Trading off Controllability and Realism in Generative Physics Emulators

Francesco Immorlano [*1,2] **Elijah Tavares** [*1] **Felix Draxler** [1] **Juan Nathaniel** [3,2] **Violette Launeau** [4]
Padhraic Smyth [1] **Pierre Gentine** [3,2] **Stephan Mandt** [1,2]

Abstract

Machine-learned emulators trained on numerical models' outputs inherit systematic biases relative to real-world observations. While adapting these emulators toward observations can enhance their realism, it often reduces their diversity from a carefully constructed ensemble of simulations. This creates a fundamental *controllability-realism tradeoff*, where gaining accuracy comes at the cost of losing the distinct identities of the underlying simulators. In this work, we formalize this tradeoff for conditional generative emulators and introduce *condition-to-condition guidance*. Our method is an inference-time approach that interpolates between simulator and observational conditional score functions controlled by a guidance scale γ. We further provide a practical procedure for selecting an optimal $\gamma^\star$. We evaluate our framework on a climate science use case and compare it with a classical bias correction method.

1. Introduction

Modern machine-learned emulators aim to reproduce the distribution of outputs from numerical simulators to enable fast sampling while maintaining acceptable accuracy. A frequently overlooked limitation of such emulators is that they learn and replicate the biases of the simulation data they are trained on (Rasp et al., 2020). When data-driven emulators are adapted toward observational data to remove these biases, a fundamental and largely unaddressed tension arises: increasing *realism* (agreement with observations) often comes at the cost of *controllability* (the ability to generate simulator-specific outputs). In this work, we identify and study this *controllability-realism tradeoff* and propose a simple inference-time method to navigate it.

A natural question is how much adaptation to observations is appropriate. While fine-tuning (Hu et al., 2021; Kumar et al., 2022) or bias correction (Xu et al., 2021) can improve agreement with observational data, excessive adaptation can overwrite simulator-specific structure and reduce the usefulness of the emulator for uncertainty quantification. This raises a key question: how can we improve realism while preserving controllability?

We focus on generative emulators, specifically on diffusion models (Sohl-Dickstein et al., 2015), because of their ability to enable uncertainty quantification through repeated sampling and to capture the full statistical structure of the simulated fields, including extremes and spatial dependencies (Bassetti et al., 2024). At a high level, our approach interpolates continuously between two modes of generation: one that faithfully reproduces the diversity of simulation outputs across multiple models, and one that matches observational data. The interpolation is controlled by a single parameter γ at sampling time, so users can balance between realism and controllability without retraining the model.

In this work, we show that the controllability-realism tradeoff can be navigated at inference time, without modifying model parameters. Our main contributions are as follows:

- We formalize the *controllability–realism tradeoff* in conditional generative emulators, defining each axis through complementary *domain-specific* and *likelihood-based* metrics.
- We introduce *condition-to-condition guidance*, an inference-time method that interpolates between simulator and observational conditional score functions through a single scalar γ.
- We propose a practical procedure for selecting γ based on a saturation point in *realism* and a phase transition in *controllability*, with no additional training required.
- We demonstrate on a climate bias-correction task that our approach improves agreement with observations while preserving inter-model diversity, outperforming the deterministic Mean Variance Trend (MVT) baseline (Xu et al., 2021).

[*]Equal contribution [1]Department of Computer Science, UC Irvine, Irvine, USA [2]LEAP Center, New York City, USA [3]Columbia University, New York City, USA [4]Celest Science, Paris, France. Correspondence to: Francesco Immorlano <fimmorla@uci.edu>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

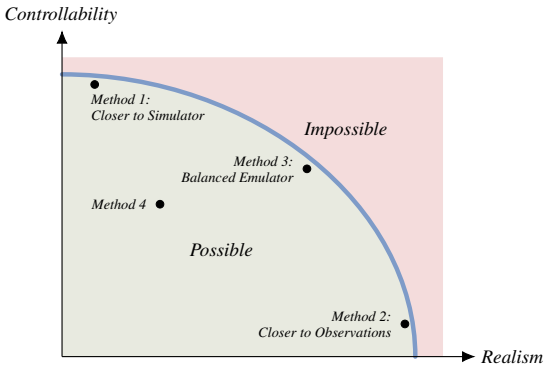


Figure 1. **Controllability-realism tradeoff (schematic).** Generative emulators cannot simultaneously maximize realism (statistical agreement with observations) and controllability (sensitivity to simulator-specific conditioning): the upper-right corner is unreachable. The frontier (here illustrative) separates the achievable region from the impossible one, and our method traces it at sampling time; an empirical trajectory is shown in Figure 2c. Inspired by Blau & Michaeli (2018).

2. Related Work

Numerical simulator outputs are routinely post-processed to remove systematic biases in the mean state, variability, extremes, and dependence structure. Classical statistical approaches include mean-variance correction, quantile mapping, trend-preserving corrections, and multivariate distributional transforms (Vrac & Friederichs, 2015; Cannon, 2018; Robin et al., 2019; Xu et al., 2021). More recent machine learning approaches replace prescribed transformations with learned nonlinear simulation-observations maps, ranging from tree-based (Watt-Meyer et al., 2021) to transfer-learning approaches (Pan et al., 2021; Kim et al., 2024; Immorlano et al., 2025).

A growing body of work treats bias correction as a generative distribution-matching problem, aiming to recover realistic joint variability across space, time, and state variables. Generative adversarial networks and image-to-image translation methods were among the first generative frameworks applied to this task (François et al., 2021; Fulton et al., 2023; McGibbon et al., 2024), followed by physically constrained adversarial networks (Hess et al., 2022; 2023) and methods that couple distributional correction with downscaling and uncertainty-aware sampling (Hess & Boers, 2024; Hess et al., 2025; Fang et al., 2025). Diffusion models have more recently gained traction in this setting (Aich et al., 2026; Springenberg et al., 2026). This line of work is closest to ours in model class, but differs in objective. Existing generative post-processing methods primarily aim to produce the best bias-corrected target-domain distribution. In contrast, we study a multi-model setting in which differences between simulators are not merely biases to remove, but encode valuable structural information about the modeled system.

Bias correction has also been framed as a fine-tuning problem, adapting pre-trained simulators to observational data via parameter-efficient methods (Hu et al., 2021; Kumar et al., 2022). These approaches optimize for alignment with the target distribution without providing explicit control over how much of each simulator's distinctive signal is retained. Our goal in this work is the opposite: to navigate the tradeoff between observational realism and preservation of inter-model diversity, rather than collapsing every simulator onto a single corrected distribution.

3. Condition-to-Condition Guidance

Because real-world training data may be scarce, for example if there is only a single realization of the system, simulations can provide the diversity needed to learn the full distribution of the data. However, if we learn a machine learning model on simulations alone, we also inherit the potential biases of the simulation. In this paper, we propose condition-to-condition guidance, a method to interpolate between models trained on simulations and real-world data.

Our goal is to explore the Pareto frontier between two conflicting goals:

- **Realism:** the extent to which generated samples replicate observational data.
- **Controllability:** the extent to which the generative model can reproduce the distinctive characteristics of different simulators when conditioned on them.

These objectives are inherently conflicting. Observational records are limited and provide a single realization of the system, while simulation ensembles are richer but biased relative to observations. Adapting an emulator toward observations to improve realism therefore tends to collapse simulator-specific differences and compromises controllability.

This tension is structurally analogous to the *perception-distortion tradeoff* in image restoration (Blau & Michaeli, 2018): distortion (i.e., closeness of a reconstruction to a reference) and perceptual quality (i.e., closeness of the reconstructed distribution to the natural-image distribution) cannot be jointly optimized below a fundamental boundary, regardless of the algorithmic scheme. The same impossibility arises in our setting. At one extreme, an emulator that matches the distribution of observations, no matter which simulator it is asked to imitate, has lost all signal that distinguishes one simulator from another. On the other hand, an emulator that perfectly preserves every simulator's distribution inherits each simulator's bias against observations. No estimator can reach both corners simultaneously, and Figure 1 sketches the resulting plane: a possible region, an

impossible one, and a convex frontier whose exact location depends on the data and model capacity but whose qualitative shape is forced by this argument. We do not provide a formal proof of convexity nor of the precise location of the impossible region. A complete characterization is the subject of ongoing work and will appear in a future extension of this paper. Figure 1 is therefore intended as a qualitative illustration.

The remainder of this section describes our method, and the structure of the above tradeoff. In Section 4, we present a concrete application to climate science.

3.1. Setup and notation

We consider the problem of modeling high-dimensional physical fields, such as daily temperature maps over a geographic region. Let $\mathbf{x}$ denote such a field (e.g., a 16×32 grid), and let c denote associated covariates such as the day of the year or a proxy for long-term forcing (e.g., Carbon Dioxide Equivalent (CO_2e) concentration). We distinguish two domains: a *source* domain (src), corresponding to simulation data, and a *target* domain (tgt), corresponding to observational data. The goal is to model the conditional distribution given the covariates in tgt, while leveraging the richer variability available in src.

Formally, our object of interest is the tgt conditional distribution

$$p_\theta(\mathbf{x} \mid \mathrm{tgt}, c), \tag{1}$$

which is the distribution of real-world observations $\mathbf{x}$ given covariates c.

Training a generative model directly on Equation (1) is limited by a fundamental data scarcity: in physical sciences, we often study chaotic and non-linear systems, yet with only a single historical realization (Eyring et al., 2016; Sippel et al., 2019; Deser et al., 2020). While such systems may be measured via diverse modalities, the lack of independent samples from the underlying stochastic process limits the distribution the model can recover. Therefore, we also use a larger pool of simulation data, sampled from

$$p_\theta(\mathbf{x} \mid \mathrm{src}, c). \tag{2}$$

This provides the necessary statistical coverage that a single historical realization lacks. While simulations contain biases, the availability of different numerical models simulating the same phenomenon multiple times enables the generative model to capture the full range of structural uncertainty inherent in the system's dynamics.

3.2. Learning distributions with diffusion models

Our approach is based on diffusion models, a powerful class of machine learning models that learn the distribution of data from samples (Sohl-Dickstein et al., 2015; Ho et al., 2020). Later, we will interpolate between the diffusion neural networks trained on the different types of data we have available to span the controllability-realism tradeoff.

In a nutshell, diffusion generate samples from a learned distribution $p_\theta(\mathbf{x})$ by modifying samples from a standard normal latent distribution $p(\mathbf{x}_T)$ through a stochastic or ordinary differential equation (ODE). A neural network $s_\theta(\mathbf{x}_t, t)$ parameterizes these differential equations, and we call the resulting distribution at time $t = 0$ the learned distribution $p_\theta(\mathbf{x})$. The ODE for sampling from a diffusion model is given by (Karras et al., 2022):

$$\frac{dx}{dt} = \frac{\dot{a}_t}{a_t} x - b_t^2 \dot{a}_t a_t \underbrace{\nabla_{x_t} \log p(x_t)}_{\text{learned } s_\theta(x_t, t)}. \tag{3}$$

We refer to Karras et al. (2022) for details on training diffusion models, choices for coefficients a_t, b_t, as well as sampling methods.

In the next section, we replace the learned network $s_\theta(x_t, t)$ with one that interpolates between network predictions on observational and simulated data.

3.3. Distribution interpolation via score guidance

The core idea of our method is to sample from a conditional diffusion model that interpolates between the src and tgt domains. We achieve this by interpolating between the score corresponding to a src domain and the tgt score corresponding to the observational data. The interpolation is controlled by *guidance scale* γ, a scalar parameter that enables continuous adjustment of the controllability-realism tradeoff without retraining.

Let i denote the index of a particular simulator among M available simulators (i.e., src_i, $i \in \{1, \ldots, M\}$; in the following we use *i* or src_i synonymously when we refer to a specific simulator). For the same covariates c, we interpolate the conditional score between src_i and tgt:

$$\begin{aligned}\nabla_{\mathbf{x}_t} \log p_{\theta,\gamma,i}(\mathbf{x}_t \mid c) = (1-\gamma)\, \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t \mid \mathrm{src}_i, c) \\ + \gamma\, \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t \mid \mathrm{tgt}, c).\end{aligned} \tag{4}$$

In other words, at every step of the reverse diffusion process, the score that drives the sampling is a combination of two conditional scores: one for simulator i and one for the observational data. Both conditional score functions must be learned by the same network to enable this sampling-time interpolation. The scalar γ tunes the strength of this adaptation: $\gamma = 0$ recovers pure src_i emulation (maximum controllability) and $\gamma = 1$ collapses to pure tgt emulation (maximum realism), thus γ is naturally bounded to $[0, 1]$. Values outside $[0, 1]$ are extrapolations beyond src_i and tgt, and lack a meaningful interpretation in our setting; we therefore restrict $\gamma \in [0, 1]$ throughout.

This formulation is conceptually related to *classifier-free guidance* (Ho & Salimans, 2021), but in this work we consider two conditional distributions rather than an unconditional and a conditional one. It is also similar to Domain Guidance (Zhong et al., 2025), which we generalize to the multi-models setting where we have access to a large pool of src-domain simulation data produced by different numerical models. Also, we are not only interested in achieving high-quality samples from tgt (realism), but also in preserving the variability across src_i (controllability). Note that a weighted sum of conditional scores is not the score of a mixture: Equation (4) samples from the geometric mean $p_{\theta,\gamma,i}(\mathbf{x}_t \mid c) \propto p_\theta(\mathbf{x}_t \mid \mathrm{src}_i, c)^{1-\gamma}\, p_\theta(\mathbf{x}_t \mid \mathrm{tgt}, c)^{\gamma}$. Because the combination is applied to the conditional scores at each noise level rather than to the clean conditionals before noising, the resulting samples do not target exactly the geometric mean. This is the same approximation underlying classifier-free guidance, with no closed-form bound.

We have left open how to achieve the aforementioned tradeoff by choosing the guidance scale γ, and address this in the next section, in which we also elaborate on the controllability-realism tradeoff that guides the selection of the scale.

3.4. Controllability-realism tradeoff

When adapting the generative emulator from src toward tgt, the Pareto frontier defined in Figure 1 shows that there's no guidance scale γ that makes the emulator fully realistic and fully controllable. Thus, making the tradeoff operational requires expressing realism and controllability as quantities that can be computed and tracked as γ varies.

We formulate the two concepts in distributional terms. Realism is the agreement of the guided distribution with the empirical observational distribution (q_{tgt}): a fully realistic guided model generates samples that are statistically indistinguishable from the observations. Controllability is the agreement of the guided distribution with each simulator-conditional distribution (q_{src_i}) that the unguided emulator was trained to reproduce. Both quantities are functions of γ for a fixed simulator i, and they move in opposite directions: increasing γ pushes the guided distribution toward q_{tgt}, raising realism and lowering controllability. The tradeoff is therefore the only basis for the choice. The operating point $\gamma^\star$ we look for is the largest γ at which realism has saturated and controllability has not yet collapsed (i.e., the elbow of the controllability-realism trajectory in Figure 1). Locating this elbow empirically requires concrete operational metrics on both axes.

We introduce two complementary families of metrics. *Likelihood-based metrics* are domain-agnostic and information-theoretic: *realism*, *controllability*, and *variability* are expressed as expectations of $\log p_{\theta,\gamma}$ under different reference distributions. *Domain-specific metrics* are physics-grounded counterparts of the same diagnostics computed directly on the generated fields, chosen to capture the phenomena that matter most in the specific application domain. The two families are designed to guide the selection of $\gamma^\star$; we present each in turn and then combine them in a single selection rule.

