Earth Systems and Environment
https://doi.org/10.1007/s41748-026-01326-w

ORIGINAL ARTICLE

ClimDiT: A Generative Latent Diffusion Transformer Framework for Multivariate Climate Downscaling

Marcos Esquivel-González[1] · Albano González[1] · Juan Pedro Díaz[1] · Juan Carlos Pérez[1] · Jose González-Abad[2] · Samuel Márquez-Cubas[1]

Received: 7 November 2025 / Revised: 10 April 2026 / Accepted: 6 July 2026
© The Author(s) 2026

Abstract Accurate, high-resolution local climate projections are essential for climate adaptation and impact assessment. Most statistical downscaling approaches rely on regression-based models that produce single deterministic estimates, thereby limiting their ability to capture the inherent variability and multivariate dependencies of the climate system. To address this limitation, ClimDiT is introduced, a generative multivariate statistical downscaling framework based on latent diffusion transformer models. Thus, instead of a single expected result, it estimates the full conditional distribution of high-resolution climate fields. The proposed system operates within a compressed, autoencoded latent space, enabling computationally efficient sampling while implicitly capturing inter-variable dependencies. In this work, ClimDiT simultaneously downscales daily maximum temperature, minimum temperature, and accumulated precipitation to approximately 5 km resolution over the Iberian Peninsula. It is trained using ERA5 predictors and ROCIO-IBEB high-resolution targets and its performance is evaluated using metrics aligned with the VALUE framework. ClimDiT attains deterministic accuracy comparable to, or moderately superior to, state-of-the-art regression networks, while improving spatial coherence. Although specialised stochastic baselines, which suffer from high deterministic errors, retain an edge in univariate and multivariate probabilistic scores, a ClimDiT ensemble generated from 50 samples demonstrates reasonable probabilistic skill and joint calibration of the three target variables (evaluated with CRPS, Brier, Energy and Variogram scores). Overall, ClimDiT demonstrates that latent diffusion models enable a generative pathway toward coherent, probabilistic, and physically consistent regional climate downscaling, advancing beyond the limitations of traditional regression-based AI systems.

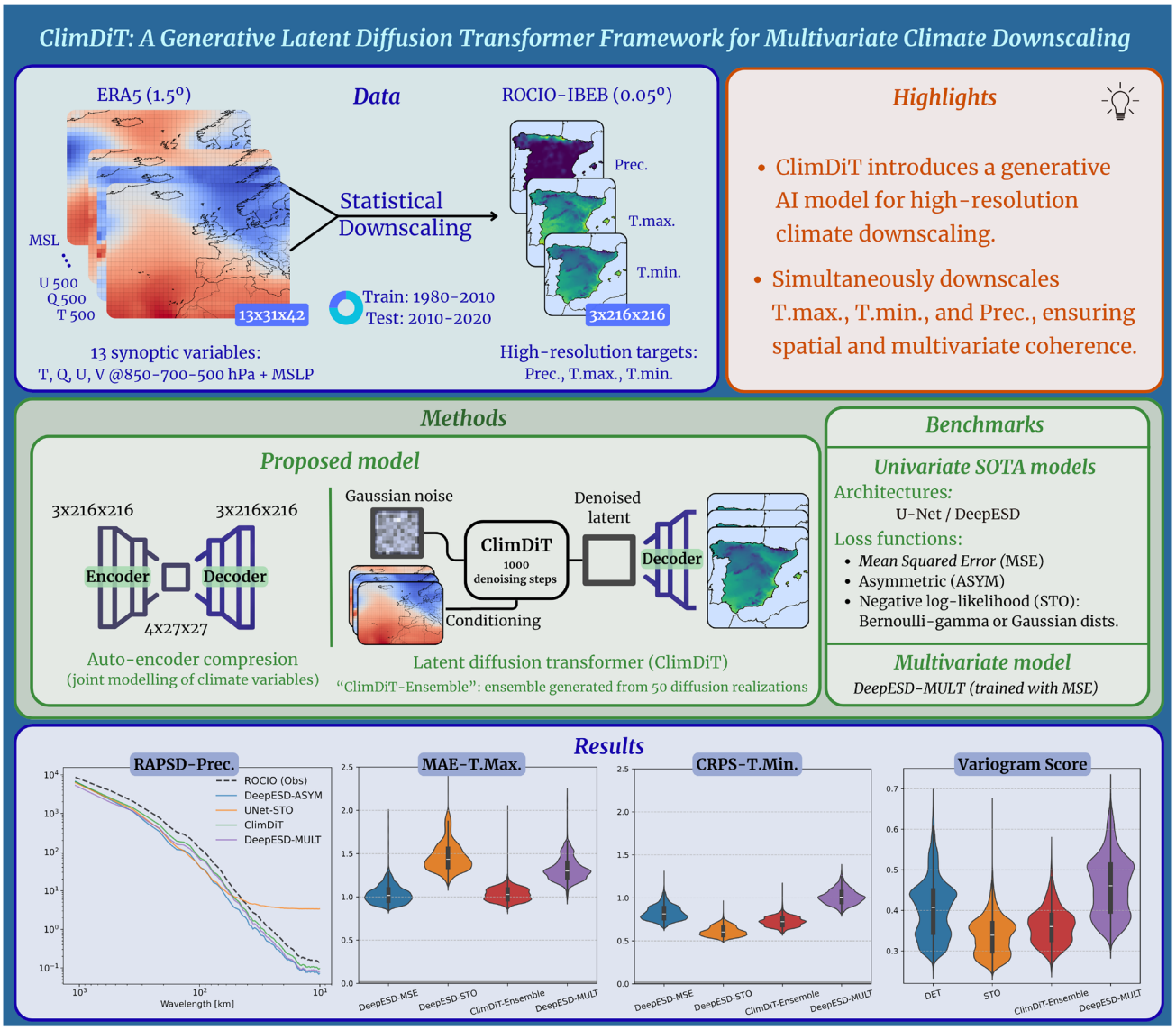
✉ Albano González
aglezf@ull.edu.es

[1] Universidad de La Laguna (ULL), San Cristóbal de La Laguna, Spain

[2] Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, Spain

Published online: 03 August 2026

Graphical Abstract



This graphical abstract provides a concise visual overview of ClimDiT: A Generative Latent Diffusion Transformer Framework for Multivariate Climate Downscaling. It summarizes the study objectives, methods, and main findings. At the top, the Data section illustrates the statistical downscaling approach, by linking coarse-resolution ERA5 predictors with high-resolution ROCIO-IBEB climate targets over the Iberian Peninsula. It highlights the three downscaled variables—maximum temperature, minimum temperature, and precipitation—as well as the corresponding training and testing periods. The Methods panel presents the core innovation: the ClimDiT model. It combines an autoencoder, which compresses climate fields into a latent space, with a latent diffusion transformer that progressively reconstructs high-resolution fields from Gaussian noise, conditioned on large-scale predictors. This reflects the proposed use of generative models, rather than the widely used regression-based AI approaches, as they allow the generation of multiple outputs for the same synoptic state. A benchmark comparison box displays the univariate state-of-the-art (SOTA) methods (DeepESD and U-Net) used as references, along with the tested loss functions (MSE, ASYM, and STO). STO refers to stochastic models, which use a Bernoulli–Gamma distribution for precipitation and a Gaussian distribution for temperature, trained using the negative log-likelihood loss. Additionally, a deterministic multivariate baseline (DeepESD-MULT, trained with MSE) is included to directly evaluate the impact of joint variable modeling within a traditional regression framework. The Results section visualizes performance across spectral, deterministic, and probabilistic metrics (RAPSD, MAE, CRPS, Variogram Score), emphasizing ClimDiT

superior spatial coherence and balanced probabilistic skill. The Highlights box succinctly reinforces these contributions: a generative, latent-space-based approach that enables coherent multivariate downscaling.

Highlights

- ClimDiT is a generative latent diffusion transformer designed for multivariate climate downscaling.
- It provides a probabilistic framework that simultaneously downscales maximum and minimum temperature as well as precipitation.
- The latent-space diffusion with an adapted DiT backbone enhances spatial coherence, as measured by SSIM and RAPSD.
- Deterministic metrics match or moderately improve results from previous deep learning methods based on U-Net and DeepESD architectures.
- A ClimDiT ensemble approaches stochastic benchmarks in probabilistic skill, while improving deterministic accuracy and spatial consistency.

Keywords Climate downscaling · Diffusion models · Transformers · Deep learning · ERA5 · ROCIO-IBEB

1 Introduction

Climate change adaptation strategies are critically dependent on accurate, high-resolution local climate projections. Although Global Climate Models (GCMs) are the primary sources for climate impact studies, their native resolutions ($\sim$ 100 km) are often too coarse for local-scale applications, particularly in regions with complex orography, which necessitates climate information at a spatial resolution on the order of a few kilometers or even less (Rössler et al. 2019). To overcome this problem, different methods of climate regionalization are used, broadly grouped into dynamic or statistical approaches. Dynamic methods employ mesoscale numerical models that take GCM outputs as boundary conditions. This makes them physically consistent under various synoptic conditions, as they are based on physical equations, but obtaining results is computationally expensive, especially at spatial scales of a few kilometers. Statistical methods are based on empirical relationships established between informative synoptic atmospheric variables (predictors) and local/regional variables of interest (predictands). They are much less computationally expensive, although they can present some problems, such as those related to transferability and extrapolation (Balmaceda-Huarte and Bettolli 2022; Baño-Medina et al. 2024; Hernanz et al. 2024).

Statistical downscaling primarily relies on two paradigms: Perfect Prognosis (PP) and Regional Climate Model (RCM) emulation. The PP framework calibrates empirical relationships between coarse-scale predictors, typically from reanalysis datasets, and fine-scale observational records. This strategy effectively decouples the downscaling process from GCM systematic biases, offering a direct link to observed climate variability. Its main drawback, however, lies in its strong dependence on historical observations, which may limit robustness when applied to future climate conditions or emission scenarios not represented in the training data. Conversely, RCM emulators approximate the output of physically based regional models using statistical or machine-learning surrogates. While they inherit the process-based coherence of RCMs and can extrapolate beyond the historical record, their fidelity is inherently tied to the quality and biases of the driving GCM–RCM simulations.

Recent studies in PP (González-Abad and Gutiérrez 2025) highlight that the current state-of-the-art in statistical downscaling using deep learning architectures primarily relies on convolutional models–either fully convolutional networks like U-Net or hybrids that combine convolutional and dense layers. These models have shown good performance in downscaling temperature and precipitation. However, they are usually trained separately for each target variable. This preference for univariate downscaling is often driven by the challenges of joint optimization; as noted by Quesada-Chacón et al. (2023), simultaneously training deep learning models for multiple variables can degrade the performance of individual variables compared to dedicated univariate networks, and may even worsen overall multivariate performance. Furthermore, joint optimization introduces additional model complexity through new hyperparameters, such as the relative weighting of different variables within the loss function. The tuning of these weights is typically performed manually in weather prediction models (Lang et al. 2024), which significantly complicates the training pipeline. On the other hand, the standard practice of using different architectures and optimization processes could also limit their ability to capture the physical and statistical interdependencies among related climate variables (González-Abad et al. 2024).

In parallel, the field of generative artificial intelligence (GenAI), particularly in image generation, has experienced unprecedented progress thanks to diffusion models (Ho et al. 2020; Nichol and Dhariwal 2021), which have surpassed previous leading generative methods (Chen et al.

2024), such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) and Variational Autoencoders (VAEs) (Kingma and Welling 2013). By iteratively denoising random noise, diffusion models are able to capture complex data distributions and generate diverse, high-fidelity samples.

These capabilities developed in GenAI for images can be extended to capture the stochastic nature of atmospheric phenomena. The underlying idea is conceptually analogous to recent advances in image generation: just as models such as Google Imagen (Saharia et al. 2022), Stable Diffusion (Rombach et al. 2023), and DALL E (Ramesh et al. 2022) can produce high-quality color images from textual prompts, it is natural to consider a model capable of generating maps of local climate conditioned on the synoptic state of the atmosphere. This approach has already inspired several applications in climate regionalization (Addison et al. 2022; Watt and Mansfield 2024). For example, Buster et al. (2024) used GANs to increase the resolution of wind and solar radiation fields to 4 km in a domain covering the United States. Other studies first employ traditional regression-based neural network systems to estimate high-resolution variables, and subsequently apply generative models, either through GANs (Rampal et al. 2025) or diffusion models (Mardani et al. 2025), to model the residuals relative to that deterministic estimate. Lopez-Gomez et al. (2025) proposed a two-step technique, in which a regional climate model provides intermediate resolution field, followed by a generative system that enhances the final output.

Furthermore, Latent Diffusion Models (LDMs) (Mármol-Rivera et al. 2025; Rombach et al. 2023) have demonstrated an optimal balance between reconstruction fidelity and computational efficiency by operating in a compressed latent space, that is learned through autoencoders. This strategy allows high-resolution problems to be addressed at a reduced computational cost and has become the state of the art in tasks such as image synthesis, super-resolution, and conditional generation. In the context of climate applications, autoencoders can also be leveraged to represent the joint distribution of target climate variables in a more natural and unified manner, further motivating their use within this framework.

Inspired by recent work (Peebles and Xie 2023), this paper proposes ClimDiT (Climate Diffusion Transformer), a statistical downscaling framework that combines the generative paradigm of latent diffusion models with the ability of transformers to model long-range spatial dependencies. ClimDiT simultaneously addresses three fundamental climate variables, daily maximum temperature, daily minimum temperature, and daily accumulated precipitation, using a single multivariable model. This design allows the system to capture both the individual variability of each target and the inter-variable dependence. Furthermore, by its stochastic nature, ClimDiT does not limit itself to producing a single smoothed scenario, but rather generates a set of plausible realizations representing the distribution of possible local climates conditioned on large-scale synoptic states. The proposed system has been designed to use the same data, both synoptic predictors and target variables, as that typically employed by deterministic statistical systems. This design facilitates methodological interchangeability and enables direct comparison between approaches.

The remainder of this paper is structured as follows. Section 2 describes the data employed in this study and outlines the methodological framework. It also introduces the benchmark models used for comparison with the proposed approach and describes the performance metrics adopted for the evaluation. Section 3 presents and discusses the results, and Sect. 4 concludes the paper by summarizing the key findings and their broader implications.

2 Methods

2.1 Data

To perform the statistical downscaling two different datasets are required: a local-scale (high-resolution) grid that serves as the ground-truth target and a large-scale (coarse-resolution or synoptic) set of atmospheric predictors. Specifically, in this work the ROCIO-IBEB dataset (Peral et al. 2017) is used as the downscaled target data and the ERA5 reanalysis dataset (Dee and Thépaut 2023) as the synoptic info. To train the machine learning models, data from the period (1980 −2010) is used as training set and data from (2011-2020) is used to assess the performance of the models (test set).

