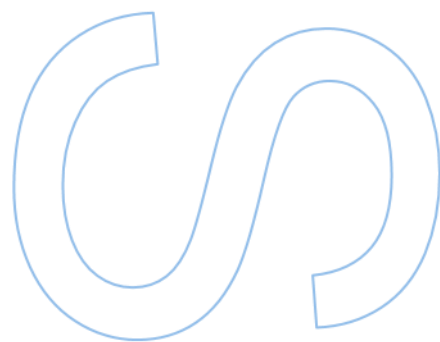
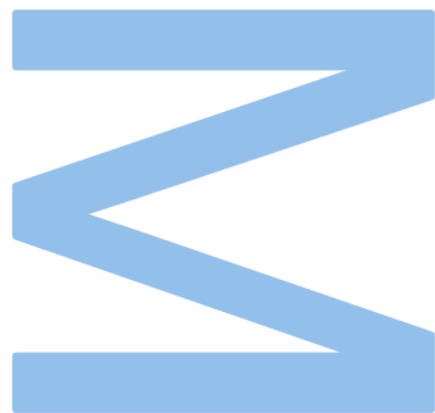
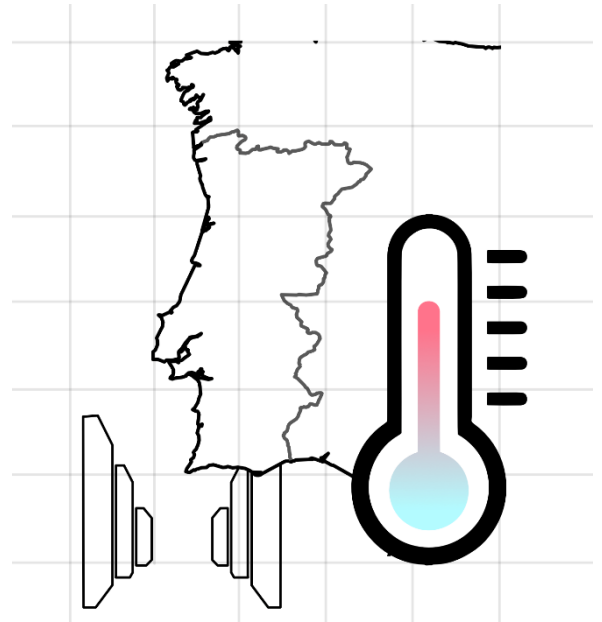


Probabilistic Spatiotemporal Downscaling of Climate Data via Conditional Diffusion Models

Diogo Filipe Gonçalves Cruz
Master in Computational Statistics and Data Analysis
Department of Mathematics
Faculty of Sciences of the University of Porto
2026





Probabilistic Spatiotemporal Downscaling of Climate Data via Conditional Diffusion Models

Diogo Filipe Gonçalves Cruz

Dissertation carried out as part of the Master in
Computational Statistics and Data Analysis
Department of Mathematics
2026

Supervisor

Prof. Carla Gonçalves, Assistant Professor, Faculty of Engineering
of the University of Porto

Co-supervisor

Prof. Ricardo J. Bessa, Centre Coordinator and Senior
Researcher, INESC TEC



*“ Whatever I do, whatever I say
Whatever may come my way
I owe it to you
You taught me to love
As long as I live, I owe it to you ”*

MARO, I OWE IT TO YOU (2026)

*Dedicated to my family, and to my grandfather José Cruz, who I will
always carry with me.*

Acknowledgements

To my supervisor, Professor Carla Silva Gonçalves, thank you for trusting me with the freedom to find my own way into this theme, without ever letting that freedom feel like being left alone. Your guidance, your suggestions, and your patience shaped this dissertation more than you probably realize, and I don't think it would have taken the form it did without you. You were always there when I needed direction. Thank you, above all, for the kindness you brought to every single meeting.

To my co-supervisor, Professor Ricardo J. Bessa, thank you for suggesting the idea that became this whole dissertation, and for every idea, suggestion, and piece of guidance you offered along the way. Thank you, also, for always making the time for my questions. Working on this theme has confirmed that this is a subject I truly love, and one I hope to keep working on for a long time.

To every teacher and professor who has crossed my path, from my very first classroom to my last lecture, thank you for filling me with knowledge and for teaching me how to think for myself.

I know this dissertation is, in the grand scheme of things, only a small step along a much longer path. But right now, at the end of it, it feels like the biggest one I have taken so far, and that is why the words that follow matter so much to me.

Aos meus amigos e pessoas com quem passei muitas tardes e noites entre cafés e sessões de música, entre eles: Afonso, Nozes, Trindade, Márcia, Toscano, Sofia e outros, obrigado. Sem vocês este caminho teria sido mais pesado de percorrer sozinho.

À minha namorada, Soraia, obrigado por estares sempre presente ao longo deste percurso, mesmo nos dias em que eu passava mais horas a olhar para um ecrã do que para ti. Obrigado pelo carinho e por me lembrares que há vida além dos resultados e gráficos.

Aos meus pais, Sónia e Carlos, não há palavras que cheguem para o que fazem por mim. Obrigado por me darem a liberdade de perseguir aquilo que amo, pela paciência de me ouvir, dia após dia, falar dos meus problemas e dos meus resultados como se fossem também vossos. Obrigado por me ajudarem a construir a pessoa que sou hoje e, acima de tudo, por nunca terem deixado de acreditar em mim, mesmo nas alturas em que eu duvidava.

À minha irmã, Catarina, obrigado por estares sempre por perto, por me fazeres rir, e por seres, mesmo sem teres noção disso, uma fonte de inspiração para mim.

Ao resto da minha família, obrigado por tudo.

Por fim, às minhas gatas, Mia e Stout, obrigado pela companhia, por vezes silenciosa, em muitas horas deste trabalho.

This work was co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project ATE, with reference 56.

Resumo

Informação climática horária de alta resolução é essencial para avaliação de impactos, planeamento de infraestruturas e adaptação climática, mas os modelos climáticos globais fornecem geralmente campos de temperatura com resolução espacial e temporal grosseira, o que torna necessário o recurso a métodos de *downscaling* capazes de os transformar em campos horários de maior resolução. Neste contexto, o problema não é de previsão determinística, mas sim de geração probabilística, uma vez que um único campo de temperatura média diária pode corresponder a várias evoluções horárias plausíveis.

Esta dissertação desenvolve uma metodologia probabilística de *downscaling* espaciotemporal baseada em modelos de difusão condicionais. O modelo gera campos horários de temperatura junto à superfície, condicionados pela temperatura diária média de baixa resolução, por variáveis geográficas estáticas, por codificações temporais e por controlos escalares de variabilidade subdiária. O modelo é treinado com dados ECMWF Reanalysis v5 (ERA5) sobre a Península Ibérica e a região adjacente do Atlântico, sendo comparado com um modelo determinístico (U-Net), uma versão sem os controlos escalares e uma referência análoga baseada em k-vizinhos mais próximos (k-NN).

Os resultados mostram que a U-Net determinística obtém o menor erro pontual, mas não representa incerteza, enquanto a referência k-NN permanece competitiva em várias métricas marginais e estruturais. Entre os modelos neuronais, o modelo de difusão completo apresenta o comportamento generativo mais equilibrado. Relativamente à versão sem os controlos escalares, o modelo proposto melhora a estrutura espaciotemporal e a calibração, e representa melhor a variabilidade subdiária e a cauda quente. Aplicado a projeções Coupled Model Intercomparison Project Phase 6 (CMIP6) com correção de viés para 2050–2059, o modelo parece preservar a condição diária imposta e produz estruturas horárias fisicamente plausíveis, embora a subdispersão e a compressão da cauda superior permaneçam limitações. Globalmente, os modelos de difusão condicionais mostram potencial para *downscaling* probabilístico e geração de campos de temperatura quando a diversidade e plausibilidade dos cenários são requisitos importantes.

Palavras-chave: Modelos de difusão condicionais, Downscaling espaciotemporal probabilístico, Geração de dados climáticos, Projeções climáticas.

Abstract

High-resolution hourly climate information is essential for impact assessment, infrastructure planning, and climate adaptation, but global climate models typically provide temperature fields at coarser spatial and temporal resolution. This limitation motivates the use of downscaling methods capable of transforming coarse climate information into higher-resolution hourly fields. In this context, the problem is not one of deterministic forecasting but of probabilistic generation, since a single daily mean temperature field can correspond to several plausible hourly evolutions.

This dissertation develops a probabilistic spatiotemporal downscaling methodology based on conditional diffusion models. The model generates hourly near-surface temperature fields, conditioned on the low-resolution daily mean temperature, static geographic variables, temporal encodings, and scalar controls of sub-daily variability. The model is trained on ERA5 data over the Iberian Peninsula and the adjacent Atlantic region, and is compared against a deterministic model (U-Net), an ablated version without the scalar controls, and an analog reference based on k-nearest neighbors (k-NN).

Results show that the deterministic U-Net achieves the lowest pointwise error but does not represent uncertainty, while the k-NN reference remains competitive across several marginal and structural metrics. Among the neural models, the full diffusion model exhibits the most balanced generative behavior. Relative to the ablated version without scalar controls, it improves spatiotemporal structure and calibration, and better represents sub-daily variability and the warm tail. Applied to bias-corrected CMIP6 projections for 2050–2059, the model appears to preserve the imposed daily condition and produces physically plausible hourly structures, although under-dispersion and upper-tail compression remain limitations. Overall, conditional diffusion models show potential for probabilistic downscaling and temperature field generation when scenario diversity and plausibility are important requirements.

Keywords: Conditional diffusion models; Probabilistic spatiotemporal downscaling; Climate data generation; Climate projections.

Table of Contents

List of Tables	ix
List of Figures	x
List of Abbreviations	xii
List of Symbols	xiv
1. Introduction	1
1.1. Context and Motivation	1
1.1.1. Climate Change and the Need for High-resolution Climate Data	2
1.1.2. The Downscaling Problem	4
1.1.3. Motivation for Deep Generative Models	5
1.2. Research Question, Objectives and Contributions	6
1.3. Structure	7
2. Climate Downscaling Background	9
2.1. The Climate Downscaling Problem	9
2.2. Climate Model Data Sources	10
2.2.1. The CNRM-CM6-1-HR Model	10
2.2.2. Shared Socioeconomic Pathways	10
2.3. ERA5 Reanalysis	11
2.4. Classical Statistical Downscaling	11
2.4.1. Bias Correction and Quantile Delta Mapping	12
2.4.2. Analog Methods and k-Nearest Neighbors	12
2.4.3. Other Classical Approaches	13
3. Generative Modeling Background	15
3.1. Generative Modeling for Probabilistic Downscaling	15
3.2. Deep Generative Models	16
3.2.1. Mathematical Formulation	16
3.2.2. Families of Deep Generative Models	16
3.3. Diffusion Models	19
3.3.1. Forward Diffusion Process	20
3.3.2. Reverse Generative Process	20
3.3.3. Training Objective	21
3.3.4. Sampling Algorithms	21
3.3.4.1. Traditional (DDPM) Sampling	21
3.3.4.2. Denoising Diffusion Implicit Models	22
3.3.4.3. DPM-Solver++	22

3.4. Conditional Diffusion Models	23
3.4.1. Conditional Reverse Process	23
3.4.1.1. Classifier Guidance	23
3.4.1.2. Classifier-Free Guidance	24
3.5. Related Work	24
3.5.1. GANs for Climate Downscaling	24
3.5.2. Diffusion Models.....	25
3.5.3. Spatiotemporal and Temporal Downscaling.....	26
4. Methodology	28
4.1. Problem Formulation	28
4.2. Data and Preprocessing	30
4.2.1. ERA5 Hourly Reference Data	30
4.2.2. Low-Resolution Daily Field Construction.....	31
4.2.3. Static Field Extraction	32
4.2.4. Data Split and Normalization	33
4.3. Target Formulation	34
4.3.1. Anomaly Target.....	34
4.3.2. Reconstruction of Absolute Temperature	35
4.4. Conditioning Design.....	35
4.4.1. Spatial and Temporal Conditions.....	35
4.4.2. Scalar Control Variables	37
4.4.3. Assembly of the Conditioning Tensor.....	38
4.5. Conditional Diffusion Models	39
4.6. Denoising Architecture.....	40
4.7. Training Procedure and Implementation Details	43
4.7.1. Implementation Details	45
4.8. Inference and Scenario Generation	45
4.9. Application to CMIP6 Projections	46
4.10. Baselines and Models for Comparison.....	47
4.10.1. Full Diffusion Model	47
4.10.2. Model without Scalar Controls	47
4.10.3. k-NN Anomaly Baseline	48
4.10.4. Deterministic U-Net.....	49
4.11. Evaluation Metrics	49
4.11.1. Evaluation Spaces	50
4.11.2. Point-wise Error Metrics.....	50
4.11.3. Probabilistic Skill and Ensemble Calibration	51
4.11.4. Diurnal and Temporal Structure.....	52

4.11.5. Spatial Structure.....	52
4.11.6. Distributional and Extreme-Temperature Diagnostics.....	53
4.11.7. Orographic and conditioning consistency.....	54
4.12. Experimental Design.....	55
4.12.1. Experiment 1 - Model Comparison	55
4.12.2. Experiment 2 - Ensemble Depth.....	56
4.12.3. Experiment 3 - Future Projection	56
5. Results and Discussion	57
5.1. Non-Stationarity of the Reference Data	57
5.2. Experiment 1: Model Comparison.....	60
5.2.1. Deterministic Accuracy and Probabilistic Skill	60
5.2.2. Temporal Structure and Sub-daily Variability	63
5.2.3. Spatial Structure and Physical Consistency	66
5.2.4. Distributional Behavior and Extremes	69
5.2.5. Summary.....	72
5.3. Experiment 2: Ensemble Depth	74
5.3.1. Ensemble Calibration and Dispersion	75
5.3.2. Distributional Behavior and Extremes	77
5.3.3. Conditioning Controllability	78
5.3.4. Summary.....	79
5.4. Experiment 3: Future Projection.....	80
5.4.1. Conditioning Consistency	80
5.4.2. Distributional and Tail Plausibility	81
5.4.3. Structural and Physical Consistency	84
5.4.4. Summary.....	87
5.5. Synthesis of Results and Discussion	88
6. Conclusions and Future Work.....	89
Bibliography	91
A. Additional Results.....	101
A.1. Experiment 1	101
A.2. Experiment 3.....	103

List of Tables

4.1.	Data partitioning for model training and testing. The dataset is split chronologically.	33
4.2.	Summary of conditioning components, channel dimensions, and their respective spatiotemporal behavior.	39
4.3.	Training configuration of the conditional diffusion model.	44
4.4.	Experimental configuration and simulation design matrix for model benchmarking, ensemble depth analysis, and climate projection runs.	55
5.1.	Comparative deterministic and probabilistic skill metrics for all models.	61
5.2.	Comparative time of maximum and minimum phase error metrics for all models.	62
5.3.	Comparative seasonal hourly temperature ramp skill and standard deviation ratios for all models.	65
5.4.	1-Wasserstein distance between land-pixel daily maximum temperature exceedances above 40 °C and ERA5.	71
5.5.	Deterministic and probabilistic skill metrics for shallow and deep configurations.	77
5.6.	1-Wasserstein distances for the full summer distribution and the extreme tail. .	78
5.7.	Correlation metrics for scalar conditioning controls.	79
A.1.	Comparison of land-pixel spatial correlation across the models.	101
A.2.	Seasonal amplitude-distribution errors across models.	102
A.3.	June–September (JJAS) $T_{\max}$ distribution shift under CMIP6 projection.	104
A.4.	Timing of daily temperature extrema under CMIP6 projection.	106

List of Figures

3.1.	Taxonomy of Deep Generative Models (DGMs).	17
3.2.	Scheme of GAN architecture	18
3.3.	Scheme of Variational autoencoder (VAE) architecture.	19
4.1.	Overview of the proposed methodology.	29
4.2.	Geographic extent of the study area covering the Iberian Peninsula and the adjacent Atlantic Ocean.	31
4.3.	Schematic of the synthetic low-resolution daily field construction.	32
4.4.	Orography field used as a static conditioning variable.	36
4.5.	Schematic representation of the conditional (2 + 1)D U-Net denoising architecture.	41
4.6.	Schematic of the residual spatiotemporal block.	42
5.1.	Land-mean ERA5 hourly temperature diurnal cycle for 1980-2009 and 2014-2023.	58
5.2.	Diurnal timing of daily temperature extremes in training and test periods.	58
5.3.	Spatial distribution of the temperature amplitude shifts between training and test periods across the land domain.	59
5.4.	Hourly deterministic and probabilistic error profiles for all models.	61
5.5.	Probabilistic calibration of the ensembles from different models.	63
5.6.	Mean diurnal temperature anomaly over land for winter (December–February (DJF)) and summer (June–August (JJA)).	63
5.7.	Within-day autocorrelation function of daily-mean temperature anomalies over land.	64
5.8.	One-week point time series for the Full diffusion ensemble.	66
5.9.	Spatial Power Spectral Density (PSD) of temperature anomalies across all models.	66
5.10.	Distribution of pairwise spatial correlation errors over land grid points relative to ERA5.	68
5.11.	Mean temperature as function of elevation over land for ERA5 and the downscaling models during 2014–2023.	68
5.12.	Seasonal probability density of sub-daily temperature anomaly amplitudes.	69
5.13.	Quantile bias of summer daily maximum temperatures at extreme percentiles for all models.	70
5.14.	Exceedance rates of summer daily maximum temperatures above 35 °C and 40 °C for all models.	71
5.15.	Model demonstration of the Full diffusion generations for a day.	73
5.16.	Rank histograms of the ERA5 reference within the generated ensembles for Full diffusion.	75
5.17.	Probabilistic calibration of the shallow and deep generated ensembles for Full diffusion.	76
5.18.	Quantile bias and exceedance rates for extended summer daily maximum temperatures for different ensemble sizes.	77
5.19.	Daily-mean conditioning consistency under CMIP6 projections.	81
5.20.	Seasonal distributions of sub-daily temperature anomalies under CMIP6 projections.	82

5.21. Upper-tail shift and extreme-heat exceedance of JJAS $T_{\max}$ under CMIP6 projection.	83
5.22. Seasonal mean diurnal cycle under CMIP6 projection.	84
5.23. Within-day autocorrelation function of sub-daily temperature anomalies under CMIP6 projection.	85
5.24. PSD under CMIP6 projection.	86
5.25. Orographic temperature profile under CMIP6 projection.	87
A.1. Spatial Root Mean Square Error (RMSE) and Continuous Ranked Probability Score (CRPS) comparison for Experiment 1.	101
A.2. Pairwise 1-Wasserstein distances of pooled JJAS daily maximum temperatures.	102
A.3. Distribution of daily mean temperature for ERA5, CMIP6, and Quantile Delta Mapping (QDM)-corrected CMIP6.	103
A.4. Seasonal absolute temperature distributions under CMIP6 projection.	104
A.5. JJAS $T_{\max}$ distribution under CMIP6 projection.	105
A.6. Absolute hourly temperature ramp distribution under CMIP6 projection.	105
A.7. Residual spatial-correlation errors under CMIP6 projection.	106

List of Abbreviations

- ACF** Autocorrelation Function. 64, 85
- AI** Artificial Intelligence. 5, 16
- CMIP6** Coupled Model Intercomparison Project Phase 6. iv, v, vii, ix–xi, 7–11, 27, 28, 33, 46, 47, 56, 80–90, 103–106
- CRPS** Continuous Ranked Probability Score. xi, 51, 52, 60–62, 65, 76, 77, 79, 88, 101
- CRPSS** Continuous Ranked Probability Skill Score. 51
- DDIM** Denoising Diffusion Implicit Model. 22, 23
- DDPM** Denoising Diffusion Probabilistic Model. vi, 19–22
- DGM** Deep Generative Model. x, 6, 8, 16, 17, 19
- DJF** December–February. x, 63, 65, 82, 84, 85, 102
- ECMWF** European Centre for Medium-Range Weather Forecasts. 11
- EDM** Elucidated Diffusion Model. 25
- EMA** Exponential Moving Average. 44
- ERA5** ECMWF Reanalysis v5. iv, v, x, 6, 8–13, 26, 28, 30–34, 37, 38, 46, 48–58, 60, 61, 63–73, 75, 77, 78, 80–88, 90, 102–106
- FiLM** Feature-wise Linear Modulation. 41, 43, 90
- GAN** Generative Adversarial Network. vii, x, 8, 17, 24–26
- GCM** Global Climate Model. 1, 2, 4, 6, 9, 10, 12
- JJA** June–August. x, 63, 65, 81, 82, 84, 85, 102
- JJAS** June–September. ix, xi, 54, 70–72, 77, 78, 83, 102, 104, 105

- k-NN** k-nearest neighbors. vii, 12, 13, 33, 47, 48, 51, 55, 60–62, 64–72, 74, 88, 89
- KL** Kullback-Leibler. 16
- KS** Kolmogorov-Smirnov. 53, 57, 59, 81
- LSM** Land-Sea Mask. 32, 33, 35, 46
- MAM** March–May. 82, 85, 102
- MLP** Multilayer Perceptron. 43
- MSE** Mean Squared Error. 21, 25, 49
- PSD** Power Spectral Density. x, xi, 53, 66, 85, 86
- QDM** Quantile Delta Mapping. xi, 12, 46, 56, 80, 103
- QM** Quantile Mapping. 12
- RCM** Regional Climate Model. 4, 5
- RMSE** Root Mean Square Error. xi, 26, 50, 54, 60–62, 67, 76, 77, 81, 86, 87, 101
- SDE** Stochastic Differential Equation. 20
- SGM** Score-Based Generative Model. 19
- SiLU** Sigmoid Linear Unit. 41, 43
- SON** September–November. 81, 85, 102
- SSP** Shared Socioeconomic Pathways. 10, 11, 56, 57, 80, 83–87, 89, 104–106
- t2m** 2-meter air temperature. 11
- VAE** Variational autoencoder. x, 8, 18, 19, 21

List of Symbols

Symbol	Dimension	Description
d	scalar	Day index.
h	scalar	Hour index.
i, j	scalar	Spatial grid indices.
k	scalar	Ensemble-member index.
K	scalar	Number of ensemble members.
y_d^{LR}	$H_{LR} \times W_{LR}$	Low-resolution daily mean temperature field.
x_d^{HR}	$24 \times H_{HR} \times W_{HR}$	High-resolution hourly temperature block.
$x_{d,h}^{HR}$	$H_{HR} \times W_{HR}$	High-resolution temperature field at hour h .
$\hat{x}_d^{HR,(k)}$	$24 \times H_{HR} \times W_{HR}$	Generated hourly temperature block for member k .
$\bar{x}_{d,h,i,j}$	scalar	Ensemble-mean generated temperature.
$x_{d,h,i,j}^E$	scalar	ERA5 reference temperature.
m_d^{HR}	$H_{HR} \times W_{HR}$	High-resolution daily mean field.
$\tilde{y}_d^{HR}$	$H_{HR} \times W_{HR}$	Low-resolution daily field projected to the high-resolution grid.
a_d	$24 \times H_{HR} \times W_{HR}$	Anomaly target relative to $\tilde{y}_d^{HR}$.
$a_{d,\text{norm}}$	$24 \times H_{HR} \times W_{HR}$	Normalized anomaly target.
$\hat{a}_d$	$24 \times H_{HR} \times W_{HR}$	Generated anomaly in physical units.
μ_a	scalar	Training-period anomaly mean.
σ_a	scalar	Training-period anomaly standard deviation.
c_d	$C_c \times 24 \times H_{HR} \times W_{HR}$	Conditioning tensor.
C_c	scalar	Number of conditioning channels.
$c_{\text{doy},1}, c_{\text{doy},2}$	scalar	Sine and cosine day-of-year encodings.
$c_{\text{hour},1}, c_{\text{hour},2}$	scalar	Sine and cosine hour-of-day encodings.
s_d	scalar	Diurnal-variability control.
r_d	scalar	Daily temperature-regime control.
$z_d = (s_d, r_d)$	2	Scalar-control vector.
u_d	$24 \times H_{HR} \times W_{HR}$	Normalized anomaly used by the diffusion model.
$u_{d,\tau}$	$24 \times H_{HR} \times W_{HR}$	Noisy anomaly at diffusion step τ .
τ	scalar	Diffusion timestep.
ϵ	$24 \times H_{HR} \times W_{HR}$	Gaussian noise.

Symbol	Dimension	Description
$\hat{\epsilon}$	$24 \times H_{HR} \times W_{HR}$	Predicted noise.
α_{τ}	scalar	Signal-retention coefficient.
$\bar{\alpha}_{\tau}$	scalar	Cumulative signal-retention coefficient.
β_{τ}	scalar	Diffusion noise variance coefficient.
θ	–	Trainable model parameters.
$\mathcal{L}_{\text{MSE}}$	scalar	Diffusion noise-prediction loss.
$v_{d,\tau}$	$(1 + C_c) \times 24 \times H_{HR} \times W_{HR}$	Denoising-network input.
$\mathcal{L}$	set	Set of land grid points.
N_L	scalar	Number of land pixels.
N_d	scalar	Number of evaluated days.
$a_{d,h,i,j}^{\text{diurn},(k)}$	scalar	Diurnal anomaly.
$A_{d,i,j}^{(k)}$	scalar	Daily temperature amplitude.
Δh	scalar	Circular phase error of daily extrema.
$\Delta_h x_{d,h,i,j}^{(k)}$	scalar	Hourly temperature ramp.
$C_{p,q}$	scalar	Spatial correlation between land pixels p and q .
$P_{d,h}^{(k)}(\zeta, \eta)$	scalar	Spatial spectral power.
D_{KS}	scalar	Kolmogorov–Smirnov statistic.
W_1	scalar	1-Wasserstein distance.
$T_{\text{max},d,i,j}^{(k)}$	scalar	Daily maximum temperature.
$Q_q(\cdot)$	scalar	Empirical quantile at level q .
Bias_q	scalar	Quantile bias.
$\text{rate}_u^{\text{gen}}$	scalar	Generated exceedance rate above threshold u .

1. Introduction

This dissertation focuses on transforming coarse daily climate information into high-resolution hourly temperature scenarios. This chapter frames the problem as a probabilistic spatiotemporal downscaling task, motivates the proposed modeling approach, and presents the objectives, contributions, and organization of the work.

1.1. Context and Motivation

Climate variability and extremes strongly affect ecosystems, economies and human societies. Critical infrastructures such as energy grids, water systems, and transportation networks were mainly designed based on historical climate conditions as a reference. As global warming intensifies, these systems are becoming increasingly exposed to high-impact events whose effects are often expressed at fine spatial and temporal scales. Therefore, assessing vulnerability and designing adaptation strategies requires high-resolution spatiotemporal climate information, often at hourly resolution and spatial scales on the order of 10-50 km, to capture processes such as heat stress, hydrological extremes and energy demand [1].

Present and future climate conditions are commonly studied using **Global Climate Model (GCM)** [2]. GCM are physically-based numerical models that solve governing equations of the atmosphere, oceans, land surface, and cryosphere at the global scale. They form the scientific foundation for coordinated international modelling efforts such as the **Coupled Model Intercomparison Project (CMIP)** [3], which provides multi-model climate projections under different greenhouse-gas forcing scenarios. GCM are indispensable for representing large-scale circulation patterns, thermodynamic balances, and long-term trends. However, due to computational constraints, GCM operate at coarse spatial resolutions (typically around 100-250 km) and commonly produce output at daily or longer temporal intervals.

This mismatch between the scale of GCM and the scale required by many climate-impact applications creates a fundamental information gap. Coarse-resolution climate projections are often insufficient for applications that depend on fine-scale, high-frequency variability. Processes such as diurnal temperature cycles, urban heat islands, and mesoscale

circulations may be weakly represented or entirely absent in raw GCM outputs. Consequently, these data cannot usually be used directly for local risk assessments or operational decision-making. Bridging this resolution gap motivates the development of climate downscaling methods, whose purpose is to transform large-scale GCM information into realistic fine-scale spatiotemporal climate fields suitable for impact modelling and adaptation planning.

1.1.1 Climate Change and the Need for High-resolution Climate Data

The consequences of inadequate climate-data resolution are not merely technical. They affect the ability to quantify risks in sectors that depend strongly on local and sub-daily weather variability. Next, energy-sector-related examples are presented, although high-resolution climate data are also relevant to sectors such as water resources, agriculture, health, and transport.

Near-surface temperature has a strong and nonlinear relationship with electricity demand. When the temperature rises above comfort thresholds, cooling demand grows rapidly, and when it falls below those thresholds, heating demand increases. Accurate modeling of the weather's temporal and spatial impact on building energy demand is critical for the planning and decarbonisation of energy systems [4]. Spatial resolution is important because coarse grid cells average heterogeneous conditions. For example, a cell that combines a warm urban center with cooler rural surroundings may underestimate both peak cooling loads in the city and the spatial extent of the exposure. Higher spatial resolution can improve the siting and operation of energy resources and may substantially reduce system costs compared with planning based on coarse-resolution inputs [5]. Similarly, the work in [6] shows that the benefits of increasing climate-data resolution are especially relevant for temperature-sensitive applications and topographically complex regions, such as the Iberian Peninsula.

Temporal resolution is equally important. A daily mean temperature field does not describe the timing, amplitude, or persistence of the diurnal cycle. However, these intra-day characteristics influence the shape and magnitude of the electricity demand curves. Therefore, models of heating and cooling demand that rely on daily aggregates systematically miss these intra-day dynamics [4, 7].

The cost of this mismatch is concrete. During the June-July 2025 European heatwave, electricity demand across Spain grew by up to 14% and peak prices nearly tripled in

Germany, with cooling demand persisting into the evening as solar generation faded. French nuclear reactors, which depend on river water for cooling, were forced to throttle or shut down output as river temperatures exceeded regulatory limits, decreasing capacity precisely when demand was highest. In Italy, heat caused widespread urban blackouts as ground cables overheated and lost their ability to dissipate the additional load from air conditioning.

This intra-day supply-demand mismatch cannot be characterized by daily temperature averages [8].

As power systems incorporate larger shares of variable renewable generation, their operation becomes increasingly sensitive to meteorological variability. Solar and wind production depend on weather conditions that vary across space and over the course of the day, while demand is also shaped by temperature, humidity, and human activity patterns [9]. High-resolution weather information is therefore needed not only to estimate demand, but also to assess renewable-resource variability, ramping events, and regional imbalances that may be smoothed out in coarse climate-model grids [10].

In April 2025, the Iberian Peninsula experienced the most severe power blackout in Europe in over two decades. The event was attributed to oscillatory instability, inadequate voltage control, and cascading generator disconnections. Although the cause was not meteorological, the blackout exposed the structural vulnerability of a grid in which solar photovoltaic generation, which is governed by sub-daily temperature and radiative dynamics, constituted the main supply source. Reliable planning for systems under future climate conditions requires probabilistic hourly climate scenarios [11].

Infrastructure resilience planning also needs fine-scale climate information. In January 2026, Storm Kristin struck central Portugal with peak gusts exceeding 170 km/h, destroying electricity poles and leaving approximately one million people without power. Recovery in the worst-affected areas took several days reflecting the scale of physical damage and the absence of resilience planning calibrated to the local intensity of extreme events. Designing infrastructure for future climates requires high-resolution hourly data and not only smoothed projections [12].

Beyond the energy sector, the need for high-resolution sub-daily temperature data extends across multiple domains. In public health, temperature related mortality depends on local exposure and on the temporal structure of heat stress, including persistent high

nighttime temperatures, while urban heat islands create fine-scale spatial temperature contrasts that are not represented by coarse climate fields [13], [14],[15]. In agriculture, crop heat stress during reproductive phases depends on the timing and duration of intra-day temperature exceedances, and reliance on daily aggregates has been identified as a key limitation of current models [16], [17]. In hydrology, sub-daily temperature drives snow-melt dynamics that daily means cannot resolve [18]. Across all these sectors, high-resolution hourly climate data are essential for reliable impact assessment.

1.1.2 The Downscaling Problem

Climate downscaling can be understood as a form of **super-resolution**, analogous to image upscaling in computer vision [19]. The objective is to transform a low-resolution data field, such as the GCM output, into a higher-resolution field that is physically plausible and statistically consistent with observed local-scale climate.

A downscaled field must satisfy two requirements. First, it must be statistically consistent with the large-scale GCM forcing. Second, it must reproduce the observed statistical properties of the regional climate, including spatial correlations, the distributions of extreme events, and the structure of the diurnal cycle, when sub-daily fields are the target.

A particularly challenging case arises when the target resolution is finer not only in space but also in time. Most downscaling studies focus on improving spatial resolution. Considerably less attention has been given to the recovery of sub-daily temporal structure from daily summary statistics. This temporal dimension is critical because a daily mean temperature field does not specify when the maximum and minimum occur, how rapidly temperature increases during the morning, or how the diurnal amplitude varies across the landscape. Many physically plausible hourly sequences can share the same daily mean. The daily-to-hourly downscaling problem is therefore underdetermined and inherently probabilistic.

Two main families of models are commonly used to address downscaling problems: **dynamical models** and **statistical models**.

Dynamical downscaling relies on physical modeling: A **Regional Climate Model (RCM)** [20] is nested within, or driven at its boundaries by, GCM output. While this approach ensures strong physical consistency by solving partial differential equations for mass, momentum,

and energy conservation, its high computational cost limits the generation of large simulation ensembles. Furthermore, RCMs contain their own systematic biases arising from the parametrization of unresolvable sub-grid processes.

Statistical downscaling provides a data-driven, computationally efficient alternative. Instead of explicitly simulating physical equations, it learns a statistical transfer function mapping large-scale predictors (X) to regional-scale observations:

$$Y_{\text{regional}} = f(X_{\text{GCM}}). \quad (1.1)$$

While suitable for producing large ensembles and exploring uncertainty, traditional statistical methods are fundamentally limited by the stationarity assumption. They assume historical relationships will remain valid under future climate change, which can cause misleading inference.

1.1.3 Motivation for Deep Generative Models

The limitations of existing downscaling methods motivate the use of flexible probabilistic models. Dynamical models are physically grounded but computationally expensive, especially when large ensembles are needed for uncertainty analysis. Classical statistical downscaling methods are faster but they may rely on restrictive assumptions, limited representations of variability, or historical analogues that cannot fully represent novel future conditions.

This motivates a transition within the statistical downscaling toward Artificial Intelligence (AI)-driven generative methods. Deep generative models provide a framework for learning complex, high-dimensional probability distributions directly from data. Unlike deterministic regression models, generative models can produce multiple plausible outputs for the same input condition, making them naturally suited to ill-posed problems such as daily-to-hourly temperature downscaling.

Classical models, whether **parametric** (e.g., regression and stochastic generators) or **non-parametric** (e.g., analog methods, k-NN), are not well suited for this task. Parametric models require strong *a priori* assumptions about the statistical structure of the data, which may not be valid. Non-parametric models, although more flexible, suffer intractable computational burdens when applied to thousands of variables and can reduce novel variability.

