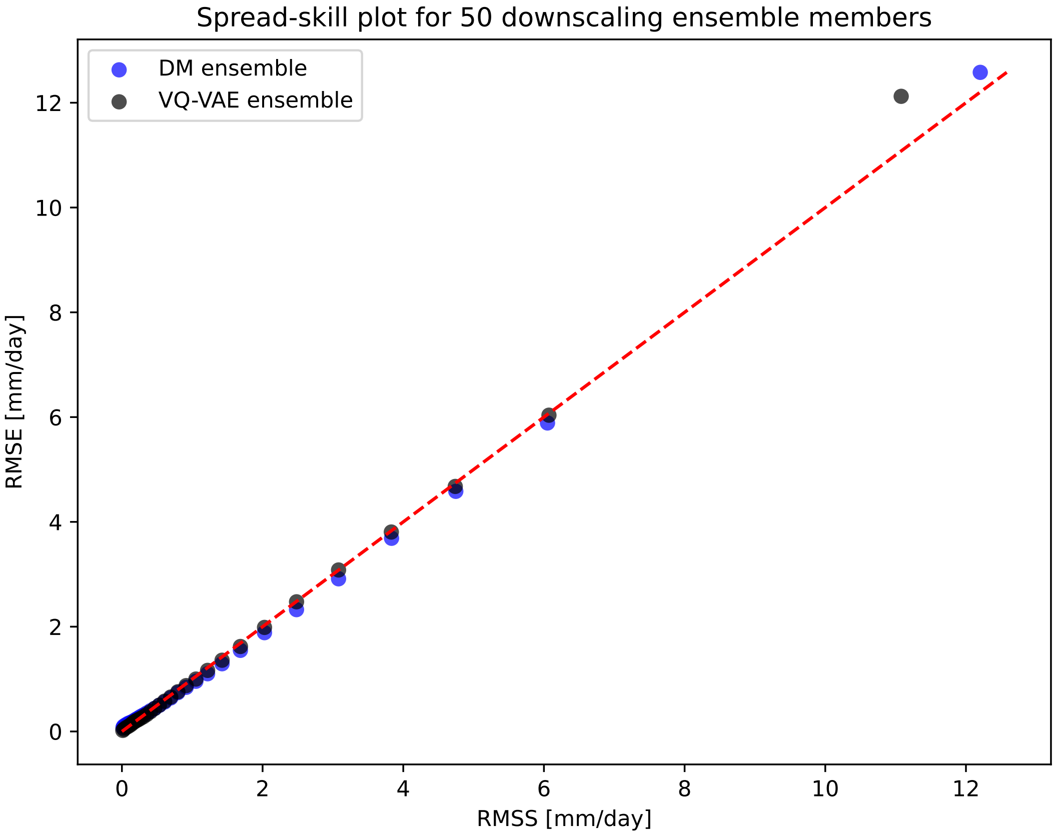

Fig. S19 Spread-skill plot for 50 downscaling ensemble members generated by conditioning on one year of low-resolution ERA5 data. The spread measures the uncertainty in terms of root-mean-squared spread (RMSS), whereas the skill is indicated by the root-mean-squared error (RMSE). The plot compares the uncertainty calibration of the DM ensemble (blue) and the VQ-VAE ensemble (black). Both models demonstrate overall good calibration, with most points aligning along the 1:1 spread-skill line (red dashed line). However, for very large spreads, the DM exhibits superior calibration, while the VQ-VAE ensemble is overconfident. The DM's close alignment with the 1:1 line indicates well-calibrated uncertainty estimates.

