Supplementary Material for "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[2,3], Samuel Márquez-Cubas[1]

[1*]Universidad de La Laguna (ULL), San Cristóbal de La Laguna, Spain.
[2*]Instituto de Física de Cantabria (IFCA), Santander, Spain.
[3*]CSIC-Universidad de Cantabria, Santander, Spain.

*Corresponding author(s). E-mail(s): fgmesquivel@ull.edu.es;

S1 Multivariate benchmark models

As mentioned in Section 2.2.3 of the main manuscript, a multivariate version of the U-Net architecture (UNet-MULT) was evaluated alongside DeepESD-MULT as a deterministic multivariate benchmark. Both models were trained to simultaneously predict daily maximum temperature, minimum temperature, and precipitation by minimizing the Mean Squared Error (MSE) across the three channels.

Figures S1, S2, and S3 compare the spatial distribution of the performance metrics over the test domain for both models. Across all three target variables, UNet-MULT exhibits performance that is generally comparable to, or worse than, DeepESD-MULT. In particular, UNet-MULT exhibits a noticeably wider spatial dispersion (i.e., longer tails in the violin plots) in most bias metrics and absolute errors (RMSE and MAE). This wider spread indicates lower spatial robustness and larger localized errors across the Iberian Peninsula compared to the DeepESD architecture. Consequently, DeepESD-MULT was selected as the reference deterministic multivariate baseline for the main analysis.

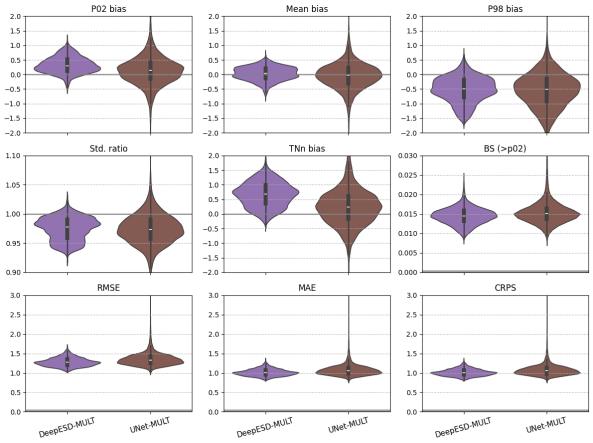

Fig. S1: Performance metrics for the test split for minimum daily temperatures (Tmin). The violin plots show the distribution of results across the Iberian Peninsula grid points for the multivariate baseline models (DeepESD-MULT and UNet-MULT).

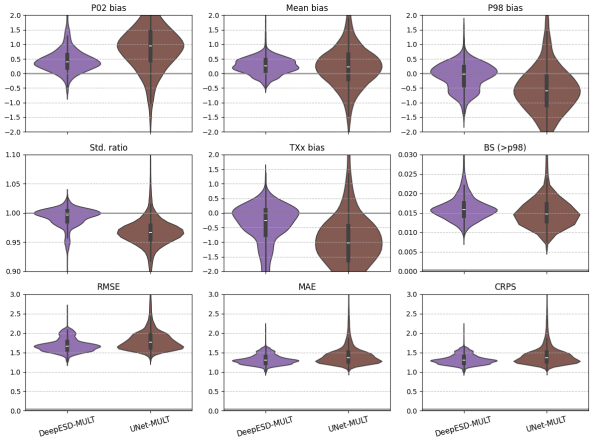

Fig. S2: Performance metrics for maximum daily temperatures (Tmax) in the test split. The violin plots show the distribution of results across the Iberian Peninsula grid points for the multivariate baseline models.

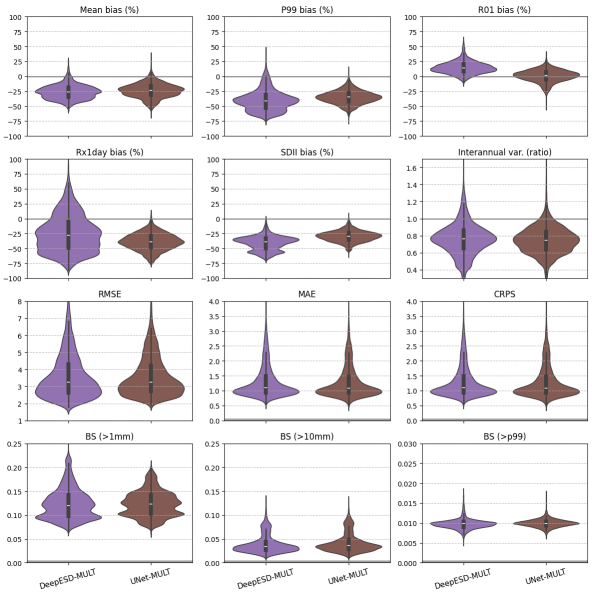

Fig. S3: Performance metrics for daily accumulated precipitation in the test split. The violin plots show the distribution of results across the Iberian Peninsula grid points for the multivariate baseline models.

S2 Training Stability and Learning Curves

S2.1 ClimDiT backbone training

As discussed in the main text, the inherent stability of the diffusion training process and the implicit regularization of the autoencoder remove the strict requirement for a validation split to monitor early stopping. To illustrate this stability, Figure S4 displays both the training loss of the ClimDiT diffusion backbone and the evolution of the spatial average of Root Mean Squared Error (RMSE) on the independent test set over the 400,000 training steps.

The left panel demonstrates smooth convergence of the objective loss without the abrupt spikes that are typical in some transformer-based architectures, highlighting the robustness of the chosen training configuration. Furthermore, the right panel corroborates this stability by tracking the test set spatial RMSE for minimum temperature, maximum temperature, and precipitation evaluated at regular 100,000-step intervals. Crucially, all test metrics decrease monotonically throughout the training process. This continuous improvement shows that the generative model does not seem to overfit to the training distribution.