Likelihood-based metrics. The natural information-theoretic measure of realism is the divergence of the guided model from the empirical distribution of tgt data,

$$D_{\mathrm{KL}}\left(q_{\mathrm{tgt}}(\cdot \mid c) \;\|\; p_{\theta,\gamma}(\cdot \mid \mathrm{src}_i, c)\right) = \underbrace{\mathbb{E}_{q_{\mathrm{tgt}}}[\log q_{\mathrm{tgt}}(\mathbf{x} \mid c)]}_{-H[q_{\mathrm{tgt}}]} - \mathbb{E}_{q_{\mathrm{tgt}}}\left[\log p_{\theta,\gamma}(\mathbf{x} \mid \mathrm{src}_i, c)\right]. \quad (5)$$

Because the entropy of q_{tgt} does not depend on γ, minimizing (5) in γ is equivalent to maximizing the cross-entropy term. We can therefore compare the realism of different models via the cross-entropy:

$$\mathcal{R}(\gamma; i) = \mathbb{E}_{q_{\mathrm{tgt}}}[\log p_{\theta,\gamma}(\mathbf{x} \mid \mathrm{src}_i, c)], \quad (6)$$

In practice, we estimate (6) by Monte Carlo sampling over held-out tgt samples, evaluating the guided log-likelihood as described in Appendix A.2. If the target distribution is given by a single realization (e.g. climate data), we need to ensure that we systematically extract a portion of the data that can be considered independent of the training data, as we detail later.

Controllability is defined by applying the same construction to the simulator *i* conditional distribution:

$$\mathcal{C}(\gamma; i) \;=\; \mathbb{E}_{q_{\mathrm{src}_i}}[\log p_{\theta,\gamma}(\mathbf{x} \mid \mathrm{src}_i, c)]\,, \quad (7)$$

Together $\mathcal{R}(\gamma; i)$ and $\mathcal{C}(\gamma; i)$ trace the controllability-realism trajectory of Figure 1.

Variability is a complementary diagnostic computed as the entropy of the guided distribution itself:

$$\begin{aligned}\mathcal{V}(\gamma; i) &= H(p_{\theta,\gamma}(\cdot \mid \mathrm{src}_i, c)) \\ &= \mathbb{E}_{p_{\theta,\gamma}}[\log p_{\theta,\gamma}(\mathbf{x} \mid \mathrm{src}_i, c)]\end{aligned} \quad (8)$$

estimated by Monte Carlo sampling over samples drawn from $p_{\theta,\gamma}$ itself. Unlike $\mathcal{R}$ and $\mathcal{C}$, which compare the guided model against external references, variability is intrinsic and measures how dispersed the guided distribution is in absolute terms. Tracking $\mathcal{V}$ matters because excessive guidance can collapse the guided distribution into a narrow mode close to a few likely observational data. Variability thus exposes mode-collapse as additional failure modes of strong guidance, orthogonal to the controllability-realism tradeoff.

Domain-specific metrics. Domain-specific metrics compute realism and controllability directly on generated samples using domain-specific quantities. Their definition is application-specific and must be chosen according to the requirements of tgt.

For the climate setup of this paper (Section 4), we choose Continuous Ranked Probability Score (CRPS) as the primary realism diagnostic (Appendix D). It penalizes both the distance between the ensemble mean of the generated samples and the observation, and the miscalibration of the ensemble spread, catching those failure modes that Root Mean Squared Error (RMSE) alone cannot detect. The *realism* minimum identifies the elbow of the CRPS curve; guidance should not be pushed beyond it.

The controllability domain-specific diagnostic we propose is the mean pattern correlation $\bar{r}(\gamma)$ of inter-model anomalies. For each simulator i, we draw samples from the guided model $p_{\theta,\gamma}(\cdot \mid \mathrm{src}_i, c)$ and average them over the evaluation period to obtain a climatology $\bar{\mathbf{x}}^{(i,\gamma)}$. The inter-model anomaly $a^{(i,\gamma)}$ is the deviation of $\bar{\mathbf{x}}^{(i,\gamma)}$ from the ensemble mean computed across all conditionings src_i of the same guided emulator. It is the spatial fingerprint by which the guided emulator, when conditioned on src_i, distinguishes itself from the ensemble average. Pattern correlation between $a^{(i,\gamma)}$ and the unguided reference $a^{(i,0)}$ is amplitude-invariant: it asks whether the shape of each conditioning's signature is preserved under guidance, independent of how its magnitude has shrunk. This is a necessary condition for controllability: conditioning the guided emulator on src_i should still produce i's spatial pattern, even after that pattern has been pushed toward tgt. Averaging $r^{(i)}(\gamma)$ over all simulator conditioning gives $\bar{r}(\gamma)$ (full derivation in Appendix D.2). However, an amplitude-aware diagnostic would quantify the magnitude of the shrinkage and complement $\bar{r}(\gamma)$, which on its own cannot distinguish a preserved fingerprint from one whose amplitude has uniformly collapsed; this is left to future work. Three regimes characterize $\bar{r}(\gamma)$:

- $\bar{r}(\gamma) \approx 1$: each simulator's spatial fingerprint has the same shape as without guidance; only the magnitude may have changed. Identity is preserved.
- $\bar{r}(\gamma) \approx 0$: deviations from the ensemble mean are uncorrelated with the unguided ones. Identity is lost.
- $\bar{r}(\gamma) \approx -1$: locations where simulator i was warmer than the ensemble at $\gamma = 0$ are now cooler, and vice versa. Identity is inverted.

CRPS and the pattern correlation of inter-model anomalies are concrete, physics-grounded instantiations of $\mathcal{R}$ and $\mathcal{C}$ for the climate setting we use to evaluate our method. The flexibility of choosing physics-grounded domain-specific diagnostics is a strength of this framework: the right diagnostics depend on what realistic and controllable mean in tgt, while the selection rule we describe in the next section remains the same.

Selection rule. Whether measured by likelihood-based or domain-specific metrics, realism improves monotonically with γ up to a saturation point, beyond which additional guidance no longer reduces error against tgt and may also increase it. Controllability instead remains close to its baseline value up to a critical γ, beyond which it drops once the tgt conditioning dominates. The value of $\gamma^\star$ to be selected is the largest one at which realism has saturated and controllability has not yet collapsed (i.e., the elbow of the controllability-realism trajectory). $\mathcal{V}(\gamma)$ is monitored simultaneously to ensure that the guidance does not cause mode collapse. Figure 2 shows the application of this rule to the climate setup; the empirical analysis is in Section 5.2.

4. Experiments: bias correction of CMIP6 climate models

In this section, we show the application of condition-to-condition guidance to a climate science use case.

The source domain consists of 2-meter temperature maps simulated by nine Coupled Model Intercomparison Project phase 6 (CMIP6) global climate models (Table B.1) (Eyring et al., 2016); the target domain corresponds to ECMWF Reanalysis v5 (ERA5) reanalysis data (Hersbach et al., 2020), used as the observational reference. Both datasets contain daily temperature maps covering the European region (16×32) over 1960–2016 (Appendix B). This is a *multi-model ensemble* setting: different numerical models (src_i) simulate the same physical process. We use a single realization per model. The goal is to generate temperature fields with statistical properties matching ERA5 (realism) while preserving each CMIP6 model's distinctive identity relative to the ensemble (controllability).

In the next sections we go through the pipeline end-to-end: a single conditional diffusion model is first trained jointly on the nine CMIP6 models and ERA5, and its unguided ($\gamma = 0$) emulation quality is assessed across all conditionings (Section 5.1). Then the guidance scale is selected from the controllability-realism tradeoff using both likelihood-based and domain-specific metrics (Section 5.2.1). Finally, we apply guidance at the selected $\gamma^\star$ to a representative simulator (MIROC6) and evaluate the resulting bias correction (Section 5.2.2).

4.1. Generative multi-model emulator

Condition-to-condition guidance requires the conditional distributions on both src and tgt to be learned, so that interpolation between them is well-defined. We therefore train a

single conditional diffusion model on

$$p_\theta(\mathbf{x} \mid s, c), \quad s \in \mathcal{S} \equiv \{\underbrace{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9}_{\mathrm{src}_1,\ldots,\mathrm{src}_M}, \underbrace{\texttt{ERA5}}_{\mathrm{tgt}}\}, \tag{9}$$

where $\mathbf{x} \in \mathbb{R}^{16\times 32}$ is the daily 2-meter temperature map, s indexes the nine CMIP6 simulators $\mathrm{src}_i = \texttt{CMIP6}_\texttt{i}, i \in \{1, \ldots, 9\}$ and $\mathrm{tgt} = \texttt{ERA5}$, and $c = \{d, \mathrm{CO_2e}(y)\}$ includes the day-of-year $d \in \{1, \ldots, 365\}$ and global-mean annual CO_2e of year y (a proxy for forced regional warming). Both endpoints required by the guidance (each src_i and tgt) are therefore learned jointly by the same parameters θ. The training schedule is implicitly 1:9 imbalanced (one ERA5 trajectory vs. nine CMIP6 trajectories of comparable length); Section 5.1 verifies that the ERA5 conditional is at least as well learned as the CMIP6 ones, so that the score interpolation in Eq. (4) is well-defined at both endpoints. We use 1960–2000 for training and 2001–2016 for testing (further details in Appendix A).

5. Results

5.1. Quality of the unguided generative emulator

In this section, we evaluate the trained diffusion model at $\gamma = 0$, conditioned on both $\mathrm{src}_i, i \in \{1, \ldots, 9\}$ and tgt, which corresponds to the unguided emulator on which *condition-to-condition guidance* subsequently operates. We sample 20 independent realizations per day from $p_\theta(\mathbf{x} \mid \mathrm{src}_i, c)$ and $p_\theta(\mathbf{x} \mid \mathrm{tgt}, c)$ over the 2001–2016 test period, and ask whether the conditional distributions are properly learned across mean, tails, distributional shape, spatial structure, and probabilistic calibration.

Mean, seasonal cycle, tails, and extremes. The regional-mean ensemble trajectory is close to the ground truth across all ten conditionals with the source curve inside the $\pm 1\sigma$ band (Figure E.8), and signed bias that fluctuates around zero (Figure E.9). Mean-bias maps (Figure E.1), std-ratio maps (Figure E.3) and seasonal-amplitude biases (Figure E.2) are close to reference for nine of ten conditionals. IPSL-CM6A-LR is the consistent outlier with cold offset, localized overdispersion, and damped continental amplitude. The monthly seasonal cycle (Figure E.10) is replicated for every conditional. Cold-tail errors dominate warm-tail ones: TN10p and Q05 biases (Figures E.5 and E.6) span a wider range than TX90p and Q95 ones (Figures E.4 and E.7), with IPSL-CM6A-LR (TN10p $=$ 3.12 pp, Q05 $=$ 2.23°C) and MPI-ESM1-2-HR (TN10p $=$ 2.64 pp, Q05 $=$ 1.46°C) the strongest outliers (Table E.1).

Distributional shape, spatial structure and calibration. Kernel Density Estimation (KDE) plots (Figure E.11) show a bimodal regional-mean distribution for every conditioning, with a winter mode near 0 °C and a summer mode near 20−25 °C; peak locations and tails align with the source. Quantile-Quantile (Q-Q) plots (Figure E.12) follow the diagonal; mean-absolute skewness bias spans 0.05–0.09 and kurtosis bias 0.07–0.20 (Table E.1). Spatial power spectra (Figure E.14) overlap all sources spectra across all eight wavenumber bins, indicating neither spectral whitening nor oversmoothing. Rank histograms (Figure E.13) are approximately uniform with Reliability Indexs (RIs) from 4×10^{-5} (MRI-ESM2-0) to 1.2×10^{-3} (EC-Earth3-Veg-LR). Crucially, ERA5 is comparable to the best CMIP6 conditionals on every tail metric: this matters for the score interpolation in Eq. (4) to be well-defined because ERA5 is the endpoint toward which guidance steers. This also confirms that, despite the 1:9 training imbalance, the ERA5 distribution is at least as well learned as CMIP6 ones.

5.2. Condition-to-Condition Guidance

We select the MIROC6 climate model simulations as the case study to assess the performance of the condition-to-condition guidance approach proposed. The model is characterized by pronounced biases over the European domain (Thorarinsdottir et al., 2020; Ossó et al., 2023; Palmer et al., 2023) and is therefore a non-trivial conditional to learn and unbias. We evaluate guidance on the test period 2001–2016, held out from training. We first use the conditional diffusion model as an emulator of MIROC6 (src_5, $\gamma = 0$) to generate 20 samples for each day in the test period. The same 20 samples are then bias-corrected with the MVT method (Xu et al., 2021) calibrated on 1960–2000 statistics to match the diffusion training period and avoid information leakage from the test period (Appendix C). Thus, both approaches are applied to an identical set of model-conditioned data. Per-model results for the remaining eight CMIP6 models are reported in Appendix F. Details of each metric can be found in Appendix D.

5.2.1. Selection of the guidance scale γ

We generate 20 samples per day over the test period (2001–2016) for $\gamma \in [0, 1]$ with step 0.125 and $\mathrm{src}_i \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9\}$.

Figure 2a shows CRPS against ERA5 as a function of γ for each model. All nine curves decrease monotonically up to $\gamma \approx 0.875$–1.0 and reach a minimum near $\gamma = 0.9$. MIROC6 (brown) starts at the highest CRPS at $\gamma = 0$ and moves toward the rest of the ensemble as γ increases. The additional realism gain from $\gamma = 0.9$ to $\gamma = 1$ is less than 1% for every model, while $\bar{r}(\gamma)$ collapses from ≈ 0.95 to ≈ 0.1 over the same interval (Figure 2b). This means that realism is saturated and controllability is the binding constraint at the operating point. Figure 2c,d show the likelihood-based $\mathcal{C}(\gamma)$-$\mathcal{R}(\gamma)$ and $\mathcal{V}(\gamma)$-$\mathcal{R}(\gamma)$ trajectories for all nine models; every model's trajectory exhibits a clear

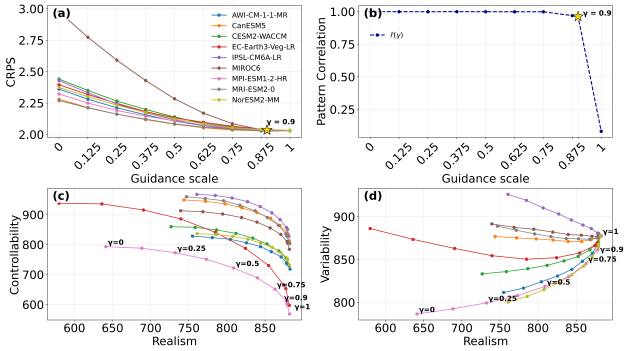


Figure 2. **Selection of the guidance scale γ (2001–2016 test period, $\gamma \in [0, 1]$, step 0.125) (a)** CRPS of the guided ensemble against ERA5 as a function of γ, for each of the nine CMIP6 models. Every curve reaches its minimum near $\gamma = 0.9$. **(b)** Mean pattern correlation $\bar{r}(\gamma)$ of inter-model anomalies (Appendix D.2): above 0.95 through $\gamma \approx 0.875$, dropping to ≈ 0.1 at $\gamma = 1$. Stars in panels (a)–(b) mark $\gamma^\star = 0.9$, the operating guidance scale. **(c)** $\mathcal{C}(\gamma)$-$\mathcal{R}(\gamma)$ and $\mathcal{V}(\gamma)$-$\mathcal{R}(\gamma)$ trajectories; $\mathcal{V}(\gamma)$ stays within a narrow band for $\gamma < 0.9$ and converges across CMIP6 models at $\gamma = 1$, with MIROC6 and CanESM5 showing the largest excursions.

elbow near $\gamma = 0.9$, which is consistent across the full ensemble. Applying the selection rule of Section 3.4, we set $\gamma^\star = 0.9$: realism has saturated and $\bar{r}(\gamma)$ has not yet collapsed, and variability remains well within the range of the unguided ensemble at this value. Throughout the paper, $\gamma^\star = 0.9$ denotes the selected operating scale. Note that although the analysis focuses on MIROC6, the same procedure applies to any model in the ensemble. Figure F.3 visualizes this progression: for every CMIP6 model the mean-bias map collapses toward zero as γ increases, with most of the residual bias removed by $\gamma \in [0.875, 0.9]$ and the $\gamma = 1$ column converging to the bias of pure ERA5 conditioning.