Deep Generative Models represent a new class of non-parametric estimators using deep neural networks to learn the underlying low-dimensional manifolds of the complex data. These models learn implicit physical and statistical correlations of the climate system directly from the data. This generative approach provides two main advantages. First, by sampling from the learned probability distribution, these models can generate a potentially large number of physically plausible synthetic spatiotemporal series (ensembles) consistent with the GCM forcing. Second, since the model learns the conditional probability distribution of the regional-scale field given the GCM state, the model can better handle non-stationarity than methods that assume a fixed parametric relationship. It should be noted, however, that DGMs are not immune to stationarity limitations. A model trained on historical data still assumes that the conditional relationship between large-scale forcing and regional climate is transferable to future states. Their advantage lies in greater flexibility and the absence of strong parametric assumptions.

Consequently, Deep Generative Models provide a promising statistical framework for producing ensembles of plausible high-resolution fields, while retaining computational advantages of data-driven methods. However, their reliability still depends on training data quality, conditioning design, and careful evaluation.

1.2. Research Question, Objectives and Contributions

This dissertation studies a probabilistic generative approach for producing high-resolution hourly near-surface temperature fields from coarse daily mean temperature inputs. The focus is on temperature over the Iberian Peninsula and the adjacent Atlantic Ocean. The core challenge is that a daily mean temperature field is an insufficient specification of the corresponding hourly evolution, as multiple physically plausible 24-hour sequences may share the same daily mean while differing in diurnal amplitude, timing of the temperature maximum and minimum, and local variability. This under-determination motivates a probabilistic formulation in which the model generates an ensemble of plausible realizations conditioned on the daily state, rather than a single deterministic prediction.

The proposed approach is based on a conditional diffusion model [21] trained on ERA5 hourly temperature data [22]. The model learns to generate anomaly fields (defined as the intra-day departures from the daily mean field) relative to the imposed daily condition, which are subsequently reconstructed into absolute hourly temperature fields. The conditioning strategy combines coarse daily mean temperature fields, static geographical

variables, temporal encodings, and scalar control variables to influence the generated ensemble's sub-daily variability characteristics.

Accordingly, this dissertation addresses the following **research question**:

To what extent can a conditional diffusion model generate physically plausible, probabilistically calibrated, high-resolution hourly temperature fields from coarse daily mean temperature conditions, while preserving spatial, temporal, distributional, and extreme-temperature characteristics?

To answer this research question, the following specific objectives are defined:

1. To construct a training framework that represents each day as a complete 24-hour spatiotemporal block, allowing the model to learn intra-day temporal structure jointly with spatial variability.
2. To design a conditioning strategy that combines coarse daily temperature information, static geographical variables, temporal encodings, and scalar control variables.
3. To implement a conditional diffusion model capable of generating multiple plausible hourly temperature scenarios for the same daily condition.
4. To apply the trained model to CMIP6 daily temperature projections, producing high-resolution hourly scenarios for future climate projections.
5. To evaluate the generated scenarios using deterministic, probabilistic, temporal, spatial, distributional, extreme-value, and physical diagnostics.

The main contribution of this work is a probabilistic downscaling framework for generating ensembles of hourly temperature fields from coarse daily inputs. The framework combines conditional diffusion modeling, anomaly-based generation, static and temporal conditioning, and scalar controls for sub-daily variability. In doing so, it aims to preserve consistency with the imposed daily condition while producing realistic sub-daily variability.

1.3. Structure

The rest of this dissertation is organized into five main chapters. They transition sequentially from theoretical foundations to the proposed methodology and empirical validation.

Chapter 2 defines the climate downscaling problem. It describes the primary data sources and reviews classical statistical downscaling methods that serve as baselines and necessary preprocessing steps.

Chapter 3 introduces generative modeling for probabilistic downscaling. It covers DGMs and their main families. The chapter outlines key architectures such as VAE and GAN. The primary focus is the theoretical and mathematical formulation of diffusion models. Finally, the chapter provides a literature review examining the applications of these specific generative architectures to climate downscaling tasks.

Chapter 4 presents the proposed methodology. It starts by formulating the probabilistic downscaling task. Then it specifies the data preprocessing, formulation and conditioning design, including the proposed scalar control variability variables. The chapter details the model architecture, training procedures, inference strategies, evaluation metrics, and experimental design.

Chapter 5 presents the experimental results and discussion. It first analyzes the differences between the ERA5 training and test periods. Next, it compares the proposed models against baselines and ablated versions. An analysis of ensemble depth is also evaluated. Then, the model is applied to future CMIP6 projections and the results are discussed.

Chapter 6 concludes the dissertation by summarizing the main findings, discussing the limitations of the proposed approach, and outlining directions for future work.

2. Climate Downscaling Background

The previous chapter established the motivation for probabilistic climate downscaling and introduced the spatiotemporal gap addressed in this work. This chapter introduces the climate-data sources and downscaling concepts used throughout this document.

Section 2.1 defines the downscaling problem and introduces the notation used throughout the thesis. Section 2.2 describes GCM and the CMIP6 framework. Section 2.3 describes the ERA5 reanalysis product used as training and reference data. Section 2.4 reviews classical statistical downscaling methods, including quantile-based bias correction and analog-based approaches, which are later used as preprocessing and baselines in the methodology.

2.1. The Climate Downscaling Problem

Climate downscaling is the task of recovering fine-scale information from a coarse-resolution representation of the climate state. Let $\mathbf{y}_d^{LR} \in \mathbb{R}^{H_{LR} \times W_{LR}}$ denote a low-resolution daily mean temperature field defined on a coarse spatial grid of height H_{LR} and width W_{LR} . Let $\mathbf{x}_d^{HR} \in \mathbb{R}^{T \times H_{HR} \times W_{HR}}$ denote the corresponding high-resolution hourly temperature field defined on a finer spatial grid of height H_{HR} and width W_{HR} , where $T = 24$ is the number of hourly time steps within a calendar day. Each individual hourly field is denoted by $\mathbf{x}_{d,h}^{HR} \in \mathbb{R}^{H_{HR} \times W_{HR}}$.

The operator that maps $\mathbf{y}_d^{LR}$ to $\mathbf{x}_d^{HR}$ through downscaling is a one-to-many projection, since many distinct hourly sequences may be consistent with the same daily mean. Recovering the hourly block from the daily mean is therefore an ill-posed inverse problem with infinitely many solutions. A deterministic formulation can provide a single representative estimate, but it cannot represent the range of plausible hourly evolutions compatible with the same daily condition.

The goal of this work is instead to learn the conditional probability distribution

$$p\left(\mathbf{x}_d^{HR} \mid \mathbf{y}_d^{LR}, \mathbf{c}_d\right), \quad (2.1)$$

where $\mathbf{c}_d$ denotes conditioning information for day d described in Section 4.4. At inference time, multiple independent samples from this distribution form an ensemble of physically plausible hourly realizations with the imposed daily condition.

2.2. Climate Model Data Sources

As introduced in Chapter 1, GCM provide the large-scale climate projections that serve as input to the downscaling framework. In this work, the conditioning field $\mathbf{y}_d^{LR}$ at inference time is provided by the daily near-surface temperature fields from the CMIP6 archive [3], using the CNRM-CM6-1-HR model and scenario, described next.

2.2.1 The CNRM-CM6-1-HR Model

The climate model used in this work is the CNRM-CM6-1-HR, developed by the *Centre National de Recherches Météorologiques* and *Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique* [23]. The suffix HR denotes increased spatial resolution relative to the standard CNRM-CM6-1 model; it should not be confused with the high-resolution ERA5 target grid used in this thesis.

The atmospheric component of CNRM-CM6-1-HR is resolved at approximately 0.5° , corresponding to roughly 50 km. This resolution is consistent with the coarse conditioning input $\mathbf{y}_d^{LR}$ considered in this work. The variable used is near-surface temperature, archived in CMIP6 as t_{as} (temperature of air at 2 meters above the surface), at daily mean resolution. The historical simulation covers the period from 1850 to 2014, while the future scenario extends from 2015 to 2100 under different Shared Socioeconomic Pathways (SSP) experiments.

2.2.2 Shared Socioeconomic Pathways

The SSP are a set of plausible trajectories of global socioeconomic development that, together with radiative forcing targets, define the forcing scenarios for CMIP6 future simulations [24]. An SSP scenario is denoted as

$$SSP_{n-m}, \quad (2.2)$$

where $n \in \{1, 2, 3, 4, 5\}$ identifies the socioeconomic narrative and m denotes the additional radiative forcing, in W m^{-2} , reached by the year 2100 relative to pre-industrial conditions.

For example, SSP1-2.6 corresponds to a low-emission world with radiative forcing limited to 2.6 W m^{-2} , while SSP5-8.5 represents a fossil-fuel-intensive trajectory reaching 8.5 W m^{-2} , implying approximately 4°C of global warming by 2100. The choice of the

SSP scenario determines the large-scale temperature trend imposed on the conditioning input $\mathbf{y}_d^{LR}$ at inference time and consequently influences the distributional shift of the generated $\mathbf{x}_d^{HR}$ fields relative to the historical training period.

2.3. ERA5 Reanalysis

Reanalysis datasets combine numerical weather prediction models with observational data by integrating diverse observational streams through data assimilation techniques. The result is a spatially complete and temporally continuous reconstruction of the atmospheric state over a historical period.

ERA5 is the fifth-generation global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) [22]. It relies on data assimilation to blend observations with physical model predictions over 12-hour assimilation windows, using the ECMWF Integrated Forecasting System. ERA5 provides global coverage at a resolution of $0.25^\circ \times 0.25^\circ$, which is approximately 25 km over the Iberian Peninsula, and hourly temporal resolution from 1940 to the present.

The main variable used in this thesis is the 2-meter air temperature (t2m), defined as the air temperature at approximately 2 meters above the surface. In addition, other variables such as geopotential and the land-sea mask are used as static conditioning variables.

ERA5 serves two distinct roles in this work. First, it provides the high-resolution hourly target $\mathbf{x}_d^{HR}$ from which the model learns the statistical structure of the diurnal cycle and its spatial organization over the domain. Second, it is used to construct the synthetic low-resolution training condition $\mathbf{y}_d^{LR}$. This construction ensures that during training, the coarse input and high-resolution target are derived from the same reference data. At inference time, the conditioning input switches from ERA5 derived daily means to bias-corrected CMIP6 daily fields.

2.4. Classical Statistical Downscaling

Statistical downscaling is a family of methods that learn an empirical mapping from large-scale climate predictors to regional-scale predictands using observed historical data. Formally, a transfer function f is estimated such that

$$\hat{\mathbf{x}}_d^{HR} \approx f(\mathbf{y}_d^{LR}). \quad (2.3)$$

where additional large-scale or static predictors may also be included, depending on the method.

Once calibrated over a historical period for which both predictors and predictands are available, the transfer function can be applied to GCM output to generate regional projections. The following subsections describe the statistical approaches most directly relevant to this thesis.

2.4.1 Bias Correction and Quantile Delta Mapping

GCM temperature fields often exhibit systematic biases relative to observational references, in this case ERA5. Their marginal distributions may be shifted in mean, skewed, or incorrectly dispersed relative to ERA5. If left uncorrected, these biases propagate into the conditioning field $\mathbf{y}_d^{LR}$ and may shift the generated $\mathbf{x}_d^{HR}$ fields away from the distribution on which the diffusion model was trained.

The simplest bias correction approach is Quantile Mapping (QM), which transforms the GCM distribution to match the observed distribution quantile by quantile. While effective at correcting historical bias, QM is limited in climate change applications as it implicitly assumes that the bias is stationary, applying the historical-derived correction unchanged to future projections. Applying a historically derived correction unchanged to future projections may suppress or distort the warming signal contained in future GCM outputs.

QDM addresses this limitation by decomposing the correction into two components: a distributional alignment with the historical observational reference, and an explicit preservation of the quantile-specific change signal between the historical and future GCM simulations. The result is a bias-corrected field that is anchored to the ERA5 historical distribution while retaining the model-projected climate change signal. The full formulation and implementation used in this thesis are given in Section 4.9.

2.4.2 Analog Methods and k-Nearest Neighbors

Analog-based downscaling [25] maps a coarse-resolution predictor to a fine-resolution predictand by searching the historical data for days whose coarse daily field most closely resembles that of the target. The corresponding observed fine-resolution fields are then used as the regional estimates. The k-NN formulation [26] formalises this search by identifying the k historical days most similar to the query field $\mathbf{y}_d^{LR}$ and sampling from the corresponding observed high-resolution blocks.

The method is non-parametric and inherits the statistical complexity of the historical reference without distributional assumptions. However, it is fundamentally constrained to reuse or recombine patterns already present in the historical archive and cannot synthesize novel structures under strong climate change conditions. Despite this limitation, it provides a strong and interpretable baseline as it draws directly from historical ERA5 samples, guaranteeing physical realism. The full formulation of the k-NN baseline method used in this thesis is given in Section 4.10.3.

2.4.3 Other Classical Approaches

Other classical statistical approaches not used in this work include [27]:

- **Regression-based methods** learn a parametric linear or nonlinear mapping from large-scale predictors to regional predictands. They are computationally efficient, but are limited by their parametric assumptions, weak representation of extremes, and distorted spatial variability.
- **Stochastic weather generators** model local climate as a stochastic process whose parameters are conditioned on large-scale predictors, producing ensembles of synthetic time-series with realistic marginal statistics. Their parametric structure can make them difficult to extend to high-dimensional spatiotemporal fields. An example related to this family is the Direct Sampling approach of [28]. Although it is not a downscaling method, it is relevant as a classical stochastic weather field generator with capability of generating extreme fields.
- **Copula-based methods** model the joint dependence structure between variables or locations separately from their marginal distributions. They have been applied to multivariate bias correction and spatially coherent field generations, but their scalability to the high-dimensional setting in this work remains limited.

Classical statistical downscaling methods provide efficient and interpretable tools for linking coarse climate information to regional-scale variables. However, their ability to represent high-dimensional spatiotemporal variability, probabilistic uncertainty, and novel future conditions is limited by assumptions about stationarity, distributional form, and the availability of historical analogs. These limitations motivate the use of more flexible probabilistic models. The next chapter introduces deep generative models, with particular

emphasis on conditional diffusion models, which serve as the generative framework for this thesis.

3. Generative Modeling Background

The previous chapter introduced the climate downscaling problem and motivated the need for probabilistic methods capable of generating multiple plausible high-resolution hourly temperature fields from the same coarse daily condition. This chapter introduces the background on generative models needed to formulate such a method.

3.1. Generative Modeling for Probabilistic Downscaling

The fundamental distinction between machine learning approaches lies in the probabilistic relationship they attempt to model. To clarify this distinction let $\mathbf{x}$ be an observation of the data (a realization of the random variable X) belonging to a high-dimensional space $\mathcal{X}$. Let y be the corresponding target value or label (a realization of the random variable Y) from the space $\mathcal{Y}$.

Discriminative models focus on modeling the conditional probability distribution $p(y|\mathbf{x})$. The main goal is to learn the decision boundary that best separates distinct classes of the data or predicts continuous values. These models do not attempt to understand the distribution of the data $p(\mathbf{x})$, rather they focus only on the relationship between the input features and the target labels. Mathematically, a discriminative model seeks to maximize the likelihood of the targets given the data. This approach is standard for tasks such as image classification, semantic segmentation and regression.

In contrast, **Generative Models** aim to model the underlying structure of the data itself. The goal is to approximate the true data distribution $p(\mathbf{x})$ from a finite set of observed samples. By learning the probability density function of the data, generative models acquire the capability to sample new instances $\mathbf{x}_{\text{new}} \sim p(\mathbf{x})$ that are statistically similar to the training data but not identical to it. According to Bayes theorem, a generative model can also be used for classification (and predictive tasks) by inverting the generative process to compute the posterior probability

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}. \quad (3.1)$$

However, despite this capability, the primary utility of generative models in modern settings lies in their ability to synthesize high-dimensional data (as images, audio or climate data) by sampling directly from the learned manifold.

3.2. Deep Generative Models

Before the widespread use of deep neural networks, generative probabilistic modeling relied mainly on classical statistical approaches such as Naive Bayes, Gaussian Mixture Models, and Hidden Markov Models [29–31]. While these models were effective for lower-dimensional problems, they struggled to model complex, high-dimensional dependencies without excessive manual feature engineering [32, 33].

Today, the field is dominated by DGMs, which leverage Deep Neural Networks to model and approximate complex probability distributions or sampling processes. This capability has made them the standard methodology for generating complex modalities. Currently DGMs represent the state-of-the-art in generative AI.

3.2.1 Mathematical Formulation

Formally, let $\mathbf{x} \in \mathcal{X}$ be an observation from an unknown distribution $p_{\text{data}}(\mathbf{x})$. The goal of DGMs is to learn a parametric approximation, $p_{\theta}(\mathbf{x})$, that is close to $p_{\text{data}}(\mathbf{x})$. This can be expressed as the minimization of a divergence between the true and learned distributions:

$$\theta^* = \arg \min_{\theta} D(p_{\text{data}} \parallel p_{\theta}), \quad (3.2)$$

where θ denotes the trainable model parameters and $\mathcal{D}$ is a divergence measure.

A commonly used divergence metric is the **Kullback-Leibler (KL) Divergence**:

$$\mathcal{D}_{\text{KL}}(p_{\text{data}} \parallel p_{\theta}) := \int p_{\text{data}}(\mathbf{x}) \log \frac{p_{\text{data}}(\mathbf{x})}{p_{\theta}(\mathbf{x})} d\mathbf{x} = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} [\log p_{\text{data}}(\mathbf{x}) - \log p_{\theta}(\mathbf{x})]. \quad (3.3)$$

Minimizing this divergence is closely related to maximizing the likelihood of the observed data under the model. However, direct maximization of likelihood is often intractable for high-dimensional data, which has motivated different classes of deep generative models.

3.2.2 Families of Deep Generative Models

DGMs are often categorized based on how they represent or approximate the data distribution. A common distinction is between *implicit density models* and *explicit density models* [34]. Figure 3.1 summarizes this taxonomy.

Implicit density models do not define an explicit probability density function $p_{\theta}(\mathbf{x})$. Instead, they learn a sampling procedure. The sampling process typically takes a form $\mathbf{x}_{\text{new}} = \mathbf{G}_{\phi}(\mathbf{z})$, where $\mathbf{z} \sim p_{\text{prior}}$ is some noise variable. These models usually generate

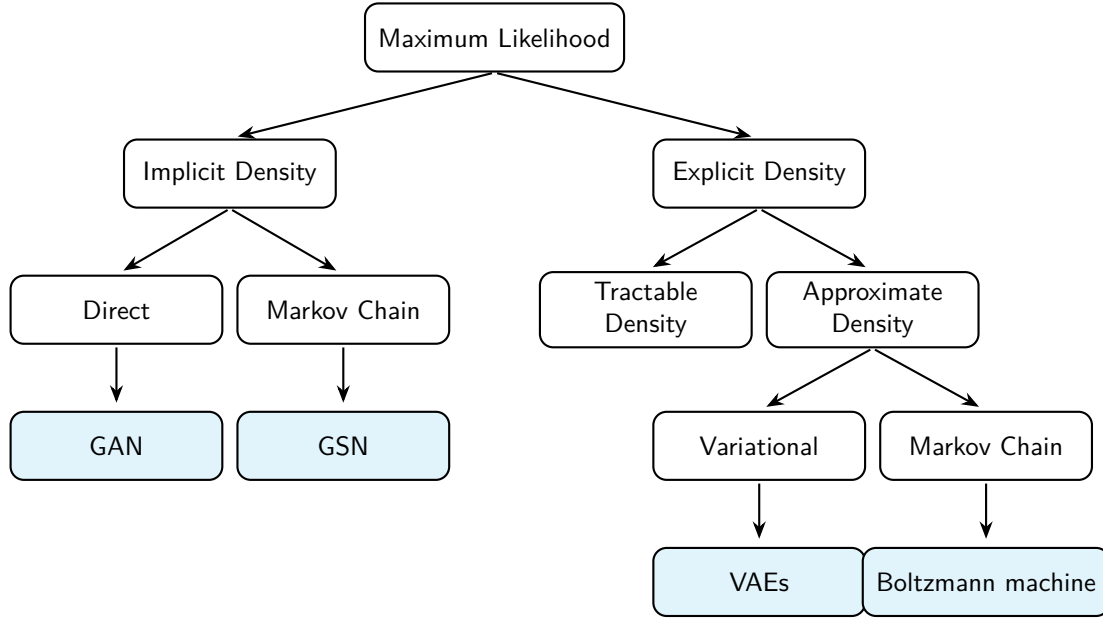


Figure 3.1: Taxonomy of DGMs based on how they represent and approximate the probability distribution, based on [34].

high-quality data but are usually unstable in the training phase. The most well-known models in this family are the GANs.

GANs, introduced in [35], formulate the generation as a two-player game between two neural networks: a **Generator** G and a **Discriminator** D . The generator tries to produce new samples that are as close as possible to real data, while the Discriminator attempts to distinguish real samples from generated ones. The training process is written as a minimax game:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log (1 - D(G(z)))] . \quad (3.4)$$

GANs have produced high-quality samples in many domains, especially computer vision. However, they are notoriously difficult to train, suffering from instability and sometimes producing a limited variety of samples, ignoring the diversity of the data (also known as mode collapse) [36]. These limitations motivate the use of alternative generative frameworks when stable training and probabilistic diversity are important.

Explicit density models define a probability distribution model $p_{\theta}(\mathbf{x})$ and are trained by maximizing the likelihood of the data or a tractable lower bound. This approach is generally more stable than the implicit one and allows us to quantitatively evaluate the model via log-likelihood scores.

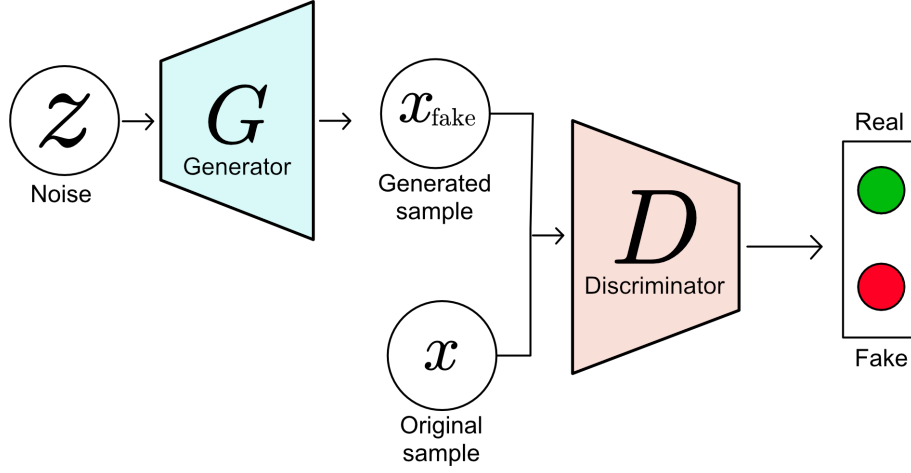


Figure 3.2: The Generator receives noise z and generates a fake sample x_{fake} . The discriminator receives either synthetic samples or real samples (x) and classifies them as real or fake.

This family includes autoregressive models, normalizing flows, variational autoencoders, and diffusion models.

Autoregressive models decompose the joint density into a product of conditional distributions:

$$p_{\theta}(\mathbf{x}) = \prod_{i=1}^n p_{\theta}(x_i \mid x_1, \dots, x_{i-1}). \quad (3.5)$$

This decomposition allows exact likelihood computation, but sampling can be slow because the variables are generated sequentially.

Variational methods introduce latent variables to represent hidden structure in the data. A central example is the VAE, introduced by Kingma and Welling [37]. A VAE consists of an encoder $q_{\phi}(\mathbf{z} \mid \mathbf{x})$, which approximates the posterior distribution over latent variables, and a decoder $p_{\theta}(\mathbf{x} \mid \mathbf{z})$, which generates data from latent samples.

The marginal likelihood involves the integral

$$p_{\theta}(\mathbf{x}) = \int p_{\theta}(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z}) d\mathbf{z}, \quad (3.6)$$

which is generally intractable. VAEs therefore maximize the evidence lower bound:

$$\log p_{\theta}(\mathbf{x}) \geq \mathcal{L}_{\text{ELBO}}(\mathbf{x}; \theta, \phi), \quad (3.7)$$

where

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_{\phi}(\mathbf{z} \mid \mathbf{x})} [\log p_{\theta}(\mathbf{x} \mid \mathbf{z})] - \mathcal{D}_{\text{KL}}(q_{\phi}(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z})). \quad (3.8)$$

The first term encourages accurate reconstruction of the input, while the second regularizes the approximate posterior toward a prior distribution, usually $\mathcal{N}(0, \mathbf{I})$. During generation, latent variables are sampled from the prior and passed through the decoder. Figure 3.3 illustrates the VAE model.

Diffusion models can be interpreted as a particular class of latent-variable generative models in which the latent variables correspond to progressively noisier versions of the data. They are especially relevant to this thesis because they provide a stable training objective and naturally support conditional sampling.

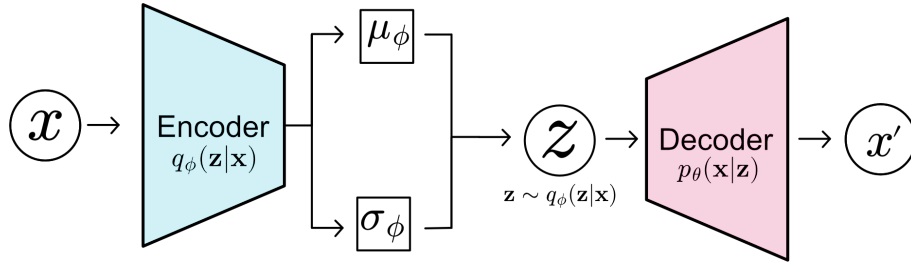


Figure 3.3: The encoder network $q_{\phi}(\mathbf{z}|\mathbf{x})$ processes the input data $\mathbf{x}$ and predicts the parameters of the latent distribution. A latent vector $\mathbf{z}$ is sampled from this distribution and passed to the Decoder network $p_{\theta}(\mathbf{x}|\mathbf{z})$, which attempts to reconstruct the original input as $\mathbf{x}'$.

3.3. Diffusion Models

Diffusion models, formally known as diffusion probabilistic models, represent a paradigm shift in DGMs. These models are grounded in non-equilibrium statistical physics. The fundamental hypothesis of these models is that the process of creating data can be modeled as the reversal of a physical diffusion process. If we can define a systematic way to slowly destroy the structure of a data distribution until only random noise remains, we can train a neural network to reverse this process.

Modern diffusion models can be formulated in three related ways [38]:

- **DDPMs**: define diffusion as a discrete-time Markov Chain with a fixed noising process and a learned reverse denoising [21, 39]. It can also be interpreted as a Hierarchical VAE with fixed encoders. It is the most common formulation and the primary focus of this work.
- **Score-Based Generative Models (SGMs)**: learn the score function $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, which points towards higher likelihood in data space [40].

- **Stochastic Differential Equations (SDEs)**: formulate diffusion in continuous time and unify discrete diffusion and score-based modeling under stochastic differential equation [41].

The methodology of this thesis is based on the DDPM formulation, so the following sub-sections focus on the discrete-time case.

The theoretical framework of diffusion models rests on two distinct Markov Processes: a **Forward pass** and a **Reverse Process**.

3.3.1 Forward Diffusion Process

Let $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ denote a clean data sample. In a DDPM, the forward diffusion process gradually corrupts $\mathbf{x}_0$ into a sequence of noisy variables $\mathbf{x}_1, \dots, \mathbf{x}_T$. The process is defined as a Markov chain:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}). \quad (3.9)$$

Here, $\{\beta_t \in (0, 1)\}_{t=1}^T$ is a variance schedule that controls the step size of the noise injection. The term $\sqrt{1 - \beta_t}$ scales the mean to prevent the variance from exploding as $t \rightarrow \infty$.

Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. We can represent the distribution $q(\mathbf{x}_t | \mathbf{x}_0)$ as:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}). \quad (3.10)$$

This allows us to express $\mathbf{x}_t$ as a linear combination of the original signal and pure noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon. \quad (3.11)$$

This equation allows training on arbitrary time steps in parallel.

3.3.2 Reverse Generative Process

The generative process aims to reverse the forward diffusion chain. Since the true reverse distribution $q(\mathbf{x}_{t-1} | \mathbf{x}_t)$ is generally intractable, it is approximated by learnable Markov Chain with learned Gaussian transitions:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \quad (3.12)$$

In the standard DDPM formulation [21], the model is simplified by fixing the covariance to a constant. The neural network is then tasked solely with predicting the mean $\mu_\theta(\mathbf{x}_t, t)$. However, instead of predicting the mean directly, it is empirically found that predicting the noise component ϵ yields better results. The mean is parametrized as:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right), \quad (3.13)$$

where ϵ_θ is a function approximator that predicts the noise present in $\mathbf{x}_t$. This is sometimes called **ϵ -Prediction**.

3.3.3 Training Objective

Similarly to VAEs the training objective is derived from the Variational Lower Bound on the negative log-likelihood. The loss function can be simplified to a weighted Mean Squared Error (MSE) between the true noise ϵ and the predicted noise ϵ_θ :

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} \left[\left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right\|^2 \right]. \quad (3.14)$$

This result implies that training a diffusion model is equivalent to training a Denoising Autoencoder at various noise levels.

3.3.4 Sampling Algorithms

Once the model is trained to approximate the noise $\epsilon_\theta(\mathbf{x}_t, t)$, generating new data requires reversing the diffusion process. This inference phase, also called sampling, iteratively denoises a random Gaussian vector $\mathbf{x}_T$ until it reaches the data space $\mathbf{x}_0$. The choice of sampling algorithm determines both inference speed and the generation's deterministic nature.

3.3.4.1 Traditional (DDPM) Sampling

The standard method proposed in [21] directly follows the reverse Markov Chain formulation. Starting from pure noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$, we sample previous steps $\mathbf{x}_{t-1}$ using the learned mean and fixed variance:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}, \quad (3.15)$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ is random Gaussian noise added at every step (except the final one $t = 0$), and σ_t is the variance schedule term, which is typically $\sigma_t^2 = \beta_t$.

The injection of noise $\mathbf{z}$ at each step is necessary since we are sampling and not simply predicting. The model predicts the mean of a distribution, however a data point can lie anywhere in the curve. Without adding noise to simulate the variance, the process would collapse to a deterministic trajectory, producing the same output at every run. The presence of this noise injection allows the model to explore the manifold effectively generating fine-grained high-frequency details that are statistically consistent with the training data. For this reason running the sampling loop twice results in slightly different results. While theoretically rigorous, this method is computationally expensive, requiring the sequential evaluation of many steps.

3.3.4.2 Denoising Diffusion Implicit Models

A major limitation of traditional sampling is its reliance on a stochastic Markov chain, which requires many sequential steps. Denoising Diffusion Implicit Model (DDIM) [42] address this by generalizing the forward process into a class of non-Markovian processes that share the same training objective. This allows a trained model on standard DDPM loss to sample using DDIM without retraining.

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(\mathbf{x}_t, t) + \sigma_t \mathbf{z}. \quad (3.16)$$

The first term reconstructs the predicted clean sample $\hat{\mathbf{x}}_0$, while the second term points in the direction of the noise. The parameter σ_t controls stochasticity. When $\sigma_t = 0$, the process becomes deterministic, removing the random noise injection entirely.

This determinism makes the consecutive steps no longer being constrained by the Markov assumption, allowing steps along the trajectory to be skipped. Sampling can therefore be performed on a sub-sequence of timesteps reducing the number of network evaluations from hundreds to as few as 20-50 while retaining competitive sampling quality.

3.3.4.3 DPM-Solver++

While DDIM accelerates sampling, its quality can still degrade at very low step counts. DPM-Solver++ [43] takes a different view: rather than a discrete Markov chain, it treats the reverse process as solving the probability flow ordinary differential equation from the continuous formulation. The key idea is that this equation has a semi-linear structure, a linear term in $\mathbf{x}_t$ and a non-linear term involving ϵ_θ . The linear part can be solved analytically, so only the non-linear part requires numerical approximation. DPM-Solver++

exploits this and uses higher-order solvers: its second-order variant accounts for the curvature of the trajectory, where a first order method like DDIM assumes it is locally straight. It is also tailored for guided sampling operating on the predicted data rather than the noise for improved stability. This sampling method yields high-quality samples in roughly 10-20 steps, with no retraining required, making it useful for large conditional models.

3.4. Conditional Diffusion Models

The diffusion models described so far are **unconditional**. They learn the marginal distribution of the data $p(\mathbf{x})$ and generate high-fidelity samples without external constraints. For the downscaling task in this thesis, unconditional generation is not sufficient. The model must generate hourly high-resolution fields that are consistent with a prescribed coarse daily temperature field and additional conditioning variables.

Conditional diffusion models address this by learning $p_\theta(\mathbf{x} | \mathbf{c})$, where $\mathbf{c}$ is the conditioning information. In this work, $\mathbf{c}$ includes the coarse daily temperature condition, static geographical information, temporal encodings, and scalar control variables.

3.4.1 Conditional Reverse Process

To achieve this, the reverse process is modified to depend on the condition $\mathbf{c}$. The neural network, which originally estimated the noise based only on the noise input $\mathbf{x}_t$, and timestep t , now takes the condition as an additional input:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t, \mathbf{c}), \sigma_t^2 \mathbf{I}). \quad (3.17)$$

There are two primary paradigms for enforcing this guidance during sampling: **Classifier Guidance** and **Classifier-Free Guidance**.

3.4.1.1 Classifier Guidance

This method, introduced in [44], leverages an external auxiliary model to guide generation. It requires training a separate noise-robust classifier $f_\phi(\mathbf{x}_t)$ that predicts the condition $\mathbf{c}$ from a noisy input.

During sampling, the gradient of this auxiliary model, $\nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{c} | \mathbf{x}_t)$, is added to the diffusion model's mean prediction. This computes a direction in the data space that maximizes the likelihood of the condition and shifts the diffusion step in that direction. This

approach is theoretically appealing, but is computationally expensive as it requires running two different neural networks at every sampling step.

3.4.1.2 Classifier-Free Guidance

Classifier-free guidance avoids the need for an external classifier [45]. Instead, a single diffusion model is trained to handle both conditional and unconditional generation. During the training, the conditioning input signal $\mathbf{c}$ is randomly dropped and replaced with a null value with a fixed probability. This forces the network to learn both the conditional distribution $p(\mathbf{x}|\mathbf{c})$ and the unconditional marginal $p(\mathbf{x})$.