S2.2 Variational Autoencoder Finetuning

In addition to the latent diffusion architecture, the performance of the Variational Autoencoder (VAE) during fine-tuning is sensitive to the choice of training hyperparameters. To evaluate this, Figure S5 illustrates the training and validation loss curves for two different fine-tuning configurations that were tested: a conservative baseline using a batch size of 64 and a learning rate of 1×10^{-4}, and an alternative setup using a batch size of 20 and a learning rate of 2×10^{-4}. The

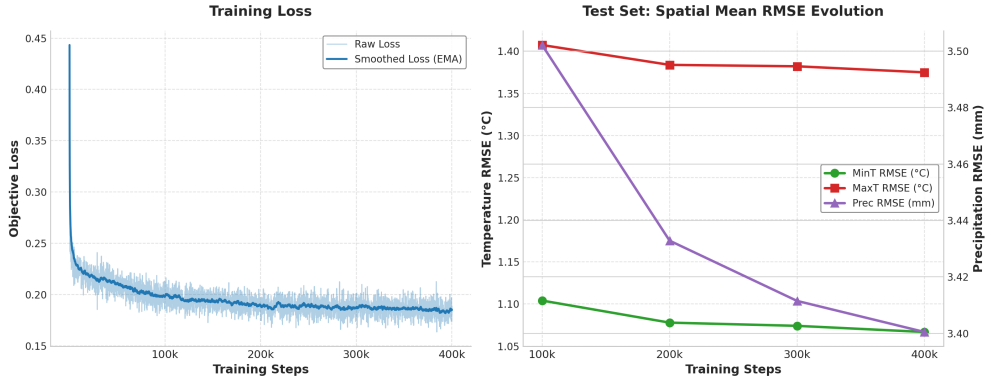

Fig. S4: Learning dynamics of the ClimDiT diffusion backbone over 400,000 training steps. Left: Evolution of the training loss step-by-step, with an overlaid exponential moving average to filter stochastic noise. Right: Evolution of the spatially averaged RMSE evaluated on the independent test set for minimum temperature, maximum temperature, and precipitation, demonstrating continuous improvement without overfitting.

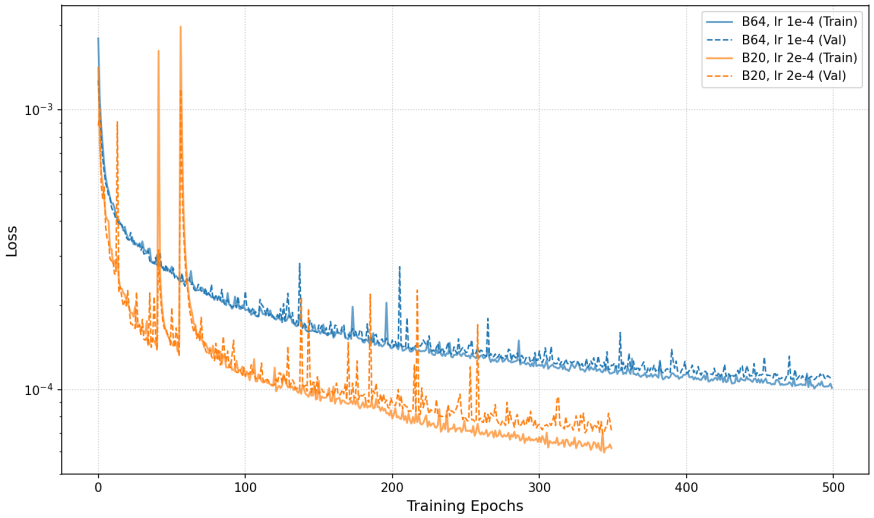

Fig. S5: Comparison of training and validation loss curves (logarithmic scale) during the VAE finetuning stage. The orange curves denote the selected optimal configuration (Batch Size 20, Learning Rate 2×10^{-4}), which slightly outperforms the baseline configuration (blue curves) by reaching a lower validation loss.

configuration with the larger batch size and lower learning rate was trained for a larger number of epochs to compensate for its inherently slower convergence rate (fewer training steps per epoch due to the larger batch size).

Empirical results indicate that the configuration with the smaller batch size and slightly higher learning rate exhibits significantly faster, albeit noisier, convergence while achieving a consistently lower validation loss. Based on these learning dynamics, this latter configuration was selected as the optimal setup for the VAE fine-tuning stage.

S3 Spatial Distribution of Errors

To complement the aggregate metric distributions shown in the main manuscript, we provide explicit spatial maps that evaluate both deterministic and probabilistic performance across the domain. Figures S6 through S8 show the spatial distribution of biases, Root Mean Squared Error (RMSE), and the Continuous Ranked Probability Score (CRPS) for the minimum temperature, the maximum temperature, and the precipitation, respectively. Additionally, Figure S9 shows the spatial distribution of the multivariate metrics.

Note that the deterministic version of ClimDiT is omitted from the deterministic metric maps (Biases and RMSE). Since its point-wise predictions are numerically very similar to the mean of the ClimDiT-Ensemble, their spatial error distributions are virtually identical, and thus only the ensemble results are shown for the sake of visual

clarity. However, both versions are included in the CRPS evaluation to highlight the probabilistic advantage of the ensemble approach.