5.2.2. MIROC6 BIAS CORRECTION

We compare condition-to-condition guidance and MVT (Xu et al., 2021) along four axes: the mean state and its spatial pattern, the variance and its calibration, the seasonal cycle, the shape and the tails of the distribution. Figure 3 is the main spatial diagnostic, while Table 1 summarizes the scalar metrics.

Mean bias. The raw MIROC6 emulator has a large warm bias relative to ERA5, exceeding $+5$°C in eastern and northern Europe (Figure 3), consistent with MIROC6's documented European warm bias (Ossó et al., 2023). MVT leaves a fairly uniform residual $\sim$1°C across central and southern Europe (Mean RMSE $= 1.02$°C, Table 1); this is the climate-change drift between the MVT calibration period (1960–2000) and the test period (2001–2016) that a stationary affine correction cannot track. Guidance halves the residual to 0.50°C, with most of the domain within ± 0.5°C because the diffusion conditioning on CO_2e(y) is year-aware and adapts to the warming trend.

Seasonal amplitude. The emulator has a positive seasonal-amplitude bias of up to $+10$ °C across most of the domain (Figure 3). MVT leaves a $+5$ to $+10$ °C residual band across northern Europe and Scandinavia with mild overcorrection in the south, partially cancelling in the regional mean. Guidance leaves a residual within $\pm 1 - 2$ °C. The seasonal cycle (Figure F.1) shows both methods tracking ERA5 on regional-averaged monthly means, but Figure 3 reveals that MVT's agreement is partly due to cancellation, while guidance is locally accurate.

Standard deviation ratio. The raw emulator is overdispersed across most of the domain); both methods reduce the residual to concentrate above 60°N. The mean absolute deviation $|\sigma_{\text{pred}}/\sigma_{\text{ERA5}} - 1|$ is 0.03 for guidance versus 0.08 for MVT (Table 1). MVT's per-cell rescaling inherits the spread of the underlying diffusion samples, while guidance restructures that spread during sampling.

Extremes. TX90p/TN10p measure the frequency at which daily temperatures cross ERA5's 90th / 10th percentiles, as a deviation from the expected 10%. The two methods are comparable on the warm tail (TX90p bias 2.03 vs. 2.70 percentage points (pp); Q95 0.58°C vs. 0.79°C) but guidance is more than twice as accurate on the cold tail (TN10p 1.50 vs. 3.55 pp; Q05 0.79 vs. 1.71°C). European cold extremes are associated with circulation regimes that pixel-independent affine transformations cannot capture.

Distributional shape and calibration. The regional-mean KDE for the raw emulator is bimodal with the summer peak several degrees warm of ERA5 (Figure 4a, left); both methods recover bimodal structure and peak locations, though MVT slightly overshoots the summer peak. Q-Q plots are close to the diagonal for guidance; MVT shows a visible cold-tail departure (Figure 4b, center), in line with the Q05/TN10p results. Both methods produce essentially flat rank histograms in Figure F.2 (RI $= 3 \times 10^{-5}$ for guidance, 9×10^{-5} for MVT, Table 1) at this sample size for MIROC6. However, MVT shows some miscalibration for other CMIP6 models (Figure F.7).

Spatial power spectrum. Guidance overlaps ERA5 across all wavenumbers; MVT sits above at intermediate wavenumbers and exceeds it by roughly $7\times$ at the highest bin, a consequence of MVT's per-cell rescaling factor injecting sampling noise that manifests as high-wavenumber power not present in ERA5 (Figure 5). The same effect appears in the Laplacian roughness of the mean-bias field (0.48 °C for guidance, 0.75 °C for MVT; Table 1).

The conclusions drawn for MIROC6 extend to the rest of

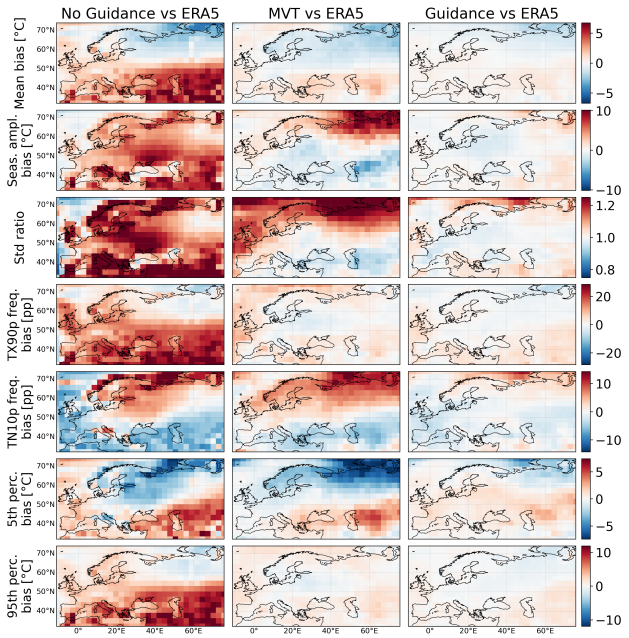
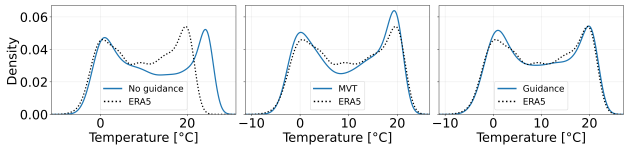
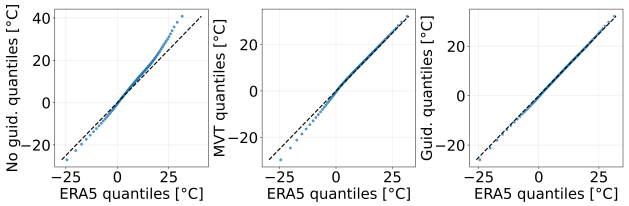


Figure 3. **Bias diagnostics for MIROC6 (2001–2016).** Rows: mean bias, seasonal-amplitude bias, std-ratio $\sigma_{\text{pred}}/\sigma_{\text{ERA5}}$, TX90p/TN10p frequency biases (pp relative to 10%), and 5th/95th-percentile biases. Columns: uncorrected diffusion ($\gamma = 0$, left), MVT-corrected (center), guidance-corrected at $\gamma^\star = 0.9$ (right). Scalars metrics are reported in Table 1.



(a) KDE of regional-mean daily temperature.



(b) Q-Q plots vs. ERA5, pooled across grid cells and days; dashed line = 1:1 reference.

Figure 4. **Distributional diagnostics for MIROC6 (2001–2016).** Columns: uncorrected ($\gamma = 0$, left), MVT (center), guidance at $\gamma^\star = 0.9$ (right). Top row (KDE): summer mode shifts from $\sim 25\,^\circ$C at $\gamma = 0$ to $\sim 20\,^\circ$C for guidance, matching ERA5. Bottom row (Q-Q): the raw S-shape is corrected to near-diagonal by guidance, while MVT retains a cold-tail departure below $\sim -20\,^\circ$C.

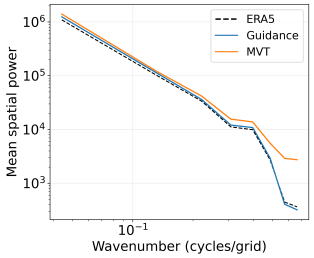


Figure 5. Radially-averaged 2D spatial power spectrum of daily 2 m temperature fields for MIROC6 (2001–2016 test period). Guidance at $\gamma^\star = 0.9$ (blue) overlaps ERA5 (dashed black) across all wavenumbers; MVT (orange) stays slightly above both at intermediate wavenumbers and diverges at the highest bins.

Table 1. **Scalar metrics for MIROC6 (2001–2016).** Lower is better. **Bold** marks the better value between guidance and MVT.

Metric	Guidance	MVT
Mean RMSE [°C]	**0.50**	1.02
Variance ratio deviation (mean)	**0.03**	0.08
Skewness bias (mean abs)	0.08	0.08
Kurtosis bias (mean abs)	**0.10**	0.20
TX90p bias [pp]	**2.03**	2.70
TN10p bias [pp]	**1.50**	3.55
Q05 bias [°C]	**0.79**	1.71
Q95 bias [°C]	**0.58**	0.79
CRPS [°C]	**1.96**	2.26
Reliability index	$\mathbf{3\times10^{-5}}$	9×10^{-5}
Laplacian roughness [°C]	**0.48**	0.75

the CMIP6 models. Repeating the same comparison for the remaining eight models (Appendix F) shows that guidance outperforms MVT on every scalar metric for every simulator.

6. Discussion and Conclusions

We introduced condition-to-condition guidance as an inference-time method for correcting systematic biases of a generative emulator relative to real-world data. A single conditional diffusion model is trained jointly on M source simulators and one target observational dataset; at sampling time a guidance scale γ continuously interpolates between the score function conditioned on simulator i and the score conditioned on observations, with no per-model fine-tuning or modification of the trained weights.

Bias correction and ensemble diversity turn out to be competing goals: differences among simulators are not merely errors but encode distinct assumptions about the same physical process, and the inter-model spread is how structural uncertainty is represented in physical phenomena. Increasing γ improves realism, the statistical agreement of generated samples with observations, but reduces controllability, the sensitivity of the emulator to simulator-specific conditioning. The selection rule we propose selects the largest

γ at which realism has saturated and controllability has not yet collapsed; on the climate use case we study, this yields $\gamma^\star = 0.9$ uniformly across all nine CMIP6 models. At this operating point, guidance substantially reduces the biases of the unguided emulator across mean, variability, extremes and spatial structure, and outperforms the deterministic MVT baseline, especially on the cold tail and in recovering the spatial power spectrum.

The framework is not specific to climate: it applies whenever a numerical simulator and a reference observational dataset can be jointly modelled by a single conditional generative model, with comparable accuracy at both endpoints. Scaling to global, higher-resolution outputs and multivariate corrections will introduce further challenges; we hope that this work stimulates further research into how diffusion guidance can address systematic bias in physics simulators.

Accessibility

Authors are kindly asked to make their submissions as accessible as possible for everyone including people with disabilities and sensory or neurological differences. Tips of how to achieve this and what to pay attention to will be provided on the conference website http://icml.cc/.

Software and Data

The CMIP6 temperature simulations used in this study were accessed through the LEAP-Pangeo platform: https://leap-stc.github.io/intro.html. ERA5 reanalysis temperature data are publicly available from the Copernicus Climate Change Service Climate Data Store at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview.

Acknowledgements

We acknowledge funding from National Science Foundation (NSF) through Learning the Earth with Artificial Intelligence and Physics (LEAP) Science and Technology Center (STC) (Award #2019625). Stephan Mandt acknowledges funding from NSF through an NSF CAREER Award IIS-2047418, and funding from the Hasso Plattner Research Center at UCI. Pierre Gentine acknowledges funding from the European Research Council Synergy Grant "Understanding and modeling the Earth System with Machine Learning" under the Horizon 2020 research and innovation programme (Grant Agreement No. 855187). Padhraic Smyth acknowledges funding from NSF through NSF award CAIG-2425932.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning for physical-science emulators. Improved bias correction of data-driven emulators could inform downstream impact assessments.

References

Aich, M., Hess, P., Pan, B., Bathiany, S., Huang, Y., and Boers, N. Conditional diffusion models for downscaling and bias correction of earth system model precipitation. *Geoscientific Model Development*, 19(4):1791–1808, 2026. doi: 10.5194/gmd-19-1791-2026. URL https://gmd.copernicus.org/articles/19/1791/2026/.

Bassetti, S., Hutchinson, B., Tebaldi, C., and Kravitz, B. Diffesm: Conditional emulation of temperature and precipitation in earth system models with 3d diffusion models. *Journal of Advances in Modeling Earth Systems*, 16(10):e2023MS004194, 2024. doi: https://doi.org/10.1029/2023MS004194. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023MS004194. e2023MS004194 2023MS004194.

Blau, Y. and Michaeli, T. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini, P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.-L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.-A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.-Y., Guenet, B., Guez, Lionel, E., Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.-B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N. Presentation and evaluation of the ipsl-cm6a-lr climate model. *Journal of Advances in Modeling Earth Systems*, 12(7):e2019MS002010, 2020. doi: https://doi.org/10.1029/2019MS002010. URL https://agupubs.onlinelibrary.wiley.

com/doi/abs/10.1029/2019MS002010. e2019MS002010 10.1029/2019MS002010.

Brenowitz, N. D., Ge, T., Subramaniam, A., Manshausen, P., Gupta, A., Hall, D. M., Mardani, M., Vahdat, A., Kashinath, K., and Pritchard, M. S. Climate in a bottle: Towards a generative foundation model for the kilometer-scale global atmosphere, 2025. URL https://arxiv.org/abs/2505.06474.

Cannon, A. J. Multivariate quantile mapping bias correction: an n-dimensional probability density function transform for climate model simulations of multiple variables. *Climate dynamics*, 50(1):31–49, 2018.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G. The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020. doi: https://doi.org/10.1029/2019MS001916. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001916. e2019MS001916 2019MS001916.

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X., and Stull, R. B. Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research: Atmospheres*, 111(D24), 2006. doi: https://doi.org/10.1029/2005JD006917. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006917.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M. Insights from earth system model initial-condition large ensembles and future prospects. 10 (4):277–286, 2020. ISSN 1758-6798. doi: 10.1038/s41558-020-0731-2. URL https://doi.org/10.1038/s41558-020-0731-2.

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., Bernardello, R., Boussetta, S., Caron, L.-P., Carver, G., Castrillo, M., Catalano, F., Cvijanovic, I., Davini, P., Dekker, E., Doblas-Reyes, F. J., Docquier, D., Echevarria, P., Fladrich, U., Fuentes-Franco, R., Gröger, M., v. Hardenberg, J., Hieronymus, J., Karami, M. P., Keskinen, J.-P., Koenigk, T., Makkonen, R., Massonnet, F., Ménégoz, M., Miller, P. A., Moreno-Chamarro, E., Nieradzik, L., van Noije, T., Nolan, P., O'Donnell, D., Ollinaho, P., van den Oord, G., Ortega, P., Prims, O. T., Ramos, A., Reerink, T., Rousset, C., Ruprich-Robert, Y., Le Sager, P., Schmith, T., Schrödner, R., Serva, F., Sicardi, V., Sloth Madsen, M., Smith, B., Tian, T., Tourigny, E., Uotila, P., Vancoppenolle, M., Wang, S., Wårlind, D., Willén, U., Wyser, K., Yang, S., Yepes-Arbós, X., and Zhang, Q. The ec-earth3 earth system model for the coupled model intercomparison project 6. *Geoscientific Model Development*, 15(7):2973–3020, 2022. doi: 10.5194/gmd-15-2973-2022. URL https://gmd.copernicus.org/articles/15/2973/2022/.

Easterling, D. R., Alexander, L. V., Mokssit, A., and Detemmerman, V. Cci/clivar workshop to develop priority climate indices. *Bulletin of the American Meteorological Society*, 84(10):1403–1407, 2003. ISSN 00030007, 15200477. URL http://www.jstor.org/stable/26216891.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016. doi: 10.5194/gmd-9-1937-2016. URL https://gmd.copernicus.org/articles/9/1937/2016/.