During sampling, the final noise prediction is computed as a linear extrapolation:

$$\tilde{\epsilon}_\theta = \epsilon_\theta(\mathbf{x}_t, t, \emptyset) + w \cdot (\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon_\theta(\mathbf{x}_t, t, \emptyset)), \quad (3.18)$$

where w is the **guidance scale**. Larger values of w force the generated sample to align more strictly with the conditioning information, which often reduces sample diversity.

3.5. Related Work

This section reviews how generative models have been applied to climate downscaling. Deep learning provided an influential early example of reframing statistical downscaling as image super-resolution [19], where a stacked super-resolution Convolutional Neural Network downscales precipitation using topography as auxiliary input. It outperforms interpolation and classical baselines but produces a single conditional-mean estimate, smoothing extremes and ignoring the one-to-many nature of the problem. This limitation motivated the development of deep generative models, which sample ensembles of plausible, high-resolution fields.

3.5.1 GANs for Climate Downscaling

GANs were among the first generative frameworks applied to downscaling, as their conditional form allows ensemble generation. Leinonen et al. [46] introduced a recurrent stochastic super-resolution GAN that generates an ensemble of time-evolving high-resolution fields from low-resolution sequences, validated with rank statistics. Harris et al. [47] mapped coarse precipitation forecasts to radar-resolution rainfall, with explicit ensemble calibration.

The method in [48] employs a GAN to simulate synthetic precipitation scenarios over India. This generation of spatiotemporal climate data was conditioned on the "extremeness" of the data, allowing the generation of diverse cases with controllable extremeness. They achieved this by discretizing the distribution of rainfall per precipitation field into n bins and using bin membership to condition the GAN. To address the high imbalance, weighted sampling was used so the model sees equal numbers of samples from each bin. Although statistical fidelity was achieved, there were some grid artifacts and extreme spatial smoothing in the generated samples. It is important to mention that the focus of this paper is not on downscaling but on the conditional generation of synthetic data.

3.5.2 Diffusion Models

Similarly to this work, DiffESM [49] proposes a conditional diffusion model to stochastically downscale and emulate climate data. Their work focuses on generating daily fields of temperature and precipitation from lower-resolution monthly mean fields, emulating Earth System Models. The presented denoising model is a 3D U-Net, similar to the ones used in video diffusion, where the convolutions are separated in spatial-only and temporal-only blocks. The generated samples under different monthly scenarios were comparable to the real models in terms of spatial and temporal correlations and behaviors. However, there was still some difficulty in the generation of extreme scenarios, and due to the model architecture the generated months are independent of one another which causes incoherence at month boundaries.

Drawing parallels to super-resolution tasks in computer vision, [50] extends video diffusion models to downscale local precipitation fields. The model uses a deterministic U-Net downscaler to produce an initial prediction of the higher-resolution field based on the low-resolution. Stochasticity and finer details are modeled using a conditional diffusion model that generates residuals. The denoising model uses a U-Net architecture with spatio-temporal attention. The model captures the statistical distribution of extreme precipitation events and spatial patterns in complex terrains. However, the authors noted a significant computational bottleneck for training and observed a perception-distortion trade-off, where optimizing perceptual realism degrades MSE of the samples.

The model presented in [51] also employs a two-step approach to estimate various climate variables at a 2 km scale using 25 km reanalysis data. The first step of this model predicts the conditional mean using a U-Net regression, while the second step uses an Elucidated

Diffusion Model (EDM) to generate residuals that are added stochastically to create the final fine-scale details of the prediction. This decomposition reduces the target variance that the diffusion model must resolve. The tests demonstrated that the model was able to recover the power-law spectra and probability distribution of wind-speed and temperature. However, the generated ensembles were significantly under-dispersed, and the variability was consistently smaller than the RMSE.

Closer to the conditioning strategy of this dissertation, [52] use a conditional diffusion model to spatial downscale coarse temperature and wind fields to ERA5 resolution. The model is conditioned on static fields (land-sea mask and orography) and scalar embeddings for month and time of day. A spectral analysis shows that it outperforms a deterministic U-Net at the finest spatial resolution. Although similar to our work, this study focuses only on spatial downscaling, leaving temporal downscaling for future work.

3.5.3 Spatiotemporal and Temporal Downscaling

The works above mainly improve spatial or spatiotemporal resolution. A smaller body of work targets the temporal dimension, which is also the focus of this dissertation.

Other works downscale in time only. Classical analog methods, such as [53], transfer the hourly profile of the most similar historical day onto a daily mean while conserving it, but reuse observed profiles and cannot synthesize novel sub-daily structure. Scher and Peßenteiner [54] train a conditional GAN on hourly radar data to disaggregate daily precipitation sums: given a daily-sum field, the model samples a 24-hour sequence of hourly fractions, with a softmax over the time dimension conserving the daily sum. The patterns are realistic, but the diurnal cycle is underestimated. Close to this dissertation in target variable, Balogun et al. [55] generate near-surface temperature over a 24-h block with a conditional GAN, framed as video generation and conditioned on region, month and period. The samples reproduce the observed diurnal cycle, but the model generates temperature directly without coarse daily conditioning, so it is not strictly a downscaling method.

Schmidt et al. [56] presented a generative framework for joint spatiotemporal downscaling. They use a score-based U-Net diffusion model trained on high-resolution reanalysis data as a physical prior that captures weather dynamics. This U-Net utilizes an encoder-decoder structure with a bottleneck layer and self-attention. Instead of integrating the conditioning into the training, this approach uses Bayesian Inference where a specific

observation model constrains the prior on CMIP6 data. To introduce continuity across days, the authors adapted [57], which processes time windows within a Markovian framework enhanced by a massively parallelized convolution operator. Empirical evaluations demonstrated that the model accurately recovers local variability and extreme events, while maintaining the statistical properties and distributions of the coarse climate input.

4. Methodology

This chapter presents the methodology used in this dissertation. It begins by formulating the daily-to-hourly downscaling problem and describing the construction of the ERA5-based training samples. It then defines the anomaly target, the conditioning variables, and the conditioning tensor used by the model. The chapter proceeds with the conditional diffusion formulation, the denoising U-Net architecture, the training procedure, implementation details, and the inference strategy used to generate ensembles. Finally, it describes the application to CMIP6 projections, the baseline models, the evaluation metrics, and the experimental design. Figure 4.1 provides a compact visual summary of the central model workflow, from preprocessing and target construction to diffusion training and inference.

4.1. Problem Formulation

This work addresses the generation of high-resolution hourly near-surface temperature fields from coarse-resolution daily mean temperature fields. For each day d , the available input is a coarse-resolution daily mean temperature field,

$$\mathbf{y}_d^{LR} \in \mathbb{R}^{H_{LR} \times W_{LR}}, \quad (4.1)$$

where H_{LR} and W_{LR} denote the spatial dimension of the low-resolution grid. The desired output is a high-resolution sequence of 24 h temperature fields,

$$\mathbf{x}_d^{HR} = (x_{d,1}^{HR}, x_{d,2}^{HR}, \dots, x_{d,24}^{HR}), \quad \mathbf{x}_d^{HR} \in \mathbb{R}^{24 \times H_{HR} \times W_{HR}}. \quad (4.2)$$

Since a daily mean field does not uniquely determine the hourly evolution, the task is formulated as a conditional generative problem:

$$p(\mathbf{x}_d^{HR} | \mathbf{y}_d^{LR}, \mathbf{c}_d), \quad (4.3)$$

where $\mathbf{c}_d$ denotes auxiliary conditioning information, including static geographical variables, temporal encodings, and scalar control variables. These conditioning variables are described in detail in the following sections.

Sampling from this conditional distribution produces an ensemble of plausible hourly realizations,

$$\left\{ \hat{\mathbf{x}}_d^{HR,(1)}, \hat{\mathbf{x}}_d^{HR,(2)}, \dots, \hat{\mathbf{x}}_d^{HR,(K)} \right\}, \quad (4.4)$$

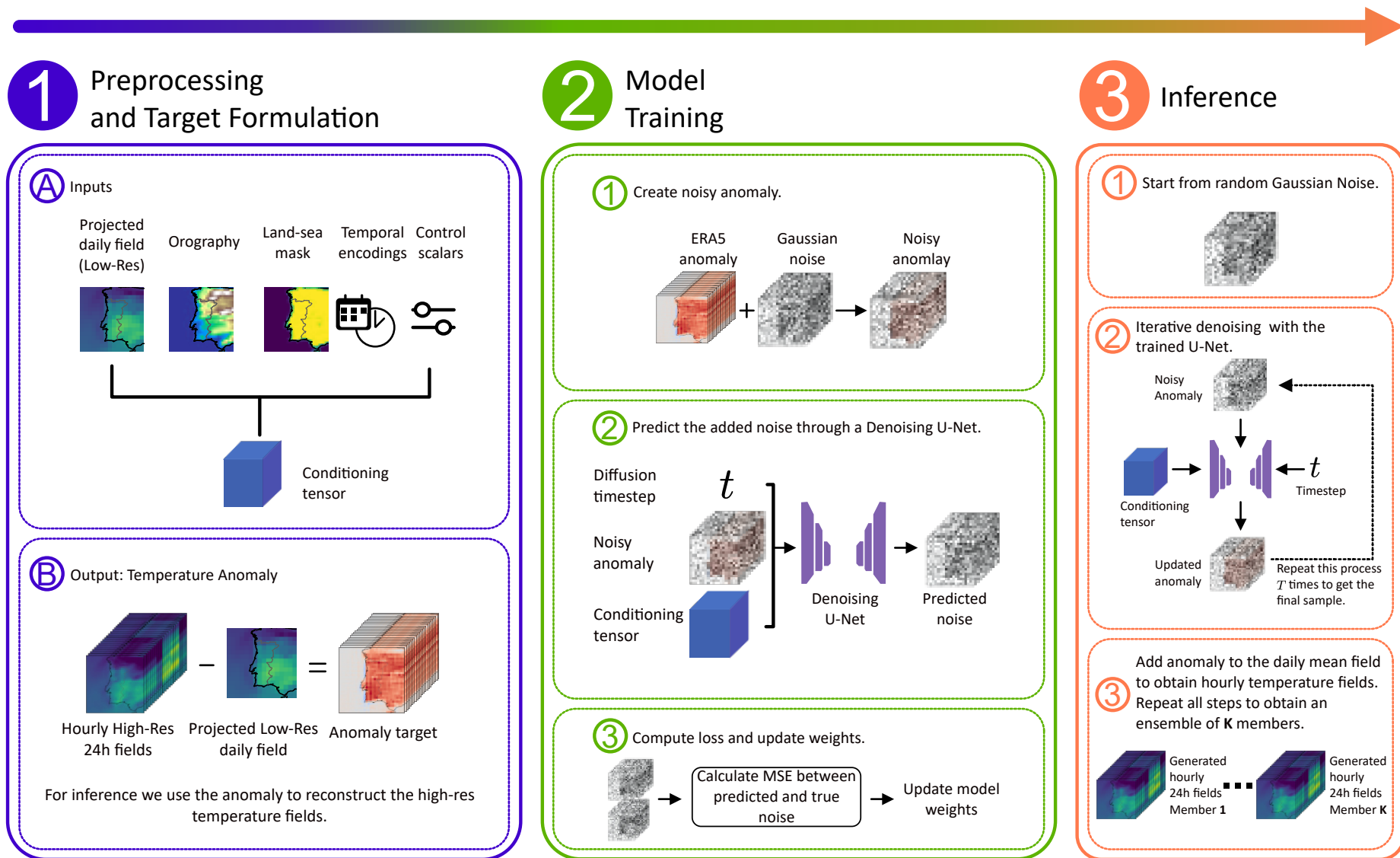


Figure 4.1: Overview of the proposed methodology. The workflow is divided into three stages: preprocessing and target formulation, model training, and inference. In the preprocessing stage, the projected low-resolution daily field, orography, land-sea mask, temporal encodings, and scalar controls are assembled into the conditioning tensor, while the anomaly target is obtained by subtracting the projected daily field from the high-resolution hourly temperature block. During training, Gaussian noise is added to ERA5 anomaly fields, and a denoising U-Net is trained to predict the added noise using an MSE loss. During inference, the trained model starts from random Gaussian noise and iteratively denoises it under the selected conditioning information. The generated anomaly is then used to reconstruct high-resolution hourly temperature fields, and the process is repeated to obtain an ensemble of K members.

where all members share the same daily condition but differ due to stochastic sampling.

The output for each input day is represented as a complete block of 24 hourly high-resolution temperature fields. This allows the model to learn the intraday evolution jointly with the spatial structure of the temperature field. It also matches the structure of the problem: the input is a daily mean field, while the desired output is the corresponding sub-daily evolution.

Daily blocks are generated independently, so continuity between consecutive days is not explicitly enforced and is treated as a limitation of the present approach.

4.2. Data and Preprocessing

The data preparation is designed to reproduce, during training, the type of information available when applying the model to climate model projections. The target application uses coarse-resolution daily mean temperature fields, whereas ERA5 provides hourly fields at a higher resolution. Therefore, the low-resolution daily condition used during training is constructed from ERA5 itself by temporal averaging and spatial coarsening. This produces paired samples in which the input resembles a coarse daily climate-model field, while the target remains the corresponding high-resolution hourly ERA5 sequence.

4.2.1 ERA5 Hourly Reference Data

The reference variable used in this work is ERA5 2-meter air temperature (t_{2m}). ERA5 is used as the reference dataset because it provides spatially complete hourly fields over the historical period, making it suitable for constructing the high-resolution daily temperature sequences defined in Section 4.1.

For each year, the raw ERA5 data are cropped to a fixed domain covering the Iberian Peninsula and the adjacent Atlantic Ocean, bounded by

- **Latitude:** 35.75°N to 43.50°N;
- **Longitude:** 12°W to 4.25°W.

At the native ERA5 spacing of 0.25°, this domain contains 32×32 grid points. Each hourly high-resolution field is therefore represented as

$$\mathbf{x}_{d,h}^{HR} \in \mathbb{R}^{32 \times 32}, \quad (4.5)$$

and each daily target block as

$$\mathbf{x}_d^{HR} \in \mathbb{R}^{24 \times 32 \times 32}. \quad (4.6)$$

The corresponding low-resolution daily condition is defined on a 16×16 grid, equivalent to an effective spacing of 0.5° :

$$\mathbf{y}_d^{LR} \in \mathbb{R}^{16 \times 16}. \quad (4.7)$$

This paired $16 \times 16 \rightarrow 32 \times 32$ setting defines the controlled downscaling task used for training and evaluation. Figure 4.2 shows the selected study region.

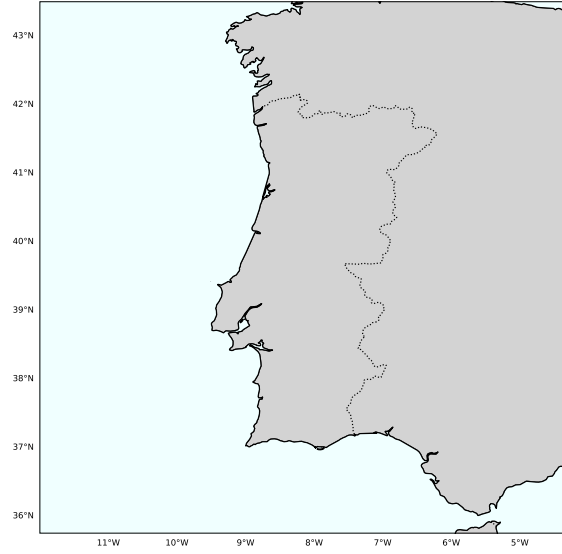


Figure 4.2: Geographic extent of the study area covering the Iberian Peninsula and the adjacent Atlantic Ocean. Domain bounded by $[35.75^\circ\text{N}, 43.50^\circ\text{N}]$ and $[12^\circ\text{W}, 4.25^\circ\text{W}]$.

The detailed construction of the low-resolution daily field from ERA5 is described in the next section. The paired grids 16×16 and 32×32 define the controlled downscaling setting used in this work: daily low-resolution fields are used to generate hourly high-resolution fields.

4.2.2 Low-Resolution Daily Field Construction

For each day d , the 24 hourly ERA5 temperature fields are first averaged to obtain a high-resolution daily mean field,

$$\mathbf{m}_d^{HR} = \frac{1}{24} \sum_{h=1}^{24} x_{d,h}^{HR}, \quad \mathbf{m}_d^{HR} \in \mathbb{R}^{32 \times 32}. \quad (4.8)$$

This daily mean field is then spatially coarsened by averaging non-overlapping 2×2 blocks, producing the synthetic low-resolution daily condition $\mathbf{y}_d^{LR} \in \mathbb{R}^{16 \times 16}$. If $B_{i,j}$ denotes

the set of four high-resolution cells associated with the low-resolution (i, j) cell, then

$$y_{d,i,j}^{LR} = \frac{1}{4} \sum_{(r,s) \in B_{i,j}} m_{d,r,s}^{HR}. \quad (4.9)$$

The resulting field contains the large-scale daily temperature state, but it no longer contains the hourly variability or the fine-scale spatial structure that the model is expected to generate. Figure 4.3 exemplifies each of the fields.

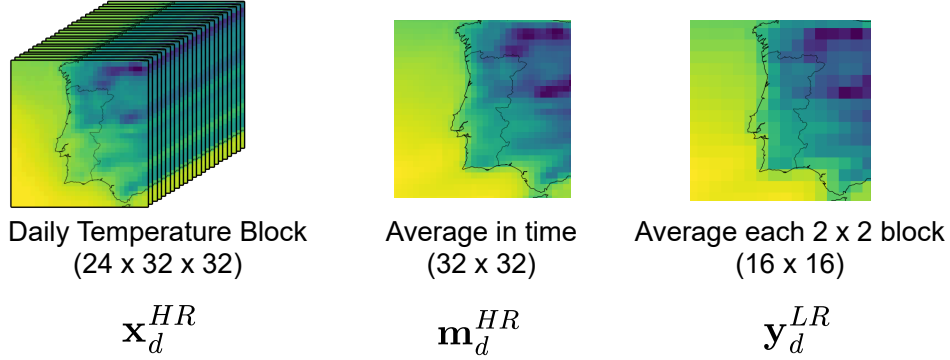


Figure 4.3: Schematic of the synthetic low-resolution daily field construction. The high-resolution hourly sequence $\mathbf{x}_d^{HR}$ is first averaged in time to obtain the daily mean field $\mathbf{m}_d^{HR}$. This field is then downsampled through 2×2 block averaging to produce the low-resolution daily condition $\mathbf{y}_d^{LR}$.

For use inside the neural network, the (16×16) low-resolution field is also projected back to the 32×32 grid by repeating each coarse value over the corresponding 2×2 block. This produces $\tilde{\mathbf{y}}_d^{HR} \in \mathbb{R}^{32 \times 32}$. This projected field preserves the information content of the coarse daily input while matching the spatial dimensions of the high-resolution target.

The physical high-resolution output is the hourly ERA5 temperature block $\mathbf{x}_d^{HR} \in \mathbb{R}^{24 \times 32 \times 32}$. The resulting daily sample is represented as

$$\mathcal{D}_d = \left(\mathbf{x}_d^{HR}, \mathbf{m}_d^{HR}, \mathbf{y}_d^{LR}, \tilde{\mathbf{y}}_d^{HR}, t_d \right), \quad (4.10)$$

where t_d denotes the calendar date associated with the sample.

4.2.3 Static Field Extraction

Two static geographical fields are extracted over the same 32×32 domain: orography and Land-Sea Mask (LSM). Orography is obtained from the ERA5 geopotential variable z , in $\text{m}^2 \text{s}^{-2}$ and converted to elevation in meters using

$$\text{orog} = \frac{z}{g}, \quad \text{orog} \in \mathbb{R}^{32 \times 32}, \quad (4.11)$$

where $g = 9.80665 \text{ m s}^{-2}$. The LSM is extracted over the same spatial window from ERA5 dataset. This field identifies the surface type at each grid point and is represented with values in the interval $[0, 1]$, where values close to 1 correspond to land and values close to 0 correspond to sea.

4.2.4 Data Split and Normalization

The processed ERA5 dataset is divided chronologically using complete calendar years. This avoids mixing neighboring days across different data splits and preserves the temporal structure of the climate record. In this work, the historical ERA5 period is divided into a training period and an independent test period. The training period is used to fit the model parameters and to compute all normalization statistics. The test period is kept separate and is used for the evaluation of the trained model. The split is described in Table 4.1.

Table 4.1: Data partitioning for model training and testing. The dataset is split chronologically.

Dataset Split	Years
Training Set	1980 – 2009
Test Set	2010 – 2024

Years 2010–2013 are reserved as a validation set used exclusively for early stopping of deterministic U-Net and hyperparameter selection for k-NN baselines; years 2014–2023 constitute the held-out test set used for model comparison and ensemble depth analysis using ERA5; and the full non-training period 2010–2024 is used only as the ERA5 reference climatology when evaluating the model under future CMIP6 projections.

All continuous variables used by the model are normalized before training. Normalization statistics are computed solely from the training period to avoid information leakage. For a generic variable v , standardization is performed as:

$$v_{\text{norm}} = \frac{v - \mu_v}{\sigma_v}, \quad (4.12)$$

where μ_v and σ_v are the mean and standard deviation estimated from the training data, respectively. The variables to which this transformation is applied are described in the following sections.

4.3. Target Formulation

Although the final output is an absolute temperature field, the diffusion model is trained to generate anomalies relative to the projected daily condition. This separates the imposed large-scale daily state from the sub-daily and fine-scale structure that must be generated.

4.3.1 Anomaly Target

As described in Section 4.2, each low-resolution daily field is projected back to the high-resolution grid, producing $\tilde{\mathbf{y}}_d^{HR}$. The anomaly target is defined as:

$$\mathbf{a}_d = \mathbf{x}_d^{HR} - \tilde{\mathbf{y}}_d^{HR}, \quad \mathbf{a}_d \in \mathbb{R}^{24 \times 32 \times 32}, \quad (4.13)$$

where $\tilde{\mathbf{y}}_d^{HR}$ is repeated across the 24 hourly steps due to its daily resolution. Equivalently, for hour h ,

$$a_{d,h} = x_{d,h}^{HR} - \tilde{y}_d^{HR}. \quad (4.14)$$

Therefore, the anomaly target has the same temporal and spatial dimensions as the original hourly block.

This anomaly contains the part of the hourly high-resolution temperature field that is not already represented by the projected coarse daily mean. It includes the diurnal evolution, local deviations from the daily state, and fine-scale spatial variability.

The diffusion model is therefore trained to model the conditional distribution of anomaly fields:

$$p_\theta \left(\mathbf{a}_d \mid \mathbf{y}_d^{LR}, \mathbf{c}_d \right). \quad (4.15)$$

Before training, the anomaly is standardized using statistics computed only from the training period. Let μ_a and σ_a denote the training-period mean and standard deviation of the anomaly field. The normalized anomaly target is then

$$\mathbf{a}_{d,\text{norm}} = \frac{\mathbf{a}_d - \mu_a}{\sigma_a}. \quad (4.16)$$

The anomaly is computed before normalization, so both the hourly ERA5 block and the projected daily field are first kept in physical temperature units.

The use of anomalies does not change the final physical quantity produced by the method. It only changes the internal representation learned by the model.

4.3.2 Reconstruction of Absolute Temperature

During inference, a generated normalized anomaly sample $\hat{\mathbf{a}}_{d,\text{norm}}$ is first transformed back to physical units,

$$\hat{\mathbf{a}}_d = \sigma_a \hat{\mathbf{a}}_{d,\text{norm}} + \mu_a. \quad (4.17)$$

and the final absolute hourly temperature block is reconstructed as:

$$\hat{\mathbf{x}}_d^{HR} = \hat{\mathbf{a}}_d + \tilde{\mathbf{y}}_d^{HR}. \quad (4.18)$$

This reconstruction produces the final generated temperature sequence in physical units. Therefore, although the diffusion model operates in anomaly space during training and sampling, the output is the reconstructed absolute hourly temperature fields.

4.4. Conditioning Design

The diffusion model is conditioned on information available for each day before generation: the imposed daily temperature state, fixed geographical properties of the domain, temporal position within the annual and daily cycles, and scalar controls used to guide scenario generation. All conditioning variables are represented on the same temporal and spatial support as the anomaly target, namely $24 \times 32 \times 32$.

4.4.1 Spatial and Temporal Conditions

The primary dynamic condition is the projected daily mean temperature field $\tilde{\mathbf{y}}_d^{HR}$. Since the diffusion model generates anomalies relative to this daily mean field, $\tilde{\mathbf{y}}_d^{HR}$ both defines the reference state used in the anomaly formulation, and provides the denoising network with the daily temperature context under which the anomaly should be generated. Before being used by the model, this field is standardized according to the normalization procedure defined in Section 4.2.4.

Static geographical information is provided through orography and the LSM, both defined on the 32×32 grid, as described in Section 4.2.3. Orography is standardized over the study domain, while the LSM is kept in its original range $[0, 1]$. These fields are repeated across the 24 hourly steps and provide fixed spatial information related to elevation, land–ocean contrast, and coastal structure. Figure 4.4 illustrates the orography field used.

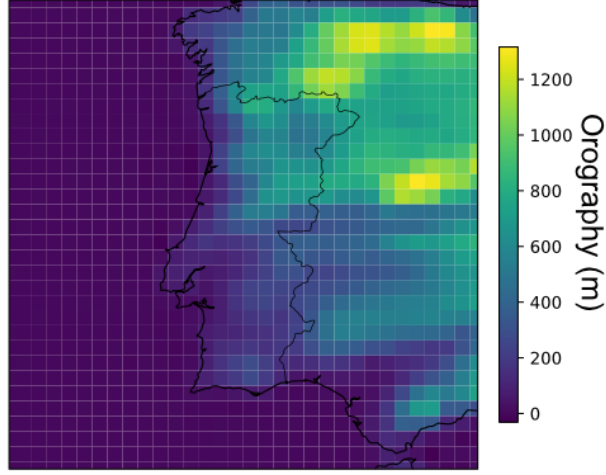


Figure 4.4: Orography field used as a static conditioning variable. Surface elevation is depicted in meters over the 32×32 study domain.

Temporal information is encoded using cyclic sine and cosine variables. The annual cycle is encoded from the day of year as

$$c_{\text{doy},1}(d) = \sin \left(2\pi \frac{\text{doy}(d)}{N_{\text{doy}}} \right) \quad (4.19)$$

$$c_{\text{doy},2}(d) = \cos \left(2\pi \frac{\text{doy}(d)}{N_{\text{doy}}} \right) \quad (4.20)$$

where $\text{doy}(d)$ is the day-of-year index for day d , and N_{doy} is the number of days in the corresponding year. This representation preserves the cyclic nature of the calendar year, so that days near the end of December and the beginning of January are represented as temporally close rather than as distant values.

The daily cycle is similarly represented using the hour of day,

$$c_{\text{hour},1}(h) = \sin \left(2\pi \frac{h}{24} \right) \quad (4.21)$$

$$c_{\text{hour},2}(h) = \cos \left(2\pi \frac{h}{24} \right) \quad (4.22)$$

where $h = 0, 1, \dots, 23$. These variables provide the model with the position of each field within the 24-hour cycle.

The day-of-year encoding is constant within each daily block, while the hour-of-day encoding varies along the temporal dimension. Since these variables are not spatial fields, they are expanded over the 32×32 grid before being passed to the model. This allows the same temporal information to be available at every spatial location. These variables

are already bounded between -1 and 1, so no additional standardization is applied.

4.4.2 Scalar Control Variables

Two scalar controls are used to guide the generated scenarios. The first scalar, s_d , measures the average sub-daily variability of the anomaly field. It is computed from the raw anomaly field before normalization. For each grid point, the standard deviation over the 24 hourly values is calculated, and these values are then averaged over the spatial domain. Let this diurnal variability scalar for day d be represented as

$$s_d = \frac{1}{32 \times 32} \sum_{i=1}^{32} \sum_{j=1}^{32} \text{std}_h(a_{d,h,i,j}) \quad (4.23)$$

where $a_{d,h,i,j}$ is the raw anomaly at hour h and grid point (i, j) at day d . Larger values of s_d correspond to days with stronger sub-daily variability, while smaller values correspond to weaker hourly deviations from the daily mean field.

The second scalar, r_d , represents the domain-mean temperature state of the daily condition. It is computed as the spatial average of the standardized projected daily field:

$$r_d = \frac{1}{32 \times 32} \sum_{i=1}^{32} \sum_{j=1}^{32} \tilde{y}_{d,\text{norm},i,j}^{HR} \quad (4.24)$$

This scalar indicates whether the coarse daily condition is generally warmer or colder than the training-period average. Although related to the coarse daily field, r_d provides an explicit domain-level control variable that can be selected or modified during inference. Together with the diurnal variability scalar, it allows the sampling process to explore different plausible versions of the same daily condition, including scenarios with stronger variability or temperature states spanning different parts of the historical distribution. The purpose of these scalars is therefore not only descriptive. They are introduced to increase controllability during scenario generation, allowing the model to explore a wider range of plausible outcomes while still remaining within a region supported by the historical data.

Together, these quantities define a two-dimensional control vector:

$$\mathbf{z}_d = (s_d, r_d). \quad (4.25)$$

During training, $\mathbf{z}_d$ is computed from each ERA5 daily sample and provided as conditioning information. During inference, the reference hourly anomaly is unknown, so the controls must be selected before generation. Different values of the diurnal variability can

be selected to generate different plausible variations of the same daily condition. In particular, the diurnal variability scalar allows the sampling process to explore scenarios with weaker or stronger variability, including more extreme sub-daily departures when values near the upper part of the historical range are used.

However, the scalar controls should not be assigned arbitrary values. Values far from the historical training distribution may lead the model into regions of the conditioning space where its behavior is less reliable. For this reason, inference control pairs are empirically selected from the training data.

Using all the ERA5 historical training period days, a control-pair bank is constructed containing the pair z_d together with the corresponding date and month.

During inference, for each target day, the second component of the control vector, r_d , is fixed to the value computed from the target daily condition. Only the diurnal-variability scalar is selected from historical analogues. Candidate analogue days are first restricted to the same month and closest neighboring months. Then, from this pool, the K , with $K = 200$ in this work, training days with the daily temperature-regime scalar closest to the target r_d are retained. To generate multiple plausible versions of the same daily condition, values of s_d are selected from empirical quantiles of this analogue pool specific to that day. The selected pairs use ten quantiles spanning low, moderate, high and upper-tail diurnal variability (quantiles $\in \{0.05, 0.15, 0.30, 0.45, 0.55, 0.70, 0.82, 0.90, 0.96, 0.99\}$). This procedure allows multiple plausible versions of the same daily condition to be generated while preserving the empirical relationship between season, domain-mean temperature, and sub-daily variability.

4.4.3 Assembly of the Conditioning Tensor

All conditioning variables are broadcast or repeated to the common support $24 \times 32 \times 32$ and concatenated along the channel dimension. The complete conditioning tensor is

$$\mathbf{c}_d \in \mathbb{R}^{C_c \times 24 \times 32 \times 32}, \quad (4.26)$$

where C_c is the number of conditioning channels. Table 4.2 describes the conditioning components included in the tensor for this configuration.

When all components are used, $C_c = 9$. The conditioning tensor is provided to the denoising network together with the noisy anomaly sample and the diffusion timestep, as explained in Section 4.6.

Table 4.2: Summary of conditioning components, channel dimensions, and their respective spatiotemporal behavior.

Conditioning component	Channels	Temporal behavior	Spatial behavior
Projected daily temperature field	1	Repeated across 24 hours	32×32 field
Orography	1	Repeated across 24 hours	32×32 field
Land-sea mask	1	Repeated across 24 hours	32×32 field
Day-of-year sine/cosine	2	Constant within the day	Broadcast over space
Hour-of-day sine/cosine	2	Varies by hour	Broadcast over space
Scalar control variables	2	Repeated across 24 hours	Broadcast over space

4.5. Conditional Diffusion Models

The generative component of this work is a conditional denoising diffusion model trained in normalized anomaly space. Let

$$\mathbf{u}_d = \mathbf{a}_{d,\text{norm}}, \quad \mathbf{u}_d \in \mathbb{R}^{24 \times 32 \times 32}. \quad (4.27)$$

The model estimates the conditional distribution

$$p_\theta(\mathbf{u}_d \mid \mathbf{c}_d), \quad (4.28)$$

where $\mathbf{c}_d$ is the conditioning tensor defined in Section 4.4. Generated samples are later mapped back to physical temperature units using the reconstruction step introduced in the previous subsection.

Diffusion time is denoted by τ to distinguish it from physical time. The forward process defined in Section 3.3.1 is applied over $T = 1000$ diffusion steps. For a clean anomaly block $\mathbf{u}_{d,0} = \mathbf{u}_d$, the noisy sample at timestep τ is

$$\mathbf{u}_{d,\tau} = \sqrt{\bar{\alpha}_\tau} \mathbf{u}_{d,0} + \sqrt{1 - \bar{\alpha}_\tau} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (4.29)$$

A cosine noise schedule [58] is used for $\bar{\alpha}_\tau$,

$$\bar{\alpha}^{\cos} = \frac{\cos^2\left(\frac{\tau/T + s_{\cos}}{1 + s_{\cos}} \frac{\pi}{2}\right)}{\cos^2\left(\frac{s_{\cos}}{1 + s_{\cos}} \frac{\pi}{2}\right)}, \quad (4.30)$$

with offset $s_{\cos} = 0.008$. The corresponding per-step variance, signal-retention factor, and cumulative coefficient in the forward process are, respectively,

$$\beta_\tau = 1 - \frac{\bar{\alpha}_\tau^{\cos}}{\bar{\alpha}_{\tau-1}^{\cos}}, \quad \alpha_\tau = 1 - \beta_\tau, \quad \bar{\alpha}_\tau = \prod_{j=1}^{\tau} \alpha_j. \quad (4.31)$$

The denoising network predicts the injected noise from the noisy anomaly block, the diffusion timestep, and the conditioning tensor:

$$\hat{\epsilon} = \epsilon_{\theta}(\mathbf{u}_{d,\tau}, \tau, \mathbf{c}_d). \quad (4.32)$$

The model is trained with the standard noise-prediction objective,

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{d,\tau,\epsilon} \left[\left\| \epsilon - \epsilon_{\theta}(\mathbf{u}_{d,\tau}, \tau, \mathbf{c}_d) \right\|_2^2 \right]. \quad (4.33)$$

The loss is averaged over the batch and over all temporal and spatial positions. No reconstruction, perceptual, or adversarial losses are used.

4.6. Denoising Architecture

The denoising function ϵ_{θ} is implemented using a U-Net architecture [59]. A U-Net is an encoder–decoder convolutional network in which the encoder progressively reduces the spatial resolution while increasing the number of feature channels, and the decoder restores the original resolution. Skip connections link encoder and decoder stages at matching resolutions, allowing high-resolution information from the encoder to be reused during reconstruction. This structure is well suited to dense field-generation tasks because it combines local detail with broader spatial context.