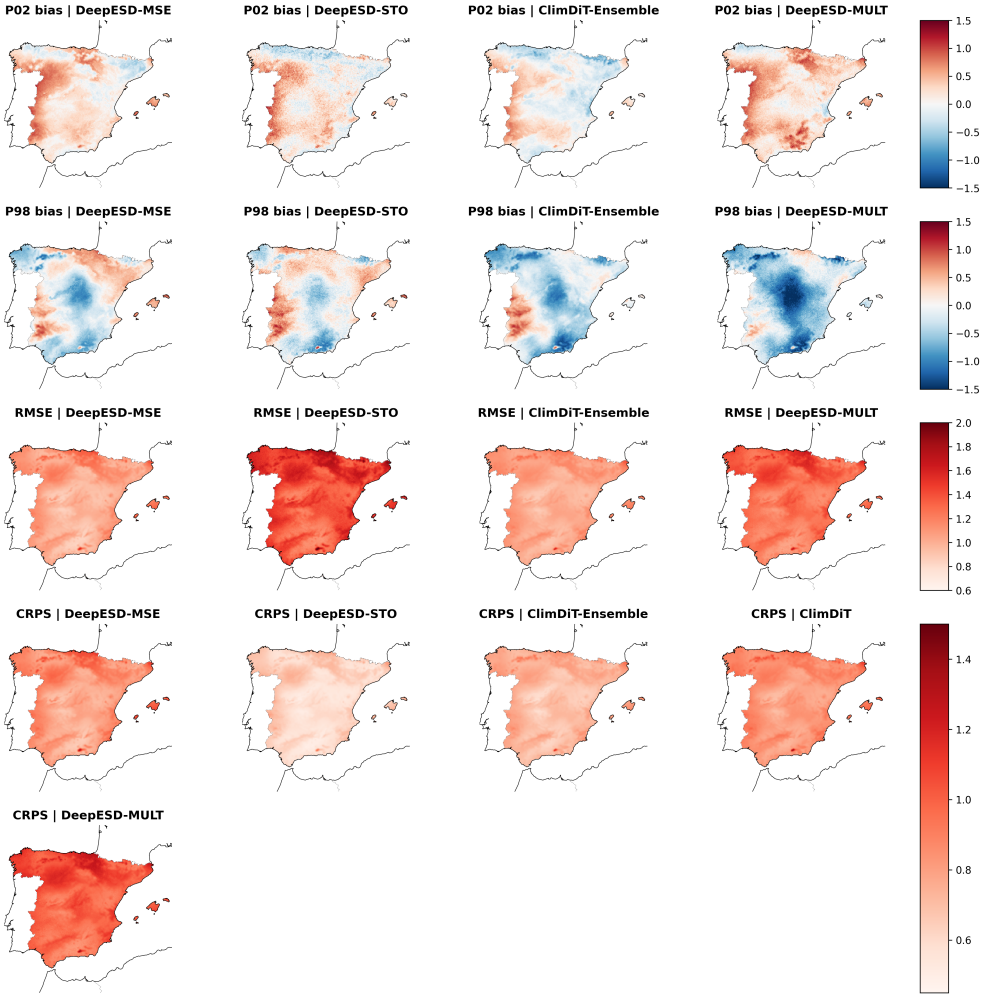

Fig. S6: Spatial distribution of deterministic extreme biases (P02 and P98), RMSE, and probabilistic CRPS for minimum daily temperature (Tmin) across the evaluated models over the Iberian Peninsula.