For the proposed diffusion model, there is no need to set aside a validation split due to the inherent stability of its training dynamics, as empirically observed during training. Its iterative denoising process, combined with the stochastically encoded inputs via the regularized autoencoder, acts as an implicit regularization mechanism that mitigates overfitting (Bonnaire et al. 2025; Kingma and Welling 2013), which has been also empirically verified in this study (see Sect. S2.1 in the Supplementary Material). In contrast, benchmark models, which will be described in Sect. 2.2, lack such built-in regularization and therefore reserve 10% of the training data for early stopping.

2.1.1 Target Variables (High-Resolution)

ROCIO-IBEB is a high-quality, gridded observational dataset covering the Spanish mainland, the Balearic Islands,

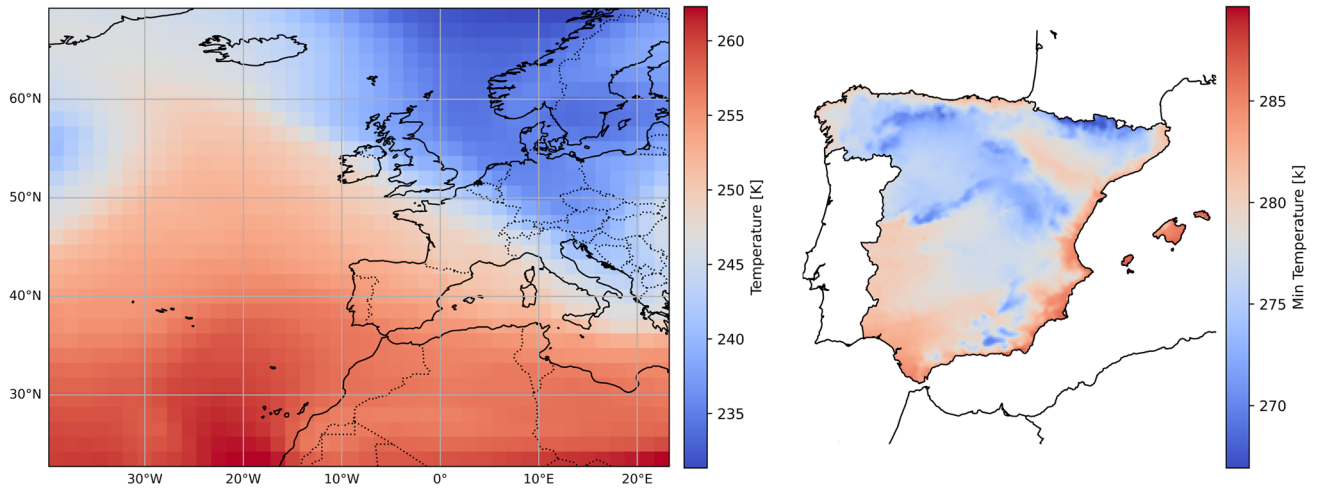


Fig. 1 Maps of two of the variables used in the study are shown for a specific day, representing the defined domains. For the low-resolution predictor domain, the ERA5 air temperature at 500 hPa is displayed (left). The high-resolution map of the ROCIO-IBEB minimum temperature (right) corresponds to the target domain

Ceuta and Melilla. It provides daily data at around 5 km spatial resolution (0.05º resolution) using rotated coordinates on a grid of 216x216 points. Figure 1 shows the target domain.

Specifically, the target variables to be downscaled are:

- Daily Maximum Temperature (Tmax)
- Daily Minimum Temperature (Tmin)
- Daily Accumulated Precipitation (Precip)

For the diffusion model, the three variables were normalized to the [0,1] interval to ensure compatibility with the pretrained autoencoder (see Sect. 2.3.2), which was originally trained to reconstruct fields within that normalized range. This guarantees consistent scaling between the encoder–decoder representation and the subsequent diffusion process.

2.1.2 Predictor Variables (Low-Resolution)

The coarse-resolution synoptic conditions used as input for the downscaling model are derived from the ERA5 reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 is a state-of-the-art global atmospheric reanalysis that provides a comprehensive and consistent picture of the Earth's climate. Based on previous works and VALUE framework (Gutiérrez et al. 2019), this work uses a set of 13 synoptic variables that are known to be physically relevant drivers of local weather patterns:

- Air temperature (K), specific humidity ($kg\,kg^{-1}$), meridional and zonal winds ($m\,s^{-1}$) at 850, 700, and 500 hPa
- Mean sea level pressure (Pa)

All predictors are standardized (subtracting the mean and dividing by the standard deviation) prior to being ingested by the models. The spatial domain of ERA5 used across all models is illustrated in Fig. 1.

2.2 Benchmark Models

The two main architectures commonly used for PP with deep learning are UNet, a fully convolutional neural network, and hybrid neural network variants, which combine both convolutional and dense layers, such as DeepESD. In this context, González-Abad and Gutiérrez (2025) homogenized various models presented in previous studies and performed a comprehensive comparison of their results on the same region used in this work.

From their extensive set of studied models, the best-performing ones were selected, including both deterministic versions and stochastic variants that produce parameters of a distribution function. These selected models, which will be briefly described below, are used in this study as benchmarks for comparing the proposed approach. For a more detailed description, readers are referred to the original work and its included references. The same spatial domain used for their predictor variables is also adopted here (Fig. 1).

2.2.1 U-Net

U-Net (Ronneberger et al. 2015) is a fully convolutional architecture extensively applied to a wide variety of tasks. Its structure is composed of an encoder path that progressively downsamples the input to capture context, and a

decoder path that upsamples the features to reconstruct the high-resolution output. Crucially, it incorporates "skip connections" between the encoder and decoder, which allow the model to combine information between the encoder and decoder, leading to sharp and accurate outputs.

2.2.2 DeepESD

DeepESD model represents a hybrid convolutional architecture, first introduced by Baño-Medina et al. (2020). Specifically, the network processes the input spatial predictors through a sequence of three convolutional layers, each followed by Rectified Linear Unit (ReLU) activations. The first two layers utilize 50 and 25 kernels, respectively. For the third and final convolutional layer, the architecture adapts to the specific target variable: it employs 10 kernels for temperature downscaling (referred to as DeepESDt architecture in this study) and a single kernel for precipitation (referred to as DeepESDpr). This distinct configuration follows the optimal architecture guidelines identified in previous studies (Baño-Medina et al. 2022). Finally, a dense layer maps these extracted spatial features to generate the specific high-resolution output at each individual grid point.

2.2.3 Loss Functions of Benchmark Models

The benchmark models rely on different loss functions depending on the climate variable being downscaled. Following the extensive review and the experiments from González-Abad and Gutiérrez (2025), two optimal combinations of model architectures and loss functions were selected for each climate variable.

For both maximum and minimum temperatures, the best-performing combinations involve DeepESD trained with Mean Squared Error (MSE) and Negative Log-Likelihood (NLL) loss functions, which correspond to DeepESD-MSE and DeepESD-STO models, respectively. While MSE directly minimizes the average squared differences between predictions and observations, the probabilistic framework minimizes the NLL by predicting the mean and standard deviation of a Gaussian distribution.

For precipitation the asymmetric loss function (ASYM), proposed in Doury et al. (2024), and stochastic NLL-based approaches consistently showed superior performance compared to others. Therefore, DeepESD with ASYM loss (DeepESD-ASYM) and UNet with a stochastic approach based on a gamma distribution (UNet-STO) were chosen as benchmarks.

The asymmetric loss function is defined as:

$$\mathcal{L} = 1/N \sum_{i=1}^{N} |y_i - \hat{y_i}| + \gamma^2 \cdot \max(0, y_i - \hat{y_i}) \tag{1}$$

where $\gamma = G(y_i)$, with G corresponding to the cumulative distribution of a gamma fitted for each grid point in the historical dataset.

On the other hand, the stochastic loss for precipitation consists on minimizing the NLL of a Bernoulli-gamma distribution:

$$P(y|X; p, \alpha, \beta) = \begin{cases} 1 - p & \text{if } y < 0.1 \\ \frac{p}{\Gamma(\alpha)\beta^{\alpha}}(y - 0.1)^{\alpha-1} e^{-(y-0.1)/b} & \text{if } y \geq 0.1 \end{cases} \tag{2}$$

where p is the probability of rain, α and β are the shape and scale parameters of a gamma distribution and Γ is the gamma function.

Furthermore, to directly address the impact of multivariate modeling within a deterministic framework and provide a fairer comparison against our proposed approach, a multivariate DeepESD model (DeepESD-MULT) is introduced as an additional baseline. This architecture shares the same backbone as the univariate DeepESDt but it outputs simultaneously maximum temperature, minimum temperature, and precipitation. To ensure some comparability with the latent diffusion model introduced in Sect. 2.3, the target variables for this baseline were preprocessed using the exact

Table 1 Summary of models evaluated. The table details the target variable(s), the underlying neural network architecture, the specific loss function optimized during training (MSE, NLL), and the total number of trainable parameters in millions (M)

Name	Variables	Architecture	Loss function	Number of parameters
DeepESD-MSE	Tmax	DeepESDt	MSE	210M
DeepESD-STO			NLL (Gaussian)	421M
DeepESD-MSE	Tmin	DeepESDt	MSE	210M
DeepESD-STO			NLL (Gaussian)	421M
DeepESD-ASYM	Prec	DeepESDpr	Asymmetric loss	21M
UNet-STO		UNet	NLL (Gamma)	8M
ClimDiT	Tmax, Tmin,	ClimDiT	Diffusion loss	72(+84) M[a]
ClimDiT-Ensemble	Prec			
DeepESD-MULT		DeepESDt	MSE	631 M

[a] For the ClimDiT models, the parameter count represents the backbone (72M) plus the Autoencoder (84M)

same normalization strategy (scaling to the [0, 1] interval and applying a $\log(1+x)$ transformation to precipitation). DeepESD-MULT was trained to minimize the Mean Squared Error (MSE) across the three channels. Although a multivariate version of the U-Net architecture was also tested, it was excluded from the final analysis because its performance was comparable to or worse than that of DeepESD-MULT. See Sect. S1 of the supplementary material for further details.

Table 1 summarizes the final set of models evaluated in this study, including the benchmark models described above and the ClimDiT model introduced in Sect. 2.3. As can be observed, for each variable to be regionalized, there is a model with a deterministic loss function and another with a stochastic one, which facilitates comparison with the proposed models, as will be explained below. This table also reports the number of trainable parameters for each model. Although DeepESD contains a large number of parameters due to its final dense layer, its computational cost during both training and inference remains remarkably low. The computational cost of each model is discussed in Sect. 2.4 following the presentation of the training hyperparameters.

2.3 Diffusion Approach

The proposed approach moves beyond deterministic prediction to a generative one. Instead of asking "what is the single most likely high-resolution prediction?", the question is "what is the distribution of all possible realistic high-resolution predictions, given the low-resolution contexts?" This is achieved using a Gaussian diffusion model (Ho et al. 2020).

The adoption of latent space for the diffusion process (Rombach et al. 2023), rather than the original data space, offers substantial advantages. This strategy not only reduces computational cost, enabling the backbone model to efficiently handle high-resolution variables, but also (with appropriate regularization constraints) facilitates the capture of more abstract and meaningful features. Therefore, the proposed system consists of three main modules: an autoencoder, which handles transformations between the original and latent data spaces; a diffusion model, operating entirely within this compressed latent representation; and the backbone, specifically designed to progressively denoise the latent states during the diffusion steps. These modules are described in this subsection.

2.3.1 Formulation of the Diffusion Process

Diffusion is modelled as a discrete-time Markov chain that progressively corrupts the data converting the original data distribution $q(x_0)$, into a Gaussian distribution $q(x_T) \sim \mathcal{N}(0, \mathrm{I})$. Therefore, given an initial clean sample x_0, the forward process is defined by $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1},\ \beta_t\mathrm{I})$ for $t = 1, 2, ..., T$ with the closed form

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\cdot \mathrm{I}) \tag{3}$$

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{I}) \tag{4}$$

where $\alpha_t = 1-\beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$ are controlled by a predefined linear variance schedule $\{\beta_t\}$.

By construction, the ground truth posterior distribution is:

$$q^*(x_{t-1} \mid x_t, x_0) = \mathcal{N}(x_{t-1};\ \tilde{\mu}_t(x_t, x_0),\ \tilde{\beta}_t\mathrm{I}), \tag{5}$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1-\bar{\alpha}_t}\,x_0 + \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\,x_t, \tag{6}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\beta_t. \tag{7}$$

The objective is to approximate this reverse Markov chain using as conditions the low-resolution maps:

$$p_\theta(x_0|\mathrm{cond}) = p_\theta(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, \mathrm{cond}) \tag{8}$$

$$\begin{aligned} &p_\theta(x_{t-1}|x_t, \mathrm{cond}) \\ &= \mathcal{N}(x_{t-1}; \mu_\theta(x_t, \mathrm{cond}, t), \mathbf{\Sigma}_\theta(x_t, \mathrm{cond}, t)) \end{aligned} \tag{9}$$

where the neural network is used to model the statistic of p_θ. In practice, the backbone network learns to predict the noise term ϵ_θ and the variance Σ_θ, where the noise is related to μ_θ with the following expression:

$$\mu_\theta(x_t, t, \mathrm{cond}) = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, \mathrm{cond}, t)\right) \tag{10}$$

The network is trained to optimize the variational lower bound of the log-likelihood of x_0. Following previous works (Nichol and Dhariwal 2021; Peebles and Xie 2023), the approach followed is to minimize $\mathcal{L}_{simple}$ for ϵ_θ and $\mathcal{L}(\theta)$ for Σ_θ:

$$\mathcal{L}_{simple}(\theta) = ||\epsilon_\theta(x_t) - \epsilon_t||^2 \tag{11}$$

$$\begin{aligned} \mathcal{L}(\theta) = &- log p_\theta(x_0|x_1) + \\ &\sum_t \mathcal{D}_{KL}(q^*(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \end{aligned} \tag{12}$$

Subsequently, at inference time, new samples are generated by starting from Gaussian noise and recursively applying the learned reverse transitions.

In this work, diffusion hyperparameters from Ablated Diffusion Model (ADM) (Dhariwal et al. 2021) and Diffusion Transformer (DiT) (Peebles and Xie 2023) are retained, using 1000 timesteps with a linear variance schedule ranging from 10^{-4} to 10^{-2}. Also the covariance Σ_θ parametrization and timestep embeddings from ADM are used.