Fang, S., Majumder, R., Hector, E., Reich, B., and Sankarasubramanian, A. A complete density correction using normalizing flows (cdc-nf) for cmip6 gcms. *Scientific Data*, 12(1):1279, 2025.

François, B., Thao, S., and Vrac, M. Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate dynamics*, 57(11):3323–3353, 2021.

Fulton, D. J., Clarke, B. J., and Hegerl, G. C. Bias correcting climate model simulations using unpaired image-to-image translation networks. *Artificial Intelligence for the Earth Systems*, 2(2):e220031, 2023.

Hamill, T. M. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550 – 560, 2001. doi: 10.1175/1520-0493(2001)129⟨0550:IORHFV⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0550_iorhfv_2.0.co_2.xml.

Hawkins, E. and Sutton, R. Time of emergence of climate signals. *Geophysical Research Letters*, 39(1), 2012. doi: https://doi.org/10.1029/2011GL050087. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2011GL050087.

Hersbach, H. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559 – 570, 2000. doi: 10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/15/5/1520-0434_2000_015_0559_dotcrp_2_0_co_2.xml.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Hess, P. and Boers, N. A generative machine learning approach for improving precipitation from earth system models. *arXiv preprint arXiv:2406.15026*, 2024.

Hess, P., Drüke, M., Petri, S., Strnad, F. M., and Boers, N. Physically constrained generative adversarial networks for improving precipitation fields from earth system models. *Nature Machine Intelligence*, 4(10):828–839, 2022.

Hess, P., Lange, S., Schötz, C., and Boers, N. Deep learning for bias-correcting cmip6-class earth system models. *Earth's Future*, 11(10):e2023EF004002, 2023.

Hess, P., Aich, M., Pan, B., and Boers, N. Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning. *Nature Machine Intelligence*, 7(3):363–373, 2025.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

Immorlano, F., Eyring, V., le Monnier de Gouville, T., Accarino, G., Elia, D., Mandt, S., Aloisio, G., and Gentine, P. Transferring climate change physical knowledge. *Proceedings of the National Academy of Sciences*, 122(15): e2413503122, 2025.

IPCC. *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2023.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf.

Kim, S., Nathaniel, J., Hou, Z., Zheng, T., and Gentine, P. Spatiotemporal upscaling of sparse air-sea pco2 data via physics-informed transfer learning. *Scientific data*, 11(1): 1098, 2024.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2202.10054, 2022. URL https://arxiv.org/abs/2202.10054.

McGibbon, J., Clark, S. K., Henn, B., Kwa, A., Watt-Meyer, O., Perkins, W. A., and Bretherton, C. S. Global precipitation correction across a range of climates using cyclegan. *Geophysical Research Letters*, 51(4):e2023GL105131, 2024.

Mital, U., Dwivedi, D., Brown, J. B., and Steefel, C. I. Downscaled hyper-resolution (400 m) gridded datasets of daily precipitation and temperature (2008–2019) for the east–taylor subbasin (western united states). *Earth System Science Data*, 14(11):4949–4966, 2022. doi: 10.5194/essd-14-4949-2022. URL https://essd.copernicus.org/articles/14/4949/2022/.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner,

E., Stemmler, I., Tian, F., and Marotzke, J. A higher-resolution version of the max planck institute earth system model (mpi-esm1.2-hr). *Journal of Advances in Modeling Earth Systems*, 10(7):1383–1413, 2018. doi: https://doi.org/10.1029/2017MS001217. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017MS001217.

Nguyen, T., Jewik, J. K., Bansal, H., Sharma, P., and Grover, A. Climatelearn: Benchmarking machine learning for weather and climate modeling. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=RZJEkLFlPx.

Ossó, A., Craig, P., and Allan, R. P. An assessment of cmip6 climate signals and biases in temperature, precipitation and soil moisture over europe. *International Journal of Climatology*, 43(12):5698–5719, 2023. doi: https://doi.org/10.1002/joc.8169. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.8169.

Palmer, T. E., McSweeney, C. F., Booth, B. B. B., Priestley, M. D. K., Davini, P., Brunner, L., Borchert, L., and Menary, M. B. Performance-based subselection of cmip6 models for impact assessments in europe. *Earth System Dynamics*, 14(2):457–483, 2023. doi: 10.5194/esd-14-457-2023. URL https://esd.copernicus.org/articles/14/457/2023/.

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., Lee, J., Tian, Y., and Ma, H.-Y. Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, 13(10):e2021MS002509, 2021.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11): e2020MS002203, 2020. ISSN 1942-2466. doi: https://doi.org/10.1029/2020MS002203. URL https://doi.org/10.1029/2020MS002203.

Robin, Y., Vrac, M., Naveau, P., and Yiou, P. Multivariate stochastic bias corrections with optimal transport. *Hydrology and Earth System Sciences*, 23(2):773–786, 2019. doi: 10.5194/hess-23-773-2019. URL https://doi.org/10.5194/hess-23-773-2019.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y.-C., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K. S., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I. H. H., Landgren, O., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T., and Schulz, M. Overview of the norwegian earth system model (noresm2) and key climate response of cmip6 deck, historical, and scenario simulations. *Geoscientific Model Development*, 13(12):6165–6200, 2020. doi: 10.5194/gmd-13-6165-2020. URL https://gmd.copernicus.org/articles/13/6165/2020/.

Semmler, T., Danilov, S., Gierz, P., Goessling, H. F., Hegewald, J., Hinrichs, C., Koldunov, N., Khosravi, N., Mu, L., Rackow, T., Sein, D. V., Sidorenko, D., Wang, Q., and Jung, T. Simulations for cmip6 with the awi climate model awi-cm-1-1. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002009, 2020. doi: https://doi.org/10.1029/2019MS002009. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002009. e2019MS002009 2019MS002009.

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A. G., Fischer, E., and Knutti, R. Uncovering the forced climate response from a single ensemble member using statistical learning. *Journal of Climate*, 32(17):5677 – 5699, 2019. doi: 10.1175/JCLI-D-18-0882.1. URL https://journals.ametsoc.org/view/journals/clim/32/17/jcli-d-18-0882.1.xml.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Springenberg, M., Otero, N., Xue, Y., and Ma, J. Diffscale: Continuous downscaling and bias correction of subseasonal wind speed forecasts using diffusion models. *Journal of Advances in Modeling Earth Systems*, 18(3): e2025MS005282, 2026.

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., von Salzen, K., Yang, D., and Winter, B. The canadian earth system model version 5 (canesm5.0.3). *Geoscientific Model Development*, 12(11):4823–4873, 2019. doi: 10.5194/gmd-12-4823-2019. URL https://gmd.copernicus.org/articles/12/4823/2019/.

Talagrand, O. Evaluation of probabilistic prediction systems. In *Workshop Proceedings" Workshop on Predictability", 20-22 October 1997, ECMWF, Reading, UK*, 1999.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizuki, T., Yoshimura, K., Takata, K., O'ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M., and Kimoto, M. Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in miroc6. *Geoscientific Model Development*, 12(7):2727–2765, 2019. doi: 10.5194/gmd-12-2727-2019. URL https://gmd.copernicus.org/articles/12/2727/2019/.

Taylor, K. E. Truly conserving with conservative remapping methods. *Geoscientific Model Development*, 17(1): 415–430, 2024. doi: 10.5194/gmd-17-415-2024. URL https://gmd.copernicus.org/articles/17/415/2024/.

Thorarinsdottir, T. L., Sillmann, J., Haugen, M., Gissibl, N., and Sandstad, M. Evaluation of cmip5 and cmip6 simulations of historical surface air temperature extremes using proper evaluation methods. *Environmental Research Letters*, 15(12):124041, dec 2020. doi: 10.1088/1748-9326/abc778. URL https://doi.org/10.1088/1748-9326/abc778.

Tsutsui, J. Minimal cmip emulator (mce v1.2): a new simplified method for probabilistic climate projections. *Geoscientific Model Development*, 15(3):951–970, 2022. doi: 10.5194/gmd-15-951-2022. URL https://gmd.copernicus.org/articles/15/951/2022/.

Vrac, M. and Friederichs, P. Multivariate—intervariable, spatial, and temporal—bias correction. *Journal of Climate*, 28(1):218–237, 2015. doi: 10.1175/JCLI-D-14-00059.1. URL https://doi.org/10.1175/JCLI-D-14-00059.1.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., and Bretherton, C. S. Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, 48(15):e2021GL092555, 2021.

Willard, J. D., Harrington, P., Subramanian, S., Mahesh, A., O'Brien, T. A., and Collins, W. D. Analyzing and exploring training recipes for large-scale transformer-based weather prediction. *Artificial Intelligence for the Earth Systems*, 4(2):240061, 2025. doi: 10.1175/AIES-D-24-0061.1. URL https://journals.ametsoc.org/view/journals/aies/4/2/AIES-D-24-0061.1.xml.

Xu, Z., Han, Y., Tam, C.-Y., Yang, Z.-L., and Fu, C. Bias-corrected CMIP6 global dataset for dynamical downscaling of the historical and future climate (1979–2100). *Scientific Data*, 8(1):293, 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-01079-3. URL https://doi.org/10.1038/s41597-021-01079-3.

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., Yabu, S., Yoshimura, H., Shindo, E., Mizuta, R., Obata, A., Asachi, Y., and Ishii, M. The meteorological research institute earth system model version 2.0, mri-esm2.0: Description and basic evaluation of the physical component. *Journal of the Meteorological Society of Japan. Ser. II*, 97(5):931–965, 2019. doi: 10.2151/jmsj.2019-051.

Zhong, J., Zhang, X., Wang, J., and Long, M. Domain guidance: A simple transfer approach for a pre-trained diffusion model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PplM2kDrl3.

A. Diffusion Model

A.1. Architecture

We use a UNet as our denoiser trained with Elucidated Diffusion Model (EDM) (Karras et al., 2022) preconditioning and loss. When sampling, we estimate the integral of the probability flow ODE with Heun's second order method. The use of an ODE makes sampling deterministic given an initial noise, which allows us to compute the log likelihood of given temperature maps for our evaluation. We use a log uniform noise schedule with $\sigma_{\text{min}} = 0.002$ and $\sigma_{\text{max}} = 200$ as per cBottle (Brenowitz et al., 2025) to better destroy the climate signal.

Since our temperature maps are standardized, we use a $\sigma_{\text{data}} = 1$.

UNet Our architecture employs a UNet backbone with progressive channel expansion (32→64→128 channels using multipliers [1,2,4]) and ResNet blocks at each resolution level. Self-attention is applied at the coarsest resolution. The conditioning vector is projected to match the timestep embedding dimension (128) and combined additively with the timestep embedding throughout the network.

Conditioning/Embedding The conditioning mechanism processes each input according to its type. Categorical variables (the simulator index s and day-of-year d) are first one-hot encoded, then transformed through independent sinusoidal positional embeddings to produce $\mathbf{e}_s$ and $\mathbf{e}_d$. CO_2e, a continuous variable, is processed through a separate learned feed-forward network to compute $\mathbf{e}_c$. These embeddings are concatenated and passed through an additional feed-forward network to create the final conditioning vector:

$$\mathbf{c}_{\text{comb}} = \text{FFN}([\mathbf{e}_s; \mathbf{e}_c; \mathbf{e}_d]). \tag{10}$$

Training During training, we maintain an exponential moving average of the model weights and use the final exponential moving average weighted model for all evaluation. We train our diffusion model to 200 epochs with an AdamW optimizer and learning rate 1e-3. We used a batch size of 1024 with 16-mixed precision training and performed all evaluation at full precision.

Although not needed for *condition-to-condition* guidance because we are interpolating between two learned conditionings, we apply dropout on the conditioning variables which serves as a form of regularization.

Loss We modified the EDM loss to weight each grid cell by the cosine of its latitude, ensuring that the loss contributions are proportional to actual geographic area. This correction is standard practice (Nguyen et al., 2023; Willard et al., 2025; Aich et al., 2026) when working with regular latitude-longitude grids to avoid over-optimization for high-latitude regions.

A.2. Log Likelihood

The metrics of controllability, variability, and realism require computing the conditional log likelihood $\log p_\theta(T \mid m, d, c)$. T is a standardized temperature map, either from the data distribution or sampled from the model.

We compute the log likelihood by integrating the EDM probability flow ODE from a data to latent noise from $\sigma_{\text{min}} = 0.002$ to $\sigma_{\text{max}} = 200.0$. We use the continuous time change of variables formula to relate the data log density to the latent log density and then accumulate the Jacobian trace. We use Hutchinson's trace estimator to estimate this with a single Rademacher vector at each EDM step. Integration uses the Dormand-Prince (dopri5) solver with absolute and relative tolerances of 5e-4. All likelihood evaluation uses full numerical precision.

Log Likelihood under Conditional-to-Conditional Guidance Under the guided model, the metrics of controllability, variability, and realism require computing the guided conditional log likelihood. The computation is evaluated the same way as the log likelihood, only that the unguided velocity field at a given noise level is replaced by guided variant of the denoiser, $D_{\theta,\gamma}(\text{src}_i, d, c) = D_\theta(\text{src}_i, d, c)(1-\gamma) + \gamma D_\theta(\text{tgt}, d, c)$, to compute the guided velocity field $v_{\theta,\gamma} = \frac{x - D_{\theta,\gamma}}{\sigma}$. In our setup $\text{src}_i \leftrightarrow \texttt{CMIP6}_i$ and $\text{tgt} \leftrightarrow \texttt{ERA5}$.

B. Data

Temperature fields We use daily 2-meter air temperature fields simulated by nine CMIP6 models (Table B.1) in 1960–2016, and ERA5 reanalysis (Hersbach et al., 2020) as real-world observations. All fields are cropped to the European domain (30°N–75°N; 18°W–77°E) and remapped to the native CanESM5 grid ($\sim 2.8°$ horizontal resolution), which is the coarsest of the nine CMIP6 models. This yields a 16×32 pixel domain. For remapping we use conservative interpolation (Taylor, 2024). Because CMIP6 historical simulations end in 2014, we use the Shared Socioeconomic Pathway 2-4.5 (SSP2-4.5) extension for 2015–2016.This is the CMIP6 scenario whose near-term radiative forcing is closest to the one observed in 2015–2016, so using it minimizes introducing scenario-specific structure into the corresponding period. For each CMIP6 model, we gather the `r1i1p1f1` member simulations since it is the primary member used in the Sixth Assessment Report of the Working Group 1 of the Intergovernmental Panel on Climate Change (IPCC, 2023).

Table B.1. CMIP6 Earth system models from the CMIP6 ensemble used in this work.

Model
AWI-CM1-1-MR (Semmler et al., 2020)
CanESM5 (Swart et al., 2019)
CESM2-WACCM (Danabasoglu et al., 2020)
EC-Earth3-Veg-LR (Döscher et al., 2022)
IPSL-CM6A-LR (Boucher et al., 2020)
MIROC6 (Tatebe et al., 2019)
MPI-ESM1-2-HR (Müller et al., 2018)
MRI-ESM2-0 (Yukimoto et al., 2019)
NorESM2-MM (Seland et al., 2020)

CO_2e The generation of temperature fields is conditioned on scalar annual global-mean CO_2e concentration c, rather than calendar years. This is a physically meaningful variable: it encodes both the temporal position within the training period (as a proxy of the year) and the long-term radiative forcing that drives the regional warming trends we aim to model. The corresponding values in 1960–2016 (considering SSP2-4.5 for 2015 and 2016) are simulated by the Minimal CMIP Emulator v1.2 (Tsutsui, 2022), a reduced-complexity model that converts emissions to radiative forcing and then to CO_2e. Annual resolution is adequate in our case since CO_2e reflects well-mixed greenhouse gases with multi-year residence times.