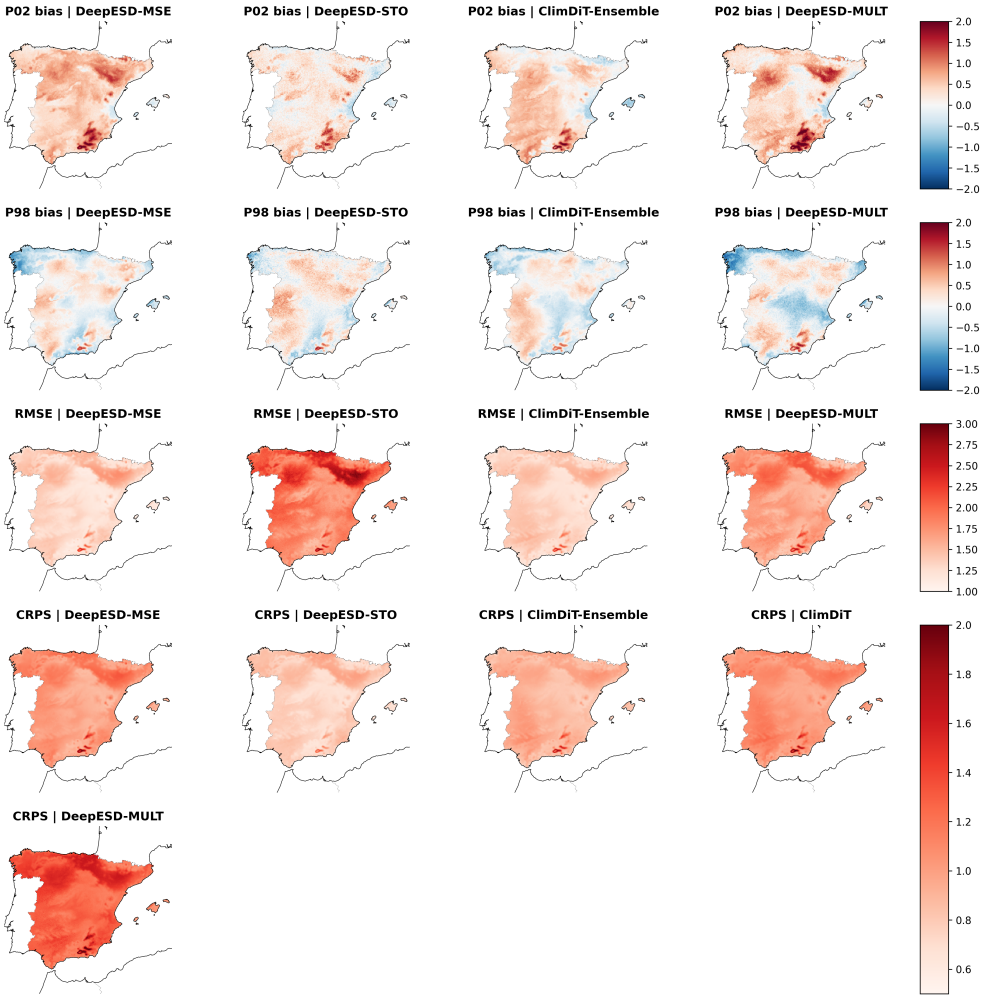

Fig. S7: Spatial distribution of deterministic extreme biases (P02 and P98), RMSE, and probabilistic CRPS for maximum daily temperature (Tmax) across the evaluated models over the Iberian Peninsula.

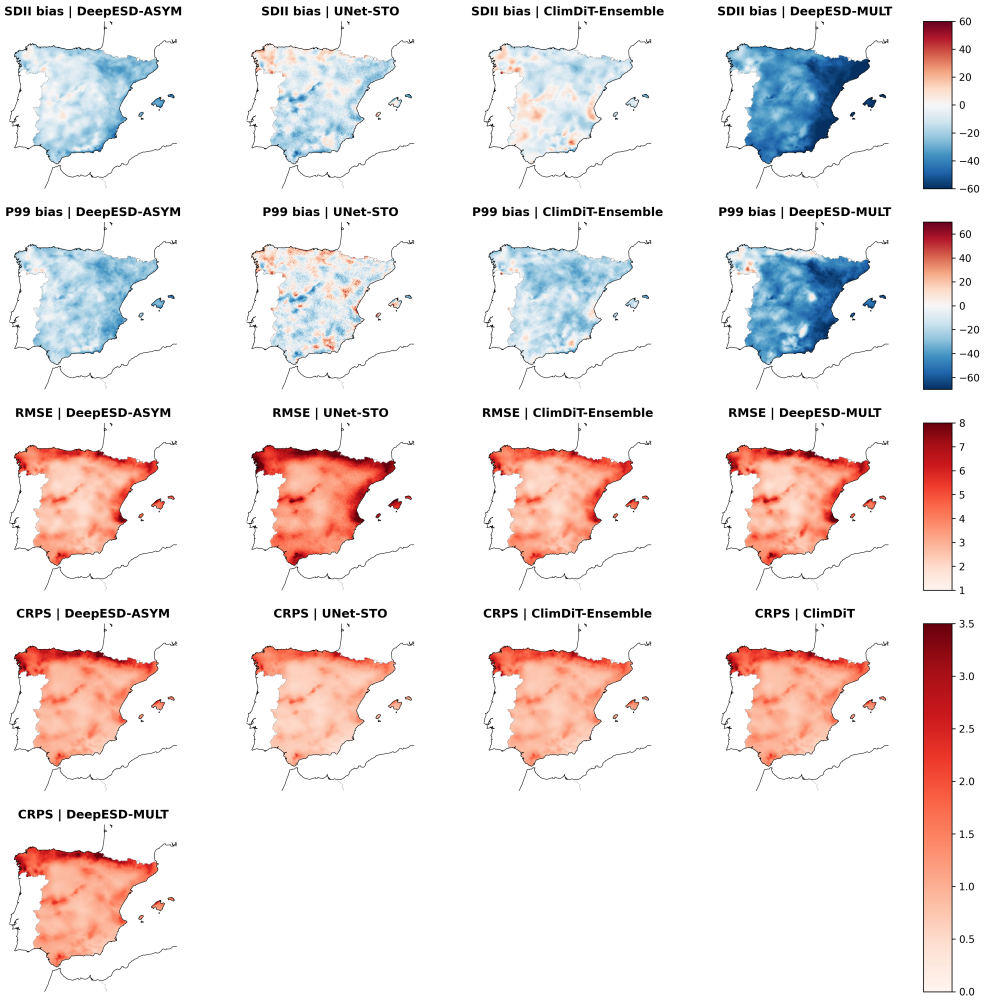

Fig. S8: Spatial distribution of deterministic intensity and extreme biases (SDII bias and P99 bias), RMSE, and probabilistic CRPS for daily accumulated precipitation across the evaluated models over the Iberian Peninsula.

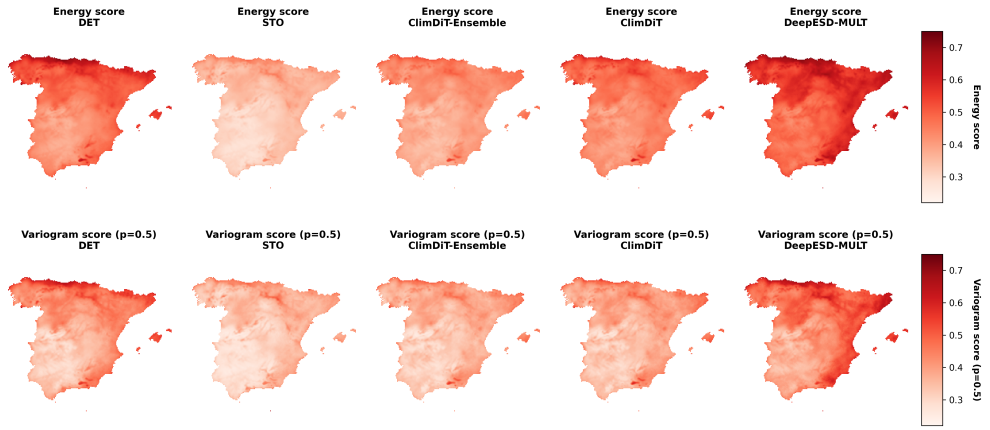

Fig. S9: Spatial distribution of the multivariate metrics, Energy Score and Variogram Score (p=0.5), across the evaluated models over the Iberian Peninsula.