In this work, the standard U-Net structure is adapted to the spatiotemporal diffusion setting. The network receives the noisy anomaly block and the conditioning tensor concatenated along the channel dimension,

$$\mathbf{v}_{d,\tau} = \text{concat}(\mathbf{u}_{d,\tau}, \mathbf{c}_d). \quad (4.34)$$

The output is a single-channel noise prediction with the same temporal and spatial dimensions as the anomaly target,

$$\hat{\epsilon}_{\theta} \in \mathbb{R}^{1 \times 24 \times 32 \times 32}. \quad (4.35)$$

The U-Net has three encoder levels with spatial resolutions

$$32 \times 32 \rightarrow 16 \times 16 \rightarrow 8 \times 8,$$

and a mirrored decoder. Only the spatial dimensions are downsampled and upsampled; the temporal dimension is preserved at all levels. This design keeps the 24 hourly positions explicitly represented throughout the network and avoids temporal interpolation

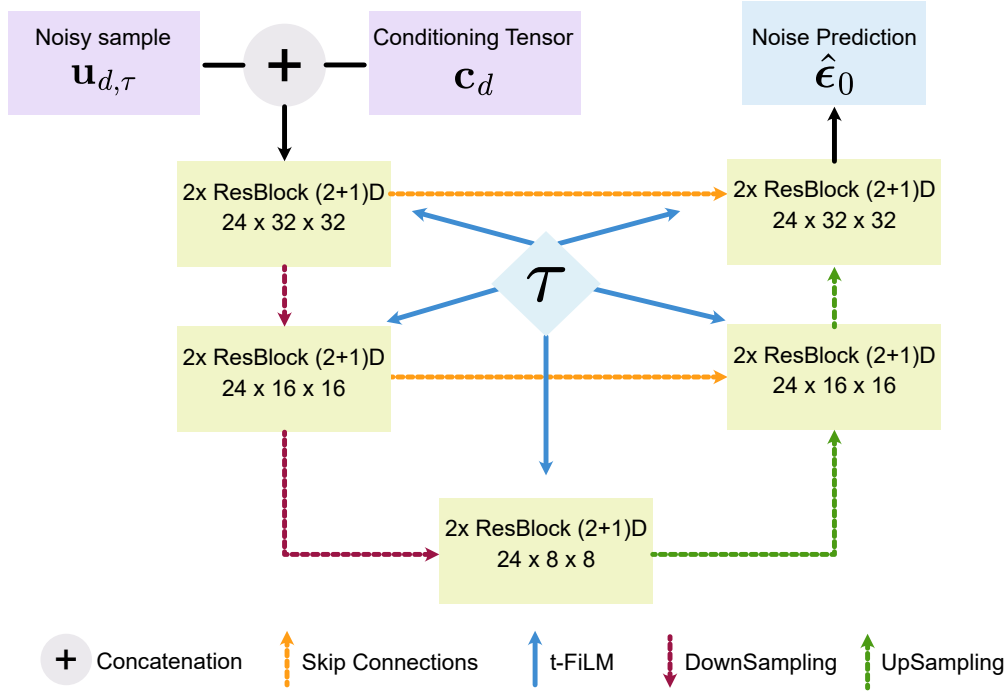


Figure 4.5: Schematic representation of the conditional $(2 + 1)$ D U-Net denoising architecture. The noisy sample $\mathbf{u}_{d,\tau}$ and conditioning tensor $\mathbf{c}_d$ are concatenated along the channel dimension and processed by a multi-resolution encoder–decoder network. Spatial resolution is reduced from 32×32 to 8×8 and then restored, while the temporal dimension is preserved. Skip connections transfer encoder features to the corresponding decoder stages. The diffusion timestep τ is injected into the residual blocks through Feature-wise Linear Modulation (FiLM) modulation.

during decoding. The feature widths are 64, 128, and 256 channels, and each level contains two residual $(2 + 1)$ D blocks.

Each residual block uses a factorized $(2 + 1)$ D convolutional design [60]. Instead of applying a full 3D convolution jointly over time and space, the block applies a spatial convolution followed by a temporal convolution. The spatial convolution has kernel size $1 \times 3 \times 3$, processing local spatial neighborhoods independently at each hour, while the temporal convolution has kernel size $3 \times 1 \times 1$, exchanging information between neighboring hourly positions at each spatial location. This factorization reduces computational cost while preserving both spatial and short-range temporal interactions.

The residual path combines group normalization [61], Sigmoid Linear Unit (SiLU) activation [62], spatial convolution, timestep modulation, temporal convolution, optional temporal attention [63], and a final spatial convolution. A skip connection is added around the

residual path; when the number of input and output channels differs, a $1 \times 1 \times 1$ convolution is used to match the channel dimension.

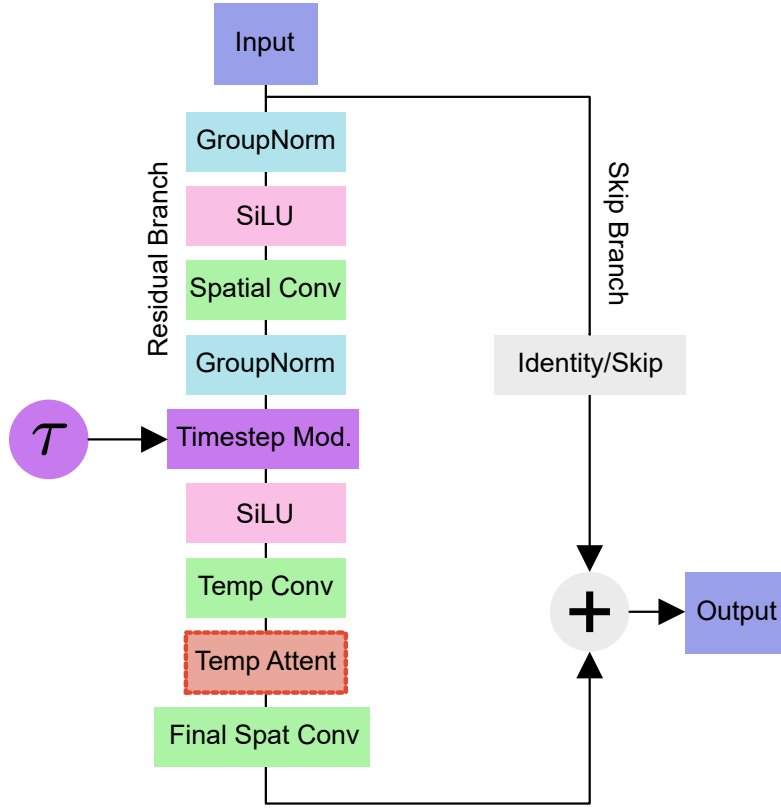


Figure 4.6: Schematic representation of the residual spatiotemporal block used in the denoising network. The residual branch applies normalization, activation, spatial convolution, timestep modulation, temporal convolution, optional temporal attention, and a final spatial convolution. The result is added to the skip branch to produce the block output.

Temporal self-attention is used to complement the local temporal convolutions. While the temporal convolution exchanges information only between neighboring hours, self-attention allows each hourly position to interact with all other hours in the same daily block. Attention is applied only along the temporal dimension and independently at each spatial location of the feature map. For an intermediate tensor

$$\mathbf{F} \in \mathbb{R}^{B \times C \times T \times H \times W}, \quad (4.36)$$

each spatial location (i, j) is treated as a temporal sequence

$$\mathbf{F}_{i,j} \in \mathbb{R}^{T \times C}. \quad (4.37)$$

Multi-head self-attention is then applied to this sequence, mixing information across the $T = 24$ hourly positions while leaving spatial interactions to the convolutional backbone.

In this implementation, temporal attention uses four heads and is enabled only at the lower spatial resolutions, 16×16 and 8×8 . This reduces computational cost relative to full-resolution attention while allowing the attention mechanism to operate on features with broader spatial context. The attention module is inserted with a residual connection, so its output is added back to the input features.

The diffusion timestep τ is encoded using a sinusoidal embedding of dimension 256, followed by a two-layer Multilayer Perceptron (MLP) with SiLU activation and hidden dimension 1024. This produces a timestep representation $\mathbf{t}_\tau \in \mathbb{R}^{256}$. The timestep representation is injected into the residual blocks using feature-wise linear modulation (FiLM) [64]. For an intermediate feature tensor $\mathbf{F}$, the timestep embedding is projected to channel-wise scale and shift parameters,

$$(\gamma_\tau, \beta_\tau^{\text{FiLM}}) = W\mathbf{t}_\tau + \mathbf{b}. \quad (4.38)$$

These parameters are broadcast over the temporal and spatial dimensions and applied as

$$\text{FiLM}(\mathbf{F}, \tau) = (1 + \gamma_\tau) \circ \mathbf{F} + \beta_\tau^{\text{FiLM}}. \quad (4.39)$$

Here γ_τ controls the timestep-dependent scaling of each feature channel, while β_τ^{FiLM} provides a timestep-dependent shift. The use of $1 + \gamma_\tau$ makes the modulation close to the identity transformation when the learned scale is near zero. This allows each residual block to adapt its computation to the current diffusion noise level.

Overall, conditioning information enters the network through channel-wise concatenation at the input, while diffusion-time information is injected internally through FiLM modulation. This separates the physical conditioning variables from the diffusion timestep while keeping both available throughout the denoising process.

4.7. Training Procedure and Implementation Details

The model is trained using the conditional noise-prediction objective in Eq. (4.33). At each optimization step, a mini-batch of daily samples is drawn from the training set, a diffusion timestep $\tau \sim \mathcal{U}\{0, \dots, T-1\}$ is sampled, Gaussian noise is added according to the forward process, and the denoising network is optimized to predict the injected noise.

The final training configuration is summarized in Table 4.3. Training uses AdamW [65] with learning rate 5×10^{-5} , batch size 8, and 500,000 optimization steps. Gradients are clipped to a maximum norm of 1.0, mixed precision is enabled, and an Exponential Moving Average (EMA) of the model weights is maintained with decay 0.999. Conditioning dropout with probability $p_{\text{drop}} = 0.1$ is applied to the projected daily temperature field and scalar-control channels, while static geographical fields and cyclic temporal encodings are always retained.

Table 4.3: Training configuration of the conditional diffusion model.

Component	Value
Optimizer	AdamW
Weight decay	0
Learning rate	5×10^{-5}
Batch size	8
Training steps	5×10^5
Diffusion steps	$T = 1000$
Gradient clipping	Maximum norm 1.0
Mixed precision	Enabled
EMA decay	$\lambda_{\text{EMA}} = 0.999$
Conditioning dropout	$p_{\text{drop}} = 0.1$
Hardware	NVIDIA RTX 3090 GPU

During training, checkpoints were saved periodically to preserve the state of the experiment and allow training to be resumed.

To make the EMA update explicit, let θ_k denote the model parameters at optimization step k . The averaged parameters $\bar{\theta}_k$ are updated as

$$\bar{\theta}_k = \lambda_{\text{EMA}} \bar{\theta}_{k-1} + (1 - \lambda_{\text{EMA}}) \theta_k, \quad (4.40)$$

with $\lambda_{\text{EMA}} = 0.999$. This averaging assigns exponentially decaying weight to past model states, so that more recent parameters contribute more than older ones while short-term fluctuations caused by stochastic mini-batch training, random diffusion timesteps, and random noise realizations are smoothed. In diffusion models, EMA weights are commonly used for sampling because they often produce more stable generations than the instantaneous training weights.

We use AdamW with weight decay set to zero, which is equivalent to Adam. The decoupled-decay setting is retained only for implementation consistency. Weight decay was not tuned and no systematic hyperparameter search was performed for the neural networks due to the large parameter count and the high training time per configuration.

4.7.1 Implementation Details

The implementation was developed in Python [66]. PyTorch [67] was used to implement the denoising network, diffusion training loop, automatic differentiation, and GPU-accelerated optimization.

Data processing was performed with NumPy [68] and xarray [69]. Processed yearly datasets were stored in Zarr format [70], enabling chunked access during training and inference. Analysis and visualization routines were implemented using SciPy [71] and Matplotlib [72].

4.8. Inference and Scenario Generation

After training, the diffusion model is used to generate high-resolution hourly temperature scenarios from a conditioning tensor. For a fixed daily condition $\mathbf{c}_d$, sampling starts from Gaussian noise,

$$\mathbf{u}_{d,T}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (4.41)$$

where k indexes the ensemble member. The trained reverse process is then applied from the final diffusion step to the clean anomaly estimate, producing a generated normalized anomaly block $\hat{\mathbf{u}}_d^{(k)}$. Repeating this process gives an ensemble

$$\left\{ \hat{\mathbf{u}}_d^{(1)}, \hat{\mathbf{u}}_d^{(2)}, \dots, \hat{\mathbf{u}}_d^{(K)} \right\}. \quad (4.42)$$

Each member is reconstructed into absolute temperature using the procedure in Section 4.3.

Two sources of ensemble diversity are used. First, different members are initialized with different Gaussian noise fields. Second, different admissible scalar-control pairs, explained in Section 4.4.2, can be selected before sampling, allowing the same daily temperature condition to produce scenarios with different plausible levels of sub-daily variability.

Sampling is performed using an accelerated DPM++ sampler [43]. This reduces the number of denoising evaluations required for ensemble generation while using the same trained denoising network. All reported inference experiments use 25 solver steps. The implementation also supports classifier-free-guidance-style conditioning through the conditioning dropout mechanism. However, after testing different values, we decided to use guidance scale $w = 1$ at inference in this work, corresponding to the fully conditional prediction without additional guidance amplification.

4.9. Application to CMIP6 Projections

The trained model is applied to daily near-surface temperature fields from CMIP6. These fields play the same role at inference time as the synthetic low-resolution daily fields derived from ERA5 during training. The CMIP6 fields are first restricted to the same geographical domain and represented on the 16×16 low-resolution conditioning grid, $\mathbf{y}_d^{LR} \in \mathbb{R}^{16 \times 16}$.

Since raw CMIP6 projections may contain systematic biases relative to ERA5, the daily fields are bias-corrected before being passed to the diffusion model. This work uses QDM, which adjusts historical model bias while preserving the model-projected quantile-specific change signal [73]. The correction is estimated separately for each calendar month and grid cell, using a common historical calibration period between CMIP6 and ERA5.

For each calendar month and grid cell, let $y_{\text{hist}}^{\text{CMIP6}}$ denote the historical CMIP6 distribution, $y_{\text{hist}}^{\text{ERA5}}$ the corresponding ERA5 reference distribution, and $y_{\text{fut}}^{\text{CMIP6}}$ the future CMIP6 distribution. For a future value $y_{\text{fut}}^{\text{CMIP6}}$, its quantile in the future CMIP6 distribution is

$$q = F_{\text{fut}}^{\text{CMIP6}} \left(y_{\text{fut}}^{\text{CMIP6}} \right). \quad (4.43)$$

For temperature, treated as an additive variable, the quantile-specific change is

$$\Delta_q = y_{\text{fut}}^{\text{CMIP6}} - \left(F_{\text{hist}}^{\text{CMIP6}} \right)^{-1} (q), \quad (4.44)$$

and the corrected value is

$$y^{\text{QDM}} = \left(F_{\text{hist}}^{\text{ERA5}} \right)^{-1} (q) + \Delta_q. \quad (4.45)$$

After QDM correction, each daily field is projected to the 32×32 model grid by repeating each coarse value over its corresponding 2×2 block. The projected field is standardized using the same training-period statistics used for the ERA5-derived daily condition. The remaining conditioning channels are constructed as in training: orography and LSM are reused, temporal encodings are computed from the target date and hour, and scalar-control values are selected from the historical admissible region defined in Section 4.4.2.

The resulting future conditioning tensor has the same structure as the training conditioning tensor,

$$\mathbf{c}_d^{\text{CMIP6}} \in \mathbb{R}^{C_c \times 24 \times 32 \times 32}. \quad (4.46)$$

It is then passed to the trained diffusion model to generate hourly temperature scenarios under the specified CMIP6 projection.

4.10. Baselines and Models for Comparison

To assess the contribution of the main modeling choices, the proposed configuration is compared with a set of baseline and ablation models. These comparisons are designed to evaluate whether the proposed method improves over simpler alternatives and to isolate the effect of specific components of the conditioning design.

The comparison includes four models. The first is the full proposed model, which uses all conditioning variables and scalar controls. The second is an ablated diffusion model in which the scalar control variables are removed. The third is a non-parametric k-NN anomaly baseline. The fourth is a deterministic U-Net baseline, included as a deterministic downscaling reference.

The objective of the comparison is not only to measure the error, but also to evaluate distributional realism, temporal coherence, and variability.

4.10.1 Full Diffusion Model

The full proposed model corresponds to the final conditional diffusion configuration described in the previous sections. It generates normalized hourly anomaly blocks conditioned on the daily temperature field, static geographical fields, temporal encodings, and scalar control variables.

This model represents the main contribution of the work. It combines stochastic generation through diffusion sampling with controllability through the scalar variables. During inference, multiple scenarios can be generated for the same daily temperature condition by changing the initial Gaussian noise and, when desired, by selecting different admissible scalar-control pairs.

The full model is used as the reference configuration against which the ablated and baseline methods are compared.

4.10.2 Model without Scalar Controls

The first ablation removes the scalar control variables from the conditioning tensor. The remaining conditioning information is unchanged: the projected daily temperature field, static geographical variables, and temporal encodings are still provided to the model.

This ablation isolates the contribution of scalar controls. Without these variables, the model can still generate stochastic samples across different initial noise fields, but it loses the explicit mechanism to guide the generated scenarios toward different levels of sub-daily variability or temperature regimes.

Comparing this variant with the full model allows the evaluation of whether scalar controls improve ensemble diversity, representation of extremes, and controlled scenario exploration without degrading consistency with the prescribed daily temperature condition.

4.10.3 k-NN Anomaly Baseline

A k-NN anomaly baseline is used as a non-parametric reference method. It evaluates how much of the task can be solved by retrieving historical hourly anomaly patterns associated with similar daily temperature conditions, rather than learning a conditional generative model [25], [26].

The baseline is constructed from ERA5 training period. For each training day, the projected daily condition is stored together with its corresponding hourly anomaly block. Given a query day d , candidate historical days n are ranked using the mean squared distance between normalized projected daily fields:

$$D(d, n) = \frac{1}{32 \times 32} || \phi(\tilde{\mathbf{y}}_d^{HR}) - \phi(\tilde{\mathbf{y}}_n^{HR}) ||_2^2, \quad (4.47)$$

where d is the query day, n is a candidate historical day, and ϕ denotes the standardization and flattening of the projected daily field.

The k historical days with the smallest distance are selected as analogue candidates. To maintain seasonal consistency, the search can be restricted to candidates from the same or neighboring calendar months. This avoids retrieving anomaly patterns from seasons with physically different diurnal behavior.

A generated scenario is obtained by selecting one of the nearest neighbors and adding its historical anomaly block to the query daily condition:

$$\hat{\mathbf{x}}_d = \tilde{\mathbf{y}}_d^{HR} + \mathbf{a}_n. \quad (4.48)$$

In sampling mode, neighbors are randomly drawn from the k closest candidates, producing an ensemble of analogue-based scenarios. The final configuration uses $k = 30$ and

a seasonal window of ± 1 month. These values were selected through a separate hyperparameter tuning procedure that tested different neighbor sizes and seasonal windows using deterministic, probabilistic, variability and extreme-temperature diagnostics.

This baseline is simple and interpretable, but it can only reuse anomaly patterns already present in the historical ERA5 bank. It therefore provides a useful reference for evaluating whether the diffusion model can generate realistic variability beyond direct historical analogue retrieval.

Since its anomaly blocks are taken directly from historical ERA5 samples, the generated structures are inherently realistic at the level of individual retrieved patterns. However, the method is limited to recombining anomaly patterns already present in the historical bank and cannot synthesize new structures conditioned on the query field. It therefore provides a strong reference for assessing whether the diffusion model can match the realism of historical analogue retrieval while offering greater flexibility, diversity, and controllability.

4.10.4 Deterministic U-Net

A deterministic U-Net baseline is included as a supervised downscaling reference. It uses the same general encoder–decoder backbone as the diffusion model but does not use the diffusion process or scalar control variables. Instead, it directly predicts a single normalized hourly anomaly block from the projected daily temperature field, static geographical fields, and temporal encodings.

The deterministic model is trained by minimizing the MSE between the predicted anomaly block and the ERA5 target anomaly block. At inference time, the predicted anomaly is added back to the projected daily condition to reconstruct the absolute hourly temperature field.

This baseline isolates the benefit of probabilistic diffusion sampling. Unlike the diffusion models, it produces only one output for a given daily condition and therefore cannot represent ensemble uncertainty.

4.11. Evaluation Metrics

The generated hourly temperature fields are evaluated using deterministic, probabilistic, temporal, spatial, distributional, and physically motivated diagnostics. When paired

ERA5 hourly fields are available, metrics are computed against the reference. For probabilistic models, the evaluation considers both the ensemble mean and the full ensemble distribution.

Let $x_{d,h,i,j}^E$ denote the ERA5 reference and let the ensemble mean be defined as

$$\bar{x}_{d,h,i,j} = \frac{1}{K} \sum_{k=1}^K \hat{x}_{d,h,i,j}^{(k)}. \quad (4.49)$$

Unless stated otherwise, metrics are computed over land grid points, denoted by $\mathcal{L}$, with $N_{\mathcal{L}}$ pixels.

4.11.1 Evaluation Spaces

Metrics are computed in three spaces. **Absolute temperature space** uses the reconstructed temperature fields directly and is used wherever the physical magnitude is the quantity of interest.

Diurnal-anomaly space removes the daily mean from each realization at each grid point:

$$a_{d,h,i,j}^{\text{diurn},(k)} = \hat{x}_{d,h,i,j}^{(k)} - \frac{1}{24} \sum_{h'=1}^{24} \hat{x}_{d,h',i,j}^{(k)}. \quad (4.50)$$

Conditioning-anomaly space subtracts the imposed projected daily condition:

$$a_{d,h,i,j}^{(k)} = \hat{x}_{d,h,i,j}^{(k)} - \tilde{y}_{d,i,j}^{HR}. \quad (4.51)$$

4.11.2 Point-wise Error Metrics

Deterministic accuracy is evaluated using the bias and RMSE of the ensemble mean in absolute temperature space:

$$\text{Bias} = \frac{1}{24N_dN_{\mathcal{L}}} \sum_{d=1}^{N_d} \sum_{h=1}^{24} \sum_{(i,j) \in \mathcal{L}} \left(\bar{x}_{d,h,i,j} - x_{d,h,i,j}^E \right), \quad (4.52)$$

which measures systematic warm or cold errors. The corresponding RMSE is

$$\text{RMSE} = \sqrt{\frac{1}{24N_dN_{\mathcal{L}}} \sum_{d=1}^{N_d} \sum_{h=1}^{24} \sum_{(i,j) \in \mathcal{L}} \left(\bar{x}_{d,h,i,j} - x_{d,h,i,j}^E \right)^2}. \quad (4.53)$$

Grid-point maps are obtained by removing the land-pixel average, while hourly versions are obtained by fixing h and averaging only over days and land pixels.

4.11.3 Probabilistic Skill and Ensemble Calibration

Probabilistic skill is assessed with the CRPS, which compares the full ensemble distribution with the reference value:

$$\text{CRPS}_{d,h,i,j} = \frac{1}{K} \sum_{k=1}^K \left| \hat{x}_{d,h,i,j}^{(k)} - x_{d,h,i,j}^E \right| - \frac{1}{2K^2} \sum_{k=1}^K \sum_{l=1}^K \left| \hat{x}_{d,h,i,j}^{(k)} - \hat{x}_{d,h,i,j}^{(l)} \right|. \quad (4.54)$$

The final CRPS score is obtained by averaging over days, hours, and land grid points. Lower CRPS indicates better probabilistic skill. When a baseline model is available, relative probabilistic skill can also be computed using Continuous Ranked Probability Skill Score (CRPSS)

$$\text{CRPSS} = 1 - \frac{\text{CRPS}_{\text{model}}}{\text{CRPS}_{\text{baseline}}}. \quad (4.55)$$

Positive values indicate improvement over the baseline. In this work k-NN is used as the baseline model.

Ensemble variability is summarized using the standard deviation across ensemble members, computed at each day, hour, and land grid point. This measures the uncertainty represented by the generated ensemble. Domain-mean ensemble spread is obtained by averaging this standard deviation over days, hours, and land grid points.

Calibration is evaluated using empirical prediction-interval coverage. For a nominal level α , the central interval is bounded by the linearly interpolated ensemble quantiles q_{low}^α and q_{high}^α , corresponding to probabilities $\frac{(1-\alpha)}{2}$ and $\frac{(1+\alpha)}{2}$. The empirical coverage is

$$\text{coverage}_{\text{emp}}(\alpha) = \frac{1}{24N_dN_{\mathcal{L}}} \sum_{d=1}^{N_d} \sum_{h=1}^{24} \sum_{(i,j) \in \mathcal{L}} \mathbf{1} \{ q_{low,d,h,i,j}^\alpha \leq x_{d,h,i,j}^E \leq q_{high,d,h,i,j}^\alpha \}. \quad (4.56)$$

For a calibrated ensemble, the empirical coverage should be close to the nominal level α .

Ensemble calibration is further assessed using rank histograms. For each verification point the ERA5 value is ranked among the sorted ensemble members. With K ensemble members, the observation can fall below all members, between two consecutive members, or above all members, resulting in $K + 1$ rank bins. A well-calibrated ensemble should produce an approximately uniform rank histogram.

4.11.4 Diurnal and Temporal Structure

Diurnal-cycle diagnostics are computed in diurnal-anomaly space. For each hour h , the generated mean diurnal anomaly is

$$\bar{a}_h^{\text{gen}} = \frac{1}{N_d K N_{\mathcal{L}}} \sum_{d=1}^{N_d} \sum_{k=1}^K \sum_{(i,j) \in \mathcal{L}} a_{d,h,i,j}^{\text{diurn},(k)}. \quad (4.57)$$

The corresponding hourly standard deviation is used to describe the variability around the mean diurnal cycle. The same quantities are computed for ERA5 and compared by hour.

Daily temperature amplitude is computed in absolute temperature space as

$$A_{d,i,j}^{(k)} = \max_h \hat{x}_{d,h,i,j}^{(k)} - \min_h \hat{x}_{d,h,i,j}^{(k)}. \quad (4.58)$$

This quantity is computed for generated and ERA5.

The timing of daily maximum and minimum temperature is evaluated using a circular phase error to account for the circular nature of the 24-hour cycle. Let h^{gen} and h^E be the generated and ERA5 hours of occurrence. The phase error is computed as

$$\Delta h = \min(|h^{\text{gen}} - h^E|, 24 - |h^{\text{gen}} - h^E|). \quad (4.59)$$

Hourly temperature ramps are also computed in absolute temperature space using consecutive hourly differences within a day:

$$\Delta_h x_{d,h,i,j}^{(k)} = \hat{x}_{d,h,i,j}^{(k)} - \hat{x}_{d,h-1,i,j}^{(k)}, \quad h = 2, \dots, 24. \quad (4.60)$$

Hourly ramps are computed as consecutive differences within each day. Their distributions are evaluated using probabilistic and distributional scores, including ramp- CRPS, ramp Wasserstein distance, and the ratio of generated to reference ramp standard deviation. Temporal persistence is assessed using the autocorrelation function of diurnal anomalies.

4.11.5 Spatial Structure

Spatial diagnostics are computed in conditioning-anomaly space, using anomalies relative to the imposed daily condition. Spatial correlation is evaluated over land pixels. For two land grid points p and q , the correlation matrix is

$$C_{p,q} = \text{corr}(a_p, a_q), \quad (4.61)$$

where a_p denotes the anomaly time series collected at grid point p over days and hours, and over ensemble members for generated fields. The generated and ERA5 spatial dependence structures are compared using pairwise correlation differences. These evaluations assess whether the generated fields reproduce the spatial covariance structure of the reference data.

Spatial variability across scales is evaluated using the radial PSD. This metric is computed on the full grid and not only on land points because the two-dimensional Fourier transform requires a complete rectangular field. For each conditioning-anomaly field, the spatial mean is removed, a two-dimensional Hann window $w(i, j)$ is applied, and the two-dimensional Fourier transform is computed as

$$\mathcal{A}_{d,h}^{(k)}(\xi, \eta) = \mathcal{F}_{2D}[w(i, j) (a_{d,h,i,j}^{(k)} - \bar{a}_{d,h}^{(k)})], \quad (4.62)$$

where ξ and η are the discrete spatial wavenumbers along the two grid axes. The spectral power, with window-energy normalization, is

$$P_{d,h}^{(k)}(\xi, \eta) = \frac{|\mathcal{A}_{d,h}^{(k)}(\xi, \eta)|^2}{\sum_{i,j} w(i, j)^2}. \quad (4.63)$$

The spectrum is then averaged radially as a function of the spatial wavenumber

$$\kappa = \sqrt{\xi^2 + \eta^2}. \quad (4.64)$$

The resulting PSD curve describes how anomaly variance is distributed across spatial scales, allowing comparison between generated fields and ERA5.

4.11.6 Distributional and Extreme-Temperature Diagnostics

Distributional agreement is assessed in absolute temperature space using the Kolmogorov-Smirnov (KS) statistic and the 1-Wasserstein distance,

$$D_{KS} = \sup_t |F_{\text{gen}}(t) - F_E(t)|. \quad (4.65)$$

and,

$$W_1(F_{\text{gen}}, F_E) = \int_{-\infty}^{\infty} |F_{\text{gen}}(t) - F_E(t)| dt. \quad (4.66)$$

These metrics compare the empirical temperature distributions of the generated fields and the ERA5 reference,

Sub-daily variability is evaluated using the distribution of daily temperature amplitudes over land points, computed as the difference between daily maximum and daily minimum temperatures.

For extremes, hourly fields are first reduced to daily maximum temperature at each land point and ensemble member, and the resulting samples are pooled over land pixels, ensemble members, and the selected season, mainly JJAS.

$$T_{\max,d,i,j}^{(k)} = \max_h \hat{x}_{d,h,i,j}^{(k)}. \quad (4.67)$$

The corresponding ERA5 quantity is computed analogously. Upper-tail errors are quantified using quantile bias at high percentiles:

$$\text{Bias}_q = Q_q(T_{\max}^{\text{gen}}) - Q_q(T_{\max}^{\text{E}}), \quad (4.68)$$

where Q_q denotes the empirical quantile at level q .

Fixed threshold exceedance rates are also computed for heat-relevant thresholds u , defined here as 35 °C and 40 °C. These thresholds were chosen because they represent very hot and severe heat conditions and are commonly used in regional studies of Iberian extreme temperatures and heat waves [74]. They are computed as

$$\text{rate}_u^{\text{gen}} = \frac{1}{N} \sum_{n=1}^N \mathbf{1} \{ T_{\max}^{\text{gen}} > u \}, \quad (4.69)$$

where N is the number of pooled samples. The same quantity is computed for ERA5. Exceedance rates are reported either directly or relative to the ERA5 exceedance rate.

4.11.7 Orographic and conditioning consistency

Orographic consistency is evaluated in absolute temperature space over land pixels. Land pixels are grouped into elevation bins. Within each elevation bin, the ensemble-mean temperature is compared against ERA5 across elevation. This assesses whether the generated fields preserve the temperature-elevation relationship of the reference data.

Conditioning consistency is evaluated by comparing the imposed conditioning scalars with the corresponding values measured from the generated fields. For the daily mean condition, the generated hourly fields are processed as in Section 4.2.2 and compared with the imposed daily condition using mean error, mean absolute error, and RMSE.

Scalar-control consistency is evaluated for the daily-temperature control r_d and the diurnal-variability control s_d , both in normalized units. The response is summarized using parity plots, Pearson and Spearman correlations, r^2 , and the fitted response slope.

4.12. Experimental Design

The trained model is evaluated through three experiments. These experiments differ in the driving data, evaluation period, ensemble size, and availability of a paired hourly reference. Together, they assess model performance under historical validation conditions, sensitivity to ensemble size, and behavior under future climate-projection input.

Table 4.4 summarizes the experiments.

Table 4.4: Experimental configuration and simulation design matrix for model benchmarking, ensemble depth analysis, and climate projection runs.

Configuration Property	Experiment 1 <i>Model comparison</i>	Experiment 2 <i>Ensemble depth</i>	Experiment 3 <i>Future projection</i>
Data	ERA5	ERA5	CNRM-CM6-1-HR
Period	2014–2023	2014	2050–2059
Simulation Days	3,652	365	3,652
Members per Day	20	20 and 200	20
Evaluated Models	Full, No-scalar, k-NN, U-Net	Full only	Full only
Primary Purpose	Benchmark vs. baselines and validation	Effect of ensemble size	Realism and signal preservation under projection

4.12.1 Experiment 1 - Model Comparison

This is the main comparative experiment, in which the full model is benchmarked against the baseline and ablation models defined in Section 4.10 on the same data. The model is driven by synthetic low-resolution daily mean field derived from ERA5 over the held-out test period 2014 to 2023. For each day an ensemble of 20 members is generated, which corresponds to generating 2 samples for the same conditioning pair every day (10 pairs $\times$ 2 noise draws).

This experiment serves two main purposes. First, because the corresponding hourly ERA5 fields are available as a paired reference, it validates the generated fields directly. Second, applying those same diagnostics to all models on the same dates allows the contribution of each design choice to be isolated and tested.

4.12.2 Experiment 2 - Ensemble Depth

This experiment quantifies how the statistical behavior of the ensemble depends on the number of generated members. The full model is run for a single year (2014) under two identical configurations except for the ensemble size, where we generated a deep ensemble of 200 members per day and a shallow ensemble of 20 members. This allows for the comparison between the two configurations and against the true ERA5 data.

4.12.3 Experiment 3 - Future Projection

The model is applied to its intended target: coarse daily mean temperature fields from the CNRM-CM6-1-HR projection under the SSP2-4.5 scenario, for 2050 to 2059. The CMIP6 fields are bias-corrected with QDM before being passed to the model. An ensemble of 20 members per day is generated (2 noise draws $\times$ 10 control pairs). No paired hourly reference exists in this setting, so the evaluation focuses on distributional and structural realism and physical consistency against ERA5 historical climatology of 2010 to 2024.

5. Results and Discussion

This chapter presents the results of the experiments defined in Section 4.12. The analysis first examines non-stationarity in the ERA5 reference data (Section 5.1), then evaluates the proposed model through historical validation (Section 5.2), ensemble-depth analysis (Section 5.3), and future projection under SSP2-4.5 (Section 5.4).

5.1. Non-Stationarity of the Reference Data

Before evaluating the downscaling models, it is useful to quantify how the reference climate changes between the training period, 1980–2009, and the more recent evaluation period, 2014–2023. This comparison provides context for the model-validation experiments and highlights the difficulty of relying on historical analogues under a changing climate.