Training / Test split We use years 1960–2000 for training and 2001–2016 for test. The anthropogenic warming signal is small relative to internal variability through the late 20th century, only emerging with greater evidence after 2000 (Hawkins & Sutton, 2012). Therefore, the training period is approximately stationary in c and the test period is characterized by a slight distributional shift (before it becomes even stronger over projection years). This is a deliberate choice: we evaluate guidance under quasi-stationary conditions so that any degradation reflects the method itself rather than the prior extrapolating far outside its training distribution. We use z-score normalization to standardize both temperature maps and CO_2e values. Each variable's respective mean and standard deviation is computed on the training data from 1960 to 2000.

C. Baseline

As a primary deterministic baseline, we use the MVT correction method (Xu et al., 2021). MVT is a classical statistical post-processing approach that corrects systematic biases in the mean state, variability, and long-term trends of numerical simulations. For a biased simulator output x_{src}, the corrected value $\hat{x}$ is obtained via a stationary affine transformation:

$$\hat{x}(t) = \mu_{\mathrm{tgt}} + \frac{\sigma_{\mathrm{tgt}}}{\sigma_{\mathrm{src}}} \left(x_{\mathrm{src}}(t) - \mu_{\mathrm{src}} \right) \tag{11}$$

where μ and σ denote the mean and standard deviation of the tgt (observational) and src (simulation) distributions, respectively. We calibrate these parameters computing the statistics for temperature fields from the training data (1960–2000) to standardize CMIP6 model outputs against the ERA5 observational record.

D. Metrics

This section describes every metric used to evaluate the quality of bias correction for the climate science case study presented in the paper. This same framework is used for all scalar metrics, spatial and distributional diagnostics, and the criterion used to select the guidance scale γ.

Notation The evaluation period is 2001–2016 (held out from training). For a given simulator conditioning i and a given correction method (no correction, MVT, or guidance at scale γ), let

$$\hat{x}^{(i)}_{y,d,s,j,k}, \qquad y = 1, \ldots, N_y,\ d = 1, \ldots, N_d,\ s = 1, \ldots, N_s,\ j = 1, \ldots, N_\phi,\ k = 1, \ldots, N_\lambda$$

denote the s-th sample of 2 m daily temperature predicted at latitude index j, longitude index k, on day-of-year d of year y, conditioned on i. The ERA5 reference field has the same spatial-temporal structure but is deterministic:

$$x^{\mathrm{ERA5}}_{y,d,j,k}.$$

For this study, $N_y = 16$ (years 2001–2016), $N_d = 365$ (days 1–365 for each year), $N_s = 20$ (samples per day), and the spatial grid is $N_\phi \times N_\lambda = 16 \times 32$ covering the European cropped domain. All temperatures are expressed in °C throughout.

Area weighted average Every spatial average computed in this work is an area-weighted average. On a regular latitude-longitude grid, areas of grid points vary as $\cos(\phi)$ because the zonal spacing shrinks toward the poles. We therefore define the area weight at grid cell (j, k) as $A_{j,k} = \cos(\phi_j)$, and the normalized weight

$$w_{j,k} = \frac{A_{j,k}}{\sum_{j'=1}^{N_\phi} \sum_{k'=1}^{N_\lambda} A_{j',k'}}, \qquad \sum_{j,k} w_{j,k} = 1.$$

Given any spatial field $f_{j,k}$, its *area-weighted spatial mean* is

$$\langle f \rangle = \sum_{j=1}^{N_\phi} \sum_{k=1}^{N_\lambda} w_{j,k}\, f_{j,k}.$$

D.1. Scalar and spatial metrics

Throughout this subsection, $\overline{\hat{x}}^{(i)}_{j,k} = \frac{1}{N_y N_d N_s} \sum_{y,d,s} \hat{x}^{(i)}_{y,d,s,j,k}$ denotes the climatological-mean prediction field, and $\overline{x}^{\mathrm{ERA5}}_{j,k} = \frac{1}{N_y N_d} \sum_{y,d} x^{\mathrm{ERA5}}_{y,d,j,k}$ the corresponding ERA5 climatology. Each metric reduces a bias map to a scalar via the area-weighted average defined in Section D.

Mean bias map The pointwise mean bias map and its area-weighted scalar summary are

$$B^{\mu}_{j,k} = \overline{\hat{x}}^{(i)}_{j,k} - \overline{x}^{\mathrm{ERA5}}_{j,k}, \qquad \mathrm{RMSE}_\mu = \sqrt{\sum_{j,k} w_{j,k} \big(B^{\mu}_{j,k}\big)^2}.$$

Lower RMSE_μ means the corrected ensemble's climatological mean is closer to ERA5 after area-weighting.

Variance-ratio deviation Let $\sigma^{\mathrm{pred}}_{j,k}$ and $\sigma^{\mathrm{ERA5}}_{j,k}$ be the standard deviations of the prediction and ERA5 at grid cell (j, k) across (y, d, s) and (y, d) respectively. The standard-deviation ratio map is $R^{\sigma}_{j,k} = \sigma^{\mathrm{pred}}_{j,k} / \sigma^{\mathrm{ERA5}}_{j,k}$, and the scalar summary is the area-weighted mean of $|R^{\sigma}_{j,k} - 1|$. It is zero when prediction and ERA5 spread match exactly, and penalizes over- and under-dispersion equally.

Skewness and kurtosis bias The pointwise standardized third and fourth central moments are

$$\mathrm{Skew}^{\mathrm{pred}}_{j,k} = \mathbb{E}\left[\left(\frac{\hat{x}^{(i)}_{\cdot,j,k} - \overline{\hat{x}}^{(i)}_{j,k}}{\sigma^{\mathrm{pred}}_{j,k}}\right)^3\right], \qquad \mathrm{Kurt}^{\mathrm{pred}}_{j,k} = \mathbb{E}\left[\left(\frac{\hat{x}^{(i)}_{\cdot,j,k} - \overline{\hat{x}}^{(i)}_{j,k}}{\sigma^{\mathrm{pred}}_{j,k}}\right)^4\right] - 3,$$

with the expectation taken empirically over (y, d, s), and analogously for ERA5. The reported scalars are the area-weighted mean absolute biases $\sum_{j,k} w_{j,k} |\mathrm{Skew}^{\mathrm{pred}}_{j,k} - \mathrm{Skew}^{\mathrm{ERA5}}_{j,k}|$ and similarly for kurtosis. They diagnose distributional shape discrepancies that the mean and variance miss.

Extreme-frequency biases (TX90p, TN10p) Let $q^{\mathrm{ERA5}}_{\alpha,j,k}$ be the empirical α-quantile of ERA5 at (j,k) across (y,d). The TX90p exceedance frequency of the prediction is

$$F^{\mathrm{TX90p,pred}}_{j,k} = \frac{1}{N_y N_d N_s} \sum_{y,d,s} \mathbf{1}\left[\hat{x}^{(i)}_{y,d,s,j,k} > q^{\mathrm{ERA5}}_{0.9,j,k}\right].$$

Because the threshold is defined from ERA5 itself, a perfectly unbiased model would exceed it 10% of the time; we therefore report the TX90p bias map as $B^{\mathrm{TX90p}}_{j,k} = F^{\mathrm{TX90p,pred}}_{j,k} - 0.1$, expressed in pp (Expert Team on Climate Change Detection and Indices (ETCCDI) convention (Easterling et al., 2003)). The reported scalar is the area-weighted mean absolute bias. TN10p is defined analogously using the 10th-quantile threshold and the indicator $\mathbf{1}[\hat{x} < q^{\mathrm{ERA5}}_{0.1,j,k}]$.

Tail-level biases (Q05, Q95) While TX90p/TN10p measure the frequency of tail-threshold exceedance, Q05 and Q95 biases measure the levels of the tails themselves. Let $q^{\mathrm{pred}}_{\alpha,j,k}$ and $q^{\mathrm{ERA5}}_{\alpha,j,k}$ denote respectively the α-quantile of predicted samples across (y,d,s) and the α-quantile of ERA5 maps across (y,d). The pointwise quantile bias map at $\alpha \in \{0.05, 0.95\}$ is

$$B^{q_\alpha}_{j,k} \;=\; q^{\mathrm{pred}}_{\alpha,j,k} - q^{\mathrm{ERA5}}_{\alpha,j,k}.$$

The scalar summary is the area-weighted mean absolute bias, reported in °C in Table 1.

CRPS CRPS (Hersbach, 2000) is a proper scoring rule for probabilistic forecasts; lower is better, and it reduces to mean absolute error for a degenerate (deterministic) forecast distribution. For a finite ensemble of N_s members $x_1, \ldots, x_{N_s}$ against scalar observation y, it has the closed form

$$\mathrm{CRPS}_{y,d,j,k} = \frac{1}{N_s}\sum_{s} |x_s - y| - \frac{1}{2N_s^2}\sum_{s,s'} |x_s - x_{s'}|,$$

which is the formula applied here. The reported scalar is the area-weighted spatiotemporal mean $\overline{\mathrm{CRPS}} = \frac{1}{N_y N_d}\sum_{y,d}\sum_{j,k} w_{j,k}\,\mathrm{CRPS}_{y,d,j,k}$. CRPS jointly penalizes the distance between the ensemble mean and the observation and miscalibration of the ensemble spread, which is why the main text uses it as the primary realism diagnostic for γ selection.

Rank histogram and RI For each (y,d,j,k), let $r_{y,d,j,k} \in \{0,1,\ldots,N_s\}$ be the rank of the ERA5 observation in the sorted ensemble $\{\hat{x}^{(i)}_{y,d,s,j,k}\}_s$. A perfectly calibrated ensemble has uniform ranks with frequency $1/(N_s+1)$ (Talagrand, 1999; Hamill, 2001). Let f_ℓ be the empirical fraction of ranks equal to ℓ, aggregated over all (y,d,j,k). The RI (Delle Monache et al., 2006) reported in Table 1 is

$$\mathrm{RI} = \sum_{\ell=0}^{N_s}\left(f_\ell - \frac{1}{N_s+1}\right)^2, \qquad \mathrm{RI} = 0 \text{ at perfect calibration.}$$

Laplacian roughness The Laplacian roughness quantifies the spatial incoherency of a mean-bias field (Mital et al., 2022). Small-scale checkerboard structure in bias fields is a sign that the spatial correlation of the underlying temperature maps has not been preserved by the bias correction. For a bias field $B^{\mu}_{j,k}$, the discrete Laplacian at an interior grid cell is

$$(\nabla^2 B^{\mu})_{j,k} \;=\; B^{\mu}_{j+1,k} + B^{\mu}_{j-1,k} + B^{\mu}_{j,k+1} + B^{\mu}_{j,k-1} - 4\,B^{\mu}_{j,k}.$$

The Laplacian roughness is the area-weighted mean of its absolute value, restricted to interior grid points $\mathcal{I}$ (boundary points are excluded because the stencil is incomplete there):

$$\mathrm{LapRough} \;=\; \frac{\sum_{(j,k)\in\mathcal{I}} w_{j,k}\left|(\nabla^2 B^{\mu})_{j,k}\right|}{\sum_{(j,k)\in\mathcal{I}} w_{j,k}}.$$

The bias field is computed across (y,d,s) before the Laplacian is applied, so the roughness in Table 1 measures systematic small-scale structure in the climatological bias rather than in any single sample.

D.2. Pattern correlation of inter-simulator anomalies

The guidance scale γ cannot be selected from the realism metrics alone (CRPS in case of domain-specific diagnostics) because these improve monotonically in $\gamma \in [0, 1]$: CRPS keeps decreasing as $\gamma \to 1$, but at $\gamma = 1$ the guided generative emulator is conditioned only on ERA5 and the CMIP6 simulator index becomes meaningless. Thus, a second measure is needed to quantify how well the guided emulator preserves the representation of each CMIP6 simulator's identity under guidance. This is the controllability measure, and the metric we use is the pattern correlation between each simulator's spatial fingerprint in the guided emulator's outputs at scale γ and the same fingerprint at $\gamma = 0$.

Setup Throughout this section, every climatological field is computed from samples drawn from the guided diffusion emulator $p_{\theta,\gamma}(\cdot \mid \mathrm{src}_i, c)$. For each scale $\gamma \in \Gamma = \{\gamma_0, \gamma_1, \ldots\}$ with $\gamma_0 = 0$ and each simulator conditioning $i \in \{1, \ldots, M\}$, we generate samples $\hat{x}^{(i,\gamma)}_{y,d,s,j,k}$ (year y, day d, sample index s, grid cell (j, k)) and compute the guided climatology

$$\overline{x}^{(i,\gamma)}_{j,k} \;=\; \frac{1}{N_y N_d N_s} \sum_{y,d,s} \hat{x}^{(i,\gamma)}_{y,d,s,j,k}.$$

In this study $M = 9$ and Γ spans values from 0 to 1 with step 0.125.

We define the ensemble mean of guided climatologies across simulator conditionings (at fixed γ):

$$\bar{\mathcal{X}}^{(\gamma)}_{j,k} \;=\; \frac{1}{M} \sum_{i=1}^{M} \overline{x}^{(i,\gamma)}_{j,k},$$

and the inter-simulator anomaly of conditioning i at scale γ:

$$a^{(i,\gamma)}_{j,k} \;=\; \overline{x}^{(i,\gamma)}_{j,k} - \bar{\mathcal{X}}^{(\gamma)}_{j,k}.$$

We emphasize that both $\overline{x}^{(i,\gamma)}_{j,k}$ and $\bar{\mathcal{X}}^{(\gamma)}_{j,k}$ are computed from the guided emulator's samples; the ensemble averaged over here is the set of M guided-emulator climatologies, one per simulator conditioning. The anomaly $a^{(i,\gamma)}$ is the spatial fingerprint by which the guided emulator, when conditioned on simulator i, distinguishes itself from the average ensemble across all M conditionings. This is the signal that bias correction should preserve across guidance scales (controllability).

Pearson correlation The Pearson correlation $\rho(u, v)$ measures the alignment of two spatial patterns independently of their absolute magnitudes. It ranges in $[-1, 1]$, equals $+1$ when v is a positive affine transformation of u and -1 when negative, and is invariant under rescaling either field by a positive constant. Consequently, the metric serves as a measure of structural or shape agreement between the two fields and is invariant to differences in amplitude. The amplitude invariance is a deliberate property the diagnostic but introduces a known limitation: a guided emulator that preserves each simulator's spatial pattern while uniformly shrinking its magnitude toward zero would still receive $\bar{r} \approx 1$, even though the inter-model differentiation has effectively been erased. A complementary amplitude-aware diagnostic would close this gap and is the subject of ongoing work. We note that on the climate use case studied here the bias maps in Figure F.3 show that residual inter-model anomaly magnitudes collapse over the same γ-interval $(0.75{-}1)$ in which $\bar{r}(\gamma)$ falls, suggesting that amplitude shrinks and shape collapses in the same interval. However, a quantitative confirmation is left to a follow-up.