2.3.2 Latent Diffusion. the Autoencoder

To make this procedure computationally feasible for high-resolution climate data, operating in a compressed latent space learned by an autoencoder is desirable. Therefore, a Latent Diffusion Model (LDM) (Rombach et al. 2023), which performs diffusion in the encoded latent space, is used. After a new 'latent' sample is created through the diffusion process, the decoder of this autoencoder is used to obtain the final maps.

Taking advantage of transfer learning, an autoencoder pre-trained from Stable Diffusion (Podell et al. 2023) with a downsample scale factor of 8 was chosen. To promote a more continuous and smooth latent space, similar to that of VAEs, this autoencoder was originally trained to have a probabilistic latent space with a Kullback-Leibler regularisation term ($\mathcal{L}_{reg}$) with a factor of 10^{-6} (see more details in Rombach et al. (2023)). This design ensures that the diffusion process can operate consistently within the latent space, as it facilitates the interpolation between states and the gradual noise reduction.

However, this model was pre-trained to predict RGB channels of color images. In order to apply it correctly to this case study, it is fine-tuned in the training set to predict *Tmax*, *Tmin* and $log1p(\text{precipitation})$. Logarithmic transformation in precipitation data is used here to mitigate skewness (Huang et al. 2023) and better match the smoothness of the values to those of the pixels in the image space in which the autoencoder was pre-trained. Also, better reconstruction results were obtained using it instead. Specifically, the autoencoder is fine-tuned with the following loss function:

$$\mathcal{L}_{AE} = MSE(\mathcal{D}(\mathcal{E}(x)), x) + 10^{-8} \cdot \mathcal{L}_{reg}(x, \mathcal{E}) \tag{13}$$

Table 2 Autoencoder results in the test set before and after being fine-tuned

	Reconstruction error (RMSE)			Latent space statistics	
	Tmax	Tmin	Prec	Z_{mean}	Z_{std}
Fine-tuned	0.13	0.10	0.16	8.4e-3	1.02
Original weights	1.06	0.77	1.37	2.0e-1	9.29

where $\mathcal{D}$ is the decoder, $\mathcal{E}$ is the encoder ($z = \mathcal{E}(x)$) and $\mathcal{L}_{reg} := \mathcal{D}_{kl}(q_{\mathcal{E}}(z|x)||\mathcal{N}(z; 0, 1))$ is the Kullback-Leibler term between the encoder distribution and a normal distribution. Moreover, MSE refers to the mean square error, which is the term responsible for reducing the reconstruction error.

This method allows the three climate variables (Tmax, Tmin and Precipitation) to be compressed from their original spatial dimensions 216x216x3 to 27x27x4 in the latent space. To ensure that the pre-trained autoencoder moderately represents the geophysical variables without introducing biases toward visual realism, its performance was evaluated before and after the fine-tuning process. Table 2 presents the reconstruction error (RMSE) and the statistics of the latent space in the test set. Remarkably, applying the original image-trained weights directly to climate data yields reasonable reconstructions. This validates the use of these pre-trained weights as a robust initialization, demonstrating that the network's generalized spatial feature extraction capabilities may be transferable to geophysical fields. However, to achieve the sub-degree precision required for climate downscaling and to properly regularize the latent space, fine-tuning is necessary. After fine-tuning, the RMSE is substantially reduced, and the latent space statistics align closely with a standard normal distribution, demonstrating that the $\mathcal{L}_{reg}$ term effectively regularizes the encoded space despite the small weighting factor (10^{-8}). Furthermore, Fig. 2 visually corroborates this improvement by comparing the spatial reconstruction bias for a randomly selected day. The fine-tuning process successfully refines the outputs, removing outlier pixels (as evidenced by the color bar value ranges) and ensuring the network is adapted to the physical properties and joint distribution of the climate variables.

2.3.3 Backbone Architecture

The core of the diffusion model is the denoiser network. Here ClimDiT is presented, a specifically designed DiT architecture (Peebles and Xie 2023). Transformers, with their self-attention mechanism, are uniquely suited to modeling complex, long-range dependencies within data (Vaswani et al. 2017). This allows ClimDiT to capture domain-wide spatial interactions and produce spatially coherent outputs. In this work, the original DiT blocks with adaLN-Zero modulation are adapted by incorporating the SwiGLU activation function (Shazeer 2020) in the Feed Forward Network (FFN), resulting in slightly better performance in internal experiments.

Figure 3 shows a schematic diagram of the proposed ClimDiT architecture with its corresponding blocks. The main denoiser employs a total of 10 (N) DiT blocks, while the conditioning encoder uses 5 (N_{cond}). Each block

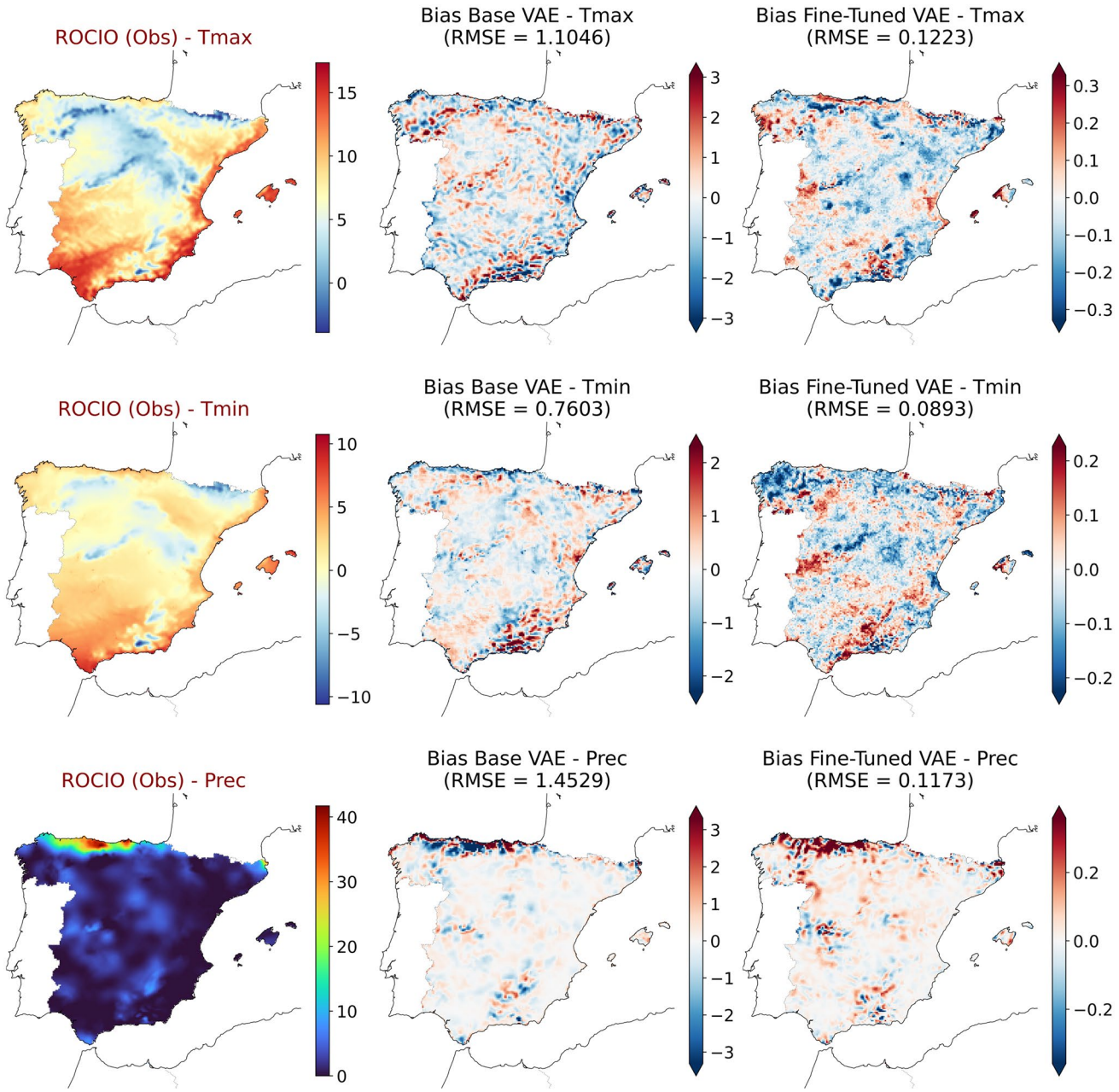


Fig. 2 Comparison of the Variational Autoencoder (VAE) reconstruction performance for a randomly selected day from the test set. The first column displays the observed high-resolution fields for maximum temperature, minimum temperature, and daily precipitation. The second and third columns show the spatial distribution of the reconstruction bias (Prediction - Observation) for the Base VAE (original pre-trained weights) and the Fine-Tuned VAE, respectively. Note that the scales are different, since the corresponding errors differ by an order of magnitude. For better visualization of spatial patterns, the bias color scales are saturated at the 98th percentile. The spatial Root Mean Squared Error (RMSE) is indicated in the title of each bias map

incorporates 8 attention heads with a hidden size of 512 (64 per head), representing an intermediate configuration between the original ViT (Dosovitskiy et al. 2021) Small (S) and Base (B) models.

Because ClimDiT is a conditional generative model, its ability to produce the specific local meteorological fields for a given day, rather than a random plausible output, depends on how effectively the synoptic predictors are encoded to modulate the denoising process. To achieve this, a specific conditioning encoder processes the low-resolution synoptic conditions together with the diffusion timestep before passing them to the main backbone. The ClimDiT conditioning mechanism is built on three key principles:

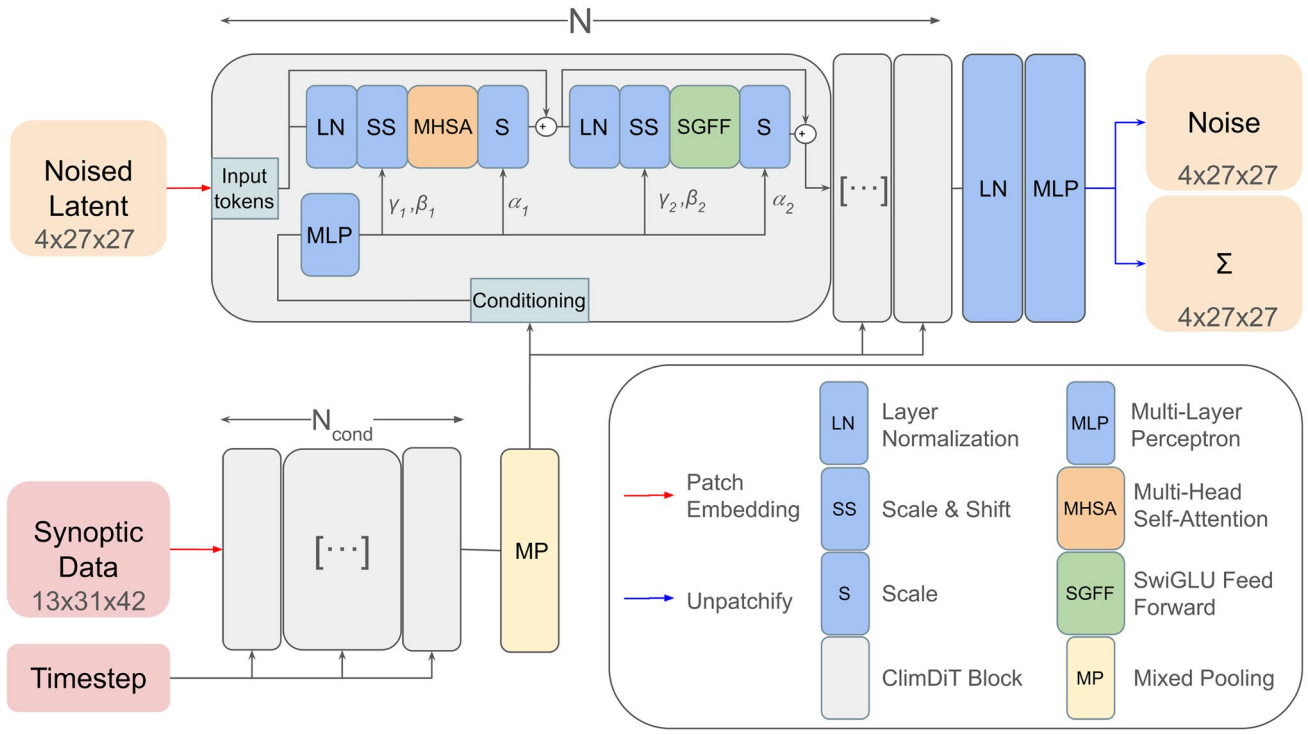


Fig. 3 ClimDiT architecture. The synoptic data enters in an specific encoder which is modulated by the timestep. Later, Mixed Pooling (MP) compress all this info in a vector which modulates the main blocks of the denoiser. The figure also zooms in an specific transformer block (ClimDiT block) in order to show specific details

- **Feature Extractor.** To process the low-resolution input maps, transformer blocks are employed to infer global relationships between synoptic variables. The low-resolution input maps (synoptic variables) are first processed by a dedicated feature extractor composed of $N_{cond} = 5$ transformer blocks. This allows the network to infer long-range spatial relationships among the synoptic drivers across the entire Iberian Peninsula before guiding the high-resolution generation performed by the main denoiser.
- **Fused Conditioning.** Unlike standard regression networks, the feature extractor is conditioned on the current diffusion timestep, meaning that the spatial context provided to the main DiT blocks varies with the noise level. This enables the model to adapt its attention to different types of conditioning information throughout the denoising process.
- **Mixed Pooling.** To convert the spatially aware synoptic features into a format suitable for modulating the main diffusion blocks with AdaLN-Zero, a mixed pooling strategy is applied. The mean and maximum values across all tokens of the feature extractor are passed through a linear layer without a bias term to obtain the final conditioning vector, which modulates the ClimDiT blocks of the denoiser. Omitting the bias prevents the network from learning a static offset, thereby ensuring that the modulation relies more on the conditioning context.

Finally, leveraging the stochastic nature of the diffusion reverse process, ClimDiT is evaluated under two configurations: (i) a single-sample configuration, and (ii) an ensemble of 50 samples. Based on a sensitivity analysis (detailed in Sect. S6 of the Supplementary Material), this ensemble size was determined to be adequate, as no significant performance improvements are observed beyond this number of members. In the ensemble configuration, the mean across members is used as the deterministic prediction (i.e., deterministic metrics are evaluated on this ensemble mean), while the empirical cumulative distribution provides probabilistic assessments (i.e., these metrics are applied to all ensemble members jointly). Throughout the paper, these two variants are referred to as ClimDiT and ClimDiT-Ensemble, respectively, as summarized in Table 1.