Figure 5.1 compares the land-mean diurnal cycle between 1980-2009 and 2014-2023, demonstrating a systematic warming across the domain. The recent period is strictly warmer during both nighttime and daytime hours, exhibiting an average systematic increase of 0.68 °C. A two-sample KS test confirms a significant difference ($D = 0.0352$, $p < 0.001$), with a Wasserstein distance of 0.6562 °C.

However the change cannot be interpreted as a simple uniform scalar displacement of the historical diurnal cycle. Instead, the mean daily cycle changes in shape, indicating modifications in both the amplitude and in the intra-day distribution of warming. An analysis of the discrete hourly rates of change confirms this structural shift. The historical baseline was characterized by abrupt and concentrated morning heating spikes (reaching up to 3.35 °C/h) and highly volatile hourly rates of change (a mean standard deviation of 1.18 °C/h). In the recent decade, this curve has structurally smoothed (mean standard deviation of 1.03 °C/h), exhibiting a more prolonged and sustained accumulation of thermal energy. Morning and early afternoon heating rates in recent years maintain elevated values continuously across a broader temporal window, which directly contributes to the delayed and intensified daily maximums observed.

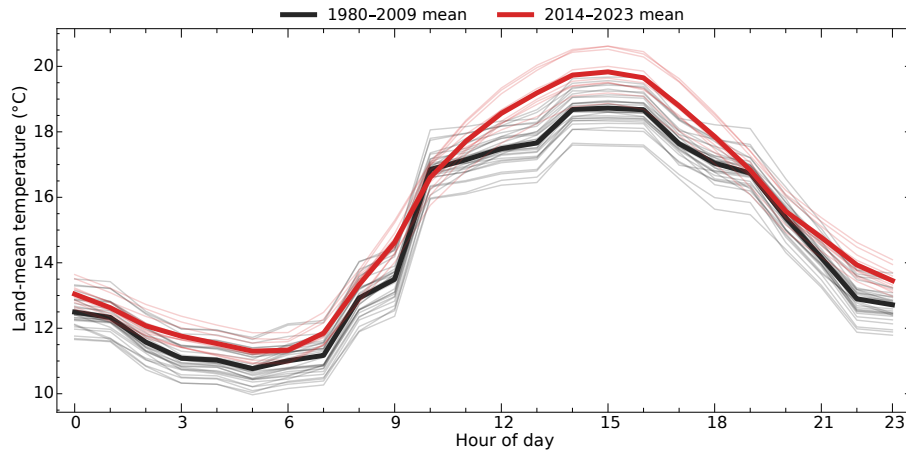


Figure 5.1: Land-mean ERA5 hourly temperature diurnal cycle for 1980-2009 and 2014-2023. Thin lines represent individual annual mean diurnal cycles, while thick lines represent period means.

The timing of daily extrema also changes between the two periods, as shown in Figure 5.2. The most evident shift occurs for daily maximum temperature: in 2014–2023, maxima are more concentrated at 15:00, whereas the training period exhibits a broader mid-afternoon distribution. Daily minimum timing shows the opposite behavior, with the historical peak around 05:00 becoming more dispersed in the recent period. These changes indicate that the sub-daily structure of temperature has evolved, not only its daily mean level.

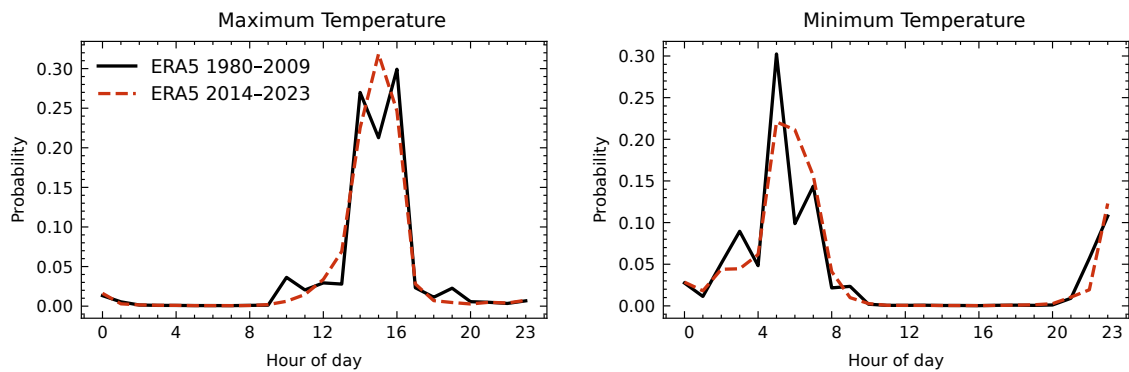


Figure 5.2: Empirical distribution of the hour of day when the daily maximum (left) and minimum (right) temperature occur in ERA5. Black solid lines indicate the training dataset period (1980 to 2009) and red dashed lines represent the test period (2014 to 2023).

In order to characterize the non-stationarity of the data, it is necessary to examine how these diurnal modifications are distributed geographically. Figure 5.3 illustrates the spatial difference in daily temperature amplitude between the historical baseline and the recent decade. The map reveals a distinct land-sea contrast. The smallest amplitude shifts (approaching -0.12°C) are concentrated along the coastline, while the largest intensification

(reaching up to $+1.09\text{ }^{\circ}\text{C}$) occur across the interior land. A two-sample KS test confirms a departure from the historical baseline ($D = 0.06$, $p < 0.001$). Given the large sample sizes, the Wasserstein distance was also calculated and it is equal to $0.5545\text{ }^{\circ}\text{C}$, confirming a fundamental restructuring of the underlying spatiotemporal data.

Beyond shifts in the mean and amplitude, the structural change in the climate is most pronounced at the upper tails of the probability distribution. Evaluating the frequency of extreme events reveals a highly asymmetric shift. Temperatures that exceed the historical 95th percentile ($q > 0.95$) during the training period are now 1.64 times more frequent in the test period. More critically, extreme occurrences exceeding the historical 99th percentile ($q > 0.99$) have multiplied by a factor of 2.22. This disproportionate inflation of the upper tail demonstrates that the recent climate is not merely warmer on average, but inherently more volatile at its upper extremes.

Consequently, downscaling methods that rely on the strict stationarity of historical analogs are inherently constrained, as they are forced to sample from a climatology that lacks these intensified amplitudes, shifted diurnal timings and amplified extremes.

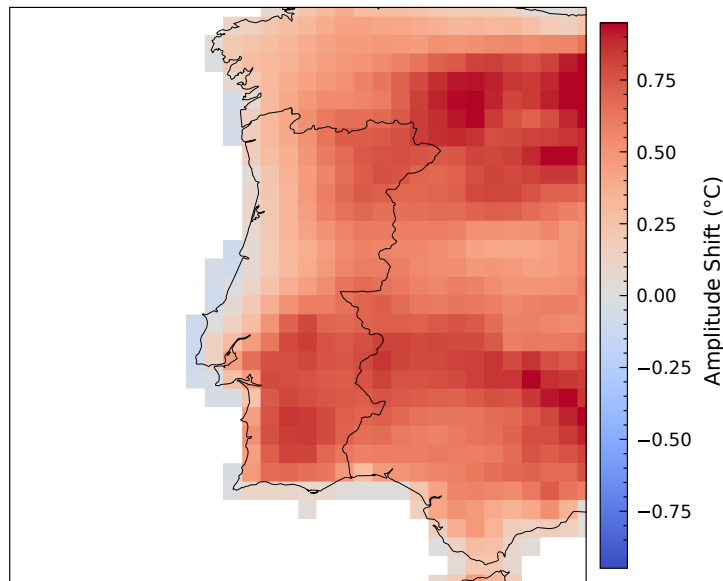


Figure 5.3: The map illustrates the absolute difference in amplitude ($^{\circ}\text{C}$) between the test evaluation period (2014-2023) and the training period (1980-2009).

5.2. Experiment 1: Model Comparison

Following the design of Section 4.12, the **Full Diffusion** model is benchmarked against the **No-Scalar Diffusion** ablation, the **k-NN** analog baseline, and the **Deterministic U-Net** on held-out ERA5 (2014-2023), with the paired hourly reference enabling direct validation.

The objective of this experiment is to evaluate the fundamental trade-offs between point-wise accuracy, probabilistic calibration, and physical realism between the models. The evaluation proceeds in four parts:

- **Deterministic Accuracy and Probabilistic Skill** (Subsection 5.2.1), evaluating standard error metrics and ensemble calibration.
- **Temporal Structure and Sub-daily Variability** (Subsection 5.2.2), evaluating diurnal amplitude, intra-day correlation, and hourly ramp rates.
- **Spatial Structure and Physical consistency** (Subsection 5.2.3), evaluating spatial power spectra, spatial correlation, and orographic lapse rates.
- **Distributional Behavior and Extremes** (Subsection 5.2.4), evaluating the distributional similarity and the structural synthesis of severe summer heat.

It ends with a summary of the discussed results (Subsection 5.2.5).

5.2.1 Deterministic Accuracy and Probabilistic Skill

Table 5.1 summarizes the main deterministic and probabilistic metrics. All models are nearly unbiased over land, with absolute biases close to zero. The deterministic U-Net achieves the lowest ensemble-mean RMSE, equal to 1.007 °C. This is expected because it is trained directly with a squared-error objective and produces a single conditional-mean estimate. However, this deterministic structure prevents it from representing ensemble uncertainty.

Among the stochastic models, both diffusion configurations outperform the k-NN baseline in RMSE and CRPS. The full diffusion model reaches an RMSE of 1.156 °C and a CRPS of 0.642 °C, while the no-scalar model obtains similar values, with RMSE of 1.120 °C and CRPS of 0.641 °C. Both improve the CRPS by approximately 15% relative to k-NN. The k-NN baseline has the largest RMSE and CRPS, but the highest 90% coverage. This

suggests that k-NN retains relatively broad ensemble spread, while the diffusion models provide ensembles with improved deterministic accuracy and probabilistic skill.

Table 5.1: Comparative deterministic and probabilistic skill metrics for all models. Metrics are aggregated over land grid points for the 2014–2023 evaluation period. **Bold** indicates the best performance, and underline indicates the second-best. Arrows in the headers indicate whether higher ($\uparrow$), lower ($\downarrow$), or closer to zero ($\rightarrow 0$) values are better.

Model	Bias (°C) ($\rightarrow 0$)	RMSE (°C) ($\downarrow$)	CRPS (°C) ($\downarrow$)	Coverage 90% (%) ($\uparrow$)	CRPSS ($\uparrow$)
Full diffusion	−0.0062	1.156	<u>0.642</u>	<u>70.7</u>	<u>0.1572</u>
No-scalar diffusion	0.0005	<u>1.120</u>	0.641	61.2	0.1583
k-NN	<u>0.0006</u>	1.354	0.762	76.7	0.0000
U-Net	−0.0103	1.007	—	—	—

Figure 5.4 decomposes RMSE and CRPS by hour of day. Errors are not uniform across the daily cycle: all models show larger errors during transition periods, when temperature changes rapidly, and lower errors during more stable parts of the day. This hourly structure is also consistent with the non-stationarity identified in Section 5.1, as the largest errors occur at hours when the test ERA5 diurnal cycle differs most from the training one. The k-NN baseline has the largest errors across most hours, while the U-Net and diffusion models form a closer group. The two diffusion models have similar hourly CRPS profiles, suggesting that scalar controls do not primarily improve average probabilistic error, but instead affect calibration and variability, as discussed below.

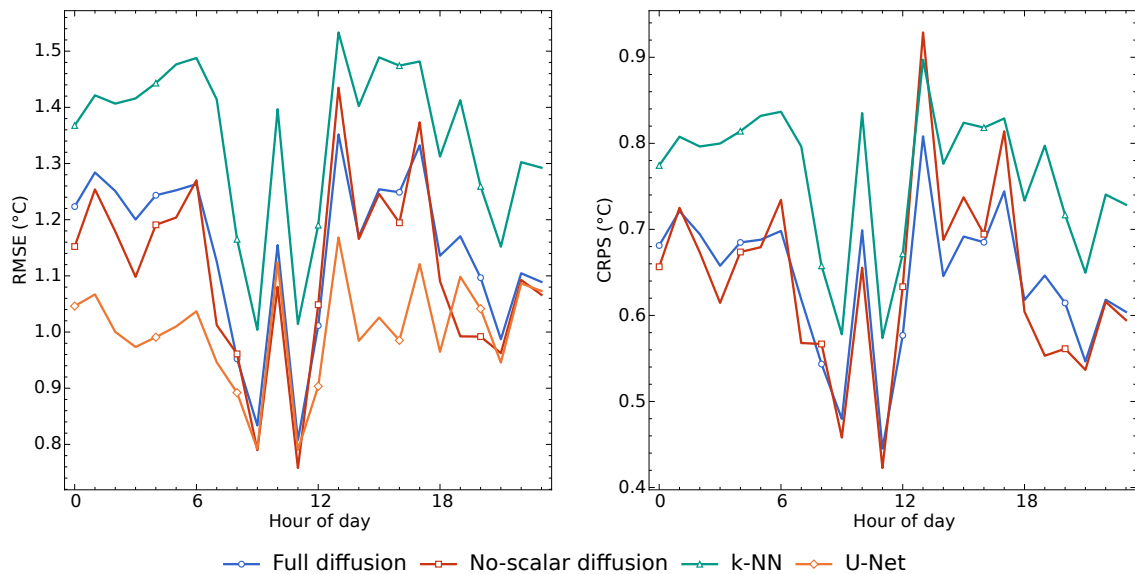


Figure 5.4: Per-hour decomposition of deterministic (RMSE, left) and probabilistic (CRPS, right) errors over land grid points for the 2014–2023 evaluation period.

Figure A.1 in the Appendix illustrates the spatial distribution of the RMSE and CRPS across the study domain. Performance is higher over coastal grid points than across the Iberian landmass for all models. While error magnitudes are spatially heterogeneous, no distinct intra-land spatial error patterns are evident.

Table 5.2 summarizes the circular phase error for the timing of daily maximum and minimum temperature. The deterministic U-Net obtains the lowest phase errors for both extrema, consistent with its conditional-mean behavior. The k-NN baseline has the largest phase errors, suggesting that historical analog timing distributions transfer imperfectly to the recent evaluation period. The two diffusion models perform similarly, with Full diffusion slightly better for $T_{\max}$ timing and No-scalar diffusion slightly better for $T_{\min}$ timing.

Across all models, phase errors are larger for $T_{\min}$ than for $T_{\max}$. This is consistent with the reference-data analysis in Figure 5.2, where recent daily maximum timing becomes more concentrated, while daily minimum timing is more dispersed.

Table 5.2: Comparative time of maximum and minimum phase error metrics for all models. Metrics are aggregated over land grid points for the 2014-2023 evaluation period. **Bold** indicates the best performance, and underline indicates the second-best. Arrows in the headers indicate whether higher ($\uparrow$) or lower ($\downarrow$) values are better.

Model	$T_{\max}$ phase error (h) ($\downarrow$)	$T_{\min}$ phase error (h) ($\downarrow$)
Full diffusion	<u>1.43</u>	2.33
No-scalar diffusion	1.47	<u>2.31</u>
k-NN	1.76	2.77
U-Net	1.19	1.85

Figure 5.5 evaluates the calibration of the ensemble by comparing empirical coverage with the nominal prediction interval. All evaluated models exhibit underdispersion, falling below the diagonal line of calibration, indicating that the generated ensembles do not capture the full extent of observed target variance. The k-NN baseline achieves the highest coverage, as the direct retrieval of historical anomalies inherently preserves spread. The Full diffusion model demonstrates substantially improved calibration over the No-scalar variant. This result isolates the primary functional benefit of the scalar control variables as they successfully drive the generative process to explore a broader, more representative range of sub-daily variability.

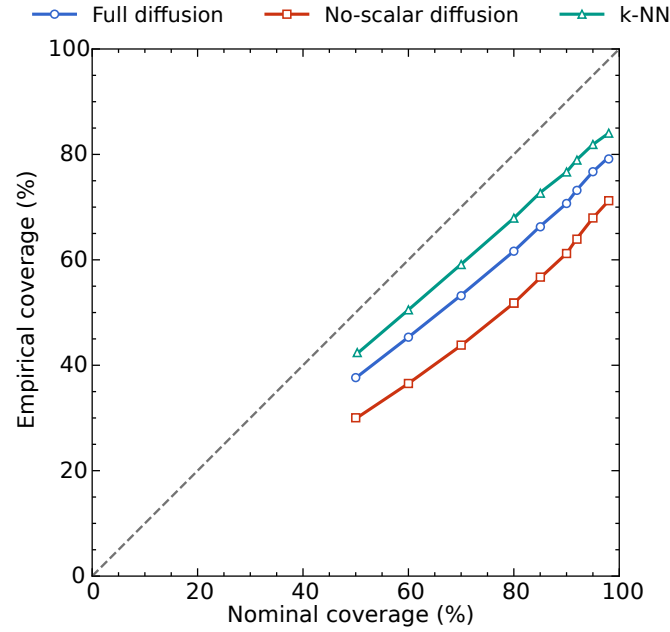


Figure 5.5: Empirical prediction-interval coverage evaluated against nominal coverage thresholds across the land domain for the 2014–2023 evaluation period. The dashed diagonal line indicates perfect statistical calibration.

5.2.2 Temporal Structure and Sub-daily Variability

Figure 5.6 compares the mean diurnal temperature anomaly and pooled variability for winter (DJF) and summer (JJA). All models reproduce the broad phase of the daily cycle, including nighttime cooling, daytime warming, and the afternoon maximum. However, all models underestimate the amplified sub-daily range of the 2014–2023 ERA5 reference,

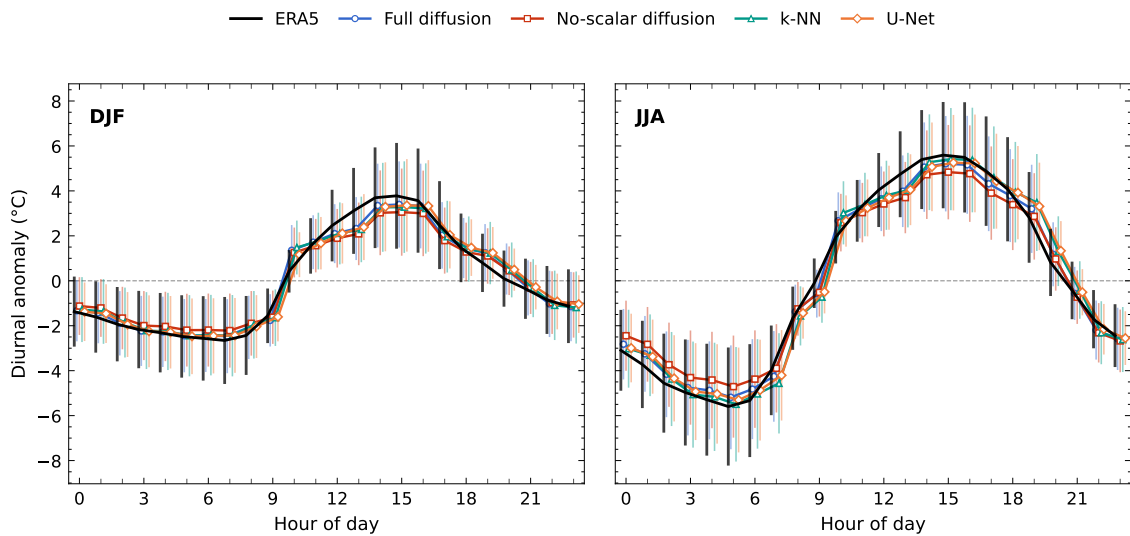


Figure 5.6: Mean diurnal temperature anomaly over land for winter (DJF) and summer (JJA) during 2014–2023 evaluation period. The vertical bars denote ± 1 pooled standard deviation across land points, days, and ensemble members where applicable.

especially in summer. This is consistent with the non-stationarity identified in Section 5.1, where the recent period exhibits stronger amplitudes and more intense warm extremes than the training period.

The No-scalar diffusion model shows the strongest amplitude compression, while the Full diffusion model recovers part of this lost variability. This suggests that the scalar controls help regulate sub-daily amplitude. Nevertheless, the vertical bars show that all stochastic models remain underdispersed relative to ERA5, consistent with the calibration results in Figure 5.5.

Figure 5.7 evaluates the within-day ACF of sub-daily temperature anomalies. All models reproduce the dominant diurnal dependence, with positive correlations at short lags, negative correlations near the half-day scale, and a return toward positive correlation close to the daily boundary. The k-NN baseline remains close to ERA5, as expected from retrieving historical anomaly sequences.

The main differences appear at longer lags. The deterministic U-Net shows the largest positive error between approximately 17 and 23 h, indicating overly persistent boundary states within the daily block. The diffusion models reduce this long-lag overcorrelation, with the Full diffusion model showing slightly smaller errors than the No-scalar variant. This suggests that scalar controls help preserve more realistic intra-day coherence.

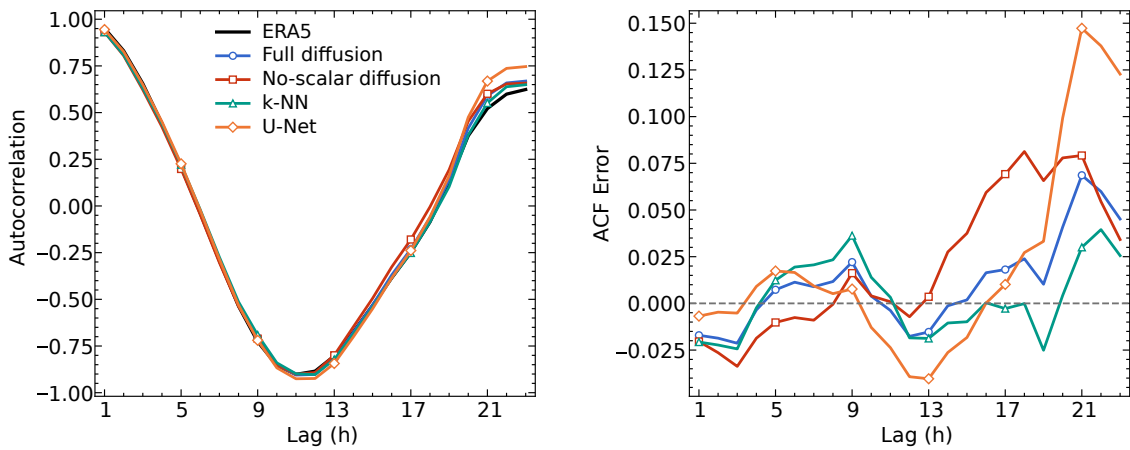


Figure 5.7: Within-day autocorrelation function of daily-mean temperature anomalies over land. The left panel shows the Autocorrelation Function (ACF) for lags 1-23 h, and the right panel shows model errors relative to ERA5 testing period (2014-2023). All models reproduce the dominant diurnal coherence, while extended lag errors indicate differences in how realistically each model preserves correlations across the daily-cycle boundaries

Table 5.3 evaluates hourly temperature-ramp skill. The diffusion models achieve the lowest Ramp- CRPS in both winter and summer, indicating better conditional skill for hour-to-hour temperature transitions. The Full diffusion model also obtains the lowest Ramp-W1 in both seasons, suggesting the closest distributional representation of sub-daily heating and cooling rates.

The deterministic U-Net has the weakest Ramp- CRPS, consistent with temporal smoothing, even though its marginal ramp variability is relatively close to reference in summer. The k-NN baseline performs competitively in winter but degrades in summer, where it overestimates ramp variability. This suggests that historical analogue ramps are less reliable under recent summer conditions. Overall, the Full diffusion model provides the best balance between conditional ramp skill and distributional realism.

Table 5.3: Comparative seasonal hourly temperature ramp skill and variability metrics for all models. Metrics evaluate within-day temperature changes aggregated over land grid points, summarized separately for winter (DJF) and summer (JJA). Ramp-CRPS evaluates conditional probabilistic skill against ERA5 hourly changes (reducing to absolute error for the deterministic U-Net), while Ramp-W1 uses the 1-Wasserstein distance to assess the overall distributional realism of sub-daily heating and cooling rates. The standard deviation ratio (Std ratio model/ERA5) evaluates the overall magnitude of generated hourly-ramps variability relative to ERA5. **Bold** indicates the best performance, and underline indicates the second-best. Arrows in the headers indicate whether lower ($\downarrow$) or closer to one ($\rightarrow 1$) values are better.

Model	Winter (DJF)			Summer (JJA)		
	Ramp-CRPS	Ramp-W1	Std ratio	Ramp-CRPS	Ramp-W1	Std ratio
	($\downarrow$)	($\downarrow$)	($\rightarrow 1$)	($\downarrow$)	($\downarrow$)	($\rightarrow 1$)
Full diffusion	0.3493	0.0928	1.089	0.4055	0.1878	1.077
No-scalar diffusion	<u>0.3496</u>	0.1144	1.030	<u>0.4115</u>	0.1938	<u>1.035</u>
k-NN	0.3733	<u>0.0991</u>	1.124	0.4592	0.2440	1.186
U-Net	0.4320	0.1495	<u>0.960</u>	0.5120	<u>0.1924</u>	1.021

To illustrate the behavior of the Full diffusion model ensemble, Figure 5.8 shows a one-week hourly trajectory at a single land grid point. The ensemble contains the ERA5 truth across most of the week, tracing the day-to-day variation in amplitude and phase, with occasional under-coverage on individual days consistent with the underdispersion identified. The trajectories are continuous within each day but break at day boundaries, a direct consequence of the model generating independent daily blocks.

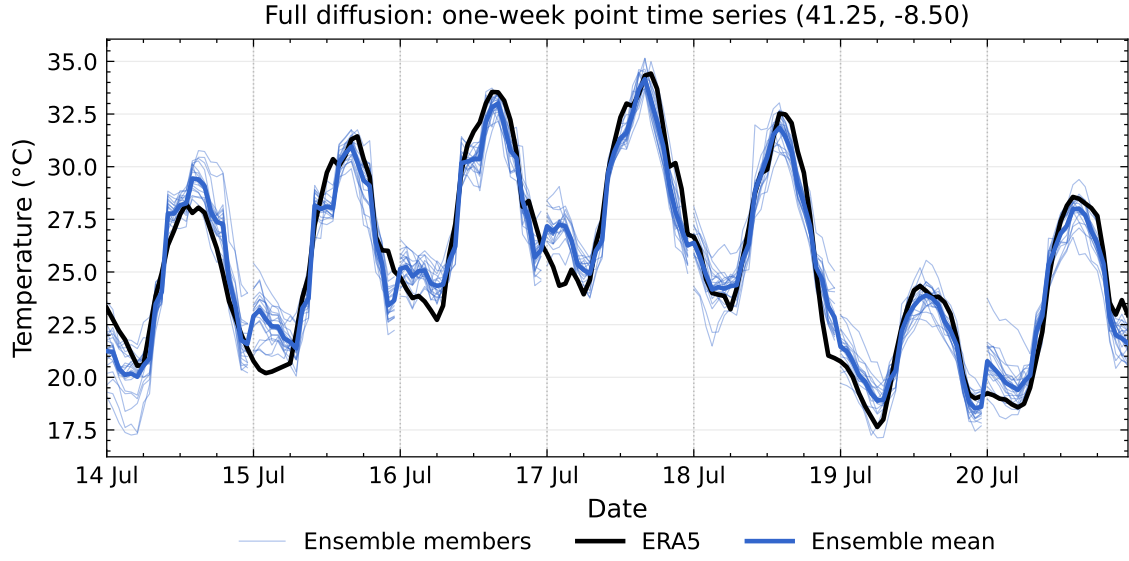


Figure 5.8: One-week hourly temperature trajectory at a single land grid point (41.25°N, 8.50°W) for the Full diffusion ensemble during July 2020, with the ERA5 truth and ensemble mean. Lines are broken at day boundaries, since each day is generated independently.

5.2.3 Spatial Structure and Physical Consistency

Figure 5.9 evaluates the spatial PSD of temperature residuals relative to the coarse conditioning field. All models broadly reproduce the decay of spectral power with increasing wavenumber, but the ratio plot reveals scale-dependent differences. The k-NN baseline is the closest to ERA5 over much of the spectrum, although its ratio remains slightly below zero. This small deficit is expected, since the anomalies retrieved by the model

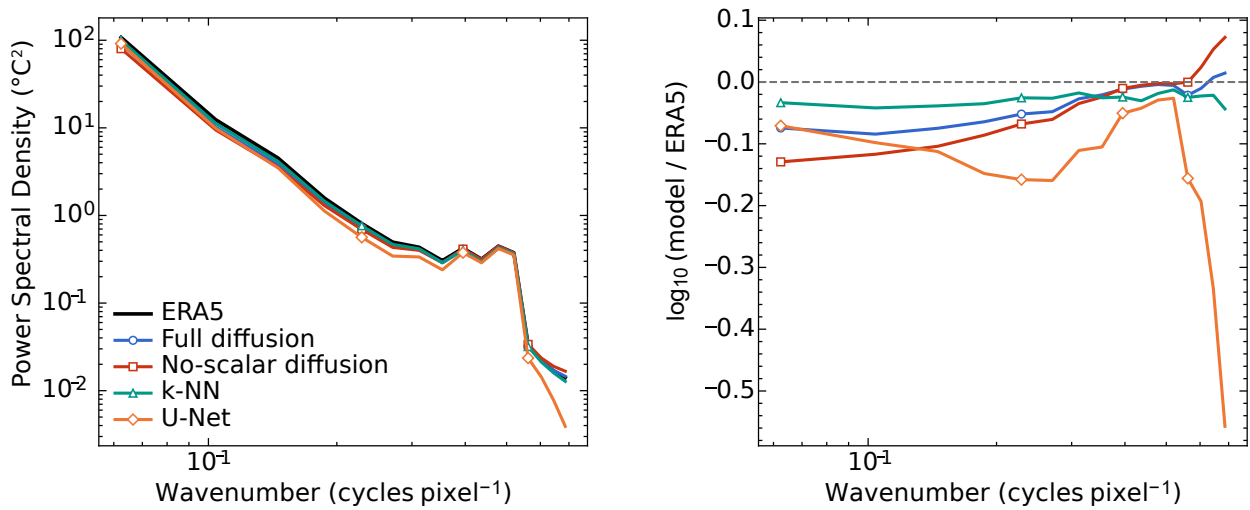


Figure 5.9: Spatial PSD of temperature anomalies across all models. The left panel shows the radially averaged PSD, and the right panel shows the logarithmic model-to-ERA5 ratio. Negative ratios indicate missing spatial variance relative to ERA5, while positive ratios indicate excess spectral energy.

are drawn from historical ERA5 data. Conversely, the U-Net exhibits a pronounced energy deficit across all scales, culminating in a sharp collapse at highest wavenumbers. This is consistent with excessive spatial smoothing and difficulty in generating fine-scale variability.

The diffusion models successfully recover the high-frequency variance suppressed by the U-Net, but diverge at the finest scales. The No-scalar model exhibits a pronounced negative ratio at low-to-intermediate wavenumbers, followed by a positive overshoot at the highest frequencies. This suggests that removing the scalar control variables does not have a uniform effect, as it leads to a coarse-scale smoothing coupled with artificial high-frequency inflation. The Full diffusion framework seems to mitigate this instability, remaining closer to ERA5 overall.

Although k-NN marginally outperforms the other models across most wavenumbers, the Full diffusion model provides the most resilient spatial generation among the neural models. It avoids the high-wavenumber collapse of the U-Net and the overshoot of the No-scalar ablation, while remaining close to the ERA5 spectrum across most spatial scales.

Figure 5.10 evaluates pairwise spatial-correlation errors over land, with numerical summaries in Appendix Table A.1. The deterministic U-Net shows the largest positive correlation errors, consistent with overly smooth fields that are too spatially uniform. The No-scalar diffusion model shows a negative bias, indicating weaker spatial coherence than ERA5. By contrast, both the k-NN baseline and the Full diffusion model are centered close to zero error. Full diffusion also achieves the lowest overall pairwise RMSE, suggesting that scalar controls help regularize the spatial organization of the generated fields.

Figure 5.11 evaluates orographic consistency by assessing mean temperature (left panel) and bias (right panel) across elevation. The left panel indicates that all models successfully reproduce the large-scale decrease in temperature with elevation, tracking closely to the ERA5 reference. For the neural models, this suggests that the learned downscaling does not disrupt the orographic structure imposed by the reference data. The k-NN baseline achieves similar consistency inherently by retrieving historical fields governed by real-world physics.

The right panel isolates the structural biases, which remain within a negligible margin with absolute bias below 0.02 °C. All the models show a small warm bias in the lowest bin,

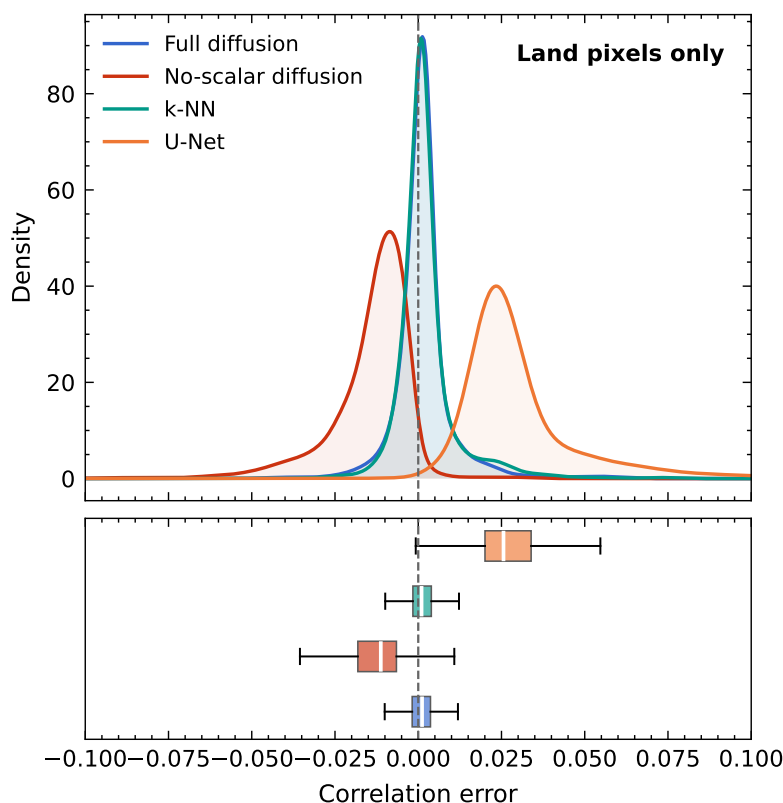


Figure 5.10: Distribution of pairwise spatial correlation errors over land grid points relative to ERA5. The upper panel shows the density of correlation errors for each model, and the lower panel summarizes the same distribution using boxplots.

followed by a transition to a weak cold-bias. The deterministic U-Net exhibits a slight cold bias at higher elevations (> 600 m). The Full diffusion model maintains a relatively stable cold bias across the profile. In contrast, both the No-scalar model and the k-NN oscillate closer to zero bias at higher elevations. Overall, the scale of these deviations confirms

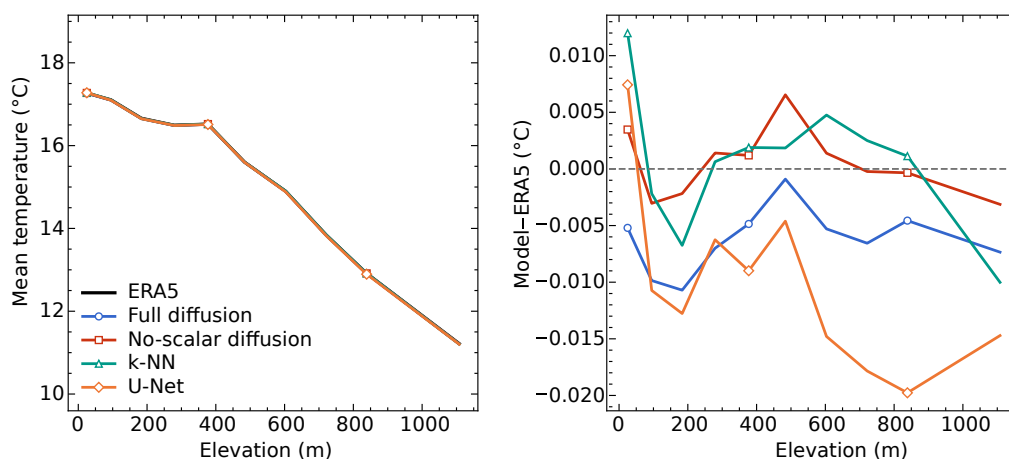


Figure 5.11: Mean temperature as function of elevation over land for ERA5 and the downscaling models during 2014–2023. Left panel shows the temperature in each elevation bin, while right panel shows the corresponding model bias relative to ERA5.

that all models successfully preserve the thermodynamic relationship between altitude and temperature.