For two spatial fields $u_{j,k}$ and $v_{j,k}$, the area-weighted means, variances and covariance are

$$\mu_u = \sum_{j,k} w_{j,k}\, u_{j,k}, \quad \mu_v = \sum_{j,k} w_{j,k}\, v_{j,k},$$

$$\mathrm{Var}_w(u) = \sum_{j,k} w_{j,k}(u_{j,k} - \mu_u)^2, \quad \mathrm{Cov}_w(u, v) = \sum_{j,k} w_{j,k}(u_{j,k} - \mu_u)(v_{j,k} - \mu_v),$$

and the area-weighted Pearson correlation is

$$\rho_w(u, v) \;=\; \frac{\mathrm{Cov}_w(u, v)}{\sqrt{\mathrm{Var}_w(u)\ \mathrm{Var}_w(v)}}.$$

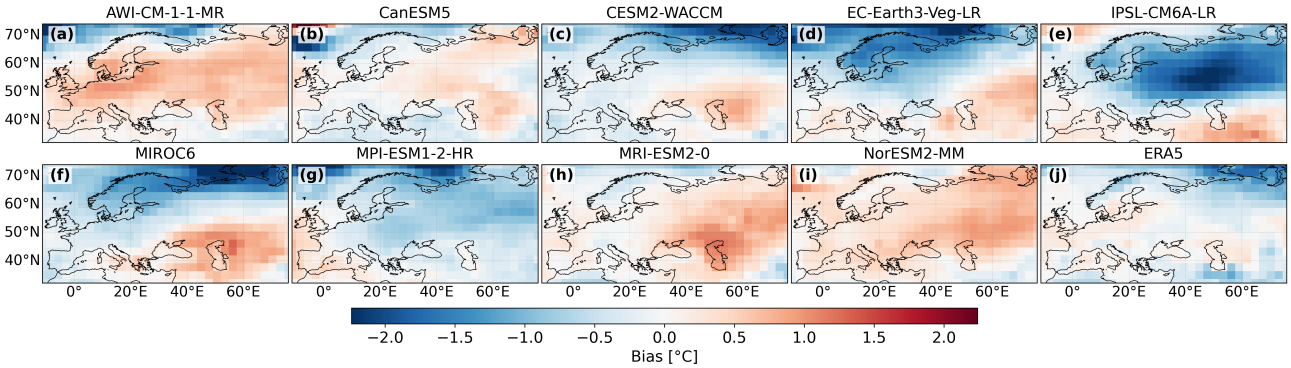


Figure E.1. Mean bias maps of the unguided diffusion ensemble ($\gamma = 0$, 20 samples per day, 2001–2016 test period) relative to the source data, for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)).

Per-conditioning pattern correlation Applied to the inter-simulator anomalies at two guidance scales, $\rho_w(a^{(i,\gamma)}, a^{(i,0)})$ describes whether the shape of simulator i's fingerprint in the guided emulator is preserved as γ varies, independent of how much its magnitude has shrunk. We define the per-conditioning pattern correlation between the anomaly at scale γ and the anomaly at $\gamma = 0$:

$$r^{(i)}(\gamma) \;=\; \rho_w\Big(a^{(i,\gamma)},\; a^{(i,0)}\Big)\,,$$

where the reference $\gamma = 0$ corresponds to the unguided emulator conditioned on simulator i, and is therefore the representation of simulator i's identity that the diffusion model actually learned at training time.

We want to evaluate whether guidance preserves this identity for each conditioning i across γ values. For a given γ we compute the mean pattern correlation across the nine CMIP6 simulator conditionings as

$$\bar{r}(\gamma) \;=\; \frac{1}{M}\sum_{i=1}^{M} r^{(i)}(\gamma).$$

Interpretation $\bar{r}(\gamma)$ guides the selection of $\gamma^\star$:

- $\bar{r}(\gamma) \approx 1$: every simulator conditioning's spatial fingerprint in the guided emulator (e.g. warmer-than-average in a region, cooler-than-average in another) has the same shape as without guidance; only the magnitude may have changed. Simulator identity is preserved.
- $\bar{r}(\gamma) \approx 0$: deviations of the guided emulator's per-conditioning climatologies from their ensemble mean are uncorrelated with the unguided deviations. Identity is lost.
- $\bar{r}(\gamma) \approx -1$: the correlation is inverted. Locations where simulator i's guided climatology was warmer than the ensemble of conditionings at $\gamma = 0$ are now cooler, and vice versa.

E. Generative emulator results

This appendix collects the per-conditional diagnostics summarized in Section 5.1: spatial bias maps, regional-mean trajectories, seasonal cycles, distributional shape, spatial power spectra, rank histograms, and scalar metrics across all ten conditionals, i.e. nine CMIP6 models and ERA5 (Figures E.1 to E.14 and Table E.1).

F. Guidance results

This appendix reports *condition-to-condition guidance* results for the eight CMIP6 models not analyzed in detail in the main text (MIROC6 is treated in Section 5.2.2). For each simulator, we apply guidance at the same operating scale $\gamma^\star = 0.9$ selected in Section 5.2.1, generate $N_s = 20$ samples per day over the 2001–2016 test period, and compare against the MVT baseline using the same diagnostics. The qualitative picture is uniform across the ensemble: guidance outperforms MVT on every metric for every CMIP6 model (Figures F.1 to F.8 and Table F.1).

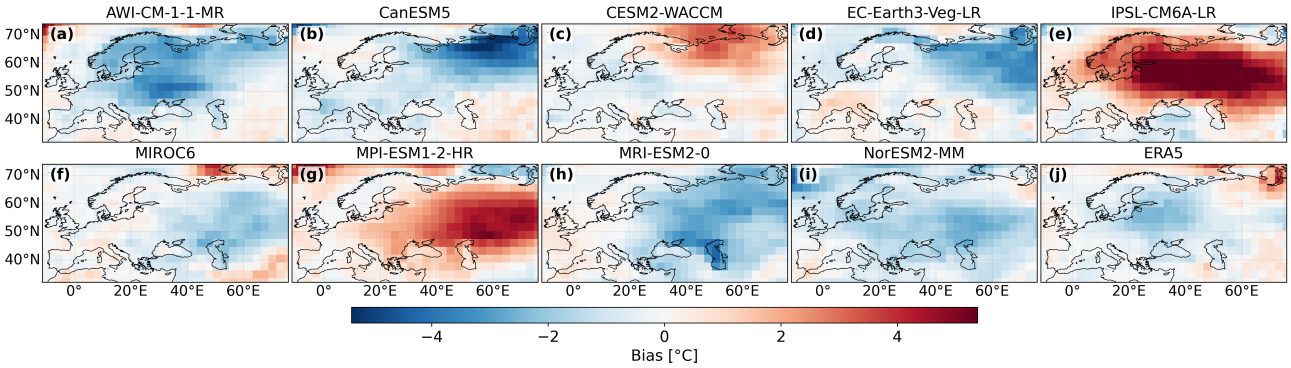


Figure E.2. Seasonal-amplitude bias maps (predicted annual maximum minus annual minimum monthly mean, minus reference) of the unguided diffusion ensemble ($\gamma = 0$, 2001–2016 test period) relative to the source data, for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)).

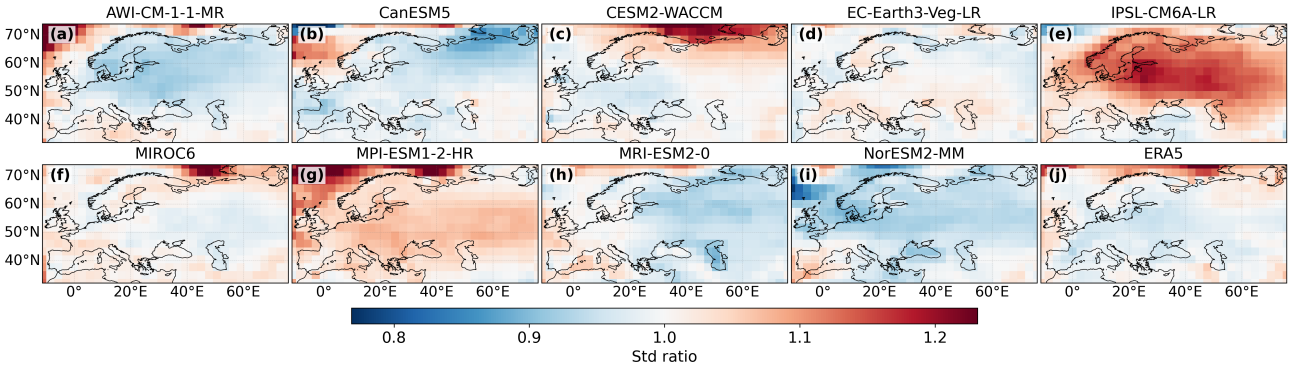


Figure E.3. Standard-deviation ratio maps $\sigma_{\text{pred}}/\sigma_{\text{ref}}$ of the unguided diffusion ensemble ($\gamma = 0$, 2001–2016 test period) relative to the source data, for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)). A value of 1 (white) indicates exact match; red is over-dispersion, blue is under-dispersion.

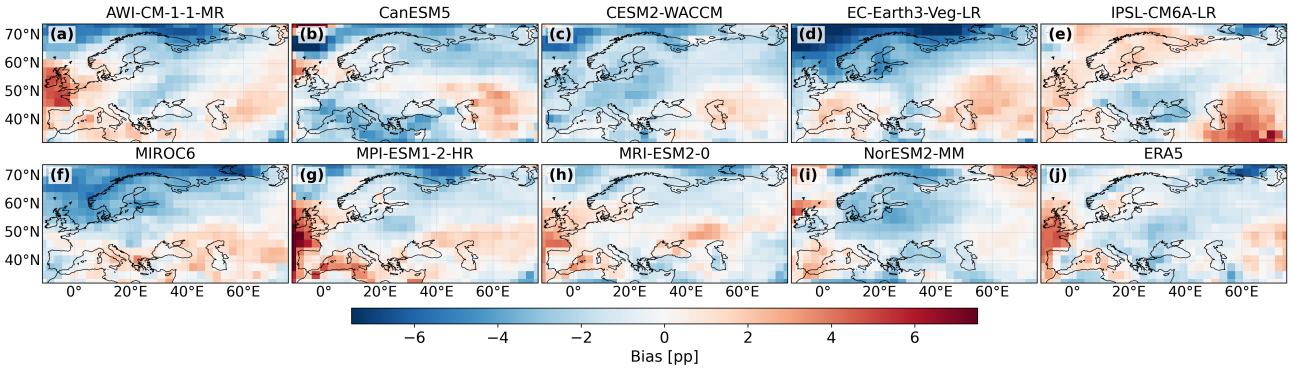


Figure E.4. TX90p frequency-bias maps (in pp relative to the expected 10% exceedance) of the unguided diffusion ensemble ($\gamma = 0$, 2001–2016 test period) relative to the source data, for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)). The TX90p threshold is computed locally from the corresponding source dataset (Appendix D).

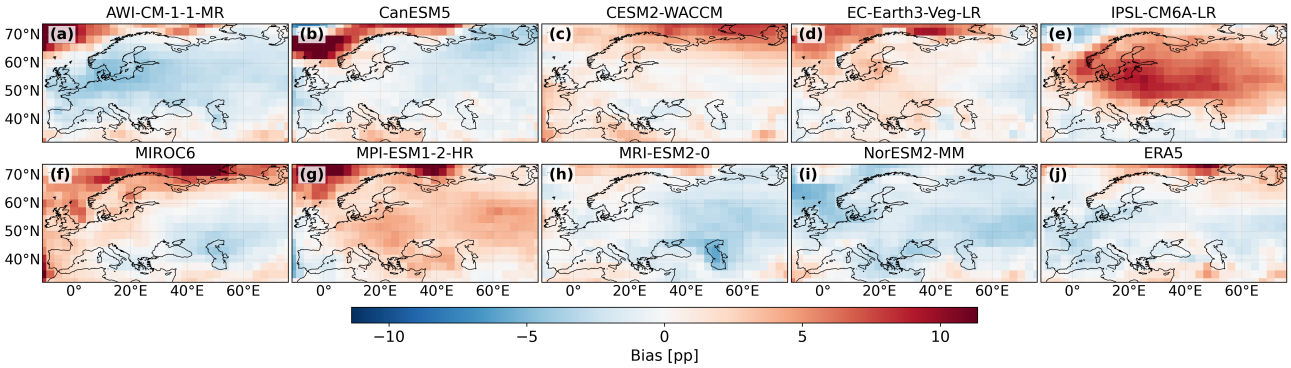


Figure E.5. TN10p frequency-bias maps (in percentage points relative to the expected 10% exceedance) of the unguided diffusion ensemble ($\gamma = 0$, 2001–2016 test period) relative to the source data, for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)). The TN10p threshold is computed locally from the corresponding source dataset (Appendix D).

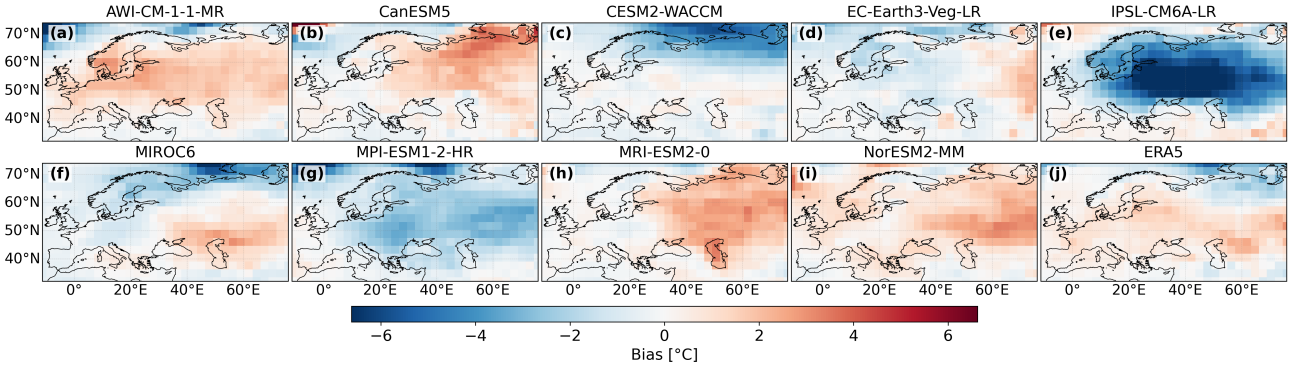


Figure E.6. 5th-percentile bias maps (predicted minus reference, in °C) of the unguided diffusion ensemble ($\gamma = 0$, 2001–2016 test period) relative to the source data, for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)).

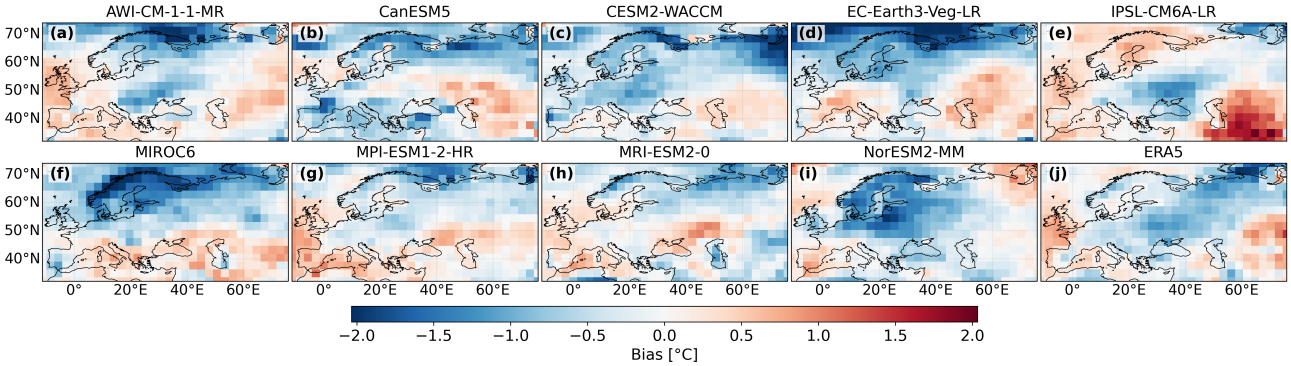


Figure E.7. 95th-percentile bias maps (predicted minus reference, in °C) of the unguided diffusion ensemble ($\gamma = 0$, 2001–2016 test period) relative to the source data, for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)).