2.4 Training Hyperparameters

For the benchmark models (DeepESD and U-Net), training followed the configuration proposed in González-Abad and Gutiérrez (2025), with a batch size of 64, the Adam optimizer with a learning rate of 1×10^{-4}, and an early-stopping

Table 3 Metrics used to compare models. For precipitation, SSIM, ES and VS use $\log(1 + \text{precip})$; other precipitation metrics use raw values

Metric	Description	Variable (Units)	Target score
Deterministic metrics			
P02 bias	Bias for the 2nd percentile	Tmin, Tmax (°C)	0
	Bias for the mean	Tmin, Tmax (°C)	0
Mean bias	Relative bias for the mean	Precipitation (%)	0
P98 bias	Bias for the 98th percentile	Tmin, Tmax (°C)	0
P99 bias	Bias for the 99th percentile	Precipitation (mm)	0
R01 bias	Relative bias for the ratio of wet days	Precipitation (%)	0
Rx1day bias	Relative bias for the climatology of max daily precipitation	Precipitation (%)	0
SDII bias	Relative bias for the simple daily intensity index	Precipitation (%)	0
TNn bias	Bias for the annual minimum of daily minimum temperatures	Tmin (°C)	0
TXx bias	Bias for the annual maximum of daily maximum temperatures	Tmax (°C)	0
		Tmin, Tmax (°C)	0
RMSE	Root mean squared error	Precipitation (mm)	0
		Tmin, Tmax (°C)	0
MAE	Mean absolute error	Precipitation (mm)	0
Std. ratio	Ratio of std. deviations	Tmin, Tmax	1
Interannual var. ratio	Ratio of interannual variability	Precipitation	1
Probabilistic metrics			
		Tmin, Tmax (°C)	0
CRPS	Continuous ranked probability score	Precipitation (mm)	0
BS	Brier score	All variables	0
Spatial fidelity metrics			
SSIM	Structural similarity index	All variables	1
RAPSD	Radially averaged power spectrum	All variables	shape match
Multivariate metrics			
ES	Energy score	Joint	0
VS	Variogram Score	Joint	0

criterion of 60 epochs, retaining the weights corresponding to the best validation performance.

In contrast, the auto-encoder component of the ClimDiT model was fine-tuned with a batch size of 20, using the AdamW optimizer, with a learning rate of 2×10^{-4} and a weight decay of 10^{-2}, for a total of 350 epochs. Analogously, the weights of the auto-encoder corresponding to the best performance on the validation split were retained. An alternative fine-tuning setup with more conservative hyperparameters was evaluated, resulting in poorer generalization on the test set (see Sect. S2.2 in the Supplementary Material). Given the robust performance already achieved by the baseline configuration (Table 2), further tuning was deemed unnecessary, prioritizing computational resources for the diffusion model.

The ClimDiT backbone followed the configuration commonly used in diffusion-based transformers (Peebles and Xie 2023), and was trained for 400,000 steps with a batch size 64 (approximately 2,300 epochs). During the training process, a dropout probability of 0.1 was used in the conditioning encoder. Optimization was performed using AdamW without weight decay and a fixed learning rate of 1×10^{-4}. Additionally, an Exponential Moving Average (EMA) version of the model was maintained during training with a decay rate of 0.9999, and this EMA model was ultimately used for evaluation. As discussed in sect. 2.1, the validation split was not required for ClimDiT training.

Regarding the computational cost of the evaluated models, training the regression baselines on a single NVIDIA A100 GPU took between 25 and 50 minutes, whereas fine-tuning the VAE required approximately 1 day, and training the diffusion backbone took 3 days. Conversely, inference for the entire 10-year test set, a crucial metric for operational applicability, was on the order of seconds for the deterministic and stochastic baselines, and approximately 9 minutes for the latent diffusion model. Although the higher computational demand of the generative framework is evident, despite the use of a compressed latent space, these requirements remain substantially lower than those needed to generate 10 years of dynamical Regional Climate Model (RCM) simulations, which typically take weeks to months.

2.5 Evaluation Metrics

To assess model performance, a comprehensive evaluation is conducted following the guidelines of the COST Action VALUE framework, widely used to validate statistical downscaling methods. Specifically, the extensive set of metrics used in this work is presented in Table 3, and the specific formulation used for all of them is provided in 5.. They can be classified into four complementary groups:

- **Deterministic metrics**. Classical gridpoint-wise metrics are considered to quantify bias, error and variability. Bias is reported in terms of mean bias (absolute for temperatures, relative for precipitation), percentile biases (P02 and P98 for temperatures; P99 for precipitation), and precipitation-specific climatological indices (R01, Rx1day, SDII). Errors are summarized using RMSE and MAE, while variability is assessed through the ratio of standard deviations for temperatures and the ratio of interannual variability for precipitation.
- **Probabilistic metrics**. To evaluate the full predictive distributions, the Continuous Ranked Probability Score (CRPS) and the Brier Score (BS) (Ferro et al. 2008) are employed. These scores quantify, respectively, the overall accuracy of predictive distributions and the reliability of event probabilities. For deterministic models, the CRPS reduces to the MAE and the Brier Score reduces to a binary squared-error measure. In the case of UNet-STO for precipitation, no explicit CRPS formulation was found in the literature; therefore, the corresponding expression for its distribution was derived (see 6.).
- **Spatial fidelity metrics**. Structural realism is quantified using the Structural Similarity Index (SSIM) (Wang et al. 2004) and the radially averaged power spectral density (RAPSD) (Sinclair and Pegram 2005). SSIM is computed daily between predicted and observed high-resolution fields and subsequently averaged; for precipitation, SSIM is applied to $log1p(\text{precip})$ to mitigate skewness and smooth the distribution values. Because SSIM relies on image variances and covariances, applying a logarithmic transformation to precipitation fields yields a more accurate similarity estimation. RAPSD, obtained from azimuthally averaged 2-D discrete Fourier spectra, characterizes the spatial distribution of energy across wavelengths and is presented as spectral curves.
- **Multivariate metrics** Cross-variable coherence is assessed using proper multivariate scores: the Energy Score (ES) (Gneiting and Raftery 2007) and the Variogram Score (VS) (Scheuerer and Hamill 2015). The joint vector $(T_{\max}, T_{\min}, \log(1 + \text{precip}))$ is standardized (using mean and variance of observations over the test period) prior to scoring to avoid unit differences (Schepen et al. 2020). ES reflects overall multivariate calibration and sharpness, generalizing CRPS to multivariate ensembles, while VS (here with exponent $p = 0.5$) emphasizes the consistency of pairwise increments. The limited sensitivity of ES (Scheuerer and Hamill 2015) to the dependency structure is complemented by the variogram-based analysis provided by VS. To facilitate comparison across paradigms, the univariate stochastic baselines (DeepESD-STO for temperatures and UNet-STO for precipitation) are grouped into a single aggregated baseline denoted as "STO". Similarly, the deterministic univariate baselines are grouped as "DET". Because the stochastic baselines output independent marginal distribution parameters rather than explicit multivariate predictive distributions, there are no closed-form analytical expressions available to directly compute their multivariate scores. Consequently, to evaluate their multivariate performance (ES and VS), an empirical ensemble of 60 members ($M = 60$) was generated by sampling from the predicted distributions. A value of 60 proved sufficient in our experiments, as we observed no improvement in the metrics when we increased the number of samples to 100. Conversely, the deterministic baselines are evaluated straightforwardly as a degenerate ensemble with a single member ($M = 1$).

Table 4 SSIM values for all models

Model	Tmin	Tmax	Log1p(Prec)
ClimDiT	0.921	0.981	0.705
ClimDiT-Ensemble	0.923	0.982	0.709
DeepESD-MULT	0.887	0.960	0.488
DeepESD-MSE	0.922	0.980	–
DeepESD-STO	0.738	0.780	–
DeepESD-ASYM	–	–	0.592
UNet-STO	–	–	0.348

3 Results and Discussion

In this section ClimDiT results are compared against DeepESD and U-Net for daily Tmax, Tmin, and precipitation variables.

3.1 Spatial Consistency

Firstly, SSIM and RAPSD are used to assess the spatial fidelity. Table 4 shows the Structural Similarity Index (SSIM) for each model and variable. ClimDiT achieves the highest SSIM values, together with DeepESD-MSE and DeepESD-MULT, for maximum temperature. However, the multivariate baseline DeepESD-MULT attains lower SSIM values than the other two models for minimum temperature, indicating a moderate degradation when moving from univariate to joint prediction. For precipitation, evaluated on log1p-transformed fields, ClimDiT achieves the best scores overall, while DeepESD-MULT improves over stochastic baselines but remains clearly below deterministic approaches specifically designed for this variable (DeepESD-ASYM). The benchmark models that use stochastic loss functions, UNet-STO and DeepESD-STO, exhibit clearly lower SSIM across all three variables, and

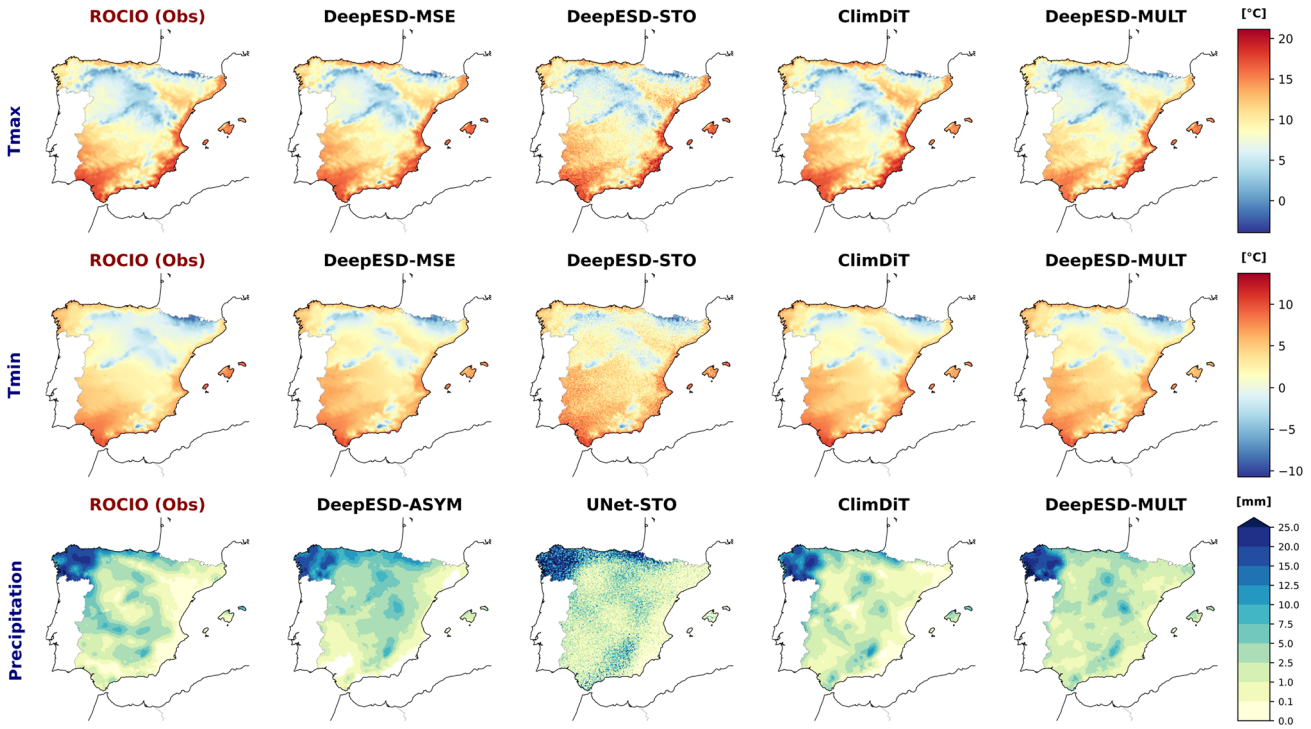


Fig. 4 Visual comparison of model outputs for a widespread precipitation event (> 1 mm over more than 20% of the domain) in the test set. The maps display maximum temperature, minimum temperature, and accumulated precipitation

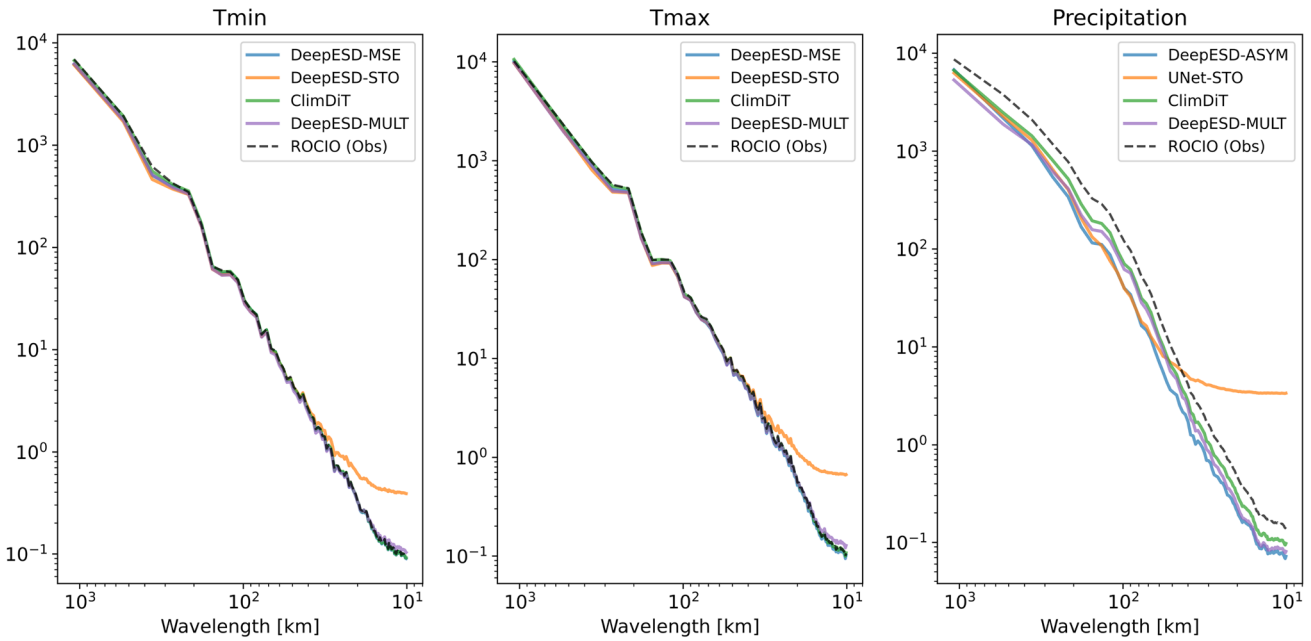


Fig. 5 Radially averaged power spectral densities as a function of wavelength

no material difference is observed between the single-sample ClimDiT and the 50-member ClimDiT-Ensemble. These results indicate that the latent diffusion transformer is effective at preserving coherent spatial patterns, in a very similar way to deterministic models for temperature and better than these for precipitation. In general, all models have more difficulty in correctly representing the spatial distribution of precipitation than that of temperature, as can be deduced from the lower SSIM values.