5.2.4 Distributional Behavior and Extremes

Figure 5.12 compares the seasonal distributions of sub-daily temperature amplitudes over land. The ERA5 reference shows broad, right-skewed distributions in all seasons, reflecting frequent moderate amplitudes and fewer high-amplitude events. All models underestimate the upper tail of the amplitude distribution, as indicated by negative P_{95} errors reported in Appendix A, Table A.2.

Among the neural models, Full diffusion provides the strongest representation of the central amplitude distribution. It reduces the median bias and W1 distance relative to the No-scalar ablation in every season, showing that scalar controls improve the representation of sub-daily variability. The improvement is especially clear in summer, where the median bias improves from -1.906 °C for No-scalar diffusion to -0.974 °C for Full diffusion. However, Full diffusion still underestimates the highest amplitudes, with P_{95} errors between approximately -1.6 °C and -1.9 °C.

The k-NN baseline achieves the lowest W1 distance in all seasons, reflecting the advantage of retrieving historical anomaly patterns. Overall, Full diffusion does not fully match the marginal amplitude distribution of the analogue baseline, but it provides the best balance among the neural models and substantially improves over the No-scalar diffusion variant.

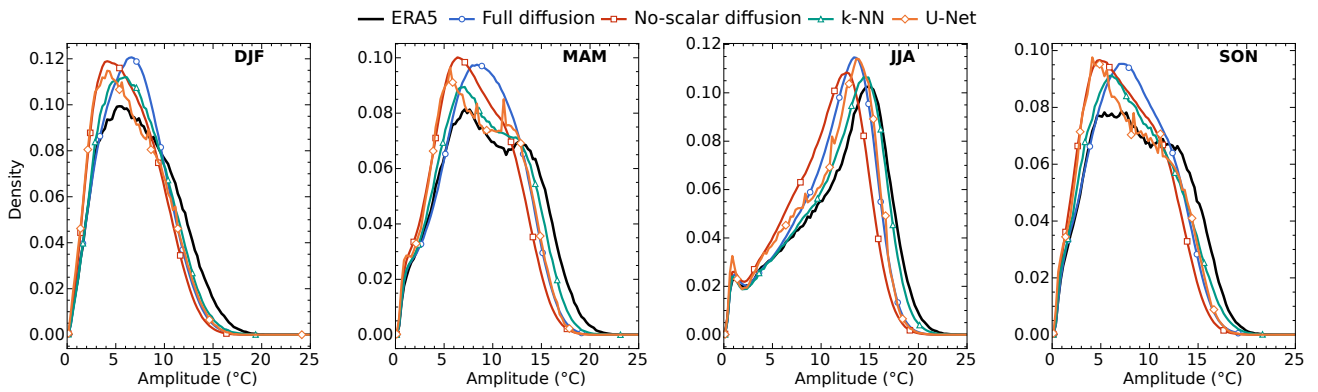


Figure 5.12: Seasonal probability density of sub-daily temperature anomaly amplitudes. The distributions evaluate the capacity of each model to synthesize realistic temperature variance relative to ERA5 reference (2014–2023).

To evaluate the synthesis of heat extremes independent of diurnal phase, the analysis uses the daily maximum temperature ($T_{\max}$) for the extended summer season (June–September). The values are pooled across test-period days, land grid points, and ensemble members.

While k-NN has the closest distributional approximation to ERA5 via 1-Wasserstein distance (0.23 °C), the Full diffusion achieves the smallest distance (0.44 °C) among the neural models. All the distances between the models and ERA5 are presented in Appendix A.

Figure 5.13 presents quantile bias at P_{95} , P_{99} and $P_{99.9}$. All models exhibit a cold bias that grows with percentile, with the k-NN baseline showing the smallest deficit at every threshold. At P_{95} and P_{99} U-Net and Full diffusion show comparable magnitudes. However, at $P_{99.9}$ Full diffusion outperforms U-Net (-0.96 °C vs -1.06 °C). No-scalar diffusion performs worst across all percentiles, with a bias of -1.40 °C at $P_{99.9}$, suggesting that the scalar control variables help the model generate extreme events.

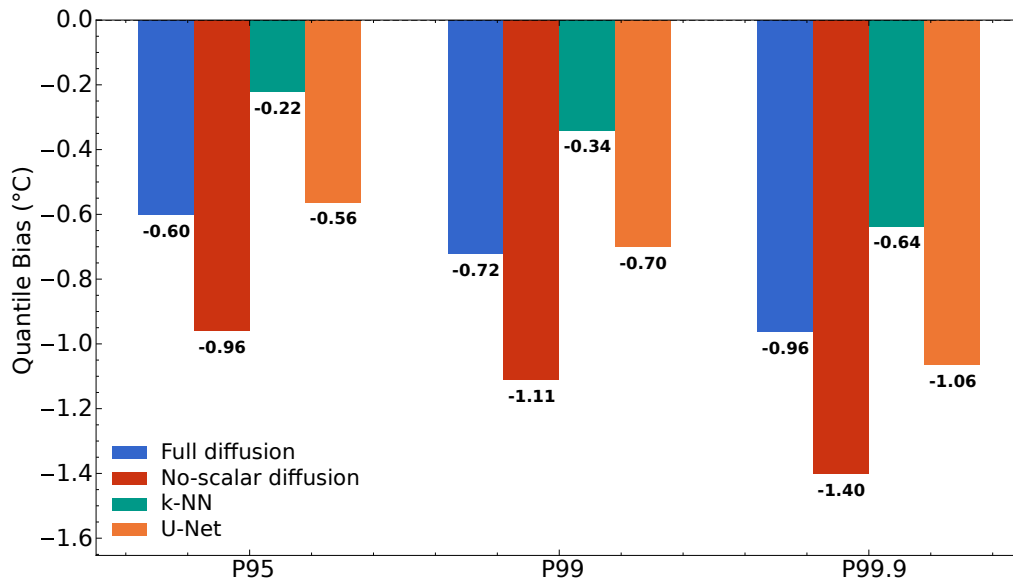


Figure 5.13: Quantile bias of pooled summer (JJAS) daily maximum temperatures over land grid points for the 2014–2023 evaluation period for all models. The bias is calculated relative to the ERA5 reference at the 95th, 99th, and 99.9th percentiles.

Figure 5.14 reports fixed-threshold exceedance rates at 35 °C and 40 °C, expressed as a percentage of the ERA5 rate. All models underproduce exceedances at both thresholds, and the deficit increases at 40 °C. This is consistent with the cold upper-tail bias: when the hottest values are shifted downward, fewer samples exceed fixed heat thresholds. The k-NN baseline remains closest to ERA5, while the No-scalar diffusion model is weakest.

Full diffusion and U-Net are similar at 40 °C, although the quantile-bias results indicate that Full diffusion better preserves the most extreme tail among the neural models.

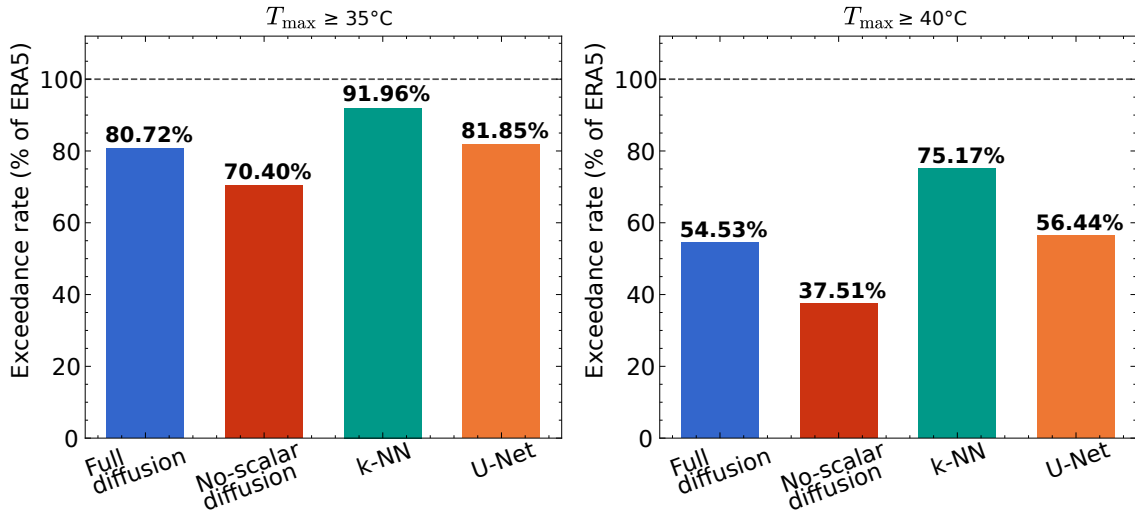


Figure 5.14: Exceedance rates of pooled summer (JJAS) daily maximum temperatures over land grid points for the 2014–2023 evaluation period, evaluated at fixed thresholds of 35°C (left) and 40°C (right) for all models. Values are expressed as a percentage of the ERA5 reference occurrences, with the dashed horizontal line indicating perfect frequency reproduction.

Table 5.4 isolates the magnitude of severe heat events by computing the 1-Wasserstein distance only among land-pixel exceedances above 40 °C. While k-NN baseline achieves the lowest distance (0.184 °C) by retrieving historical extremes, Full diffusion yields the closest structural match among the neural architectures (0.224 °C), markedly outperforming both the U-Net and the No-scalar ablation, which are themselves similar to each other. The shared trait between the two models is the absence of scalar conditioning, suggesting that, for this tail-restricted measure, scalar conditioning is what separates Full diffusion from the other two neural models.

Comparing this tail-restricted distance to the bulk-distribution distance reported in Appendix Figure A.2 reveals a further asymmetry. For every neural model, the tail-restricted

Table 5.4: 1-Wasserstein distance between each model's pooled JJAS daily maximum temperature exceedances above 40 °C and the corresponding ERA5 exceedances, land grid points, 2014-2023.

Source	Wasserstein distance (°C) (↓)
k-NN	0.184
Full diffusion	0.224
U-Net	0.285
No-scalar diffusion	0.299

distance is a substantially smaller fraction of the model's own bulk distance (51% for full diffusion, 54% for U-Net and 38% for No-scalar diffusion) than is the case for k-NN (80%). This indicates that, for neural models, distributional disagreement with ERA5 is comparatively smaller among the most severe heat events than across the fuller range of daily maxima, which includes more moderate hot days as well. k-NN shows no such asymmetry, as its distance among severe events tracks its distance across the full range in roughly the same proportion, indicating no comparative concentration of error at either end of the daily-maximum spectrum.

Taken together, the distributional and extreme-temperature diagnostics show a consistent upper-tail deficit. All models underestimate the most severe JJAS daily maximum temperatures and underproduce exceedances above 35 °C and 40 °C. The k-NN baseline remains strongest for marginal and tail distributions because it retrieves physically realistic historical patterns. Among the neural models, Full diffusion provides the most balanced representation: it improves the amplitude distributions relative to the No-scalar ablation, reduces the most severe quantile bias relative to U-Net, and achieves the lowest neural tail-restricted Wasserstein distance. Nevertheless, upper-tail compression remains a central limitation.

5.2.5 Summary

Figure 5.15 illustrates the output of the Full diffusion model for a representative summer day. The ensemble members share the same large-scale daily structure imposed by the conditioning field while differing in their sub-daily evolution and fine-scale spatial detail. The land-mean diurnal cycle shows the members distributed around the ERA5 truth.

Overall, Experiment 1 shows that each modeling strategy exhibits a distinct trade-off. The deterministic U-Net achieves the strongest pointwise accuracy, but its deterministic structure limits its capacity to represent probabilistic variability and produces smoother temporal and spatial behavior. The k-NN baseline remains highly competitive in marginal distributional diagnostics, particularly for amplitude and hot extremes, because it retrieves physically consistent historical fields. However, its stochastic representation is constrained by the analogue bank and the choice of K . Reducing K increases conditional similarity but limits ensemble diversity, while increasing K broadens the ensemble at the cost of retrieving less similar analogues and generating less plausible scenarios. The No-scalar

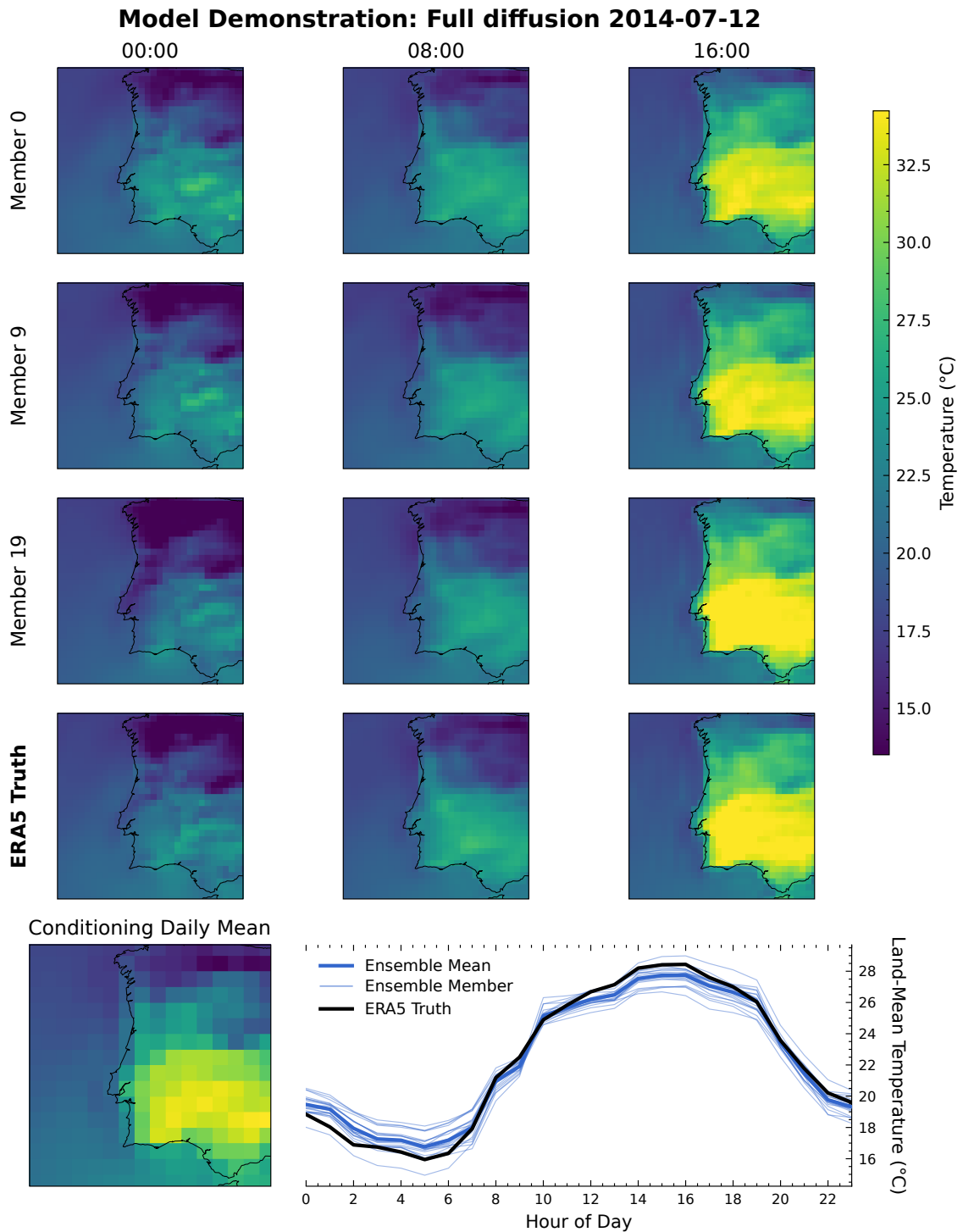


Figure 5.15: Example Full diffusion generations for a representative summer day (2014-07-12). The first three rows show three ensemble members at 00:00, 08:00, and 16:00; the fourth row shows the corresponding ERA5 truth, and the bottom-left panel shows the coarse conditioning daily-mean field. The bottom-right panel shows the land-mean diurnal cycle of the full ensemble against the ERA5 truth.

diffusion model demonstrates that removing scalar controls weakens calibration, amplitude representation, spatial organization, and extreme-event behavior. In contrast, Full diffusion provides the most consistent balance among the neural generative models. It improves calibration relative to the ablated variant, better preserves temporal and spatial structure, and provides the strongest neural representation of sub-daily transitions and hot-tail behavior.

The generation cost reflects the same trade-off seen in the quality metrics. Both diffusion models require, on average, roughly 0.14 s per generated 24-hour daily block due to their iterative sampling, which is about $18\times$ the cost of the k-NN inference (0.008 s) and substantially more than a single-pass deterministic U-Net (0.06 s). Training the diffusion models is also demanding, taking roughly 113 h for the full diffusion and 95 h for the No-scalar variant over 500k steps on a single NVIDIA RTX 3090. The deterministic U-Net was trained with early stopping and converged in approximately 6 h. The expense is the price of the diffusion model's controllability and scenario diversity, and remains modest in absolute terms, being readily parallelized across days and members.

5.3. Experiment 2: Ensemble Depth

This section isolates the effect of ensemble depth on the Full diffusion. For the year 2014, the generative behavior is compared between a shallow ensemble ($K = 20$ members per day) and a deep ensemble ($K = 200$ members per day), holding the model and the daily control pairs fixed so that the only variable is the number of stochastic draws per control pair.

The objective of this experiment is to determine how ensemble depth affects the results, and to verify how the model responds to its conditioning scalars. Structural properties that remain essentially invariant to ensemble size, such as spatial and temporal structure were excluded from this analysis. The evaluation proceeds in three parts:

- **Ensemble Calibration and Dispersion** (Subsection 5.3.1), evaluating how close the generated ensemble is to the reference and how its spread responds to ensemble size.
- **Distributional Behavior and Extremes** (Subsection 5.3.2), evaluating the representation of severe summer heat and how the ensemble depth influences it.

- **Conditioning Controllability** (Subsection 5.3.3), evaluating whether the imposed scalar controls exert their intended influence on the generated fields.

It ends with a summary of the discussed results (Subsection 5.3.4).

5.3.1 Ensemble Calibration and Dispersion

Figure 5.16 shows the rank distribution of the ERA5 reference within the generated ensembles. Both configurations exhibit a pronounced U-shape, with peak frequencies at the extreme ranks. This indicates underdispersion, as the observation falls at or beyond the ensemble bounds more often than expected. The U-shape persists at both ensemble sizes, suggesting it is not a small-sample artifact, but instead is a property of the learned distribution. Both ensembles show the interior ranks to be roughly flat but lying below the expected frequency, with the missing mass concentrated at the outermost ranks. The ensembles are therefore correctly centered, and the deficit is related to spread alone.

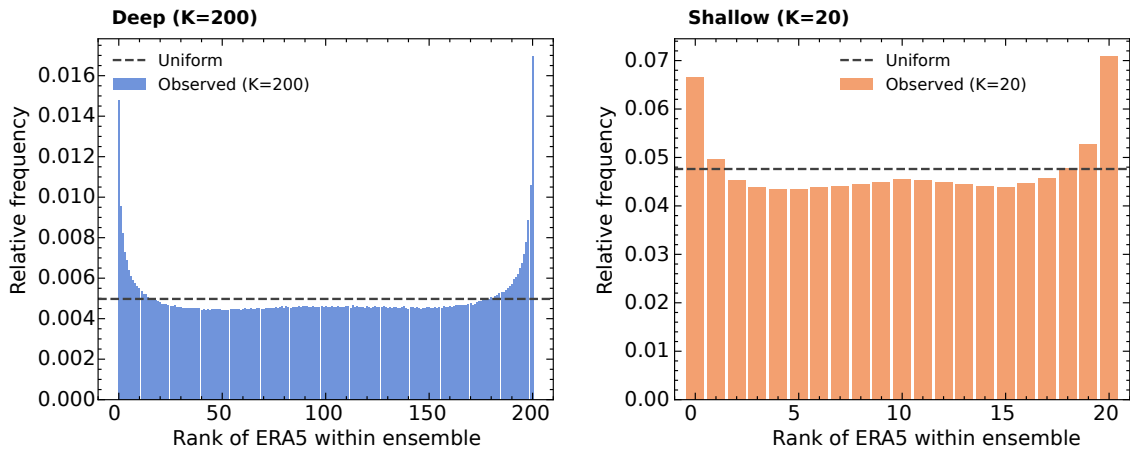


Figure 5.16: Rank histograms evaluating the position of the ERA5 reference observations within the sorted generated ensembles over land grid points for the year 2014. The left panel displays the deep configuration ($K = 200$) and the right panel displays the shallow configuration ($K = 20$). The dashed horizontal line indicates the theoretical uniform distribution expected from a perfectly calibrated model. The pronounced U-shape in both panels confirms a structural underdispersion that persists independent of ensemble depth.

Figure 5.17 evaluates the ensemble calibration and corroborates this result. Both ensembles fall below the diagonal of perfect calibration at every nominal level assessed. The deep ensemble shifts the empirical coverage curve upwards. For nominal coverage of 50% the shallow ensemble reaches only 42.47 % while the deep ensemble reaches 45.23 %; for a nominal coverage of 90% the values are 76.75% and 83.01% for the shallow and deep ensemble respectively. The improvement reflects the larger ensemble sampling the

model's variability more completely, which leads to a more accurate estimation of the interval edges. The persistence of substantial undercoverage at $K = 200$ suggests that the conservative spread is not only a consequence of limited sampling, but reflects a narrow predictive distribution. Separating the sampling and model contributions to undercoverage would require a finer sweep of ensemble sizes, which is left for future work.

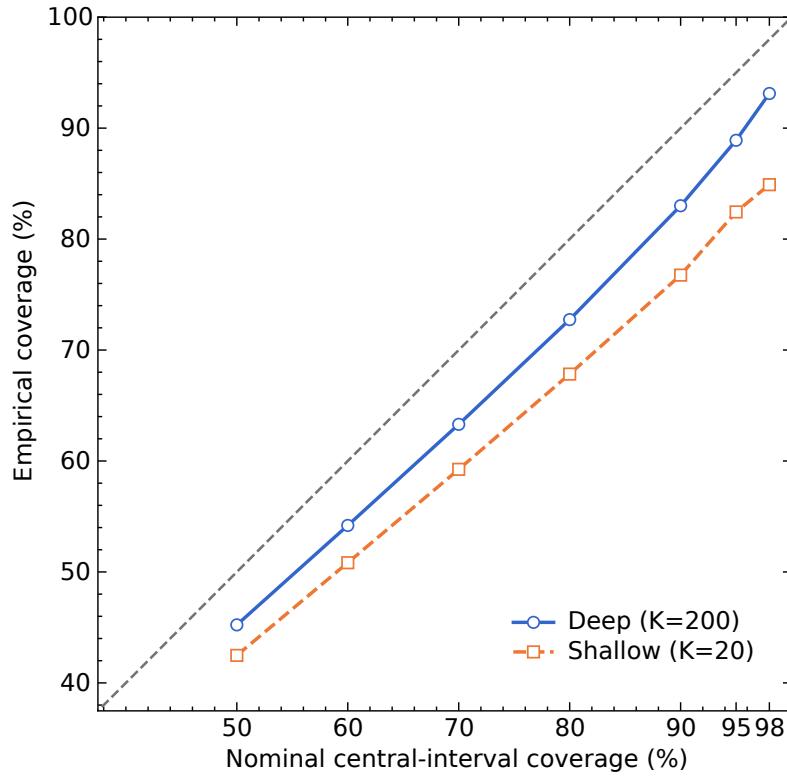


Figure 5.17: Probabilistic calibration of the shallow ($K = 20$) and deep ($K = 200$) generated ensembles for Full diffusion. Empirical interval coverage is evaluated against nominal coverage thresholds across the land domain for the 2014 evaluation year. The dashed diagonal line indicates perfect statistical calibration.

Table 5.5 summarizes deterministic and probabilistic metrics for the two ensemble sizes. The bias and RMSE are nearly unchanged, confirming that increasing K does not substantially alter the ensemble mean. The main differences appear in probabilistic quantities: the deep ensemble has lower CRPS, higher mean ensemble spread, and better empirical coverage.

These changes indicate that an ensemble with more scenarios better characterizes the model distribution, but does not fundamentally change the learned behavior. The $K = 200$ configuration is therefore preferable when estimating distributional quantities, while $K = 20$ remains a useful lower-cost configuration for representative scenario generation.

Table 5.5: Deterministic and probabilistic skill metrics evaluated over land grid points for the year 2014, comparing the shallow ($K = 20$) and deep ($K = 200$) configurations. Mean ensemble spread denotes the standard deviation across ensemble members, averaged over days, hours, and land points.

Configuration	Bias (°C) ($\rightarrow 0$)	RMSE (°C) ($\downarrow$)	CRPS (°C) ($\downarrow$)	Mean ensemble spread (°C)
Deep ($K = 200$)	-0.0057	1.0059	0.5420	0.8667
Shallow ($K = 20$)	-0.0056	1.0149	0.5581	0.8481

5.3.2 Distributional Behavior and Extremes

Figure 5.18 compares the shallow and deep ensemble through the upper-tail quantile bias and exceedance rate of the JJAS daily maximum temperature ($T_{\max}$), pooled over land grid points, days, and ensemble members. Unlike in Experiment 1, 40 °C threshold was not included because in year 2014 only a single day exceeds this threshold which would cause a very unstable estimation. Therefore we only evaluate the 35 °C threshold. Both ensembles exhibit the same cold-tail bias that deepens toward the rarest percentiles, with effectively identical values at P_{95} and P_{99} . A small divergence occurs at $P_{99.9}$, where the deep ensemble shows a slightly smaller deficit than the shallow ensemble. This is coherent with the fact that with $K = 200$ the extreme margins of the distribution are more explored than with $K = 20$ samples. The exceedance rates reinforce this near-invariance,

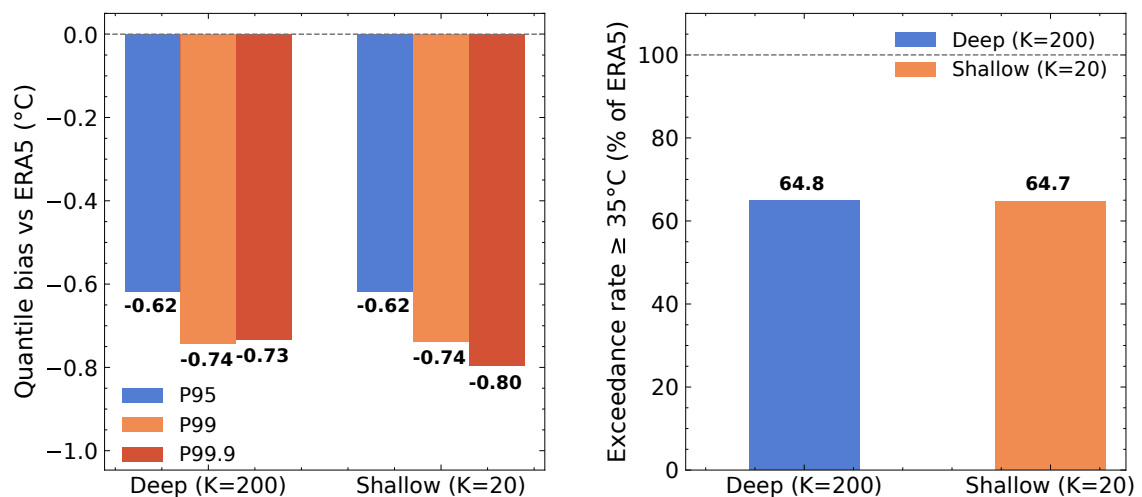


Figure 5.18: Evaluation of summer (JJAS) daily maximum temperature extremes over land for the 2014 evaluation year for deep ($K = 200$) and shallow ($K = 20$) ensembles. Left shows the quantile bias (generated minus ERA5) at the 95th, 99th, and 99.9th percentiles. Right panel shows the exceedance rate above 35°C, expressed as a percentage of the ERA5 reference exceedance rate.

with both configurations reproducing the same fraction of ERA5 exceedance rate above 35 °C. Since the deep ensemble draws ten times as many samples yet allocates the same relative mass to the warm tail the deficit is not a finite-sample artifact, but a property of the model.

The 1-Wasserstein distances in Table 5.6 agree. Over the full extended summer $T_{\max}$ distribution the two ensembles are almost indistinguishable, with the shallow ensemble marginally closer to ERA5. On the other hand, above the 35 °C threshold, the deep ensemble is marginally closer. These differences are too small to indicate a real performance gap.

In summary, deeper ensembles improve the empirical characterization of the generated extremes but do not remove the upper-tail compression identified in Experiment 1. The $K = 200$ ensemble is preferable for estimating rare-event distributions as it provides far more tail samples. However, the hot-tail deficit is a limitation of the learned distribution rather than of the sample size.

Table 5.6: 1-Wasserstein (W1) distances between the generated ensembles and the ERA5 reference for pooled JJAS daily maximum temperatures over land (2014). The metric is evaluated across the full seasonal distribution and strictly isolated to the extreme tail ($> 35^\circ\text{C}$).

Configuration	W1 Full JJAS (°C) (↓)	W1 Tail $> 35^\circ\text{C}$ (°C) (↓)
Deep ($K = 200$)	0.2708	0.147
Shallow ($K = 20$)	0.2658	0.153

5.3.3 Conditioning Controllability

Table 5.7 assesses whether the Full diffusion model responds to its two conditioning scalars, the daily-temperature r_d and the sub-daily variability control s_d , evaluated on both ensemble products. The daily-temperature control is a summary of the conditioning field itself, and since generated daily mean is anchored by that same conditioning, the realized r_d reproduces the imposed value almost exactly. This reflects the fidelity of the conditioning pathway rather than an independent control, as r_d was not varied independently of the conditioning in the present generation. An analysis of these values is left for future work.

The sub-daily variability control s_d is selected for each day from regime-matched historical

Table 5.7: Controllability of the sub-daily variability (s_d) and daily-temperature (r_d) scalar controls. For s_d , metrics are computed within each day (removing the daily mean) to isolate the model’s response from the temperature regime, and are shown for both ensemble sizes. For r_d , the realized value reproduces the imposed value through the conditioning pathway. The table reports the Pearson correlation (r), Spearman rank correlation (ρ), the coefficient of determination (r^2), and the response slope between the imposed control and its realized effect.

Control	Pearson (r)	Spearman (ρ)	r^2	Slope
s_d ($K = 20$)	0.91	0.92	0.83	1.01
s_d ($K = 200$)	0.91	0.91	0.82	1.00
r_d	~ 1.00	~ 1.00	~ 1.00	~ 1.00

analogues, so the imposed s_d is correlated with the daily temperature state by construction. To isolate the model’s response to the control from this underlying regime relationship, s_d is assessed within each day, removing the daily mean from both the imposed and realized values so that the temperature regime is held fixed and only the control varies. Under this analysis the generated diurnal variability tracks the imposed s_d strongly, with a Spearman correlation of $\rho \approx 0.91$, a coefficient of determination of $r^2 \approx 0.82$, and a response slope of ≈ 1.0 , meaning that within a day, a given increase in the imposed control produces an equal increase in the realized variability. The relationship is positive in every day of the evaluation year for both ensemble sizes, and the metrics are essentially identical at $K = 20$ and $K = 200$, confirming that responsiveness to the control is a property of the model and not a property of ensemble size.

5.3.4 Summary

In summary, ensemble depth affects the probabilistic characterization of the generated fields but not their underlying behavior. Increasing K from 20 to 200 improves coverage, lowers CRPS, and yields more stable estimates of the rare-event tail, while leaving the ensemble mean, the calibration shape, and the extreme-heat deficit essentially unchanged. The persistent underdispersion and upper-tail compression at $K = 200$ confirms that these are properties of the learned distribution rather than artifacts of finite sampling. The scalar controls remain effective regardless of ensemble size, with the sub-daily variability control tracking its imposed target almost identically at both depths. The deep ensemble is therefore preferable to characterize the full distribution or to sample rare events. The shallow ensemble remains adequate when a smaller set of representative scenarios is sufficient, since the underlying behavior is almost unchanged by ensemble size.

The per-sample generation cost is invariant to ensemble size. Therefore, the generation of $K = 200$ members costs roughly ten times more than generating $K = 20$. The cost scales linearly with K .