S4 Seasonal Performance Analysis

Climate downscaling performance can vary significantly depending on the prevailing seasonal synoptic conditions. To ensure a robust evaluation of ClimDiT's general performance and multivariate skill across different climatological regimes, we provide a seasonal breakdown of the evaluation metrics based on the spatial median. This approach mitigates the risk that the apparent multivariate performance is merely driven by the annual cycle (e.g., models systematically predicting hotter and drier summers). By isolating the analysis into boireal Winter (DJF: December, January, February), Spring (MAM: March, April, May), Summer (JJA: June, July, August) and Autumn (SON: September, October, November), the models ability to capture intra-seasonal daily weather dynamics is evaluated.

Table S1 presents the univariate deterministic and probabilistic scores. For temperature variables (Tmax and Tmin), DeepESD-STO and ClimDiT-Ensemble consistently show the highest probabilistic skill (CRPS), while DeepESD-MSE and ClimDiT-Ensemble exhibit the strongest deterministic performance (RMSE). The median bias remains generally low across these models, with ClimDiT-Ensemble and DeepESD-MULT frequently ranking among the best. Regarding precipitation (Prec), the mean bias is dominated by DeepESD-MULT and DeepESD-ASYM and the RMSE by ClimDiT-Ensemble and DeepESD-ASYM. In CRPS, UNet-STO is the top-performing model, followed by ClimDiT-Ensemble.

To further assess the multivariate skill, Table S2 details the Energy Score (ES) and Variogram Score (VS) by season. The results demonstrate that the stochastic baseline (STO) achieves the most robust multivariate performance across all seasons, followed by ClimDiT-Ensemble and ClimDiT.

Table S1: Seasonal univariate performance analysis: spatial median of CRPS, mean bias, and RMSE by Winter (DJF), Spring (MAM), Summer (JJA), and Autumn (SON) across all variables. The top two results per metric and season are highlighted in bold.

Variable	Metric	Architecture	DJF	MAM	JJA	SON
Tmax	CRPS	ClimDiT	1.10	1.04	0.95	1.08
		ClimDiT-Ensemble	**0.95**	**0.90**	**0.81**	**0.92**
		DeepESD-MSE	1.08	1.05	0.91	1.00
		DeepESD-STO	**0.83**	**0.81**	**0.67**	**0.75**
		DeepESD-MULT	1.36	1.32	1.21	1.28
	Mean Bias	ClimDiT	**0.11**	**0.03**	0.37	0.36
		ClimDiT-Ensemble	**0.11**	**0.03**	0.37	**0.35**
		DeepESD-MSE	0.20	0.14	**0.23**	**0.35**
		DeepESD-STO	0.38	0.06	**0.21**	**0.35**
		DeepESD-MULT	0.37	**0.01**	0.39	**0.34**
	RMSE	ClimDiT	1.43	**1.34**	1.24	1.39
		ClimDiT-Ensemble	**1.41**	**1.32**	**1.21**	**1.36**
		DeepESD-MSE	**1.39**	**1.34**	**1.18**	**1.28**
		DeepESD-STO	2.00	1.91	1.64	1.77
		DeepESD-MULT	1.74	1.68	1.55	1.63
Tmin	CRPS	ClimDiT	0.90	0.81	0.78	0.85
		ClimDiT-Ensemble	**0.79**	**0.70**	**0.67**	**0.73**
		DeepESD-MSE	0.87	0.78	0.79	0.81
		DeepESD-STO	**0.66**	**0.57**	**0.58**	**0.61**
		DeepESD-MULT	1.08	0.94	0.97	1.02
	Mean Bias	ClimDiT	**-0.08**	-0.12	0.10	**0.04**
		ClimDiT-Ensemble	**-0.07**	-0.12	0.10	**0.04**
		DeepESD-MSE	**0.08**	**0.02**	0.20	0.14
		DeepESD-STO	0.17	-0.03	**0.08**	**0.11**
		DeepESD-MULT	**0.08**	**0.01**	**-0.07**	0.14
	RMSE	ClimDiT	1.14	1.02	**0.99**	1.08
		ClimDiT-Ensemble	**1.13**	**1.00**	**0.97**	**1.06**
		DeepESD-MSE	**1.11**	**0.98**	1.00	**1.01**
		DeepESD-STO	1.61	1.33	1.31	1.40
		DeepESD-MULT	1.37	1.19	1.22	1.28
Prec	CRPS	ClimDiT	0.91	1.31	0.55	1.41
		ClimDiT-Ensemble	**0.82**	**1.18**	**0.51**	**1.28**
		DeepESD-ASYM	0.96	1.27	0.56	1.39
		UNet-STO	**0.67**	**0.97**	**0.44**	**1.04**
		DeepESD-MULT	1.03	1.33	0.66	1.44
	Mean Bias	ClimDiT	-0.16	**-0.13**	-0.24	-0.17
		ClimDiT-Ensemble	-0.16	**-0.13**	-0.25	-0.18
		DeepESD-ASYM	**-0.12**	-0.21	**-0.21**	**-0.15**
		UNet-STO	**-0.02**	**-0.14**	**0.01**	**0.02**
		DeepESD-MULT	-0.19	-0.50	-0.25	-0.55
	RMSE	ClimDiT	2.65	3.27	2.17	3.99
		ClimDiT-Ensemble	**2.62**	**3.22**	2.16	**3.93**
		DeepESD-ASYM	**2.64**	**3.12**	**2.05**	**3.75**
		UNet-STO	3.49	4.25	2.75	5.11
		DeepESD-MULT	3.00	3.28	**2.12**	4.07

Table S2: Seasonal Multivariate Performance: Spatial median of the Energy Score (ES) and Variogram Score ($p = 0.5$, VS) by Winter (DJF), Spring (MAM), Summer (JJA), and Autumn (SON). Top-2 results per metric and season are highlighted in bold.