To illustrate these structural characteristics, Fig. 4 presents a comparison of the generated fields for a selected day characterized by a widespread precipitation event. As

observed in the maps, the lower SSIM and spatial performance of traditional stochastic models are driven by a characteristic high-frequency noise pattern. This artifact is a direct consequence of the independent point-by-point sampling used in these models (González-Abad and Gutiérrez 2025). While this approach allows them to accurately capture distributional properties, as shown in Sects. 3.2 and 3.3 for the bias metrics, it simultaneously compromises spatial consistency and leads to point-wise errors. Although using the mean of the predicted distributions would smooth these fields and improve pixel-wise errors, it would compromise their probabilistic advantages by degrading their calibrated biases. In contrast, for precipitation, the ClimDiT model shows a slightly better fit to the spatial pattern of the observations than the other models.

Figure 5 compares the RAPSD of the predicted daily fields with observations, providing a scale-by-scale view of spatial realism. ClimDiT-Ensemble has been omitted from the figure for the sake of simplicity, as its results are indistinguishable from those of ClimDiT. For temperature (Tmin, Tmax), spectral differences are modest across methods. At synoptic scales (wavelengths $\sim 10^3$ km), all methods align closely with the observed spectrum, indicating that large-scale variability is robustly captured. In the mesoscale band of tens of kilometers (roughly 20–80 km), the stochastic baselines (DeepESD-STO) exhibit a noticeable excess of power relative to observations, with spectra surpassing the observational curve. This behavior is consistent with pixel-wise stochastic sampling discussed previously and illustrated in Fig. 4, which weakens short-range spatial correlations. By contrast, the deterministic baselines and the ClimDiT variants closely track the observed spectral slope.

For precipitation, where scale interactions are stronger, all models underestimate the observed spectrum, indicating smoother fields than observed. A plausible cause is that per-pixel training losses, when applied to the zero-censored and skewed distributions typical of precipitation data, discourage the complex high-variance small-scale structures. However, ClimDiT, followed by DeepESD-MULT as the second-best performing model, mitigate this loss of variance, particularly at spatial scales of approximately 100 km and 10 km, yielding a precipitation spectrum that more closely matches observations. This phenomenon may be attributed to ClimDiT's operation within the latent space, which, due to the autoencoder's original training formulation, combining perceptual loss with a patch-based adversarial objective (Rombach et al., 2022), inherently preserves fine-scale variability. This feature fidelity within the compressed representation directly translates to a minimization of reconstruction artifacts and, consequently, a reduction in pixel-wise errors. Furthermore, the use of the log1p transformation for rainfall in both DeepESD-MULT and ClimDiT may also contribute to improved preservation of spatial patterns. In contrast, the stochastic decoder of UNet-STO degrades variability at smaller scales, a behavior that is also evident in the temperature spectra of the stochastic variants and is attributable to its sampling methodology.

3.2 Univariate Performance

In this Subsect. 3.2 and in the following one, 3.3, violin plots illustrate the spatial distribution of metrics explained in Sect. 2.5, showing not only the median performance for the whole domain, but also the consistency of the model across every high-resolution grid point of the Iberian Peninsula. To provide spatial context, maps of selected metrics are included in Sect. S3 of the Supplementary Material.

3.2.1 Temperature Downscaling (Tmax,Tmin)

Regarding temperature downscaling, the results presented in Fig. 6 (Tmin) and Fig. 7 (Tmax) demonstrate that ClimDiT consistently matches or slightly outperforms the benchmarks in terms of both median values and distribution of errors across the spatial domain.

For minimum temperature, ClimDiT reduces the bias in the cold tail (P02) relative to baselines (DeepESD-MSE, DeepESD-STO, DeepESD-MULT). For the annual minimum (TNn), ClimDiT and ClimDiT-Ensemble shows a slightly positive median bias, whereas DeepESD-STO is closer to zero but with a much larger spatial spread, indicating larger errors in certain regions (see the lower end of the DeepESD-STO violin plot, around −1.5 °C). The standard-deviation ratio of ClimDiT and its ensemble is very close to 1, indicating that it preserves the observed variability. By contrast, both univariate benchmark models are mildly under-dispersed, with median values below 1 while DeepESD-MULT exhibits nearly its entire spatial distribution below 1. In the warm tail of Tmin (P98), both ClimDiT variants perform slightly worse than the univariate benchmarks in the P98 bias, although this upper percentile is arguably less relevant for minimum temperatures, where the focus is typically on cold extremes. The mean bias shows similar values for all models; however, ClimDiT and its ensemble are marginally closer to zero.

Looking at RMSE and MAE, DeepESD-STO, followed by DeepESD-MULT, exhibits larger values than the other models, which perform comparably. On the other hand, in the probabilistic metrics, DeepESD-STO attains the lowest CRPS and the best Brier Score for cold events (BS(>p02)), indicating that although its distributional shape is well captured, its point-wise predictive accuracy is weaker. Regarding ClimDiT-Ensemble, it matches single-sample ClimDiT on deterministic metrics but improves, as expected, the

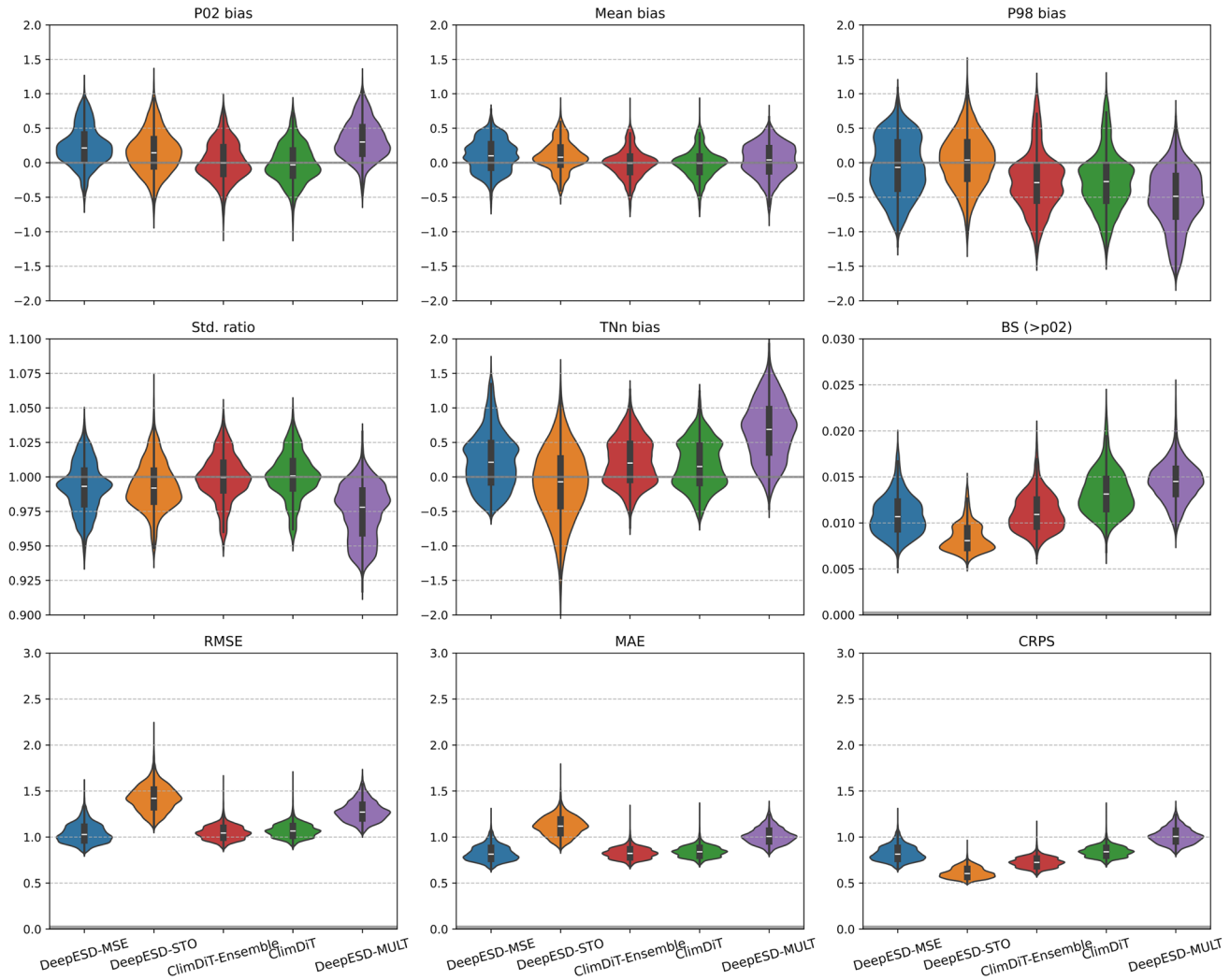


Fig. 6 Results in the test split of the deep learning methods for minimum daily temperatures. Violins indicate the results distribution across Iberian Peninsula grid points

probabilistic ones (lower CRPS and BS (>p02)), although not enough to surpass DeepESD-STO.

A detailed analysis across various grid points and multiple days (See Figs. S10 and S11 in Supplementary Material) revealed that DeepESD-STO consistently predicts broad probability distributions. Consequently, even when the distribution's mean deviates significantly from the observation, the observation often falls within the predicted distribution with relatively high probability.

Figure 7 displays the results for maximum temperatures. The differences among models are less pronounced than for minimum temperatures. Both ClimDiT variants and DeepESD-MSE achieve lower RMSE and MAE than DeepESD-STO and DeepESD-MULT, which perform considerably worse on these deterministic errors. Mean bias and P98 bias are similar across models, and the standard-deviation ratio is slightly better for DeepESD-STO. Additionally, ClimDiT variants and DeepESD-MULT tend to underestimate TXx (negative bias). However, the ClimDiT configurations exhibit substantially smaller dispersion than the benchmarks, suggesting improved spatial robustness in capturing annual maxima.

The probabilistic metrics follow the same behavior than for minimum temperatures: DeepESD-STO provides the strongest probabilistic calibration, with the lowest CRPS and the best Brier score for hot extremes (BS, Tmax > p98), indicating superior probabilistic reliability. ClimDiT-Ensemble improves over single-sample ClimDiT in these metrics, but not enough to surpass DeepESD-STO.

Regarding the spatial distribution of errors (Figs. S6 and S7), for cold extremes, P02 of minimum temperature, the different models tend to overestimate temperatures in the western side and, to a lesser extent, in the mountain ranges of the southeastern Iberian Peninsula. For warm extremes (P98), derived from maximum temperature, the models show an overestimation in the same areas, as well as

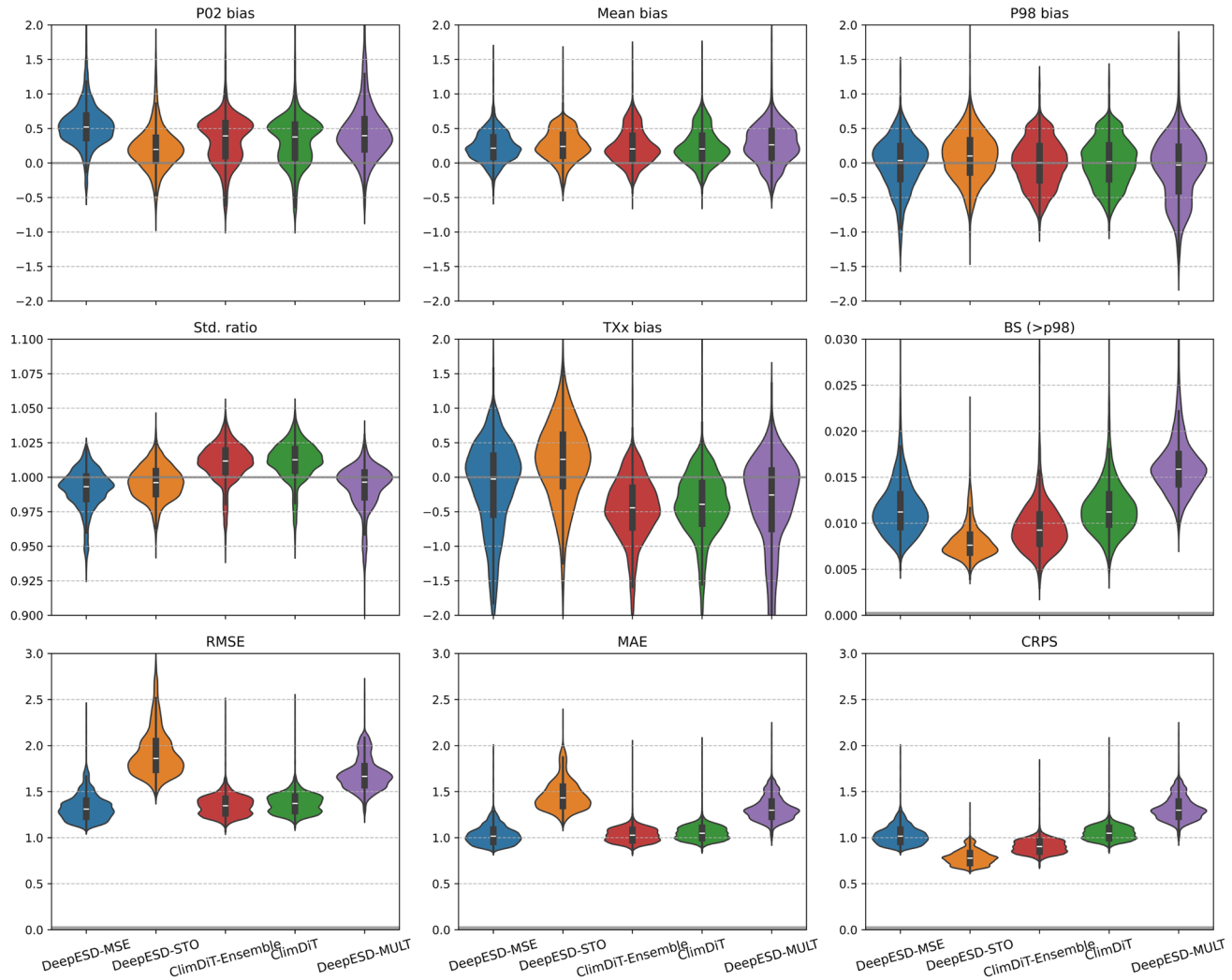


Fig. 7 Same as Fig. 6 but for maximum daily temperature

in the Ebro River basin, located in the northeastern part of the peninsula. It is precisely in this basin where the highest RMSE values occur, notably for DeepESD-STO and DeepESD-MULT. This region has a complex continental Mediterranean climate with strong temperature variability, low rainfall, and the influence of the cold, dry Cierzo wind.