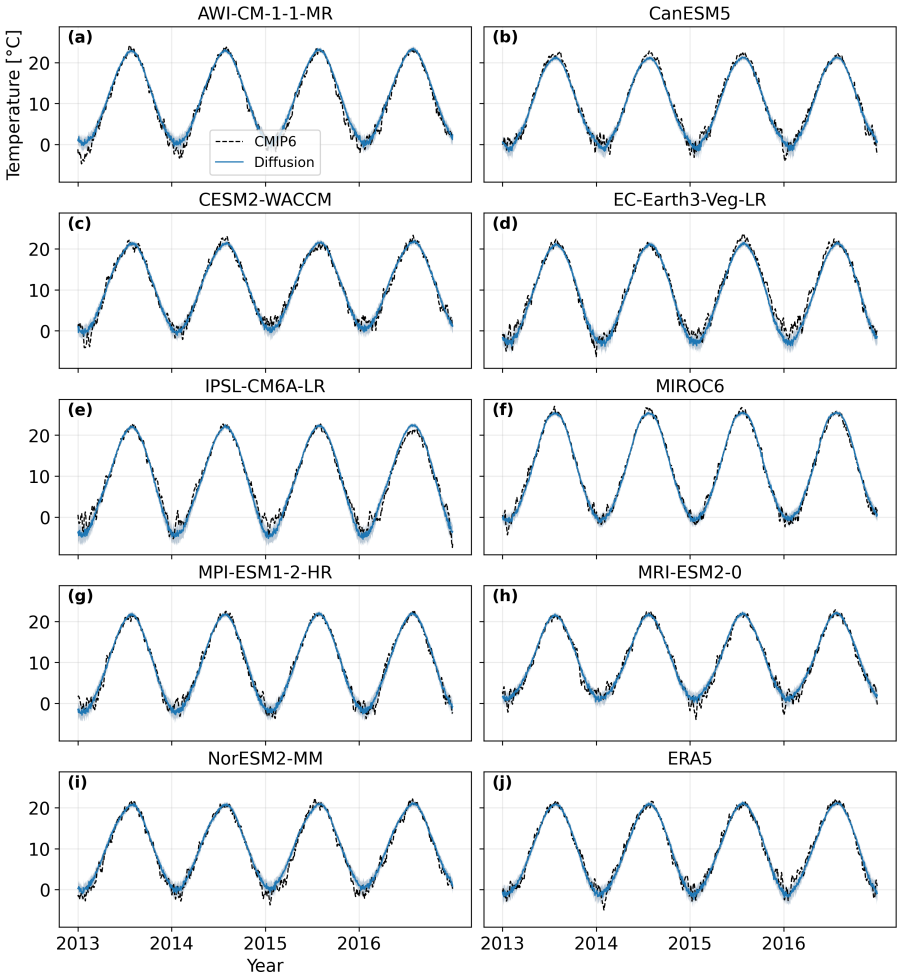


Figure E.8. Daily regional temperature of the unguided diffusion ensemble (solid blue, $\gamma = 0$, 20 samples per day) over 2013–2016, conditioned on $s = \{\texttt{CMIP6}_1, \ldots, \texttt{ERA5}\}$ (panels (a)–(j)); ; the corresponding source curve is shown in dashed black. The light-blue band is the $\pm 1\sigma$ across the 20 daily samples. The four-year window shown is a representative subset of the full 2001–2016 evaluation period.

Table E.1. Scalar metrics of the unguided diffusion emulator ($\gamma = 0$) for all ten conditionings $\{\texttt{CMIP6}_1, \ldots, \texttt{ERA5}\}$, ranked by climatological-mean error. Metric definitions are in Appendix D.

Model	**Mean RMSE**	**Var. Ratio Deviation**	**Skewness**	**Kurtosis**	**TX90p**	**TN10p**	**Q05**	**Q95**	**CRPS**	**RI**	**Laplacian Roughness**
CanESM5	0.417	0.034	0.073	0.158	1.947	1.861	0.874	0.523	2.053	2.4×10^{-4}	0.413
MRI-ESM2-0	0.453	0.036	0.066	0.135	1.276	1.555	0.903	0.398	2.001	4.0×10^{-5}	0.304
NorESM2-MM	0.455	0.047	0.092	0.198	1.570	1.943	0.902	0.482	2.016	2.4×10^{-4}	0.242
ERA5	0.456	0.026	0.071	0.097	1.425	1.437	0.766	0.444	1.954	3.2×10^{-4}	0.402
AWI-CM-1-1-MR	0.557	0.040	0.069	0.145	1.545	2.166	1.036	0.432	2.055	8.0×10^{-5}	0.397
CESM2-WACCM	0.594	0.028	0.080	0.138	1.607	1.583	0.870	0.476	1.943	5.4×10^{-4}	0.358
MPI-ESM1-2-HR	0.642	0.054	0.076	0.128	1.617	2.639	1.456	0.378	2.085	4.8×10^{-4}	0.424
EC-Earth3-Veg-LR	0.749	0.016	0.055	0.122	2.123	1.442	0.651	0.560	2.199	1.2×10^{-3}	0.448
MIROC6	0.769	0.022	0.047	0.074	1.787	2.355	0.930	0.563	1.959	9.4×10^{-4}	0.448
IPSL-CM6A-LR	0.900	0.068	0.093	0.145	1.481	3.120	2.233	0.487	2.081	2.3×10^{-4}	0.343

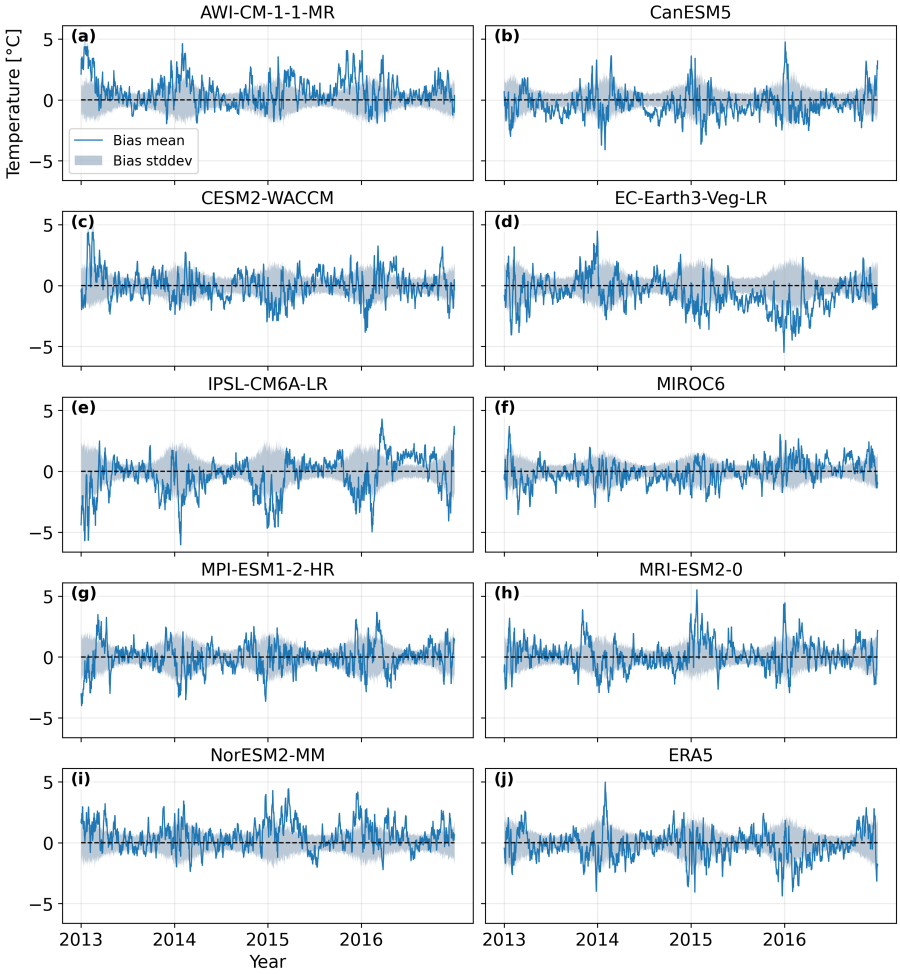


Figure E.9. Daily regional-mean signed bias of the unguided diffusion ensemble ($\gamma = 0$) relative to the source data, over 2013–2016 and for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)). Blue line: ensemble-mean bias on each day. Light-blue band: $\pm 1\sigma$ across the 20 daily samples. The dashed black line is the no-bias reference. Bias fluctuates around zero with no systematic drift for any conditioning.

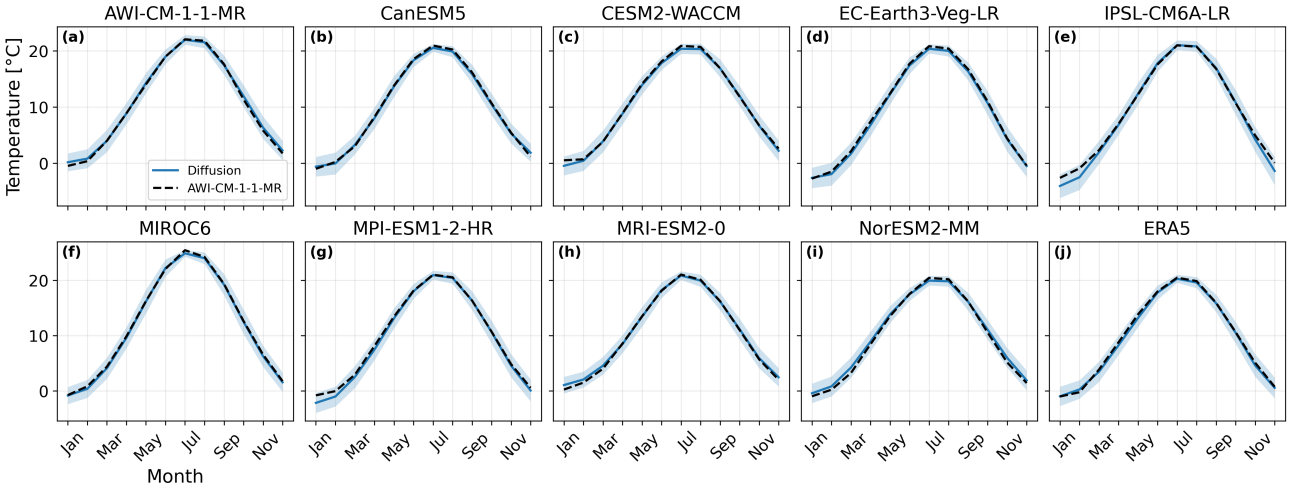


Figure E.10. Regional-mean monthly seasonal cycle of the unguided diffusion ensemble ($\gamma = 0$, solid blue) and the corresponding source data (dashed black) for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)), averaged over the 2001–2016 test period. The light-blue band is the $\pm 1\sigma$ inter-annual spread of the predicted monthly mean.

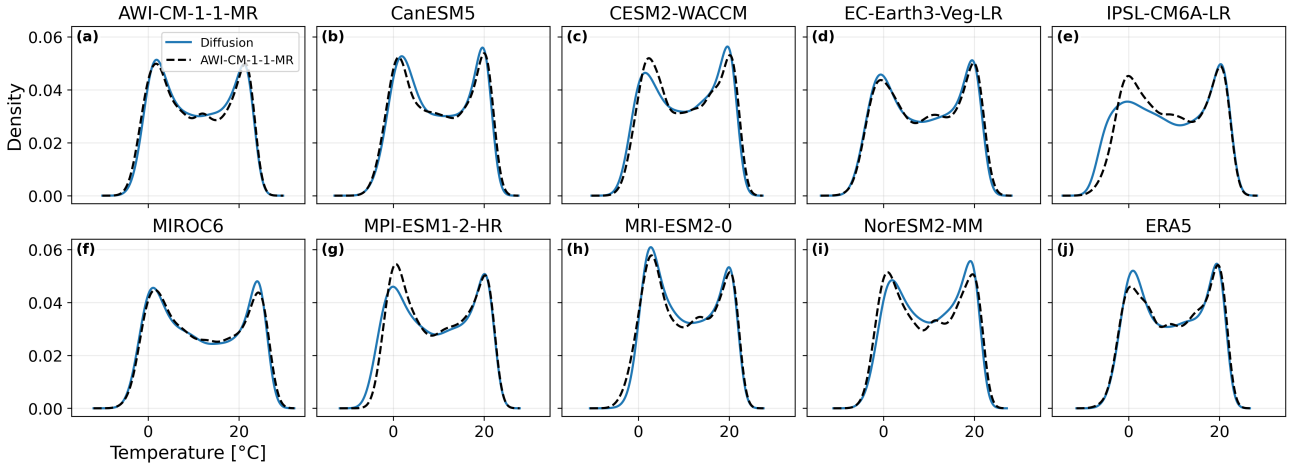


Figure E.11. KDE of regional-mean daily temperature for the unguided diffusion ensemble ($\gamma = 0$, solid blue) and the corresponding source data (dashed black), pooled across years, days, and (for the diffusion) 20 daily samples over 2001–2016. Each panel corresponds to a conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)). All panels use a shared kernel bandwidth.

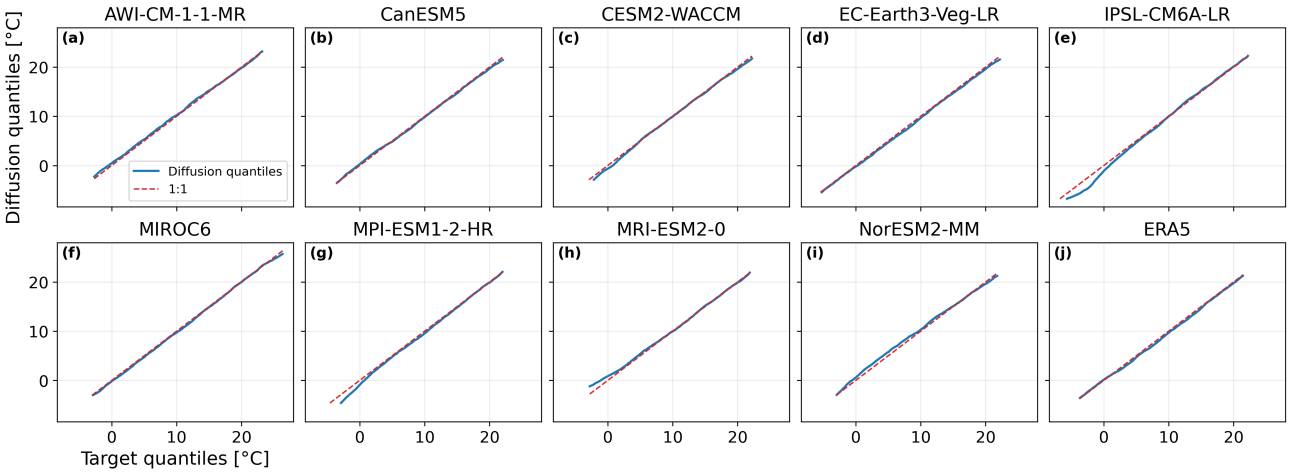


Figure E.12. Q-Q plots of daily temperature fields of the diffusion ensemble (solid blue line) and the corresponding ground truth data (dashed red line), pooled across all grid cells, years, days, and (for the diffusion) samples. The diffusion ensemble is made up of 20 samples per day, for every day, in 2001–2016. Each panel corresponds to the ensemble generated by the diffusion conditioned respectively on $s = \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$.