Model	Energy Score				Variogram Score			
	DJF	**MAM**	**JJA**	**SON**	**DJF**	**MAM**	**JJA**	**SON**
ClimDiT	0.44	0.50	0.36	0.50	0.34	0.45	0.26	0.42
ClimDiT-Ensemble	**0.39**	**0.45**	**0.33**	**0.44**	**0.33**	**0.43**	**0.25**	**0.41**
DET	0.47	0.52	0.38	0.51	0.38	0.48	0.28	0.44
STO	**0.33**	**0.38**	**0.28**	**0.37**	**0.33**	**0.41**	**0.21**	**0.37**
DeepESD-MULT	0.53	0.58	0.47	0.57	0.42	0.53	0.34	0.51

S5 Comparison of Predictive Distributions

To provide a deeper understanding of the probabilistic performance of the models, Figures S10, S11, and S12 compare the empirical histograms generated by the 50-member ClimDiT-Ensemble against the theoretical Probability Density Functions (PDFs) predicted by the stochastic baseline models. For each variable, 20 different cases (each representing a random day and a grid point from the test set) were selected. For precipitation (Figure S12), a stratified sampling was applied to ensure the inclusion of a representative number (10 cases) of rainy days (observed precipitation $> 1mm$), given the zero-inflated nature of the variable.

- **Temperatures (Figures S10 and S11):** The theoretical distributions for the stochastic baseline (DeepESD-STO) are Gaussian. The plots show that DeepESD-STO generally predicts broad distributions, with standard deviations frequently ranging from 0.7 °C to over 1.6 °C. This spread allows the stochastic model to include the true observation (red dashed line) with reasonable probability even when its mean (center of the orange curve) is biased. Conversely, the ClimDiT-Ensemble (blue histogram) produces narrower predictive distributions, indicating higher confidence but resulting in heavier penalties in continuous probabilistic metrics, such as CRPS, when the ensemble mean deviates from the observation.
- **Precipitation (Figure S12):** The theoretical distribution for the stochastic baseline (UNet-STO) is a Bernoulli-Gamma mixture. The probability of no rain is explicitly shown in the text boxes for both models. As in the temperature results, UNet-STO tends to generate broader continuous distributions (green curve) for the precipitation amount compared to the empirical spread of the ClimDiT-Ensemble, which concentrates its members more tightly around its mean. Notably, despite the difference in units, which makes comparison difficult, this trend towards a broader distribution appears to be more pronounced in the case of precipitation than in that of temperature.

Overall, these visualizations corroborate the grid-point metric results discussed in the main manuscript, highlighting the trade-off between the sharp, spatially coherent

fields generated by the diffusion ensemble and the broader, well-calibrated marginal uncertainties explicitly modeled by the stochastic baselines.