Fig. S20

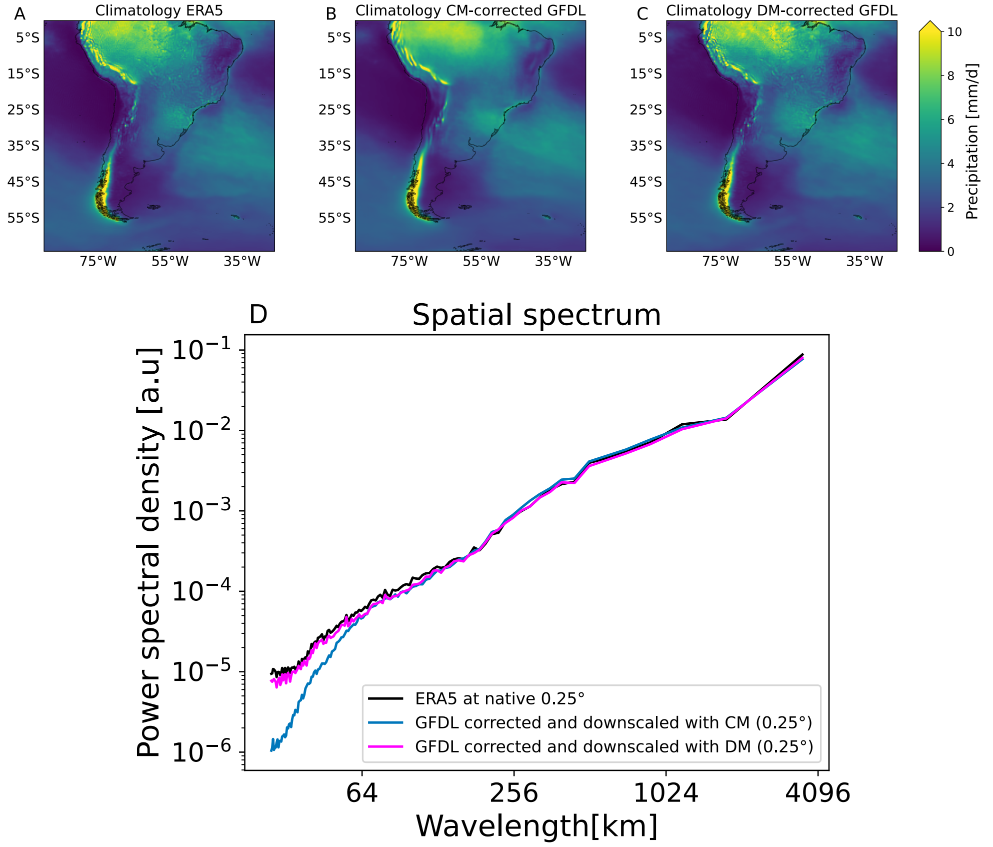

Fig. S20 Comparison of climatology to Consistency Model. Comparing the 0.25° climatologies of ERA5 **(A)**, the consistency model **(B)** and our diffusion model correction **(C)**, we can see that the consistency model struggles to learn the target distribution accurately, leading to substantial blurring compared to ERA5 and our DM. The lack of detail in the small spatial scales is also apparent in the spatial PSD **(D)**.

Fig. S21

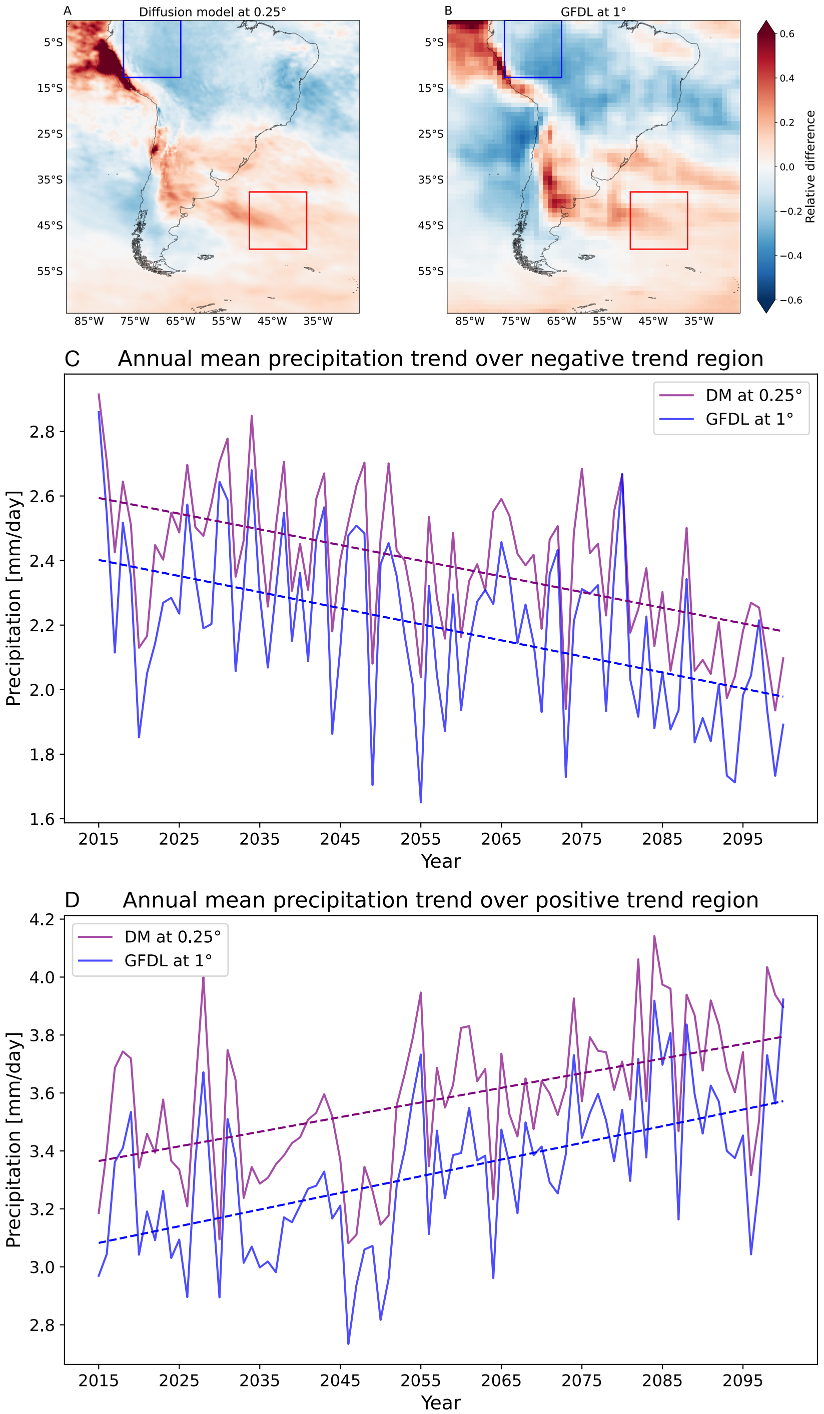

Fig. S21 Comparison of the trends for the SSP5-8.5 scenario over different sub-regions. **(A)** shows the relative climate change signal between the late 21st century (2081-2100) under the DM-corrected GFDL SSP5-8.5 scenario and the historical DM-corrected GFDL period (1995–2014). **(B)** shows the same for relative climate change signal for GFDL without the correction. We specifically chose one region with a strong negative trend (blue box) and one with a strong positive trend (red box). **(C)** The annually averaged trend over the region with negative trend between GFDL and DM-corrected GFDL is consistent. **(D)** The annually averaged trend over the region with positive trend also shows that the DM-corrected GFDL preserves the trend in GFDL. Note that the DM-corrected trend is bias corrected and hence a deviation in mean is to be expected.