5.4. Experiment 3: Future Projection

This section applies the Full diffusion model to its intended target: the coarse daily-mean temperature fields from the CNRM-CM6-1-HR CMIP6 projection under the SSP2-4.5 scenario for 2050–2059, bias corrected with QDM. An ensemble of 20 members per day is generated. Unlike previous experiments, no paired hourly reference exists for the future period. Therefore, the evaluation assesses whether the model continues to generate realistic and physically plausible fields. The generated fields are compared against recent ERA5 data (2010–2024), which serves as a reference for realistic temperature behavior. The evaluation proceeds in three parts:

- **Conditioning Consistency** (Subsection 5.4.1), verifying that the generated fields reproduce the imposed daily-mean condition under projection inputs.
- **Distributional and Tail Plausibility** (Subsection 5.4.2), evaluating whether the seasonal temperature and summer hot-tail distributions remain plausible relative to the recent observed climate.
- **Structural and Physical Consistency** (Subsection 5.4.3), evaluating whether the spatial, temporal, and orographic structure of the generated fields is preserved under projection-based conditioning.

Following the procedure described in Section 4.9, QDM was applied to the CMIP6 projection for 2050–2059 to correct its bias relative to the ERA5. The comparison of the daily mean distribution between the reference period ERA5 and CMIP6 before and after applying the QDM bias correction is presented in the appendix A, Figure A.3.

It ends with a summary of the discussed results (Subsection 5.4.4).

5.4.1 Conditioning Consistency

Figure 5.19 evaluates the daily-mean consistency error, defined as the difference between the daily mean of the generated hourly fields, computed using the aggregation explained in Section 4.2.2, and the imposed CMIP6 daily-mean condition. The spatial

map shows a low-magnitude, slightly negative error across the entire land domain, with no pronounced spatial pattern. The boxplot of pooled land-pixels confirms that the individual daily, member, and grid-cell errors are tightly centered around zero, with a small spread. Quantitatively, the mean error is $-0.0062\text{ }^{\circ}\text{C}$, while the mean absolute error and RMSE are only $0.0144\text{ }^{\circ}\text{C}$ and $0.0183\text{ }^{\circ}\text{C}$, respectively. These errors are negligible relative to the magnitude of daily-mean temperature. This indicates that, when driven by bias-corrected CMIP6 inputs, the Full diffusion model preserves the imposed daily condition with very high fidelity.

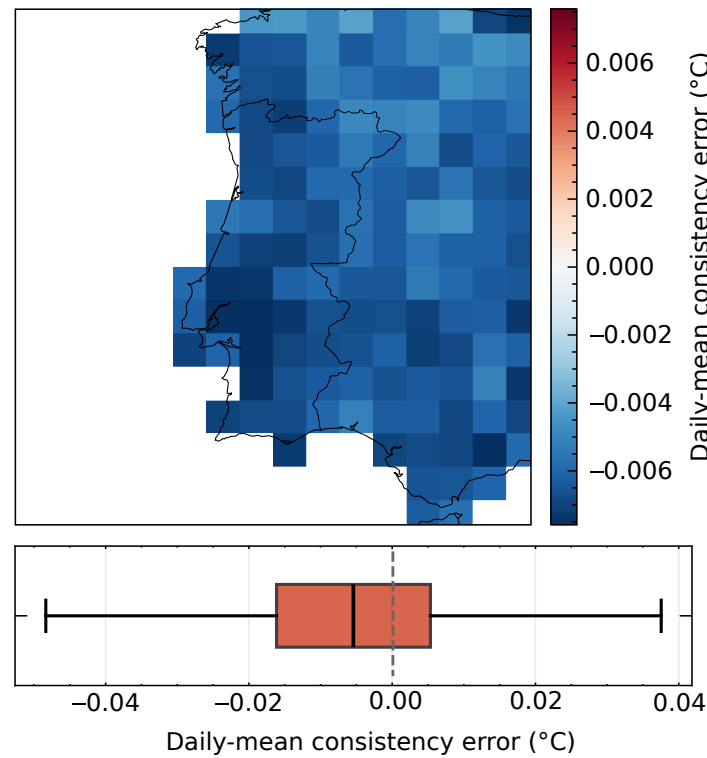


Figure 5.19: Daily-mean conditioning consistency error of the fields generated by the Full diffusion model under CMIP6 projections for 2050–2059. Errors are defined as the difference between the daily mean of the generated hourly fields, after aggregation to the conditioning resolution, and the imposed CMIP6 daily-mean condition. Top panel shows the spatial mean error over land, while bottom panel shows the pooled distribution of individual land-cell errors.

5.4.2 Distributional and Tail Plausibility

Figure 5.20 compares the seasonal sub-daily anomaly distributions of the generated fields against the ERA5 (2010–2024). The generated anomalies are correctly centered in all seasons with mean differences below $0.011\text{ }^{\circ}\text{C}$. Distributional distances to ERA5 are low, with 1-Wasserstein distance between $0.181\text{ }^{\circ}\text{C}$ (for September–November (SON)) and $0.227\text{ }^{\circ}\text{C}$ (for JJA) and KS below 0.027.

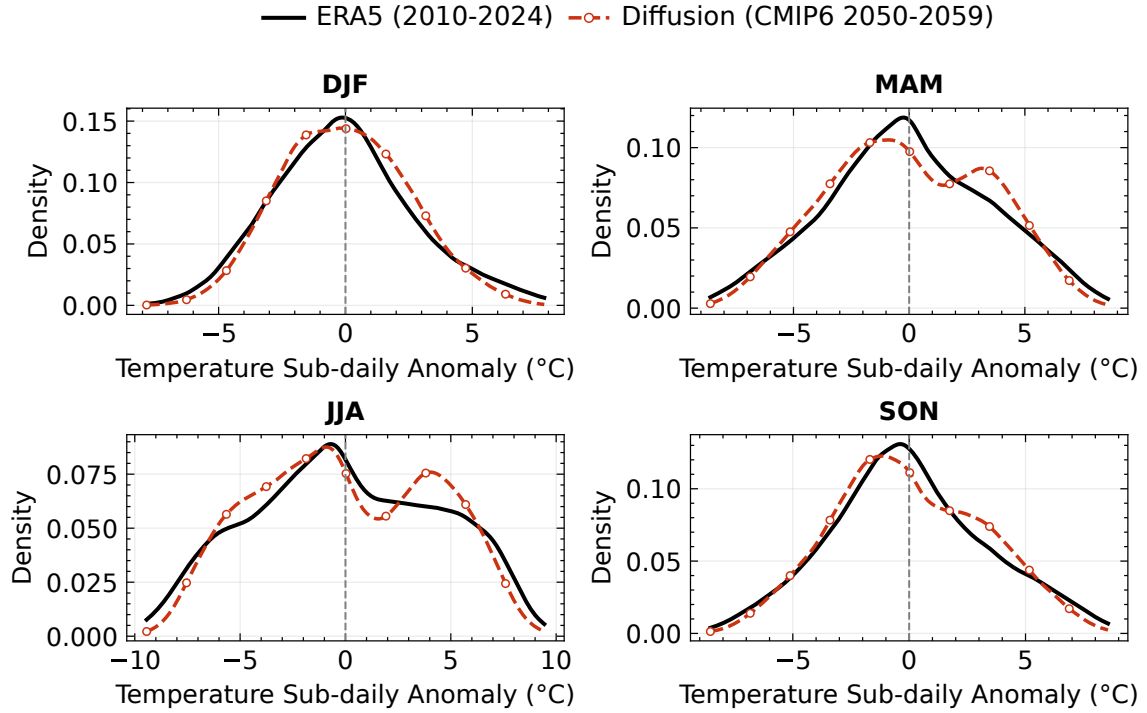


Figure 5.20: Seasonal distributions of sub-daily temperature anomalies for ERA5 during 2010–2024 and for the CMIP6-conditioned Full diffusion generations during 2050–2059. Generated anomalies are defined as the generated hourly temperature minus the imposed daily-mean condition.

The standard-deviation ordering by season is preserved in the generated days. The generated distributions are mildly compressed, with standard-deviation ratios between 0.892 (DJF) and 0.966 (March–May (MAM)). This is consistent with the underdispersion identified in Experiments 1 and 2. As no paired future reference exists, the underdispersion cannot be uniquely attributed to the model, as a real change in future sub-daily variability is also possible.

In MAM and JJA the generated density shows a warm-side secondary mode that is absent or less pronounced in ERA5. Elsewhere the seasonal shapes track the ERA5.

Despite the mild compression and the warm-side mode, the sub-daily anomaly structure under CMIP6 is therefore distributionally plausible as it is correctly centered and presents low distance to ERA5.

The corresponding absolute temperature distributions are shown in Appendix A Figure A.4. As expected, the generated densities are shifted warmer than ERA5 in all seasons, reflecting the daily-mean warming imposed by the CMIP6 projection, while retaining a smooth and unimodal shape consistent with the observed 2010–2024 climate.

Appendix A, Figure A.5 shows the full JJAS daily-maximum temperature distribution under the CMIP6 projections compared with the recent ERA5 reference period. The distribution exhibits a systematic warm shift quantified by a 1-Wasserstein distance of 2.15 °C, consistent with the warming daily conditions imposed by the SSP2-4.5 scenario. The generated future density remains smooth and unimodal with continuous tails, suggesting the model remains stable under the shifted future inputs.

Figure 5.21 shows the upper-tail shift and fixed-threshold exceedance of the same distribution. The warming propagates consistently into the upper tail, with shifts of +1.80 °C at P_{95} and +1.60 °C at both P_{99} and $P_{99.9}$. The similarity of these shifts suggests that the robust upper tail mainly follows the broader rightward shift of the distribution, rather than showing a strong additional reshaping.

The shift produces a large increase in extreme-heat frequency. The probability of a land location exceeding 35 °C in daily maximum temperature nearly doubles ($1.91\times$), while exceedance of 40 °C more than triples ($3.24\times$). The stronger relative increase at 40 °C is expected, since rare exceedances are highly sensitive to a rightward shift in the temperature distribution. The near-uniform quantile shifts support this interpretation. A genuine intensification of warming in the upper tail cannot be excluded, but if present it is likely underestimated here, given the model's tail conservatism identified in Experiments

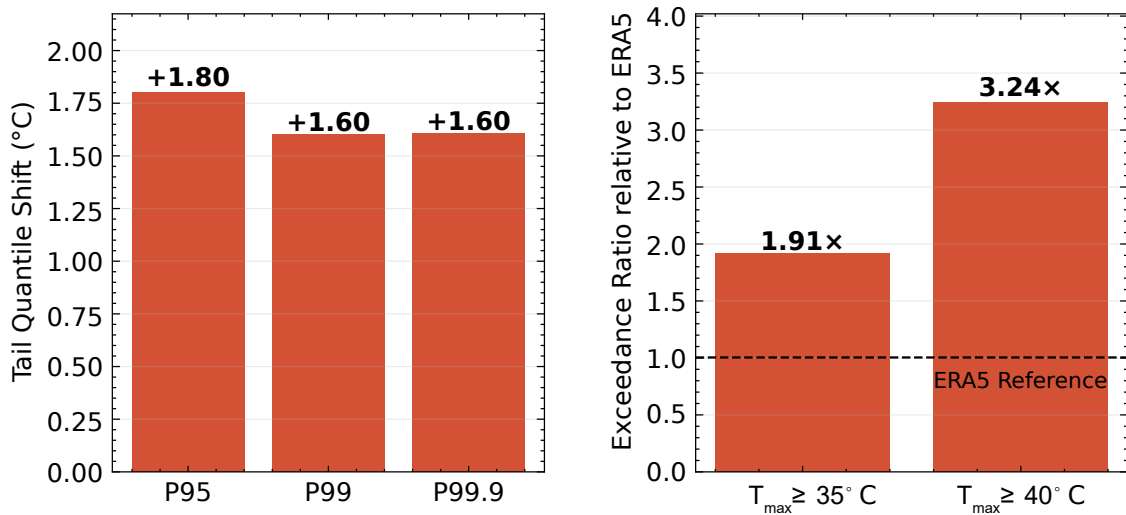


Figure 5.21: Upper-tail warming and extreme-heat frequency of JJAS daily-maximum temperature under the CMIP6 SSP2-4.5 projection (2050–2059) relative to ERA5 (2010–2024). Left plot shows the future shift of the upper-tail quantiles (P_{95} , P_{99} , $P_{99.9}$), in °C. Right plot shows the exceedance ratio relative to ERA5 for fixed thresholds of 35 °C and 40 °C, where the exceedance rate is the probability that a land location reaches a daily maximum above the threshold on a given JJAS day. The dashed line marks the ERA5 reference ($1\times$).

1 and 2. The generated extremes are therefore best read as a plausible lower bound on the projected change. The absolute quantiles and exceedance rates are reported in Appendix A Table A.3.

5.4.3 Structural and Physical Consistency

Figure 5.22 compares the seasonal mean diurnal cycle of daily-mean temperature anomalies. The generated fields reproduce the main ERA5 structure in all seasons, with negative anomalies at night and in the early morning, warming after sunrise, positive anomalies in the afternoon, and evening cooling. The phase is preserved, with the daily minimum occurring in the early morning and the maximum in the afternoon. The seasonal amplitude ordering is also maintained, with the weakest cycle in DJF and the strongest in JJA.

The generated cycle is slightly less smooth than the ERA5 reference, particularly around the morning transition and afternoon maximum. This behavior also appeared under historical ERA5 forcing in Experiment 1, suggesting that it is a property of the learned sub-daily

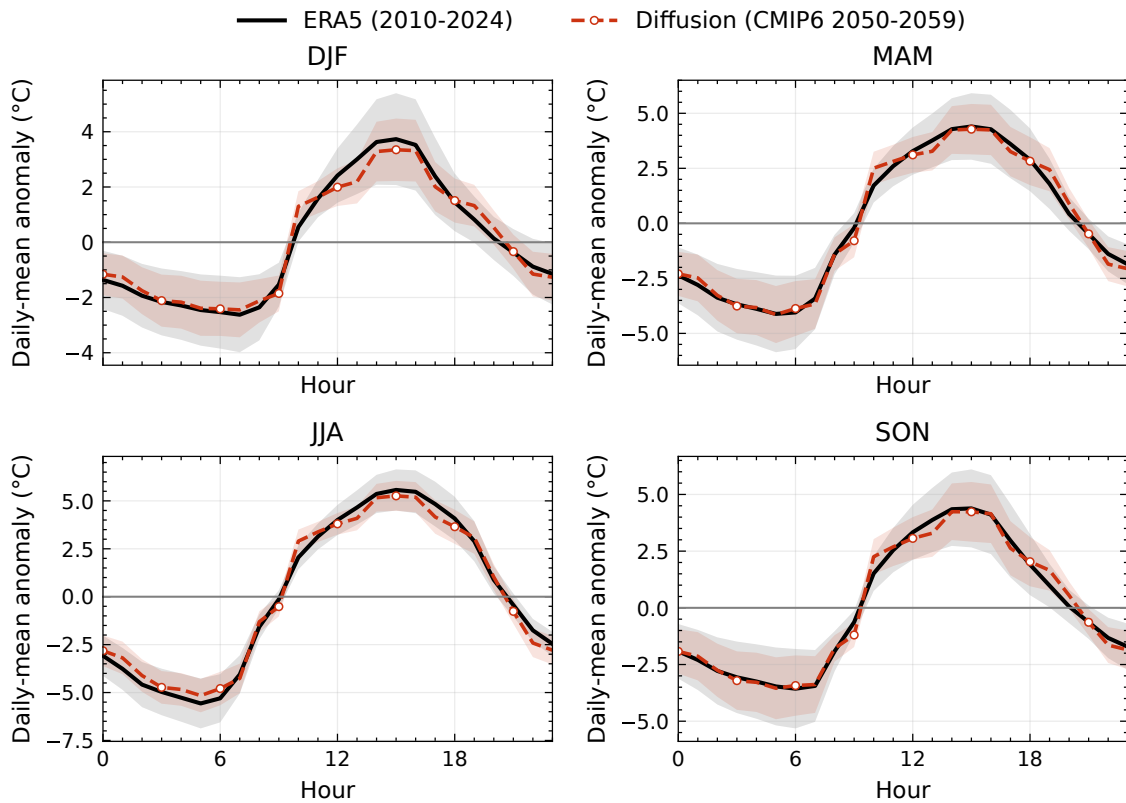


Figure 5.22: Seasonal mean diurnal cycle of the daily-mean temperature anomaly over land grid points, comparing ERA5 (2010–2024) with the fields generated by the Full diffusion model under the CMIP6 SSP2-4.5 projection (2050–2059). Solid and dashed lines denote the mean cycle; shaded bands denote ± 1 standard deviation across days and ensemble members.

representation rather than a specific effect of the CMIP6 input. Agreement is strongest in MAM and SON, while DJF slightly underestimates the afternoon peak and JJA shows narrower spread around the daytime maximum. Overall, the ensemble bands overlap the ERA5 reference in all seasons, indicating a physically plausible diurnal cycle under projection forcing.

Figure 5.23 evaluates the within-day ACF of sub-daily temperature anomalies. Both ERA5 and the generated fields show the expected diurnal dependence, with high short-lag correlation, negative correlation near 11–12 h, and recovery toward positive correlation near the daily boundary. Differences are small during the first half of the day, remaining within approximately ± 0.03 .

At longer lags, from about 14 h onward, the generated ACF is systematically higher than ERA5, reaching about +0.09 near lag 21. This indicates that late-day anomalies remain slightly too correlated with early-day anomalies within the generated daily block. Since a similar pattern was observed under historical forcing in Experiment 1, this is more likely a model characteristic than a projection-specific effect. Nevertheless, the main within-day temporal dependence is reproduced.

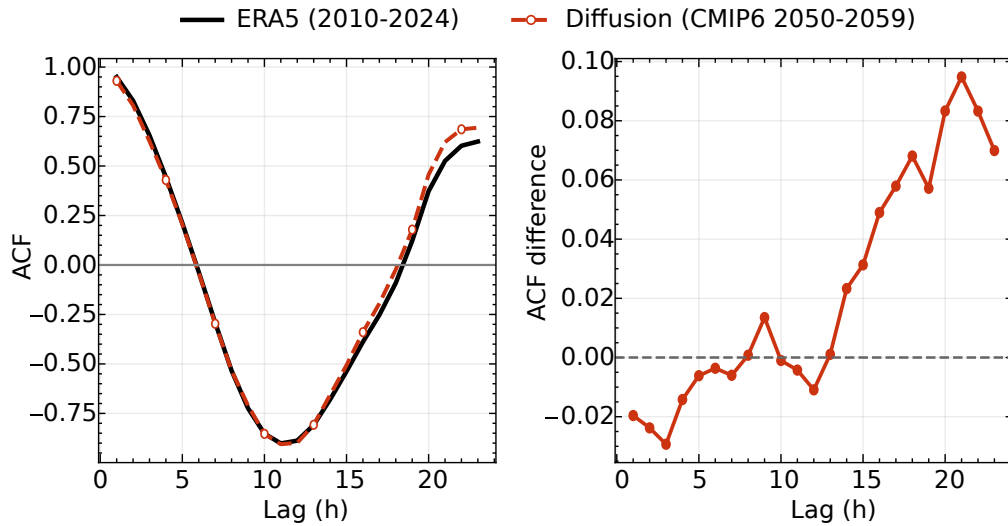


Figure 5.23: Within-day autocorrelation function of sub-daily temperature anomalies over land, with lagged products restricted to individual 24-hour cycles to isolate intra-day coherence. The left panel shows the ACF for lags 1–23 h, comparing ERA5 (2010–2024) with the generated fields under the CMIP6 SSP2-4.5 projection (2050–2059). The right panel shows the ACF difference (generated minus ERA5).

Figure 5.24 shows the radially averaged spatial PSD of the generated fields. The generated PSD follows ERA5 closely across the full range of spatial scales, reproducing the dominant concentration of variance at large scales and the same decay towards finer

scales. The spectral ratio remains close to zero throughout, with all scale-dependent deviations within roughly ± 0.06 in $\log_{10}$ ratio.

The deviations are small and follow opposite trends. At large and intermediate scales the generated fields hold slightly less variance than ERA5, consistent with the mild under-dispersion identified earlier, since most anomaly variance resides at these scales. At the finest resolved scales the generated fields hold slightly more variance, indicating a small excess of fine-scale variability.

The transition between the two regimes is smooth, with no sharp peak that would suggest grid-scale artifacts. There is no evidence of over-smoothing, as the generated fields retain fine-scale variability, and no evidence of strong high-frequency noise, as the fine-scale excess is small and localized. Overall, the model preserves the ERA5 spatial structure of the temperature fields under CMIP6 conditioning, with only minor scale-dependent deviations.

This is corroborated by the residual spatial-correlation errors, shown in Appendix A Figure A.7, which are tightly centered near zero (median of 0.0024 and RMSE of 0.022), confirming the generated fields reproduce the spatial covariance structure of ERA5.

This matches the spatial behavior of the model under ERA5 forcing in Experiment 1, which suggests that the structure is retained under CMIP6 conditioning.

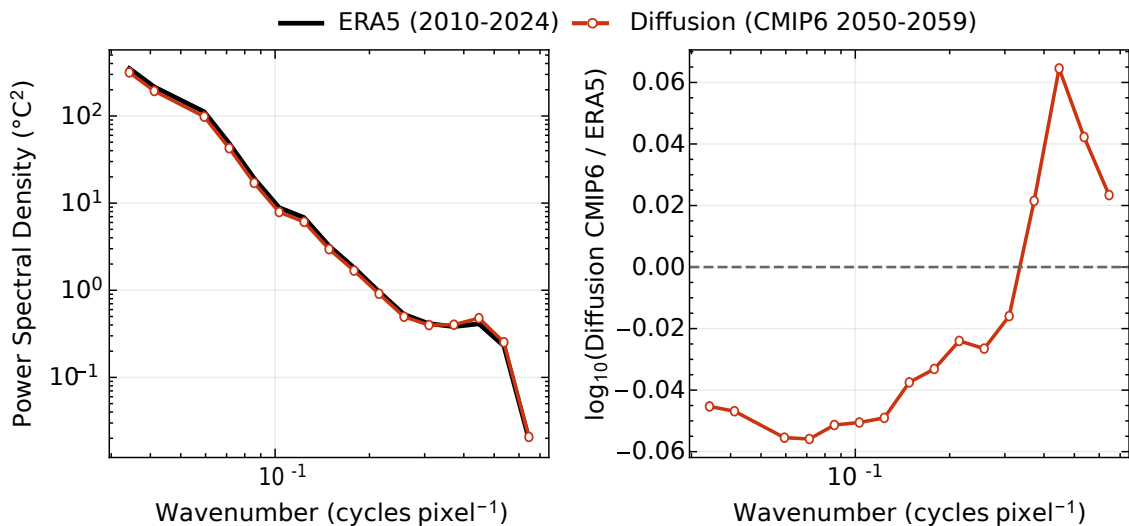


Figure 5.24: Radially-averaged spatial PSD of the temperature anomaly fields, comparing ERA5 (2010–2024) with the Full diffusion fields under the CMIP6 SSP2-4.5 projection (2050–2059). Left plot shows PSD as a function of spatial wavenumber. Right plot shows $\log_{10}$ ratio of the generated to ERA5 PSD. The dashed line marks perfect agreement.

Figure 5.25 shows that the generated fields are warmer than ERA5 at all elevations, with a land-mean shift of $+1.73\text{ }^{\circ}\text{C}$, reflecting the warming inherited from the CMIP6 conditioning. Once this land-mean offset is removed, the relative temperature-elevation structure overlays ERA5 closely (RMSE of $0.187\text{ }^{\circ}\text{C}$) with the maximum deviation of $0.507\text{ }^{\circ}\text{C}$ at the lowest elevations, where the generated profile is marginally cooler than ERA5, and shows no broader elevation-dependent distortion.

The generated fields therefore seem to retain the ERA5 lapse rate. This relative structure is governed by physics and is largely independent of the imposed warming, so its preservation is strong evidence of physical realism, consistent with that observed in Experiment 1.

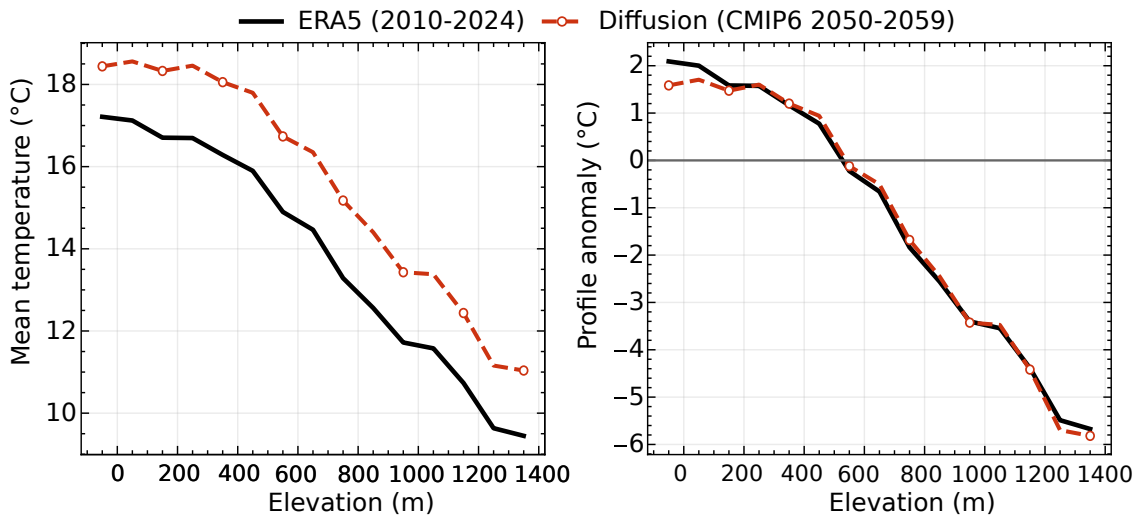


Figure 5.25: Mean temperature as a function of elevation over land grid points, comparing ERA5 (2010–2024) with the Full diffusion fields under the CMIP6 SSP2-4.5 projection (2050–2059). Left plot shows absolute mean temperature profile. Right plot shows profile with the land-mean temperature removed, isolating the relative temperature–elevation structure. Land pixels are grouped into elevation bins.

5.4.4 Summary

Overall, Experiment 3 shows that the Full diffusion model continues to generate realistic and physically plausible fields when driven by bias-corrected CMIP6 projections. The model reproduces the imposed daily-mean condition with negligible error, and the generated sub-daily structure remains distributionally close to the observed climate while inheriting the projected warming as a coherent level shift. The temporal, spatial and orographic diagnostics are all preserved and consistent with the behavior observed under ERA5 forcing in Experiment 1. The underdispersion and the conservative upper-tail make the

projected extremes a plausible lower bound. These results indicate the model generalizes well to the future-projection setting and can serve as a tool for generating physically consistent hourly fields from coarse climate-model output.

5.5. Synthesis of Results and Discussion

Across the three experiments, the Full diffusion model provides the most consistent neural generative compromise in the historical validation setting. Although the deterministic U-Net achieves the lowest pointwise error and the baseline k-NN remains competitive in marginal distributional diagnostics, the Full diffusion model combines stochastic generation, scalar controllability, and realistic temporal and spatial structure. These properties motivate its use in the projection-driven experiment.

The results also show that the main limitations of the model are structural rather than purely sampling-related. Increasing ensemble size from $K = 20$ to $K = 200$ improves coverage, CRPS, and rare-event sampling, but does not remove the underdispersion and the hot-tail compression. These limitations therefore reflect the learned conditional distribution and remain important when interpreting generated scenarios.

When applied to bias-corrected CMIP6 fields, the Full diffusion model preserves the imposed daily-mean condition and produces fields with plausible temporal and spatial structure relative to recent ERA5. Because no paired future hourly reference exists, the diagnostics are not verification scores for future hourly temperature. Overall, the experiments support conditional downscaling, while identifying calibration, upper-tail representation, and inter-day continuity as the main areas for future improvement.

6. Conclusions and Future Work

This dissertation developed and evaluated a probabilistic generative framework for spatiotemporal downscaling of temperature fields. Instead of producing a single deterministic forecast, the central goal was to generate diverse, physically plausible high-resolution hourly scenarios conditioned on a coarse daily-mean field, capturing the full range of high-resolution spatial and temporal structure consistent with that daily condition. The problem was treated as a one-to-many mapping, since the same daily mean can correspond to multiple plausible hourly evolutions with different amplitudes, phases, and spatial structures. To address this, a conditional diffusion model was implemented and compared against deterministic, ablated, and analogue-based baselines.

The research objectives were fulfilled through five main steps. First, the downscaling task was formulated using complete 24-hour spatiotemporal blocks, allowing the model to learn intra-day temporal structure and spatial variability jointly. Second, a conditioning strategy was developed combining the coarse daily temperature field, static geographical variables, cyclic temporal encodings, and scalar controls. Third, the model was used to generate probabilistic ensembles of hourly temperature scenarios from a single daily state. Fourth, the framework was applied to bias-corrected CMIP6 daily projections under SSP2-4.5 for the 2050–2059 period. Fifth, the generated fields were evaluated using deterministic, probabilistic, temporal, spatial, distributional, extreme-value, and physical-consistency diagnostics.

In the historical validation setting, the deterministic U-Net achieved the lowest pointwise error but could not represent uncertainty, while the k -NN baseline remained competitive on several marginal distributional diagnostics and performed strongly on temporal and spatial structural measures, by virtue of retrieving historical sub-daily anomaly fields directly. The Full diffusion model, by contrast, offered the most balanced generative behavior, combining stochastic generation with realistic temporal and spatial structure. The comparison with the No-scalar ablation confirmed that scalar controls meaningfully improve calibration, sub-daily variability, spatial organization, and hot-tail behavior.

The ensemble-depth experiment showed that increasing the number of generated members improves the empirical characterization of the distribution without changing the learned distribution itself, as deeper ensembles provide a denser estimate but do not resolve its

limitations. Applying the model to CMIP6 projections showed that it can be operationalized under projection-driven conditioning, as the generated fields preserved the imposed daily mean and produced physically plausible hourly structures relative to ERA5. Since no paired future hourly reference exists, this experiment is best read as a plausibility and consistency check rather than verification against future hourly reference data.

The research question posed in Chapter 1 can therefore be answered positively: conditional diffusion models provide a promising framework for probabilistic spatiotemporal downscaling, generating diverse and physically plausible hourly temperature scenarios while preserving the imposed daily condition and reproducing relevant temporal, spatial, and distributional characteristics.

The main limitations of the framework are the conservative ensemble spread, the compressed upper tail of hot extremes, the independent generation of daily blocks, and the scalar-control selection mechanism for future scenarios, in which the sub-daily variability control s_d is selected from historical analogues conditioned on r_d . Because this procedure restricts s_d to values already observed historically for similar regimes, it limits the range of sub-daily variability that can be explored at inference time, which may be particularly limiting for extreme or under-sampled regimes. Similar limitations have been reported in related studies [49, 51, 54]. This motivates the five future work directions described below.

First, calibration and extreme-temperature representation should be improved, through targeted sampling of rare regimes, modified training objectives, post-processing calibration or loss formulations that weight physically important hot extremes more heavily or explicitly encourage greater ensemble spread.

Second, the conditioning mechanism should be strengthened in two ways. Architecturally, feature-wise modulation, such as FiLM conditioning [64], could be used so that temporal encodings and scalar-control information modulate internal U-Net features directly rather than entering only as input channels. More fundamentally, the scalar-control selection procedure itself should be revised to allow broader exploration of plausible s_d values beyond the historical analogue pool, while still respecting the joint relationship between r_d and s_d . This is particularly important for hot extremes, where the admissible range of sub-daily variability may not be well represented by historical analogues alone.

Third, two-stage residual generation should be explored, following regression-plus-diffusion

approaches such as [51]. A deterministic U-Net could first generate a high-resolution 24-hour baseline from the coarse daily field. A diffusion model would then build on this baseline, either generating stochastic residual corrections around it or using it directly as an added conditioning field for generating diverse hourly scenarios.

Fourth, temporal continuity and broader generalization should be addressed. The current model generates independent daily blocks, with no mechanism enforcing consistency across day boundaries. Future work could explore multi-day or autoregressive generation, or post-hoc methods for linking ensemble members across consecutive days according to transition consistency [56].

Fifth, the framework's generalization and scope should be extended. This includes testing across additional climate models, emission scenarios, and spatial domains, including more strongly non-stationary futures, as well as extending to other climate variables or multivariate generation, aiming to jointly produce variables such as temperature, humidity, wind and precipitation. Scaling to larger domains and additional variables may in turn require more efficient architectures, such as latent diffusion models or flow matching [75, 76].

These directions would help make the framework more robust and easier to extend. Overall, the results show that conditional diffusion is a promising approach for probabilistic spatiotemporal temperature downscaling, while also highlighting areas for improvement.

References

- [1] H. C. Bloomfield, D. J. Brayshaw, M. Deakin, and D. Greenwood, “Hourly historical and near-future weather and climate variables for energy system modelling,” *Earth System Science Data*, vol. 14, no. 6, pp. 2749–2766, 2022. [Cited on page 1.]
- [2] K. Calvin, D. Dasgupta, G. Krinner, A. Mukherji, P. W. Thorne, C. Trisos *et al.*, “IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.” Intergovernmental Panel on Climate Change (IPCC), Tech. Rep. [Cited on page 1.]
- [3] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, “Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization,” *Geoscientific Model Development*, vol. 9, no. 5, pp. 1937–1958, 2016. [Cited on pages 1 and 10.]
- [4] I. Staffell, S. Pfenninger, and N. Johnson, “A global model of hourly space heating and cooling demand at multiple spatial scales,” *Nature Energy*, vol. 8, no. 12, pp. 1328–1344, 2023. [Cited on page 2.]
- [5] L. Qiu, R. Khorramfar, S. Amin, and M. F. Howland, “Decarbonized energy system planning with high-resolution spatial representation of renewables lowers cost,” *Cell Reports Sustainability*, vol. 1, no. 12, 2024. [Cited on page 2.]
- [6] J. T. Malle, C. P. Reyer, Y. Amitai, A. L. Augustynczyk, Y. Be’eri-Shlevin, E. Ben-Zur, P. Burek, T. Chaudhari, J. Chang, A. Collalti *et al.*, “When and where higher-resolution climate data improve impact model performance,” *arXiv preprint arXiv:2512.17739*, 2025. [Cited on page 2.]
- [7] L. Fischer and D. Diaz, “READi insights: Extreme heat events and impacts to the electric system,” Electric Power Research Institute (EPRI), Palo Alto, CA, Tech. Rep. 3002025522, Sep. 2022. [Cited on page 2.]
- [8] P. Czyzak, “Heat and power: Impacts of the 2025 heatwave in europe,” Jul. 2025. [Online]. Available: <https://ember-energy.org/latest-insights/heat-and-power-impacts-of-the-2025-heatwave-in-europe/> [Cited on page 3.]