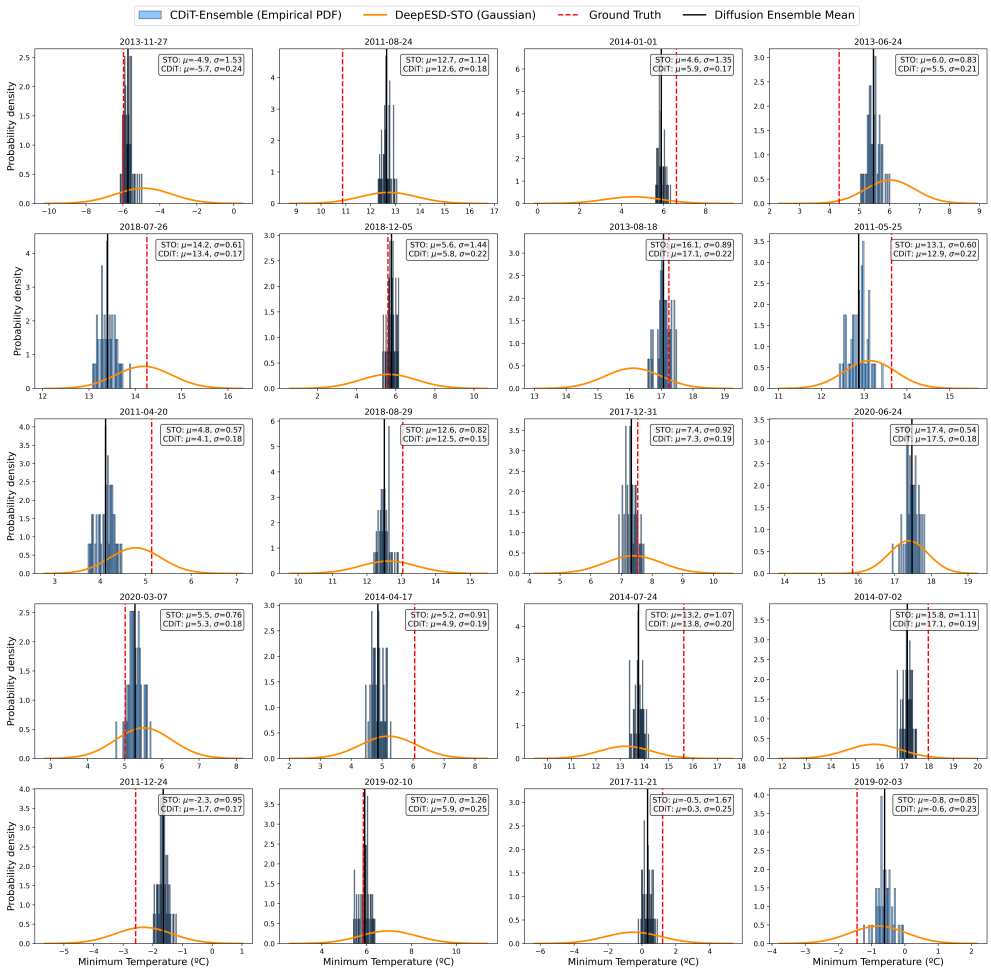

Fig. S10: Histograms of the 50-member ClimDiT-Ensemble (blue histograms) versus the theoretical Gaussian PDFs predicted by the DeepESD-STO baseline (orange curves) for **Minimum Temperature**. Red dashed lines indicate the observed values, and black solid lines denote the ClimDiT-Ensemble mean. The text boxes indicate the mean (μ) and standard deviation (σ) for both models. Each subplot represents a randomly selected day and grid point from the test set.

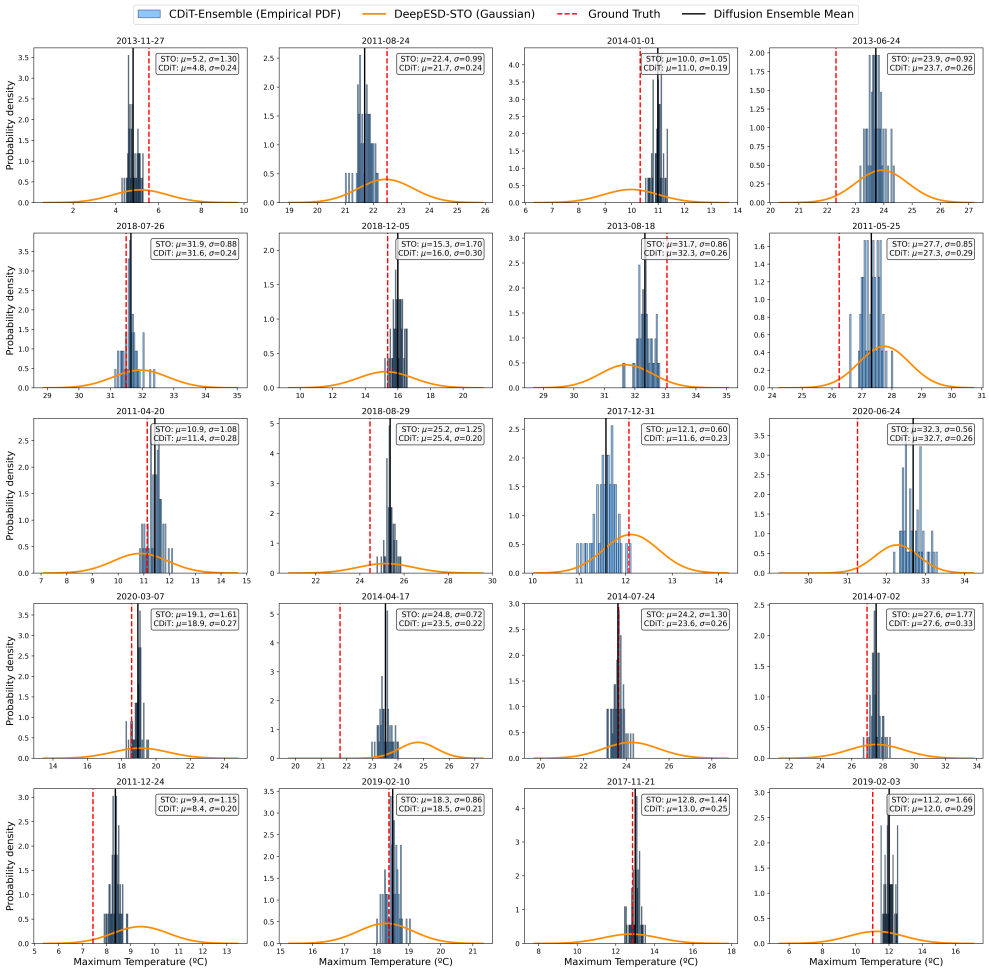

Fig. S11: Histograms of the 50-member ClimDiT-Ensemble (blue histograms) versus the theoretical Gaussian PDFs predicted by the DeepESD-STO baseline (orange curves) for **Maximum Temperature**. Red dashed lines indicate the observed value, and black solid lines denote the ClimDiT-Ensemble mean. Each subplot represents a randomly selected day and grid point from the test set.