3.2.2 Precipitation Downscaling

Figure 8 shows the results for precipitation downscaling. DeepESD-ASYM and ClimDiT display overall similar behavior, although ClimDiT achieves the best performance in SDII bias and presents the narrowest distributions in R01 bias and the interannual variance ratio, indicating larger spatial robustness.

In contrast, the multivariate deterministic baseline, DeepESD-MULT, struggles significantly to capture precipitation patterns. It exhibits severe systematic underestimations of precipitation intensity and extremes, evidenced by its pronounced negative biases in mean bias, P99, Rx1day, and SDII, while simultaneously overestimating the occurrence of wet days (positive R01 bias).

On the other hand, UNet-STO produces smaller biases in Rx1day, P99, and mean precipitation compared with the other models. This improvement, however, comes at the cost of higher deterministic absolute error, with RMSE and MAE approximately $1mm$ higher on average, and with wider dispersion across nearly all metrics. The largest errors generally occur in the northern part of the Iberian Peninsula, the eastern coastal region, and the mountain ranges located in the southernmost and central-western regions (Fig. S8). These regions are characterized by frequent precipitation or by intense rain events.

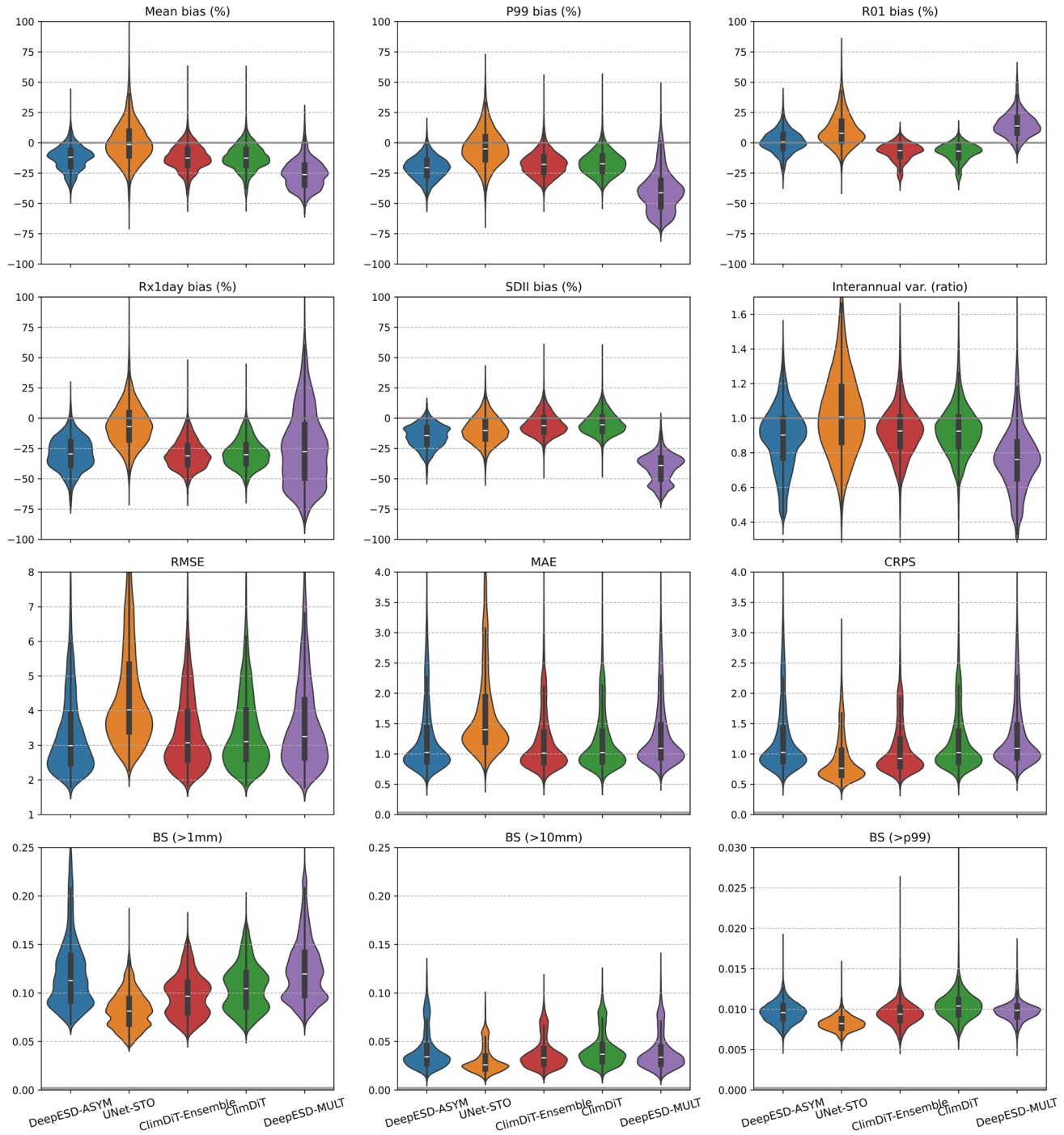


Fig. 8 Same as Fig. 6 but for precipitation

Regarding the probabilistic metrics, UNet-STO shows the lowest CRPS, suggesting that the shifted gamma distribution used may better capture precipitation uncertainty than the ClimDiT-Ensemble. However, the MAE metric, closely related to CRPS in its deterministic form, favors ClimDiT variants. The similarity between MAE and CRPS results for ClimDiT-Ensemble suggests the ensemble approach does not substantially improve the uncertainty estimation compared to UNet-STO, which has a specifically designed distribution. This behavior is corroborated by the visual analysis of the predictive distributions included in the supplementary material (Fig. S12). UNet-STO tends to generate very broad continuous distributions for precipitation, which spread the probability density but allow it to cover the observations more reliably than the ClimDiT-Ensemble, whose members do not deviate as far from their mean. The

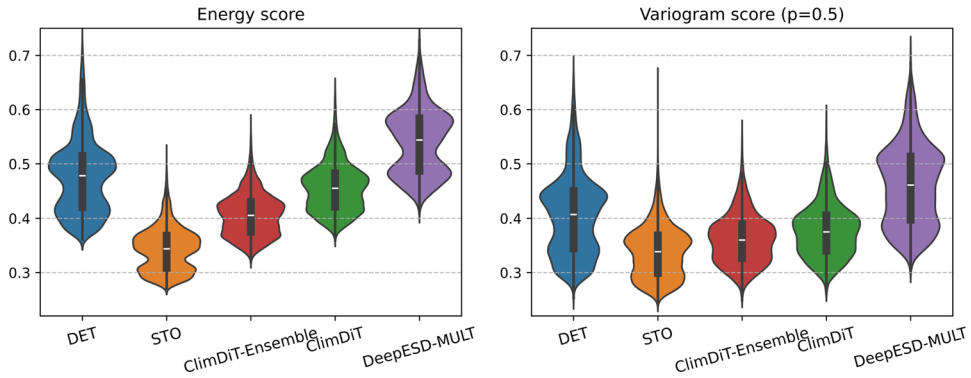


Fig. 9 Energy score and variogram score using the three target variables: maximum temperature, minimum temperature and precipitation. The metrics are displayed using the same scheme that in Fig. 6

same hierarchy holds for the Brier Scores at 1 mm, 10 mm, and p99 thresholds, with UNet-STO performing best, followed by the ClimDiT-Ensemble.

3.3 Multivariate Performance

A distinctive feature of ClimDiT is the joint prediction of the three target variables with the same model, allowing the preservation of inter-variable relationships. Thus, to evaluate the multivariable performance, Fig. 9 reports the Energy Score (ES) and the Variogram Score (VS) metrics. As mentioned in Subsect. 2.5, the aggregation of all stochastic models is denoted “STO”, and that of deterministic models “DET”. Additionally, the corresponding maps of these metrics, which provide a more detailed view of their spatial structure, are presented in Sect. S3 of the Supplementary Material.

Figure 9 summarizes the behavior of $(T_{\max}, T_{\min}, \log(1+\text{precip}))$ according to ES and VS metrics. The bimodal distribution, or "double hump", frequently observed in these violin plots reflects a distinct spatial divide in performance across the Iberian Peninsula. Specifically, lower error magnitudes generally correspond to the central plateau and the south-western region. In contrast, the largest errors correspond to areas where precipitation errors are higher, namely the northern coast and the Mediterranean coast, and to regions where temperature shows the highest errors, such as the Ebro River basin, and the south-eastern mountain ranges (see Fig. S9 of the Supplementary Material).

Across the domain, the aggregated STO baselines (UNet-STO, DeepESD-STO) achieve the lowest (best) ES values, reflecting their strong marginal calibration (CRPS) and its impact on the overall multivariate score. However, ClimDiT-Ensemble performs competitively, particularly in terms of the Variogram Score, which specifically evaluates inter-variable dependence. Notably, the benefit of the ClimDiT-Ensemble over the single-sample ClimDiT is more evident in ES than in previous metrics, highlighting a sharper yet calibrated multivariate predictive distribution. Deterministic baselines (DET and DeepESD-MULT) yield intermediate VS and ES: they perform the multivariate prediction reasonably well but show a higher violin dispersion toward larger (worse) values than the other degenerate ensemble (M=1) ClimDiT. The improvement of the ClimDiT-Ensemble over these two deterministic benchmark models is more evident, especially in ES.

Furthermore, to assess the physical consistency of the generated multivariate fields, the mean probability of satisfying the condition $T_{\min} < T_{\max}$ is evaluated across the entire test set and at all points on the grid. All evaluated models exhibit near-perfect physical consistency (over 99.9% probability). These results are encouraging, as previous studies have shown that multivariate deep learning models can struggle to satisfy hard physical constraints (González-Abad et al. 2024). However, in this study, both the traditional regression-based architectures and the latent diffusion generative framework inherently adhere to this fundamental thermodynamic limit almost perfectly, without the need for explicit architectural constraints.

3.4 Seasonal Analysis

To ensure that the performance of models is not partially masked by the strong annual cycle of the variables, a comprehensive seasonal evaluation (DJF, MAM, JJA, and SON) is presented in Sect. S4 of the Supplementary Material. The results show that ClimDiT-Ensemble consistently ranks among the top-performing models in all seasons for the univariate metrics CRPS, mean bias and RMSE across the three variables. Moreover, the multivariate metrics (Energy Score and Variogram Score) of the ensemble remain superior to those of the deterministic baselines across all climatological regimes, and slightly behind the aggregated stochastic (STO) baseline. This demonstrates that the model accurately captures intra-seasonal weather dynamics, indicating an optimal balance between probabilistic calibration (CRPS, ES, and VS) and point-wise errors (RMSE) regardless of the prevailing season.

4 Conclusions

This study introduced ClimDiT, a generative approach for multivariate climate downscaling based on latent diffusion transformer models. ClimDiT effectively downscales multiple climate variables simultaneously, namely daily maximum temperature, minimum temperature, and accumulated precipitation.The results demonstrate that ClimDiT achieves deterministic accuracy comparable to, or moderately superior to, state-of-the-art single-target benchmarks (DeepESD and U-Net) while effectively capturing extremes, such as p98 bias for Tmax and the p02 bias for Tmin.

ClimDiT preserves spatial structures effectively, as evidenced by high SSIM values and realistic spectral slopes in the RAPSD analysis. In particular, it mitigates the loss of power at small scales observed in other stochastic models and achieves the best spectral match for precipitation among all the evaluated methods. These findings indicate that its latent representation, combined with the diffusion-based sampling, contributes to maintaining spatial coherence without introducing artificial noise.

Regarding probabilistic performance, stochastic baselines (UNet-STO, DeepESD-STO) achieve the best marginal and joint calibration in terms of CRPS, Energy, Variogram, and Brier scores, but at the cost of higher pointwise errors (RMSE, MAE) and weaker spatial fidelity (lower SSIM and less faithful RAPSD). On the other hand, the 50-member ClimDiT-Ensemble achieves a competitive performance regarding these probabilistic metrics while maintaining clearly lower point-wise errors and stronger spatial fidelity. This suggests that the latent diffusion framework attains a robust balance between sharpness and reliability, successfully preserving inter-variable dependencies, which is an essential property for physically coherent climate projections.

Overall, ClimDiT introduces a promising generative and probabilistic framework for multivariate downscaling that extends beyond single deterministic estimates by representing the variability and uncertainty inherent in the climate system through ensemble generation. By modeling multiple variables jointly, it is able to capture dependencies between variables and produce spatially coherent, high-resolution climate fields, while demonstrating better performance for complex variables, such as precipitation, compared to traditional deterministic approaches.

However, the representation of highly intermittent and zero-inflated variables such as precipitation remains challenging, highlighting the potential for future research into tailored latent approaches or decoding strategies. The obtained results are particularly relevant for downstream impact-oriented applications, including hydrological modeling, flood risk assessment, and climate adaptation planning. In regions like the Iberian Peninsula, characterized by spatially variable extreme precipitation driven by complex processes (Vicente-Serrano et al. 2025), reliable high-resolution data is essential for hazard and impact characterization. ClimDiT probabilistic and spatially coherent outputs can be used in extreme-event analyses (Zhao et al. 2023) and used in frameworks linking climate projections to risk and adaptation strategies (Zhao et al. 2022).