- [9] H. Bloomfield, P. Gonzalez, J. K. Lundquist, L. P. Stoop, J. Browell, R. Dargaville, M. De Felice, K. Gruber, A. Hilbers, A. Kies *et al.*, “The importance of weather and climate to energy systems: a workshop on next generation challenges in energy–climate modeling,” *Bulletin of the American Meteorological Society*, vol. 102, no. 1, pp. E159–E167, 2021. [Cited on page 3.]
- [10] H. C. Bloomfield, D. J. Brayshaw, L. C. Shaffrey, P. J. Coker, and H. E. Thornton, “Quantifying the increasing sensitivity of power systems to climate variability,” *Environmental Research Letters*, vol. 11, no. 12, p. 124025, 2016. [Cited on page 3.]
- [11] ENTSO-E Expert Panel, “Final Report on the Grid Incident in Spain and Portugal on 28 April 2025,” European Network of Transmission System Operators for Electricity (ENTSO-E), Brussels, Belgium, Tech. Rep., Mar. 2026. [Online]. Available: <https://www.entsoe.eu/news/2026/03/20/entso-e-publishes-expert-panel-final-report-on-28-april-2025-blackout-in-spain-and-portugal/> [Cited on page 3.]
- [12] Instituto Português do Mar e da Atmosfera, “Boletim climatológico mensal – janeiro 2026, portugal continental,” IPMA, Divisão de Clima e Alterações Climáticas, Lisboa, Portugal, Tech. Rep., Mar. 2026, ISSN 2183-1076. [Online]. Available: https://www.ipma.pt/resources.www/docs/im.publicacoes/edicoes.online/20260313/FvTHXCNIkkXzTHuuxaHk/cli_20260101_20260131_pcl_mm_co_pt.pdf [Cited on page 3.]
- [13] Q. Zhao, Y. Guo, T. Ye, A. Gasparrini, S. Tong, A. Overcenco, A. Urban, A. Schneider, A. Entezari, A. M. Vicedo-Cabrera *et al.*, “Global, regional, and national burden of mortality associated with non-optimal ambient temperatures from 2000 to 2019: a three-stage modelling study,” *The Lancet Planetary Health*, vol. 5, no. 7, pp. e415–e425, 2021. [Cited on page 4.]
- [14] P. Murage, S. Hajat, and R. S. Kovats, “Effect of night-time temperatures on cause and age-specific mortality in London,” *Environmental epidemiology*, vol. 1, no. 2, p. e005, 2017. [Cited on page 4.]
- [15] W. T. K. Huang, P. Masselot, E. Bou-Zeid, S. Fatichi, A. Paschalis, T. Sun, A. Gasparrini, and G. Manoli, “Economic valuation of temperature-related mortality attributed to urban heat islands in European cities,” *Nature communications*, vol. 14, no. 1, p. 7438, 2023. [Cited on page 4.]

- [16] D. Deryng, D. Conway, N. Ramankutty, J. Price, and R. Warren, “Global crop yield response to extreme heat stress under multiple climate change futures,” *Environmental Research Letters*, vol. 9, no. 3, p. 034011, 2014. [Cited on page 4.]
- [17] J. Richetti, V. O. Sadras, D. He, B. Leske, P. Hu, Y. Beletse, C. M. Cossani, H. Nguyen, B. Zheng, D. M. Deery *et al.*, “Challenges in modelling the impact of frost and heat stress on the yield of cool-season annual grain crops,” *Frontiers in Plant Science*, vol. 16, p. 1613432, 2025. [Cited on page 4.]
- [18] R. Hock, “Temperature index melt modelling in mountain areas,” *Journal of hydrology*, vol. 282, no. 1-4, pp. 104–115, 2003. [Cited on page 4.]
- [19] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, “DeepSD: Generating high resolution climate change projections through single image super-resolution,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1663–1672. [Cited on pages 4 and 24.]
- [20] R. Laprise, “Regional climate modelling,” *Journal of Computational Physics*, vol. 227, no. 7, pp. 3641–3666, 2008. [Cited on page 4.]
- [21] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. [Cited on pages 6, 19, and 21.]
- [22] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut, “The ERA5 global reanalysis,” *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020. [Cited on pages 6 and 11.]
- [23] A. Voldoire, “CNRM-CERFACS CNRM-CM6-1-HR model output prepared for CMIP6 HighResMIP,” 2019. [Online]. Available: <https://doi.org/10.22033/ESGF/CMIP6.1387> [Cited on page 10.]

- [24] K. Riahi, D. P. Van Vuuren, E. Kriegler, J. Edmonds, B. C. O'Neill, S. Fujimori, N. Bauer, K. Calvin, R. Dellink, O. Fricko *et al.*, “The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview,” *Global Environmental Change*, vol. 42, pp. 153–168, 2017. [Cited on page 10.]
- [25] E. Zorita and H. Von Storch, “The analog method as a simple statistical downscaling technique: comparison with more complicated methods,” *Journal of climate*, vol. 12, no. 8, pp. 2474–2489, 1999. [Cited on pages 12 and 48.]
- [26] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967. [Cited on pages 12 and 48.]
- [27] S. Hadjipetrou, “A review of statistical methods for climate downscaling: the under-explored potential of geostatistical simulation,” *Theoretical and Applied Climatology*, vol. 157, no. 4, p. 216, 2026. [Cited on page 13.]
- [28] J. Guevara, M. Garcia, P. Avegliano, D. Lima, D. Queiroz, M. Macedo, L. Tizzei, D. Szwarcman, B. Zadrozny, C. Watson *et al.*, “Direct sampling for spatially variable extreme event generation in resampling-based stochastic weather generators,” *Journal of Advances in Modeling Earth Systems*, vol. 15, no. 11, p. e2022MS003542, 2023. [Cited on page 13.]
- [29] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4. [Cited on page 16.]
- [30] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
- [31] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. [Cited on page 16.]
- [32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>. [Cited on page 16.]
- [33] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013. [Cited on page 16.]

- [34] I. Goodfellow, “NIPS 2016 tutorial: Generative adversarial networks,” 2017. [Online]. Available: <https://arxiv.org/abs/1701.00160> [Cited on pages 16 and 17.]
- [35] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014. [Online]. Available: <https://arxiv.org/abs/1406.2661> [Cited on page 17.]
- [36] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training GANs,” *Advances in neural information processing systems*, vol. 29, 2016. [Cited on page 17.]
- [37] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” *arXiv preprint arXiv:1312.6114*, 2013. [Cited on page 18.]
- [38] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, “Diffusion models: A comprehensive survey of methods and applications,” *ACM computing surveys*, vol. 56, no. 4, pp. 1–39, 2023. [Cited on page 19.]
- [39] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265. [Cited on page 19.]
- [40] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in neural information processing systems*, vol. 32, 2019. [Cited on page 19.]
- [41] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” *arXiv preprint arXiv:2011.13456*, 2020. [Cited on page 20.]
- [42] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020. [Cited on page 22.]
- [43] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, “DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models,” *Machine Intelligence Research*, vol. 22, no. 4, pp. 730–751, 2025. [Cited on pages 22 and 45.]
- [44] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021. [Cited on page 23.]

- [45] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022. [Cited on page 24.]
- [46] J. Leinonen, D. Nerini, and A. Berne, “Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7211–7223, 2020. [Cited on page 24.]
- [47] L. Harris, A. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer, “A generative deep learning approach to stochastic downscaling of precipitation forecasts,” *Journal of Advances in Modeling Earth Systems*, vol. 14, no. 10, p. e2022MS003120, 2022. [Cited on page 24.]
- [48] M. Choraria, D. Szwarcman, B. Zadrozny, C. D. Watson, and L. R. Varshney, “Controllable generation for climate modeling,” in *NeurIPS 2022 workshop on tackling climate change with machine learning*, 2022. [Cited on page 25.]
- [49] S. Bassetti, B. Hutchinson, C. Tebaldi, and B. Kravitz, “DiffESM: Conditional emulation of temperature and precipitation in Earth System Models with 3D diffusion models,” *Journal of Advances in Modeling Earth Systems*, vol. 16, no. 10, p. e2023MS004194, 2024, e2023MS004194 2023MS004194. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023MS004194> [Cited on pages 25 and 90.]
- [50] P. Srivastava, R. Yang, G. Kerrigan, G. Dresdner, J. McGibbon, C. Bretherton, and S. Mandt, “Precipitation downscaling with spatiotemporal video diffusion,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 56 374–56 400, 2024. [Cited on page 25.]
- [51] M. Mardani, N. Brenowitz, Y. Cohen, J. Pathak, C.-Y. Chen, C.-C. Liu, A. Vahdat, M. A. Nabian, T. Ge, A. Subramaniam *et al.*, “Residual corrective diffusion modeling for km-scale atmospheric downscaling,” *Communications Earth & Environment*, vol. 6, no. 1, p. 124, 2025. [Cited on pages 25, 90, and 91.]
- [52] R. A. Watt and L. A. Mansfield, “Generative diffusion-based downscaling for climate,” *arXiv preprint arXiv:2404.17752*, 2024. [Cited on page 26.]
- [53] F. Zabel and B. Poschlod, “The Teddy tool v1.1: temporal disaggregation of daily climate model data for climate impact analysis,” *Geoscientific Model Development*, vol. 16, no. 18, pp. 5383–5399, 2023. [Cited on page 26.]

- [54] S. Scher and S. Peßenteiner, “Temporal disaggregation of spatial rainfall fields with generative adversarial networks,” *Hydrology and Earth System Sciences*, vol. 25, no. 6, pp. 3207–3225, 2021. [Cited on pages 26 and 90.]
- [55] E. Balogun, R. Rajagopal, and A. Majumdar, “TemperatureGAN: generative modeling of regional atmospheric temperatures,” *Environmental Data Science*, vol. 3, p. e33, 2024. [Cited on page 26.]
- [56] J. Schmidt, L. Schmidt, F. M. Strnad, N. Ludwig, and P. Hennig, “A generative framework for probabilistic, spatiotemporally coherent downscaling of climate simulation,” *npj Climate and Atmospheric Science*, vol. 8, no. 1, p. 270, 2025. [Cited on pages 26 and 91.]
- [57] F. Rozet and G. Louppe, “Score-based data assimilation,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 40 521–40 541, 2023. [Cited on page 27.]
- [58] A. Q. Nichol and P. Dhariwal, “Improved denoising diffusion probabilistic models,” in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171. [Cited on page 39.]
- [59] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. [Cited on page 40.]
- [60] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459. [Cited on page 41.]
- [61] Y. Wu and K. He, “Group normalization,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19. [Cited on page 41.]
- [62] S. Elfving, E. Uchibe, and K. Doya, “Sigmoid-weighted linear units for neural network function approximation in reinforcement learning,” *Neural networks*, vol. 107, pp. 3–11, 2018. [Cited on page 41.]
- [63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017. [Cited on page 41.]

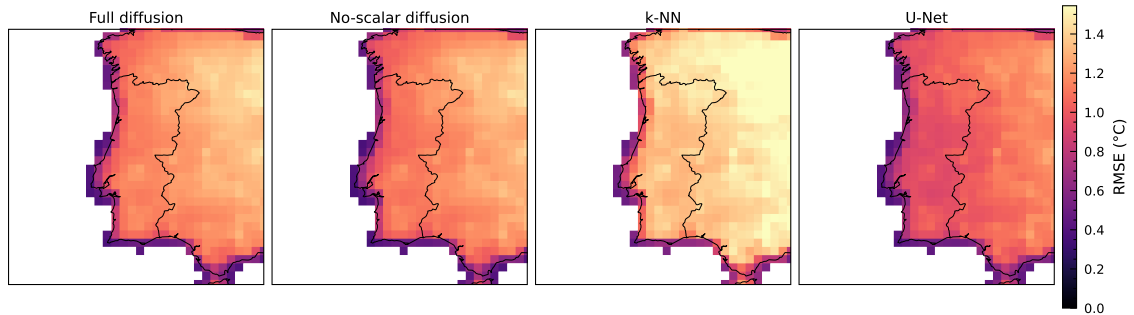
- [64] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018. [Cited on pages 43 and 90.]
- [65] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017. [Cited on page 44.]
- [66] Python Software Foundation, “Python programming language,” 2026, version 3.12, Available at python.org. [Online]. Available: <https://www.python.org> [Cited on page 45.]
- [67] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “PyTorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019. [Cited on page 45.]
- [68] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: <https://doi.org/10.1038/s41586-020-2649-2> [Cited on page 45.]
- [69] S. Hoyer and J. Hamman, “xarray: Nd labeled arrays and datasets in python,” *Journal of Open Research Software*, vol. 5, no. 1, p. 10, 2017. [Cited on page 45.]
- [70] A. Miles, jakirkham, J. Hamman, D. P. Orfanos, D. Stansby, M. Bussonnier, J. Moore, D. Bennett, T. Augspurger, N. Rzepka, D. Cherian, S. Verma, J. Bourbeau, A. Fulton, H. Spitz, R. Abernathey, G. Lee, M. Jones, M. R. B. Kristensen, Z. Patel, S. Chopra, M. Rocklin, N. Zimmerman, A. B. AWA, E. S. de Andrade, M. Durant, J. Nunez-Iglesias, and V. Schut, “zarr-developers/zarr-python: v3.0.7,” Apr. 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.15255977> [Cited on page 45.]
- [71] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H.

Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nature Methods*, vol. 17, pp. 261–272, 2020. [Cited on page 45.]

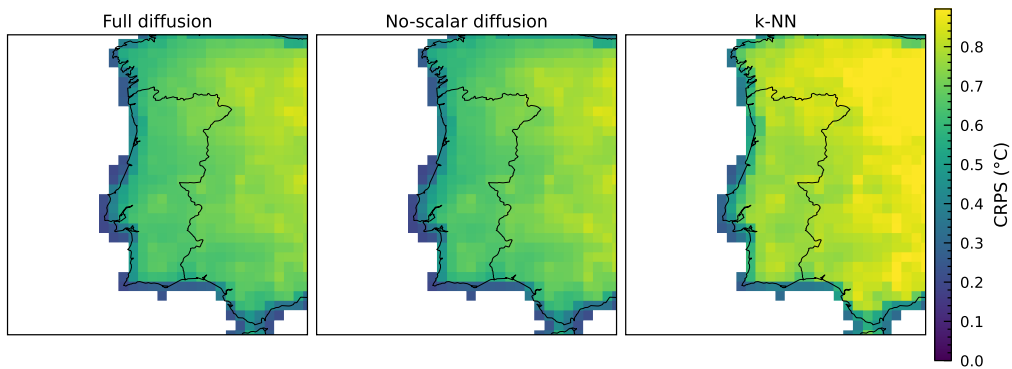
- [72] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. [Cited on page 45.]
- [73] A. J. Cannon, S. R. Sobie, and T. Q. Murdock, “Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes?” *Journal of climate*, vol. 28, no. 17, pp. 6938–6959, 2015. [Cited on page 46.]
- [74] C. Viceto, S. Cardoso Pereira, and A. Rocha, “Climate change projections of extreme temperatures for the Iberian Peninsula,” *Atmosphere*, vol. 10, no. 5, p. 229, 2019. [Cited on page 54.]
- [75] Y. Zhuang and K. Duraisamy, “LaDCast: A latent diffusion model for medium-range ensemble weather forecasting,” *arXiv preprint arXiv:2506.09193*, 2025. [Cited on page 91.]
- [76] F. Strnad, J. Schmidt, F. Mockert, P. Hennig, and N. Ludwig, “Assessing the risk of future Dunkelflaute events for Germany using generative deep learning,” *Environmental Data Science*, vol. 5, p. e11, 2026. [Cited on page 91.]

A. Additional Results

A.1. Experiment 1



(a) Spatial RMSE



(b) Spatial CRPS

Figure A.1: Spatial RMSE and CRPS comparison for Experiment 1.

Table A.1: Comparison of land-pixel spatial correlation across the models for the 2014–2023 evaluation period. **Bold** indicates the best performance, and underline indicates the second-best. Arrows in the headers indicate whether lower ($\downarrow$) or closer to zero ($\rightarrow 0$) values are better.

Model	Median error ($\rightarrow 0$)	IQR ($\downarrow$)	RMSE ($\downarrow$)
Full diffusion	<u>0.0011</u>	0.0055	0.0094
No-scalar diffusion	−0.0112	<u>0.0116</u>	0.0193
k-NN	0.0010	0.0055	<u>0.0100</u>
U-Net	0.0256	0.0139	0.0374

Table A.2: Seasonal amplitude-distribution errors across models for the 2014–2023 evaluation period. Metrics are computed over land grid points from daily sub-daily temperature amplitudes, defined as the daily maximum minus minimum hourly temperature. W1 denotes the 1-Wasserstein distance between each model and ERA5. Median bias, P05 error, and P95 error are computed as model minus ERA5. **Bold** indicates the best performance, and underline indicates the second-best. Arrows in the headers indicate whether lower ($\downarrow$) or closer to zero ($\rightarrow 0$) values are better.

(a) Winter (DJF)					(b) Spring (MAM)				
Model	W1 ($\downarrow$)	Median bias ($\rightarrow 0$)	P05 error ($\rightarrow 0$)	P95 error ($\rightarrow 0$)	Model	W1 ($\downarrow$)	Median bias ($\rightarrow 0$)	P05 error ($\rightarrow 0$)	P95 error ($\rightarrow 0$)
Full diffusion	<u>0.606</u>	-0.441	0.009	-1.629	Full diffusion	<u>0.743</u>	-0.519	-0.037	<u>-1.863</u>
No-scalar diffusion	1.155	-1.148	-0.260	-2.095	No-scalar diffusion	1.481	-1.482	-0.306	-2.488
k-NN	0.540	<u>-0.500</u>	<u>0.014</u>	-1.213	k-NN	0.578	<u>-0.584</u>	<u>-0.071</u>	-1.057
U-Net	0.976	-1.037	-0.338	<u>-1.495</u>	U-Net	1.127	-1.125	-0.311	-1.879

(c) Summer (JJA)					(d) Autumn (SON)				
Model	W1 ($\downarrow$)	Median bias ($\rightarrow 0$)	P05 error ($\rightarrow 0$)	P95 error ($\rightarrow 0$)	Model	W1 ($\downarrow$)	Median bias ($\rightarrow 0$)	P05 error ($\rightarrow 0$)	P95 error ($\rightarrow 0$)
Full diffusion	<u>0.840</u>	-0.974	<u>-0.074</u>	<u>-1.608</u>	Full diffusion	<u>0.717</u>	-0.625	-0.044	-1.600
No-scalar diffusion	1.583	-1.906	-0.259	-2.239	No-scalar diffusion	1.413	-1.453	-0.301	-2.263
k-NN	0.261	-0.292	0.058	-0.582	k-NN	0.668	<u>-0.765</u>	<u>-0.126</u>	-0.833
U-Net	0.963	<u>-0.965</u>	-0.250	-1.640	U-Net	1.079	-1.216	-0.327	<u>-1.468</u>

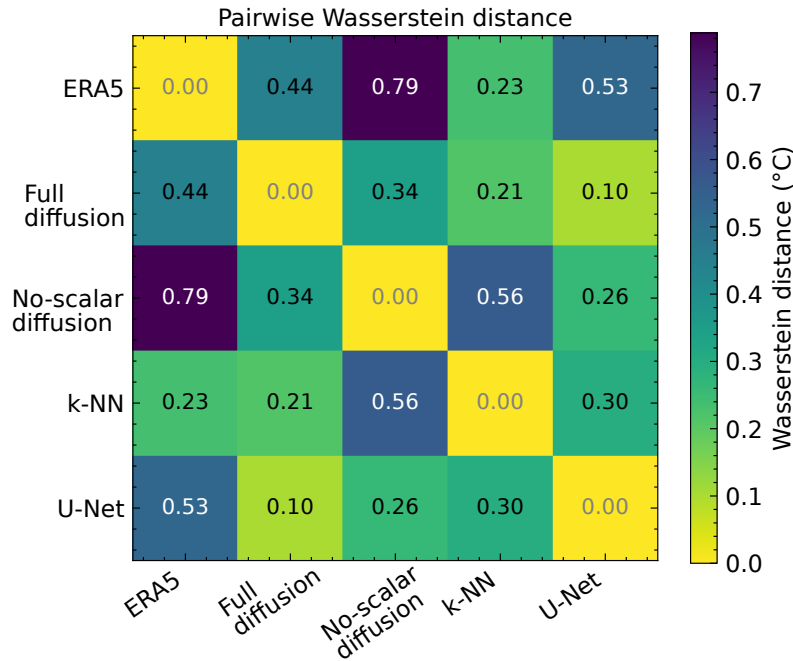


Figure A.2: Pairwise 1-Wasserstein distance matrix for pooled summer (JJAS) daily maximum temperature ($T_{\max}$) distributions. Values represent the structural distributional distance (°C) between the ERA5 reference and the evaluated downscaling models, computed over all land grid points and ensemble members for the 2014–2023 test period. Lower values indicate higher overall distributional similarity between the paired sources.

A.2. Experiment 3

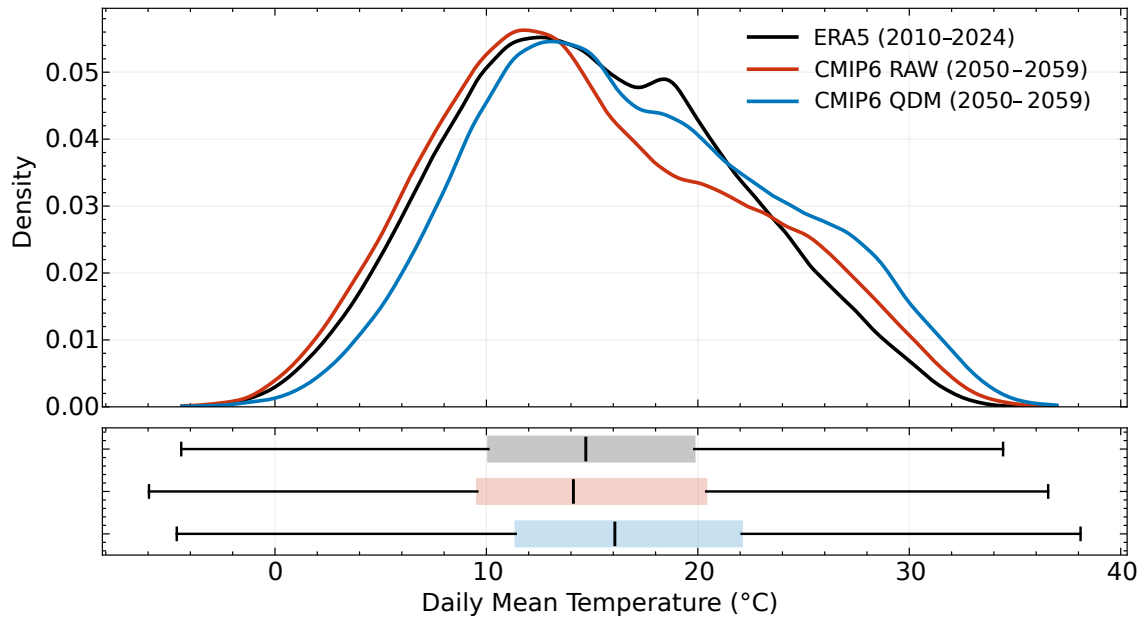


Figure A.3: Distribution of daily mean near-surface temperature over land for Experiment 3. The upper panel shows kernel density estimates of pooled land-grid points daily temperatures, while the lower panel shows the corresponding boxplots. ERA5 daily means are computed for 2010–2024 and compared with raw CMIP6 SSP2-4.5 temperatures and the corresponding QDM bias-corrected CMIP6 temperatures for 2050–2059.

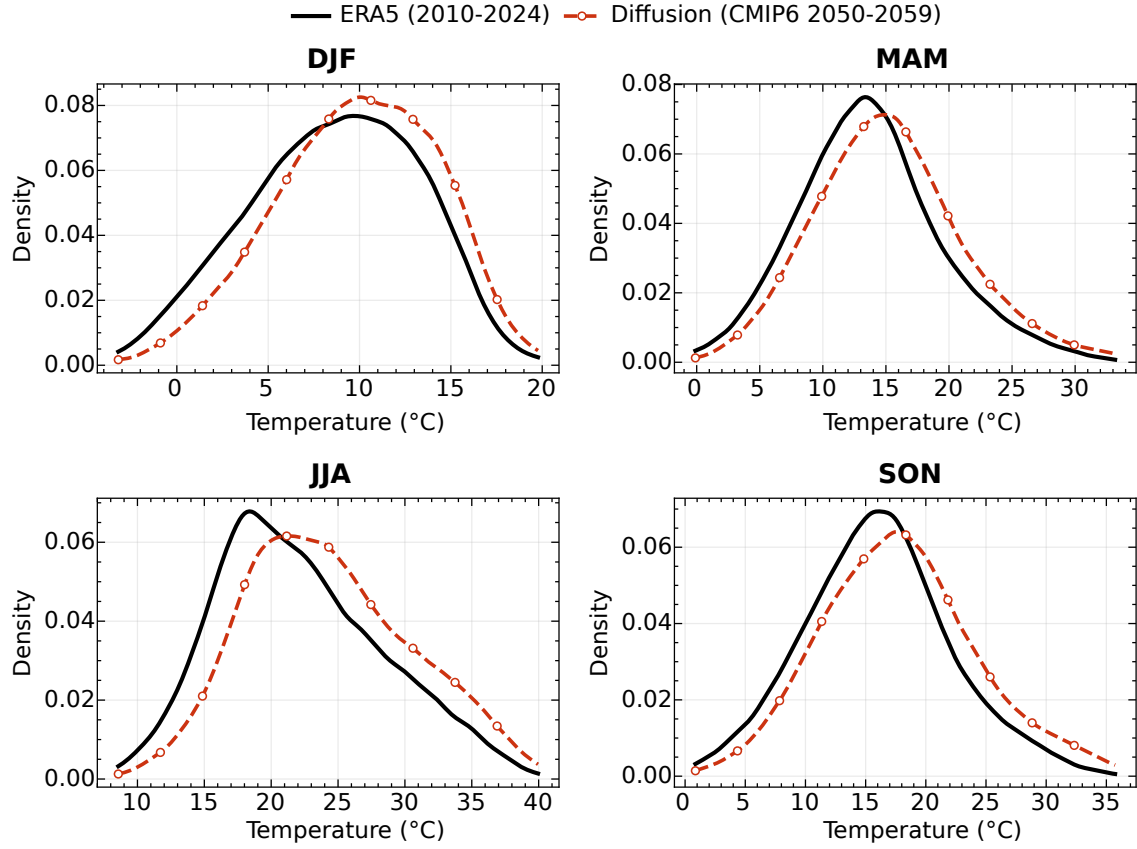


Figure A.4: Seasonal probability density of absolute temperature over land grid points, comparing ERA5 (2010–2024) with the fields generated by the Full diffusion model under the CMIP6 SSP2-4.5 projection (2050–2059). Densities are pooled over days, land pixels, and ensemble members.

Table A.3: JJAS daily-maximum temperature ($T_{\max}$) under the CMIP6 projection (2050–2059) compared against ERA5 (2010–2024). Quantiles report the future shift. Fixed thresholds report the exceedance ratio relative to ERA5. The exceedance rate is the probability that a land location reaches a daily maximum above the threshold on a given JJAS day, estimated over all days, land pixels, and ensemble members. W_1 is the 1-Wasserstein distance between the full distributions.

	ERA5	CMIP6	Shift / Ratio
P95 (°C)	36.92	38.72	+1.80
P99 (°C)	39.58	41.19	+1.60
P99.9 (°C)	42.18	43.79	+1.60
$\geq 35^\circ\text{C}$ rate (%)	10.78	20.62	1.91×
$\geq 40^\circ\text{C}$ rate (%)	0.73	2.37	3.24×
W_1 (°C)	2.15		

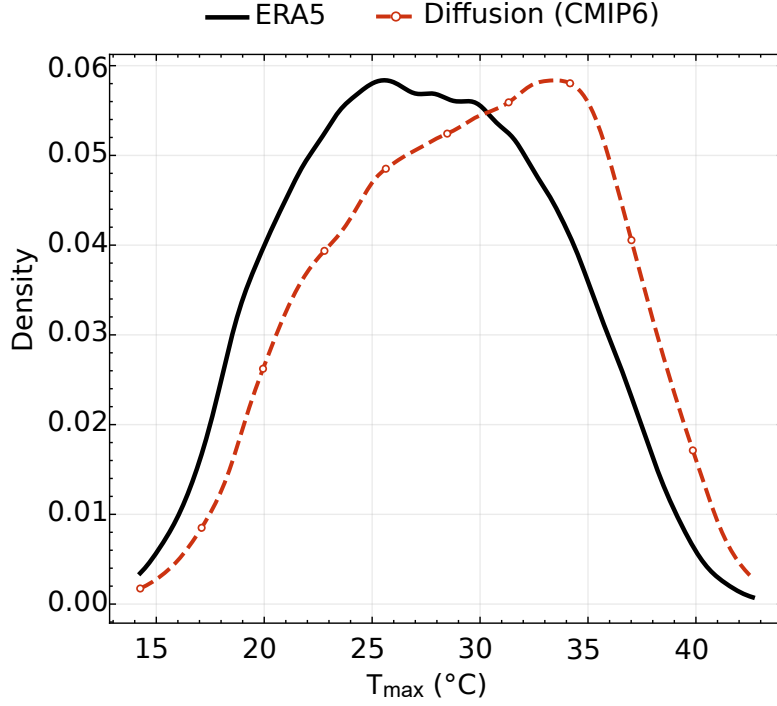


Figure A.5: Probability density of JJAS daily-maximum temperature ($T_{\max}$) over land grid points, comparing ERA5 (2010–2024) with the fields generated by the Full diffusion model under the CMIP6 SSP2-4.5 projection (2050–2059). Densities are pooled over days, land pixels, and ensemble members.

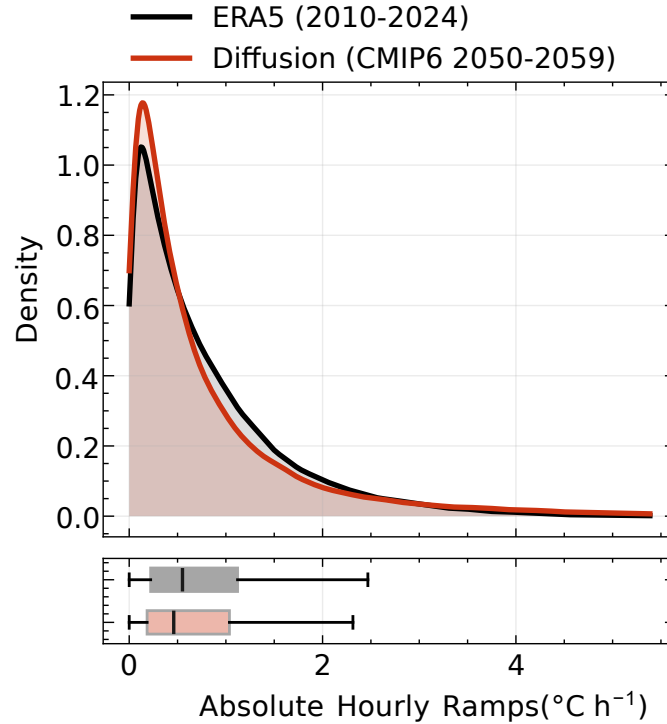


Figure A.6: Distribution of absolute hourly temperature ramps $|T_{h+1} - T_h|$ over land grid points, comparing ERA5 (2010–2024) with the Full diffusion fields under the CMIP6 SSP2-4.5 projection (2050–2059). The generated ramps closely match ERA5 in the bulk (mean absolute ramp ratio 1.005, $W_1 = 0.121$ $^{\circ}\text{C h}^{-1}$, KS = 0.048) with a slightly heavier tail (P99 of 4.79 versus 3.75 $^{\circ}\text{C h}^{-1}$). Boxplots summarize the pooled land-point distributions.

Table A.4: Circular statistics of the hour of daily maximum and minimum temperature over land grid points, comparing ERA5 (2010–2024) with the Full diffusion fields under the CMIP6 SSP2-4.5 projection (2050–2059). The circular mean gives the average hour of occurrence and the timing shift is the generated minus ERA5 difference. Arrows in the headers indicate that values closer to zero ($\rightarrow 0$) are better.

Metric	Circular mean hour (h)		Timing shift (h) ($\rightarrow 0$)	Modal hour (h)	
	ERA5	Diffusion (CMIP6)		ERA5	Diffusion (CMIP6)
$T_{\min}$ timing	4.45	4.33	−0.12	5	5
$T_{\max}$ timing	14.92	14.96	0.04	15	16

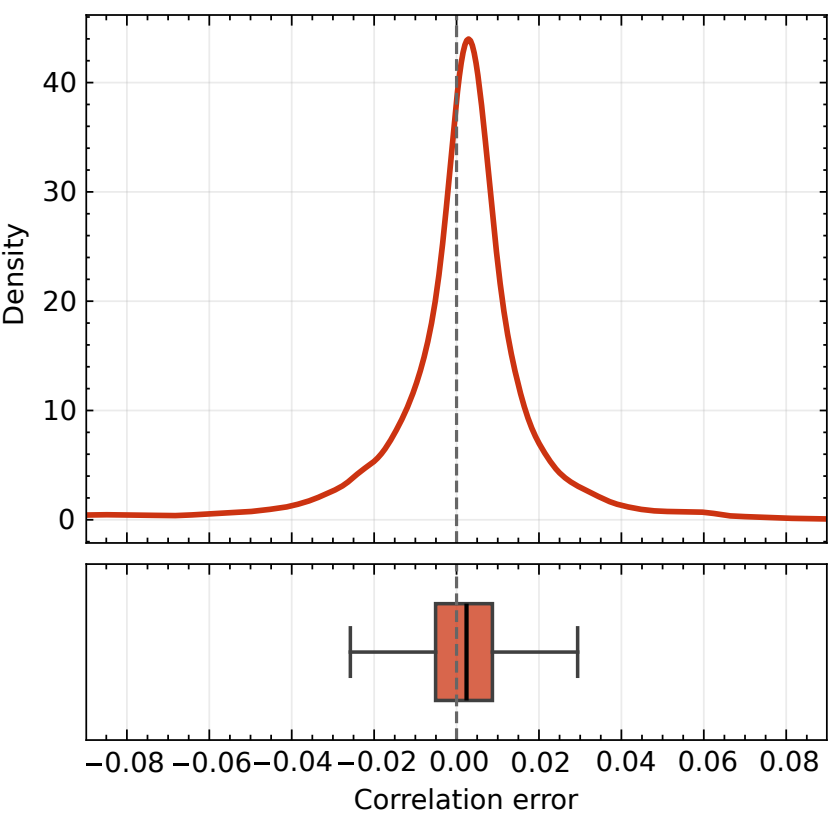


Figure A.7: Distribution of pairwise spatial-correlation errors (generated minus ERA5 (2010-2024)) over land grid points, for the generated fields under the CMIP6 SSP2-4.5 projection (2050–2059). Errors are tightly centered near zero (median 0.0024, RMSE 0.022). The boxplot summarizes the pooled distribution.