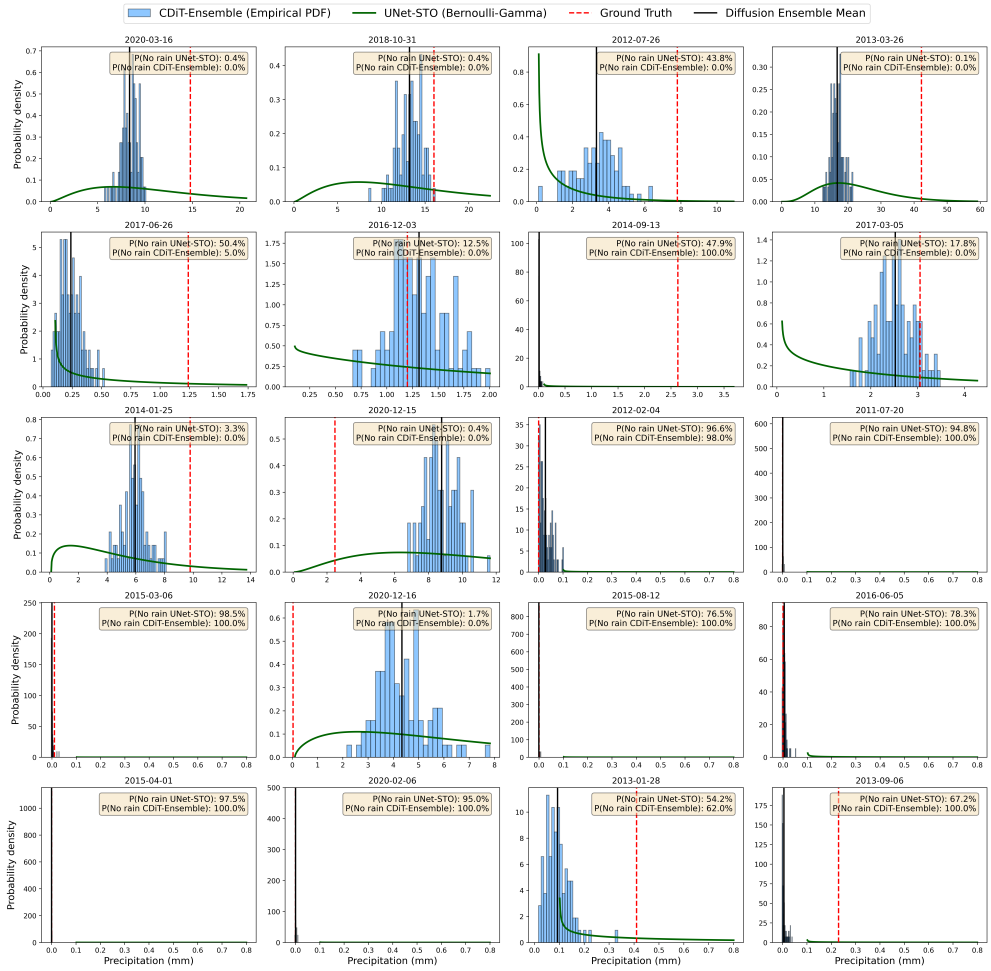

Fig. S12: Histogramas of the 50-member ClimDiT-Ensemble (blue histograms) versus the theoretical Bernoulli-Gamma mixture PDFs predicted by the UNet-STO baseline (green curves) for **Daily Accumulated Precipitation**. Red dashed lines indicate the observed value, and black solid lines denote the ClimDiT-Ensemble mean. The text boxes display the explicitly modeled or empirically derived probability of no rain for each method. The 20 cases were selected using a stratified approach to highlight behavior on both dry and rainy days.

S6 Ensemble Size Sensitivity Analysis

To evaluate the robustness and sensitivity of the ClimDiT-Ensemble to the number of generated members (m), the Continuous Ranked Probability Score (CRPS), the Brier Score (BS) for extreme percentiles, and the multivariate Energy Score were computed for different ensemble sizes ($m \in 1, 20, 50, 100$). Figure S13 illustrates the spatial distribution of these metrics over the Iberian Peninsula domain. The results demonstrate a substantial improvement in predictive skill and uncertainty calibration when increasing the ensemble size from a single deterministic pass ($M = 1$) to $M = 20$. However, from this point on, performance improvements become stable; the improvement observed when moving from $M = 20$ to $M = 50$ is negligible, and the spatial error distributions for $M = 100$ are virtually identical to those for $M = 50$.

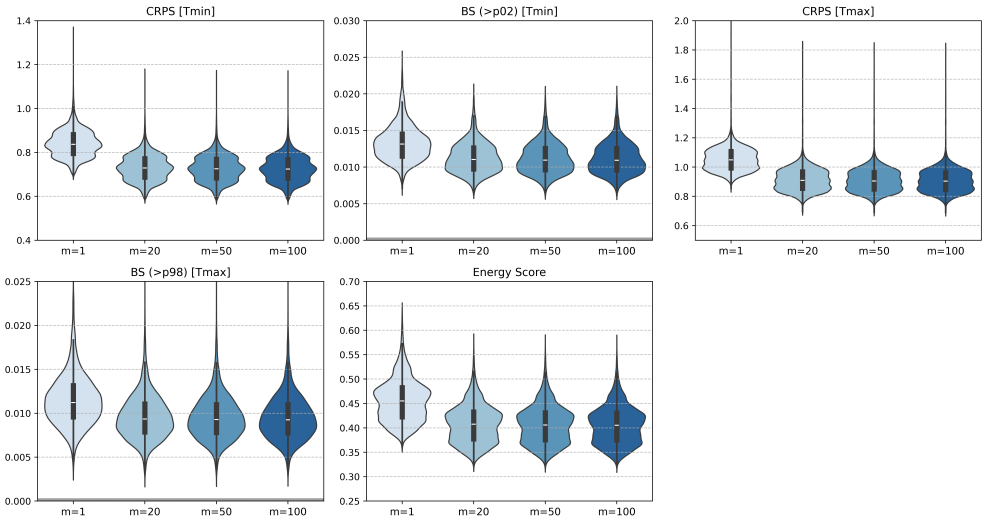

Fig. S13: Sensitivity of ClimDiT-Ensemble predictive skill to the number of ensemble members (m). The violin plots show the spatial distribution of CRPS for minimum and maximum temperature, the Brier Score for extreme percentiles (p02 for Tmin, p98 for Tmax), and the multivariate Energy Score. Performance improves significantly from $M = 1$ to $M = 20$ and stabilizes for larger ensemble sizes.