Future research may focus on exploring alternative diffusion architectures, including U-Net–based variants, together with refined training strategies and different loss functions, such as CRPS-based objectives, aimed at strengthening probabilistic performance. One possible direction is to explore end-to-end training to jointly optimize the autoencoder and the diffusion model (Leng et al. 2025). Additionally, future work could assess the impact of training domain-specific autoencoders from scratch, exploring different compression factors and latent space dimensionalities to further optimize the representation of multivariate climate fields. An interesting approach would be to explore the use of these regularized autoencoders specifically in conjunction with classical regression models, which would help to isolate the specific contribution of the latent representation and the diffusion process. Similarly, some alternatives to the log1p function could be evaluated to transform the precipitation data, in order to examine potential improvements in some metrics related to this variable. Furthermore, in future work, the model will be trained using high-resolution data derived from regionalizations obtained from numerical models driven by different GCMs, thereby creating an emulator of these models, which would also allow for the inclusion of additional high-resolution variables typically not found in observational datasets. This will enable the evaluation of the generative model's transferability to different time periods or emission scenarios (soft transferability), as well as to changes in the driving GCMs (hard transferability), and its ability to capture the internal variability of each model.

Appendix A Glossary of Metrics and Formulas

Notation

Let $i \in \Omega$ index grid cells (domain), $t = 1, \ldots, T$ index time (days), and $\mathcal{T}_g$ denote the set of days in year g (with G total years). Let $y_t(i)$ denote the observed value at (t, i). A deterministic model provides $\hat{y}_t(i)$, and an ensemble model provides values for each member $x_{t,m}(i)$, $m = 1, \ldots, M$. For precipitation, $r_t(i)$ denotes daily precipitation; a wet-day threshold is r_0 ($1\,\mathrm{mm\,day^{-1}}$ in this case). $T_{\min,t}(i)$ and

$T_{\max,t}(i)$ specify minimum and maximum daily temperature, respectively. Formulas are given per (t, i) or per year as specified.

Averaging Operators

Time average: $\bar{a} = \frac{1}{T}\sum_{t=1}^{T} a_t$; sample standard deviation: $\mathrm{SD}(a) = \sqrt{\frac{1}{T}\sum_t (a_t - \bar{a})^2}$. Absolute value $|\cdot|$, Euclidean norm $\|\cdot\|_2$.

Deterministic Metrics

- **Mean bias (MB).**
  For temperature (absolute):

$$\mathrm{MB}(i) = \overline{\hat{y}(i)} - \overline{y(i)}.$$

  For precipitation (relative):

$$\mathrm{MB}_{\%}(i) = 100 \cdot \frac{\overline{\hat{r}(i)} - \overline{r(i)}}{\overline{r(i)}}.$$

- **Percentile bias (PB).** Let $y_{(p)}(i)$ denote the p-th percentile over time of the observed values at grid cell *i*. For a given percentile $p \in (0, 1)$:

$$\mathrm{PB}_p(i) = \hat{y}_{(p)}(i) - y_{(p)}(i),$$

  For precipitation, the $p = 0.99$ bias is expressed in relative terms:

$$\mathrm{PB}_{99,\%}(i) = 100 \cdot \frac{\hat{r}_{(99)}(i) - r_{(99)}(i)}{r_{(99)}(i)}.$$

- **Relative bias for the ratio of wet days (R01 bias).**
  Wet-day ratio:

$$\mathrm{R01}(i) = \frac{1}{T}\sum_t 1\{r_t(i) \geq r_0\}.$$

  Relative bias:

$$\mathrm{Bias}_{\%}(\mathrm{R01})(i) = 100 \cdot \frac{\widehat{\mathrm{R01}}(i) - \mathrm{R01}(i)}{\mathrm{R01}(i)}.$$

- **Relative bias for the Simple Daily Intensity Index (SDII).**

$$\mathrm{SDII}(i) = \frac{\sum_t r_t(i)\, 1\{r_t(i) \geq r_0\}}{\sum_t 1\{r_t(i) \geq r_0\}},$$

$$\mathrm{Bias}_{\%}(\mathrm{SDII})(i) = 100 \cdot \frac{\widehat{\mathrm{SDII}}(i) - \mathrm{SDII}(i)}{\mathrm{SDII}(i)}.$$

- **Relative bias for the climatology of maximum daily precipitation (Rx1day).**

$$\mathrm{Rx1day}_g(i) = \max_{t \in \mathcal{T}_g} r_t(i), \quad \overline{\mathrm{Rx1day}}(i) = \frac{1}{G}\sum_g \mathrm{Rx1day}_g(i).$$

  Relative bias:

$$100 \cdot \frac{\widehat{\overline{\mathrm{Rx1day}}}(i) - \overline{\mathrm{Rx1day}}(i)}{\overline{\mathrm{Rx1day}}(i)}.$$

- **Bias for annual temperature extremes (TNn/TXx).**

$$\mathrm{TNn}_g(i) = \min_{t \in \mathcal{T}_g} T_{\min,t}(i), \qquad \mathrm{TXx}_g(i) = \max_{t \in \mathcal{T}_g} T_{\max,t}(i).$$

  Biases (averaged across years):

$$\overline{\mathrm{TNn}}^{\mathrm{bias}}(i) = \frac{1}{G}\sum_g (\widehat{\mathrm{TNn}}_g(i) - \mathrm{TNn}_g(i)), \quad \text{and analogously for TXx.}$$

- **Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).**

$$\mathrm{RMSE}(i) = \sqrt{\frac{1}{T}\sum_t (\hat{y}_t(i) - y_t(i))^2},$$

$$\mathrm{MAE}(i) = \frac{1}{T}\sum_t |\hat{y}_t(i) - y_t(i)|.$$

- **Standard-deviation ratio (temperatures).**

$$\mathrm{StdRatio}(i) = \frac{\mathrm{SD}(\hat{y}(i))}{\mathrm{SD}(y(i))}.$$

- **Interannual variability ratio (precipitation).** For each year y, the annual mean precipitation is

$$\overline{r}_g(i) = \frac{1}{|\mathcal{T}_g|}\sum_{t \in \mathcal{T}_g} r_t(i).$$

  The ratio of interannual variability is then

$$\mathrm{IAVR}(i) = \frac{\mathrm{SD}_g(\widehat{\overline{r}}_g(i))}{\mathrm{SD}_g(\overline{r}_g(i))}.$$

Probabilistic Metrics

- **Continuous Ranked Probability Score (CRPS)**. Let $F_{t,i}(z) = \Pr(X_t(i) \le z)$ denote the predictive cumulative distribution function (CDF) for day t and grid cell i.

$$\mathrm{CRPS}_t(i) = \int_{-\infty}^{\infty} \big(F_{t,i}(z) - 1\{y_t(i) \le z\}\big)^2 \, \mathrm{d}z.$$

Ensemble estimator (members $x_{t,m}(i) \sim F_{t,i}$):

$$\mathrm{CRPS}_t(i) = \frac{1}{M}\sum_{m=1}^{M} |x_{t,m}(i) - y_t(i)| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M} |x_{t,m}(i) - x_{t,m'}(i)|.$$

The CRPS for each grid cell is then averaged over the time period:

$$\mathrm{CRPS}(i) = \frac{1}{T}\sum_{t=1}^{T} \mathrm{CRPS}_t(i).$$

For deterministic forecasts, CRPS = MAE. For UNet-STO precipitation, a closed-form expression for the Bernoulli–Gamma predictive law is provided in 6..

- **Brier Score (BS)**. For an event E (e.g., $T_{\min} < p02$, $T_{\max} > p98$, $r \ge u$) with forecast probability $p_t(i)$ and outcome $o_t(i) = 1\{E \text{ occurs}\}$:

$$\mathrm{BS}(i) = \frac{1}{T}\sum_{t=1}^{T} \big(p_t(i) - o_t(i)\big)^2.$$

For an ensemble, $p_t(i)$ is the fraction of members satisfying E.

Spatial Fidelity Metrics

- **Structural Similarity Index (SSIM)**. For two fields A, B on a given day (then averaged over the test period):

$$\mathrm{SSIM}(A,B) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)}.$$

Here μ_A, μ_B are means, σ_A^2, σ_B^2 variances and σ_{AB} the covariance over spatial Gaussian windows with $\sigma = 1.5$, as implemented in Wang et al. (2004).

- **Radially averaged power spectral density (RAPSD)** (Sinclair and Pegram 2005). This method computes the two-dimensional discrete Fourier transform (DFT) of a field $f(h, w)$ with dimensions $H \times W$ and grid spacing s, obtaining

$$F(u,v) = \frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f(h,w)\, e^{-j2\pi\left(\frac{uh}{M} + \frac{vw}{N}\right)}.$$

The power spectral density is defined as $P(f) = |F(u,v)|^2$. Then, the spectrum is averaged in concentric rings. Each ring represents a specific range of frequencies, and the average power within that ring is calculated:

$$P^r(f_r) = \frac{1}{N_r(f_r)}\sum_{i=1}^{N_r(f_r)} P(f_{r,i}),$$

where $N_r(f_r)$ denotes the number of discrete frequencies within each ring. The resulting one-dimensional spectrum $P^r(f_r)$ describes the distribution of spatial variance across scales, which can also be expressed in terms of the spatial wavelength $\lambda = s/f_r$. $P^r(f_r)$ is calculated, for a given variable and day, for the prediction and the observation and then averaged over time. In this study, RAPSD is computed using the `PySTEPS` package.

Multivariate Metrics

Standardization and Transformation

Let the multivariate observation at time t and grid cell i be

$$\mathrm{y}_t(i) = \big[T_{\max,t}(i),\, T_{\min,t}(i),\, \log(1 + r_t(i))\big]^\top \in \mathbb{R}^d,$$

with $d = 3$ in our case. For each grid cell i and component $k \in \{1, \ldots, d\}$, let $\mu^k(i)$ and $\sigma^k(i)$ denote the mean and standard deviation of the observations over the test period. Observations and forecasts are standarized, before scoring, componentwise as

$$\tilde{y}_t^k(i) = \frac{\big(y_t^k(i) - \mu^k(i)\big)}{\sigma^k(i)}, \qquad \tilde{x}_{t,m}^k(i) = \frac{\big(x_{t,m}^k(i) - \mu^k(i)\big)}{\sigma^k(i)},$$

- **Energy Score (ES).** For a multivariate predictive distribution $F_{t,i}$ on $\mathbb{R}^d$ and observation $\tilde{\mathrm{y}}_t(i)$, the Energy Score at (t,i) is

$$\mathrm{ES}_t(i) = \mathbb{E}\big\|\mathrm{X} - \tilde{\mathrm{y}}_t(i)\big\|_2 - \tfrac{1}{2}\,\mathbb{E}\big\|\mathrm{X} - \mathrm{X}'\big\|_2, \quad \mathrm{X}, \mathrm{X}' \overset{\text{iid}}{\sim} F_{t,i}.$$

Ensemble estimator:

$$\widehat{\mathrm{ES}}_t(i) = \frac{1}{M}\sum_{m=1}^{M} \big\|\tilde{\mathrm{x}}_{t,m}(i) - \tilde{\mathrm{y}}_t(i)\big\|_2 - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M} \big\|\tilde{\mathrm{x}}_{t,m}(i) - \tilde{\mathrm{x}}_{t,m'}(i)\big\|_2.$$

Time-averaged per grid cell:

$$\mathrm{ES}(i) = \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{ES}}_t(i).$$

- **Variogram Score (VS).** For exponent $p \in (0, 2]$ and nonnegative weights $w^{k\ell}$ on component pairs $(k < \ell)$, the Variogram Score at (t, i) is

$$\mathrm{VS}_{p,t}(i) = \sum_{k<\ell} w^{k\ell} \left(|\tilde{y}_t^k(i) - \tilde{y}_t^\ell(i)|^p - \mathbb{E}|\mathrm{X}^k - \mathrm{X}^\ell|^p \right)^2.$$

Ensemble estimator:

$$\widehat{\mathrm{VS}}_{p,t}(i) = \sum_{k<\ell} w^{k\ell} \left(|\tilde{y}_t^k(i) - \tilde{y}_t^\ell(i)|^p - \frac{1}{M}\sum_{m=1}^{M} |\tilde{x}_{t,m}^k(i) - \tilde{x}_{t,m}^\ell(i)|^p \right)^2.$$

Time-averaged per grid cell:

$$\mathrm{VS}_p(i) = \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{VS}}_{p,t}(i).$$

In this study, uniform weights ($w^{k\ell} = 1$) and an exponent of $p = 0.5$ are used. The specific parameterization of $p = 0.5$ was selected following the recommendations of Scheuerer and Hamill (2015) and Alexander et al. (2024), as it provides the score with a higher discrimination ability to distinguish between different inter-variable dependence structures compared to higher exponents (e.g., $p = 1$ or $p = 2$), which tend to over-penalize large individual errors and mask systematic differences. This theoretical loss of model discernibility at higher exponent values was empirically corroborated in our own sensitivity evaluations.

Appendix B CRPS for the Bernoulli–Gamma Distribution

Definition 1 *(Bernoulli–Gamma distribution)* Equivalent to the definition of Eq. 2, let

$$X = \begin{cases} 0, & \text{with probability } 1-p, \\ 0.1 + Z, & \text{with probability } p, \end{cases} \qquad Z \sim \mathrm{Gamma}(\alpha, \beta),$$

where we use the shape–scale parametrization, i.e., $\mathbb{E}[Z] = \alpha/\beta$. By construction, the cumulative distribution function (CDF) of X is

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1-p, & 0 \le x \le 0.1, \\ 1 - p + p\,G_{\alpha,\beta}(x - 0.1), & x > 0.1, \end{cases}$$

where $G_{\alpha,\beta}(x)$ is the Gamma CDF with parameters (α, β).

For any distribution F with finite first moment, the Continuous Ranked Probability Score (CRPS) admits the representation (Scheuerer and Möller 2015)

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F|X - y| - \tfrac{1}{2}\,\mathbb{E}_F|X - X'|, \qquad X, X' \overset{\mathrm{iid}}{\sim} F. \tag{B1}$$

Theorem 1 *(CRPS formula) Define $\mu_Z := \alpha/\beta$. Then, the CRPS of the Bernoulli-Gamma CDF (Fx) is given by*

$$\boxed{\mathrm{CRPS}(F_X, y) = (1-p)\,|y| \;+\; p\,\eta(y - 0.1) \;-\; p(1-p)\,(0.1 + \mu_Z) \;-\; p^2\,\mu_Z\,\mathcal{G}(\alpha)} \tag{B2}$$

where

$$\eta(y) = \begin{cases} w\big(2G_{\alpha,\beta}(y) - 1\big) - \dfrac{\alpha}{\beta}\big(2G_{\alpha+1,\beta}(y) - 1\big), & if\, y \ge 0, \\ \dfrac{\alpha}{\beta} - y, & if\, y < 0, \end{cases}$$

and

$$\mathcal{G}(\alpha) = \frac{\Gamma\big(\alpha + \frac{1}{2}\big)}{\sqrt{\pi}\,\Gamma(\alpha + 1)}$$

is the Gini ratio (Scheuerer and Möller 2015) of the gamma distribution with shape α.

Proof From (B1), it suffices to compute $\mathbb{E}|X - y|$ and $\mathbb{E}|X - X'|$ separately.

Step 1: $\mathbb{E}|X - y|$. Define $w := y - 0.1$. By mixture decomposition,

For $w < 0$: since $Z \ge 0$ a.s., $|Z - w| = Z - w$ and

$$\mathbb{E}|Z - w| = \mathbb{E}Z - w = \frac{\alpha}{\beta} - w.$$

For $w \ge 0$, splitting the integral at w, letting $f_{\alpha,\beta}$ and $G_{\alpha,\beta}$ denote the gamma PDF and CDF, and using results from Appendix A of Scheuerer and Möller (2015),

$$\mathbb{E}|Z - w| = \int_0^w (w - z) f_{\alpha,\beta}(z)\,dz + \int_w^\infty (z - w) f_{\alpha,\beta}(z)\,dz = w\big(2G_{\alpha,\beta}(w) - 1\big) - 2\int_0^w z f_{\alpha,\beta}(z)\,dz + \mathbb{E}Z.$$

Using

$$\int_0^w z f_{\alpha,\beta}(z)\,dz = \frac{\alpha}{\beta}\,G_{\alpha+1,\beta}(w),$$

we obtain

$$\mathbb{E}|Z - w| = w\big(2G_{\alpha,\beta}(w) - 1\big) - \frac{\alpha}{\beta}\big(2G_{\alpha+1,\beta}(w) - 1\big),$$

i.e., the stated expression of $\eta(w)$.

Step 2:$\mathbb{E}|X - X'|$The four combinations in the mixture yield:

(i) $X = X' = 0$ with prob. $(1-p)^2$: contribution 0.
(ii) $X = 0,\ X' = 0.1 + Z$ or vice versa, prob. $2p(1-p)$: $\mathbb{E}|0 - (0.1 + Z)| = 0.1 + \mu_Z$.
(iii) $X = 0.1 + Z,\ X' = 0.1 + Z'$ with prob. p^2 : $\mathbb{E}|Z - Z'| = 2\mu_Z\,\mathcal{G}(\alpha)$.

Therefore,

$$\mathbb{E}|X - X'| = 2p(1-p)\big(0.1 + \mu_Z\big) + 2p^2\,\mu_Z\,\mathcal{G}(\alpha),$$

and substituting these expressions from Steps 1 and 2 into (B1) gives (B2).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41748-026-01326-w.

Acknowledgements This work was supported by the PLANCLIMAC2 Project (1/MAC/2/2.4/0006), financed through the European Union's INTERREG MAC 2021–2027 Program. Additional support was provided by the Government of the Canary Islands, through the Consejería de Transición Ecológica y Energía, as part of the cooperation agreement signed on November 8, 2024. The methods developed in this work will be used at higher spatial resolution in archipelagos within the CORDEX Flagship Pilot Study: Macaronesian Archipelagos: Convection Permitting projections focused on island processes (FPS-I-Mac). The authors have no relevant financial or non-financial interests to disclose.

Author Contributions M. E-G. was responsible for Methodology, Investigation, Validation, Visualization, Software, Writing – original draft. A.G. contributed to the Conceptualization, Methodology, Formal analysis, Validation, Writing – original draft. J.P.D was in charge of Supervision, Funding acquisition, Resources, Project administration, Writing – review & editing. J.C.P contributed to the Supervision, Methodology, Formal analysis, Resources, Writing – review & editing. J. G-A. was in charge of Methodology, Software, Resources, Validation, Writing – review & editing. S. M-C contributed to Methodology, Software.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availibility All datasets used in this study are publicly available. The ERA5 reanalysis data can be obtained from the Copernicus Climate Data Store (CDS) upon registration and acceptance of the license agreement (https://doi.org/10.24381/cds.50314f4c). The ROCIO-IBEB observational dataset is hosted by AEMET and can be accessed through its website (https://www.aemet.es/en/serviciosclimaticos/cambio_climat/datos_diarios?w=2).

Declarations

Competing Interests On behalf of all authors, the corresponding author states that there is 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

Alexander C, Coulon M, Han Y, Meng X (2024) Evaluating the discrimination ability of proper multi-variate scoring rules. Ann Oper Res 334(1):857–883. https://doi.org/10.1007/s10479-022-04611-9

Addison H, Kendon E, Ravuri S, Aitchison L, Watson P (2022) Machine Learning Emulation of a Local-scale UK Climate Model. Tackling Climate Change with Machine Learning workshop at NeurIPS 2022; Tackling Climate Change with Machine Learning workshop at NeurIPS 2022; Conference date: 09-12-2022 Through 09-12-2022. https://www.climatechange.ai/events/neurips2022

Balmaceda-Huarte R, Bettolli ML (2022) Assessing statistical downscaling in argentina: Daily maximum and minimum temperatures. Int J Climatol 42(16):8423–8445. https://doi.org/10.1002/joc.7733

Baño-Medina J, Iturbide M, Fernández J, Gutiérrez JM (2024) Transferability and explainability of deep learning emulators for regional climate model projections: Perspectives for future applications. Artificial Intell Earth Syst 3(4):230099. https://doi.org/10.1175/AIES-D-23-0099.1

Baño-Medina J, Manzanas R, Gutiérrez JM (2020) Configuration and intercomparison of deep learning neural models for statistical downscaling. Geoscientific Model Development 13(4):2109–2124. https://doi.org/10.5194/gmd-13-2109-2020

Baño-Medina J, Manzanas R, Cimadevilla E, Fernández J, González-Abad J, Cofiño AS, Gutiérrez JM (2022) Downscaling multimodel climate projection ensembles with deep learning (deepesd): contribution to cordex eur-44. Geoscientific Model Development 15(17):6747–6758. https://doi.org/10.5194/gmd-15-6747-2022

Bonnaire T, Urfin R, Biroli G, Mézard M (2025) Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training. arXiv:2505.17638

Buster G, Benton BN, Glaws A, King RN (2024) High-resolution meteorology with climate change impacts from global climate model data using generative machine learning. Nat Energy 9(7):894–906. https://doi.org/10.1038/s41560-024-01507-9

Chen M, Mei S, Fan J, Wang M (2024) Opportunities and challenges of diffusion models for generative ai. Natl Sci Rev 11(12):348. https://doi.org/10.1093/nsr/nwae348

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. https://doi.org/10.1007/s00382-024-07350-8

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929

Dee D, Thépaut JN (2023) Era5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.adbb2d47 (Technical report)

Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW (eds.) Advances in Neural Information Processing Systems. Curran Associates, Inc 34:8780–8794 https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

Ferro CAT, Richardson DS, Weigel AP (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. Meteorol Appl 15(1):19–24. https://doi.org/10.1002/met.45

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437

González-Abad J, Hernández-García Á, Harder P, Rolnick D, Gutiérrez J (2024) Multi-variable hard physical constraints for climate model downscaling. Proceedings of the AAAI Symposium Series 2(1), 62–67. https://doi.org/10.1609/aaaiss.v2i1.27650. Chap. Artificial Intelligence and Climate: The Role of AI in a Climate-Smart Sustainable Future

González-Abad J, Gutiérrez JM (2025) Are deep learning methods suitable for downscaling global climate projections? an intercomparison for temperature and precipitation over spain. Artificial Intell Earth Syst 4(4):240121. https://doi.org/10.1175/AIES-D-24-0121.1

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. https://doi.org/10.1145/3422622

Gutiérrez JM, Maraun D, Widmann M, Huth R, Hertig E, Benestad R, Roessler O, Wibig J, Wilcke R, Kotlarski S, San Martín D, Herrera S, Bedia J, Casanueva A, Manzanas R, Iturbide M, Vrac M, Dubrovsky M, Ribalaygua J, Pórtoles J, Räty O, Räisänen J, Hingray B, Raynaud D, Casado MJ, Ramos P, Zerenner T, Turco M, Bosshard T, Štěpánek P, Bartholy J, Pongracz R, Keller DE, Fischer AM, Cardoso RM, Soares PMM, Czernecki B, Pagé C (2019) An intercomparison of a large ensemble of statistical downscaling methods over europe: Results from the value perfect predictor cross-validation experiment. Int J Climatol 39(9):3750–3785. https://doi.org/10.1002/joc.5462

Hernanz A, Correa C, Sánchez-Perrino J-C, Prieto-Rico I, Rodríguez-Guisado E, Domínguez M, Rodríguez-Camino E (2024) On the limitations of deep learning for statistical downscaling of climate change projections: The transferability and the extrapolation issues. Atmospheric Sci Letters 25(2):1195. https://doi.org/10.1002/asl.1195

Ho J, Jain A, Abbeel P (2020) Denoising Diffusion Probabilistic Models. arXiv:2006.11239

Huang Z, Zhao T, Lai R, Tian Y, Yang F (2023) A comprehensive implementation of the log, box-cox and log-sinh transformations for skewed and censored precipitation data. J Hydrol 620:129347. https://doi.org/10.1016/j.jhydrol.2023.129347

Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. https://doi.org/10.48550/arXiv.1312.6114

Lang S, Alexe M, Chantry M, Dramsch J, Pinault F, Raoult B, Clare MCA, Lessig C, Maier-Gerber M, Magnusson L, Bouallègue ZB, Nemesio AP, Dueben PD, Brown A, Pappenberger F, Rabier F (2024) AIFS – ECMWF's data-driven forecasting system. arXiv:2406.01465

Leng X, Singh J, Hou Y, Xing Z, Xie S, Zheng L (2025) REPA-E: Unlocking VAE for End-to-End Tuning with Latent Diffusion Transformers. arXiv:2504.10483

Lopez-Gomez I, Wan ZY, Zepeda-Núñez L, Schneider T, Anderson J, Sha F (2025) Dynamical-generative downscaling of climate model ensembles. Proc Natl Acad Sci 122(17):2420288122. https://doi.org/10.1073/pnas.2420288122

Mardani M, Brenowitz N, Cohen Y, Pathak J, Chen C-Y, Liu C-C, Vahdat A, Nabian MA, Ge T, Subramaniam A, Kashinath K, Kautz J, Pritchard M (2025) Residual corrective diffusion modeling for km-scale atmospheric downscaling. Commun Earth Environ 6(1):124. https://doi.org/10.1038/s43247-025-02042-5

Mármol-Rivera JA, Fernández-Rodríguez JD, Asenjo B, López-Rubio E (2025) Latent diffusion for arbitrary zoom mri super-resolution. Expert Syst Appl 284:127970. https://doi.org/10.1016/j.eswa.2025.127970

Nichol AQ, Dhariwal P (2021) Improved denoising diffusion probabilistic models. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research. PMLR, Cambridge 139:8162–8171. https://proceedings.mlr.press/v139/nichol21a.html

Peebles W, Xie S (2023) Scalable diffusion models with transformers. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp 4172–4182. https://doi.org/10.1109/ICCV51070.2023.00387

Peral C, Navascués B, Ramos-Calzado P (2017) Serie de precipitación diaria en rejilla con fines climáticos. aemet. Technical report. http://hdl.handle.net/20.500.11765/7573

Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, Penna J, Rombach R (2023) SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952

Quesada-Chacón D, Baño-Medina J, Barfus K, Bernhofer C (2023) Downscaling cordex through deep learning to daily 1 km multivariate ensemble in complex terrain. Earth's Future 11(8):2023–003531. https://doi.org/10.1029/2023EF003531

Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. ArXiv arXiv 2204:06125

Rampal N, Gibson PB, Sherwood S, Abramowitz G, Hobeichi S (2025) A reliable generative adversarial network approach for climate downscaling and weather generation. J Adv Modeling Earth Syst 17(1):2024–004668. https://doi.org/10.1029/2024MS004668

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2023) High-resolution image synthesis with latent diffusion models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Los Alamitos, CA, USA x, pp 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. Springer, pp 234–241

Rössler O, Fischer AM, Huebener H, Maraun D, Benestad RE, Christodoulides P (2019) Challenges to link climate change data provision and user needs: Perspective from the COST action VALUE. Int J Climatol 39(9):3704–3716. https://doi.org/10.1002/joc.5060

Schepen A, Everingham Y, Wang QJ (2020) On the joint calibration of multivariate seasonal climate forecasts from gcms. Mon Weather Rev 148(1):437–456. https://doi.org/10.1175/MWR-D-19-0046.1

Scheuerer M, Hamill T (2015) Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities*. Mon Weather Rev 143:1321–1334. https://doi.org/10.1175/MWR-D-14-00269.1

Scheuerer M, Möller D (2015) Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. Ann Appl Stat. https://doi.org/10.1214/15-aoas843

Shazeer N (2020) GLU Variants Improve Transformer. arXiv:2002.05202

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 DJ, Norouzi M (2022) Photorealistic text-to-image diffusion models with deep language understanding. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A (eds) Advances in Neural Information Processing Systems, vol 35. Curran Associates, Inc., New York, NY, USA, pp 36479–36494 https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf

Sinclair S, Pegram GGS (2005) Empirical mode decomposition in 2-d space and time: a tool for space-time rainfall analysis and nowcasting. Hydrol Earth Syst Sci 9(3):127–137. https://doi.org/10.5194/hess-9-127-2005

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. CoRR arXiv 1706:03762

Vicente-Serrano SM, Garrido-Perez JM, Fernández-Álvarez JC, Gimeno-Sotelo L, Beguería S, Halifa-Marín A, Latorre B, El Kenawy AM, Franquesa M, Adell-Michavila M, Domínguez-Castro F, Barriopedro D, Gimeno-Presa L, Nieto R, Azorin-Molina C, Andres-Martin M, Gutiérrez JM, García-Herrera R (2025) Characteristics of widespread extreme precipitation events in peninsular spain and the balearic islands: spatio-temporal dynamics and driving mechanisms. Clim Dyn 63(9):340. https://doi.org/10.1007/s00382-025-07829-y

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861

Watt RA, Mansfield LA (2024) Generative Diffusion-based Downscaling for Climate. arXiv:2404.17752

Zhao W, Abhishek Kinouchi T, Ang R, Zhuang Q (2022) A framework for quantifying climate informed heavy rainfall change Implications for adaptation strategies. Sci Total Environ 835:155553. https://doi.org/10.1016/j.scitotenv.2022.155553

Zhao W, Abhishek Takhellambam BS, Zhang J, Zhao Y, Kinouchi T (2023) Spatiotemporal variability of current and future sub-daily rainfall in japan using state-of-the-art high-quality data sets. Water Resour Res 59(4):2022–034305. https://doi.org/10.1029/2022WR034305

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