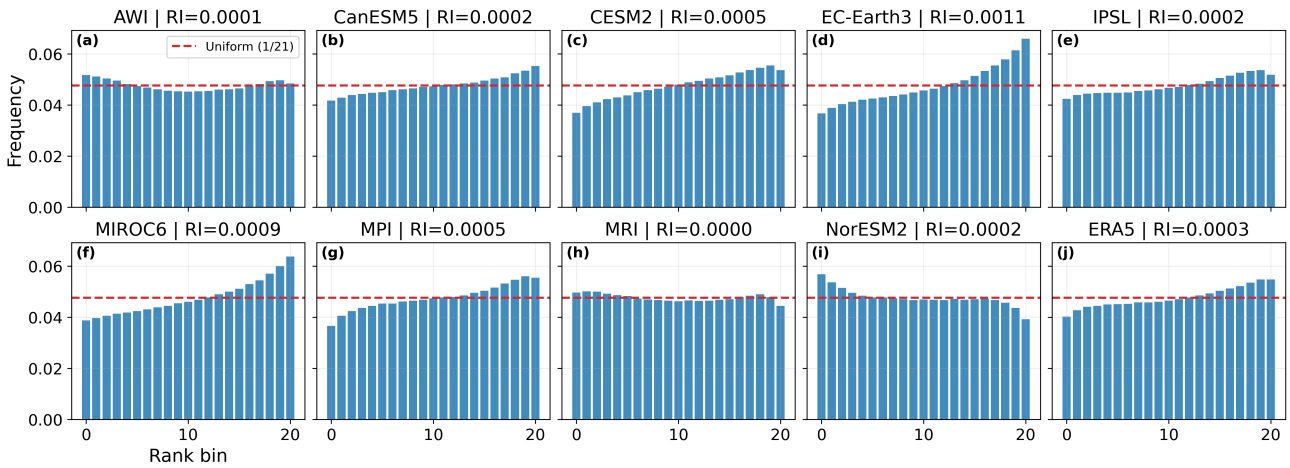


Figure E.13. Rank histograms of the source data within the unguided diffusion ensemble ($N_s = 20$ samples per day, 2001–2016 test period). Each panel corresponds to a conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)); the dashed red line marks the uniform calibration level $1/(N_s + 1)$. The RI (lower is better) is reported in each panel title.

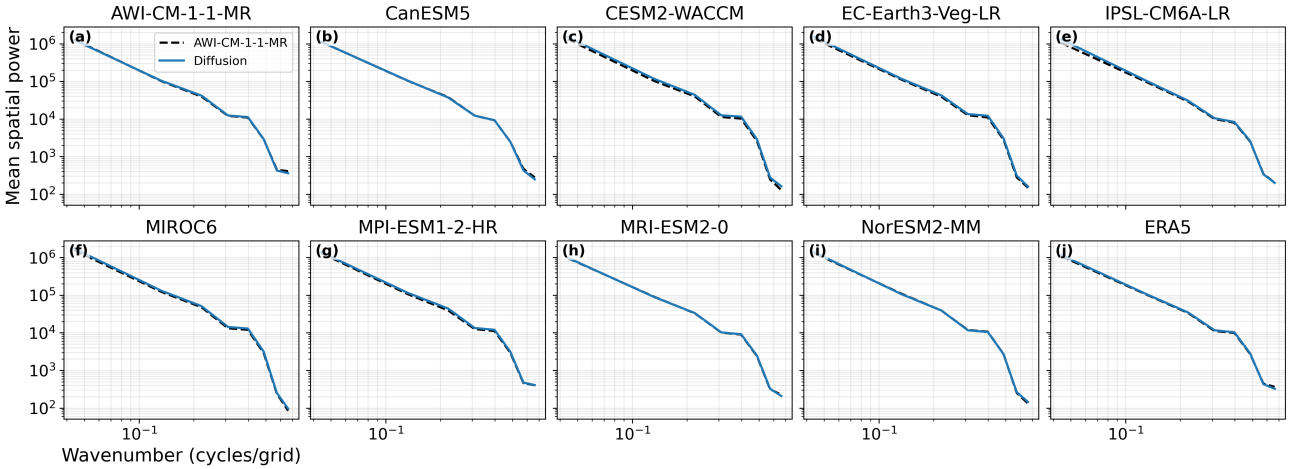


Figure E.14. Radially-averaged 2D spatial power spectra of the source data (dashed black) and the unguided diffusion ensemble ($\gamma = 0$, solid blue) for each conditioning $s \in \{\texttt{CMIP6}_1, \ldots, \texttt{CMIP6}_9, \texttt{ERA5}\}$ (panels (a)–(j)), 2001–2016 test period. Spectra overlap across all eight wavenumber bins for every conditioning, indicating neither spectral whitening nor oversmoothing of the diffusion ensemble.

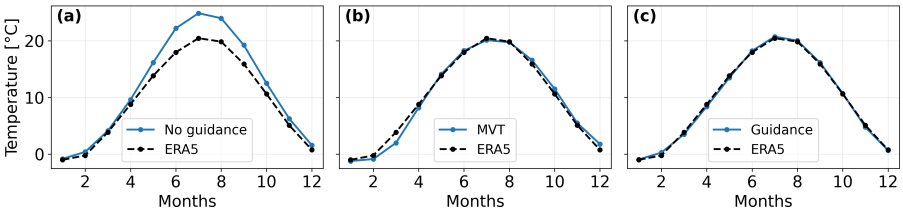


Figure F.1. Regional-mean monthly seasonal cycle for MIROC6 over the European domain (2001–2016 test period), compared with ERA5 (dashed black). **(a)** Uncorrected diffusion samples ($\gamma = 0$), with the July peak near 25°C against ERA5's 21°C. **(b)** MVT-corrected. **(c)** Guidance-corrected at $\gamma^\star = 0.9$. Both methods recover the regional-mean seasonal-cycle amplitude, but Figure 3 (row 2) shows that MVT achieves this through partial cancellation between an over-corrected southern region and an under-corrected northern region, whereas guidance is locally accurate.

Table F.1. **Scalar metrics for all CMIP6 models except MIROC6 (European domain, 2001–2016).** Lower is better. **Bold** marks the better value between guidance ($\gamma^\star = 0.9$) and MVT. The bottom row reports the mean across all eight models.

Model	Method	Central tendency		Higher moments		Extremes [pp]		Quantile bias [°C]		CRPS [°C]	RI	Lap. Rough. [°C]
		RMSE [°C]	VRD	Skew	Kurt	TX90p	TN10p	Q05	Q95			
AWI-CM-1-1-MR	Guidance	**0.45**	**0.03**	**0.07**	**0.10**	**1.63**	**1.56**	**0.83**	**0.50**	**1.95**	$\mathbf{1.2\times10^{-4}}$	**0.45**
	MVT	1.36	0.06	0.14	0.19	5.47	3.45	1.88	1.72	2.36	7.3×10^{-3}	0.61
CanESM5	Guidance	**0.44**	**0.03**	**0.08**	**0.10**	**1.49**	**1.52**	**0.79**	**0.46**	**1.95**	$\mathbf{2.6\times10^{-4}}$	**0.43**
	MVT	0.97	0.06	0.11	0.16	3.49	2.53	1.42	1.24	2.25	9.2×10^{-4}	0.76
CESM2-WACCM	Guidance	**0.51**	**0.03**	**0.07**	**0.11**	**1.63**	**1.66**	**0.87**	**0.50**	**1.96**	$\mathbf{2.0\times10^{-4}}$	**0.45**
	MVT	1.02	0.08	0.15	**0.11**	4.37	3.50	1.94	1.55	2.29	3.0×10^{-3}	0.78
EC-Earth3-Veg-LR	Guidance	**0.56**	**0.03**	**0.07**	**0.10**	**1.57**	**1.54**	**0.76**	**0.50**	**1.96**	$\mathbf{4.9\times10^{-4}}$	**0.45**
	MVT	0.71	0.06	0.09	0.17	3.51	2.42	1.13	1.31	2.27	2.1×10^{-3}	0.57
IPSL-CM6A-LR	Guidance	**0.48**	**0.02**	**0.07**	**0.09**	**1.45**	**1.36**	**0.65**	**0.45**	**1.95**	$\mathbf{3.2\times10^{-4}}$	**0.42**
	MVT	2.43	0.08	0.17	0.25	6.48	5.29	2.26	2.46	2.68	4.4×10^{-3}	1.40
MPI-ESM1-2-HR	Guidance	**0.49**	**0.02**	**0.07**	**0.10**	**1.56**	**1.40**	**0.71**	**0.49**	**1.95**	$\mathbf{4.0\times10^{-4}}$	**0.45**
	MVT	1.12	0.11	0.13	0.18	5.04	2.98	1.64	2.01	2.39	1.2×10^{-3}	0.80
MRI-ESM2-0	Guidance	**0.43**	**0.03**	**0.08**	**0.10**	**1.52**	**1.66**	**0.90**	**0.48**	**1.95**	$\mathbf{1.8\times10^{-4}}$	**0.42**
	MVT	1.98	0.12	0.11	0.18	6.67	4.70	2.40	2.41	2.54	3.2×10^{-3}	1.12
NorESM2-MM	Guidance	**0.50**	**0.03**	**0.08**	**0.11**	**1.71**	**1.69**	**0.88**	**0.52**	**1.96**	$\mathbf{2.8\times10^{-4}}$	**0.44**
	MVT	0.98	0.07	0.15	0.13	2.70	3.81	2.25	1.01	2.23	2.1×10^{-3}	0.62
Mean	Guidance	**0.48**	**0.03**	**0.07**	**0.10**	**1.58**	**1.55**	**0.80**	**0.49**	**1.95**	$\mathbf{2.8\times10^{-4}}$	**0.44**
	MVT	1.32	0.08	0.13	0.17	4.72	3.59	1.87	1.71	2.38	3.0×10^{-3}	0.83

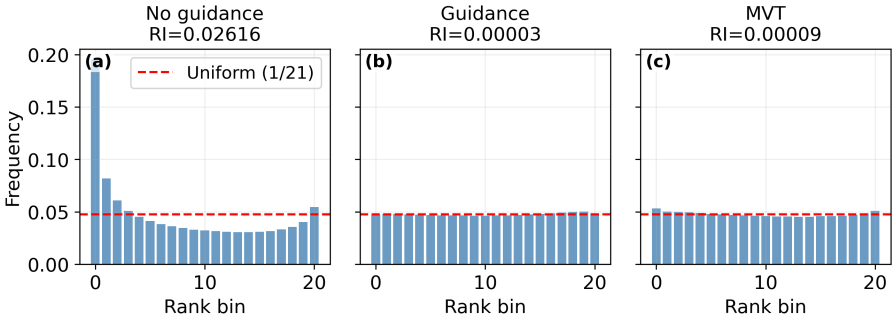


Figure F.2. Empirical rank histograms for MIROC6 over the European domain (2001–2016 test period), with $N_s = 20$ samples per day. The dashed red line at $1/(N_s + 1) \approx 0.048$ is the uniform calibration reference; the RI (lower is better) is reported in each panel title. **(a)** Uncorrected ($\gamma = 0$): pronounced spike at rank 0, indicating systematic under-prediction relative to ERA5 (RI $= 2.6 \times 10^{-2}$). **(b)** Guidance-corrected at $\gamma^\star = 0.9$: nearly flat (RI $= 3 \times 10^{-5}$). **(c)** MVT-corrected: also nearly flat, with mild residual rank-0 excess (RI $= 9 \times 10^{-5}$). Both correction methods produce well-calibrated ensembles; guidance is roughly $3\times$ tighter at this sample size.

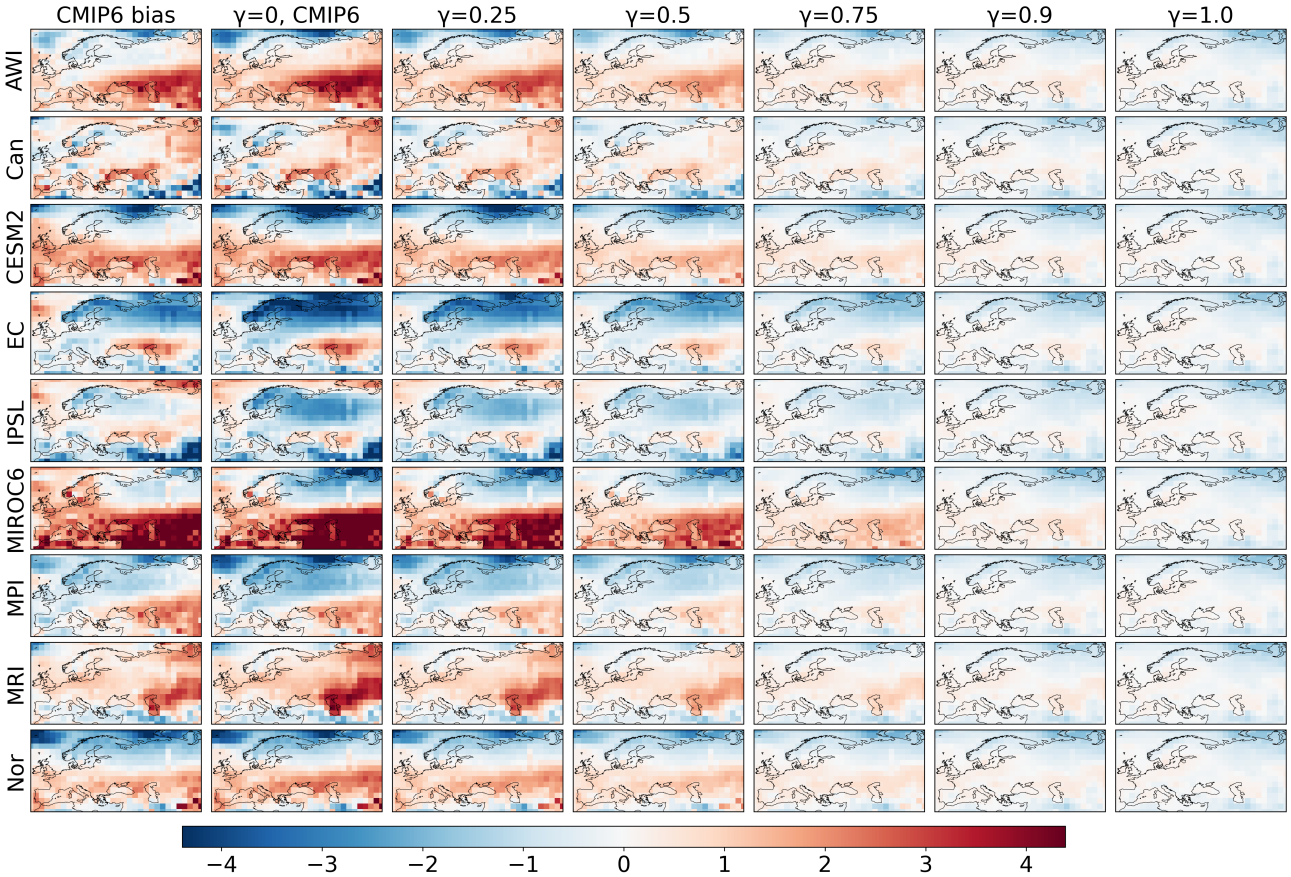


Figure F.3. Mean-bias evolution under condition-to-condition guidance for all nine CMIP6 models against ERA5 (2001–2016). Each row corresponds to one simulator (top to bottom: AWI-CM-1-1-MR, CanESM5, CESM2-WACCM, EC-Earth3-Veg-LR, IPSL-CM6A-LR, MIROC6, MPI-ESM1-2-HR, MRI-ESM2-0, NorESM2-MM). The leftmost column shows the raw CMIP6 models' mean bias against ERA5; the second column shows the unguided emulator at the corresponding CMIP6 conditioning ($\gamma = 0$, $s = \texttt{CMIP6}_i$) versus ERA5. Intermediate columns span the guidance scale $\gamma \in \{0.25, 0.5, 0.75, 0.9, 1.0\}$.

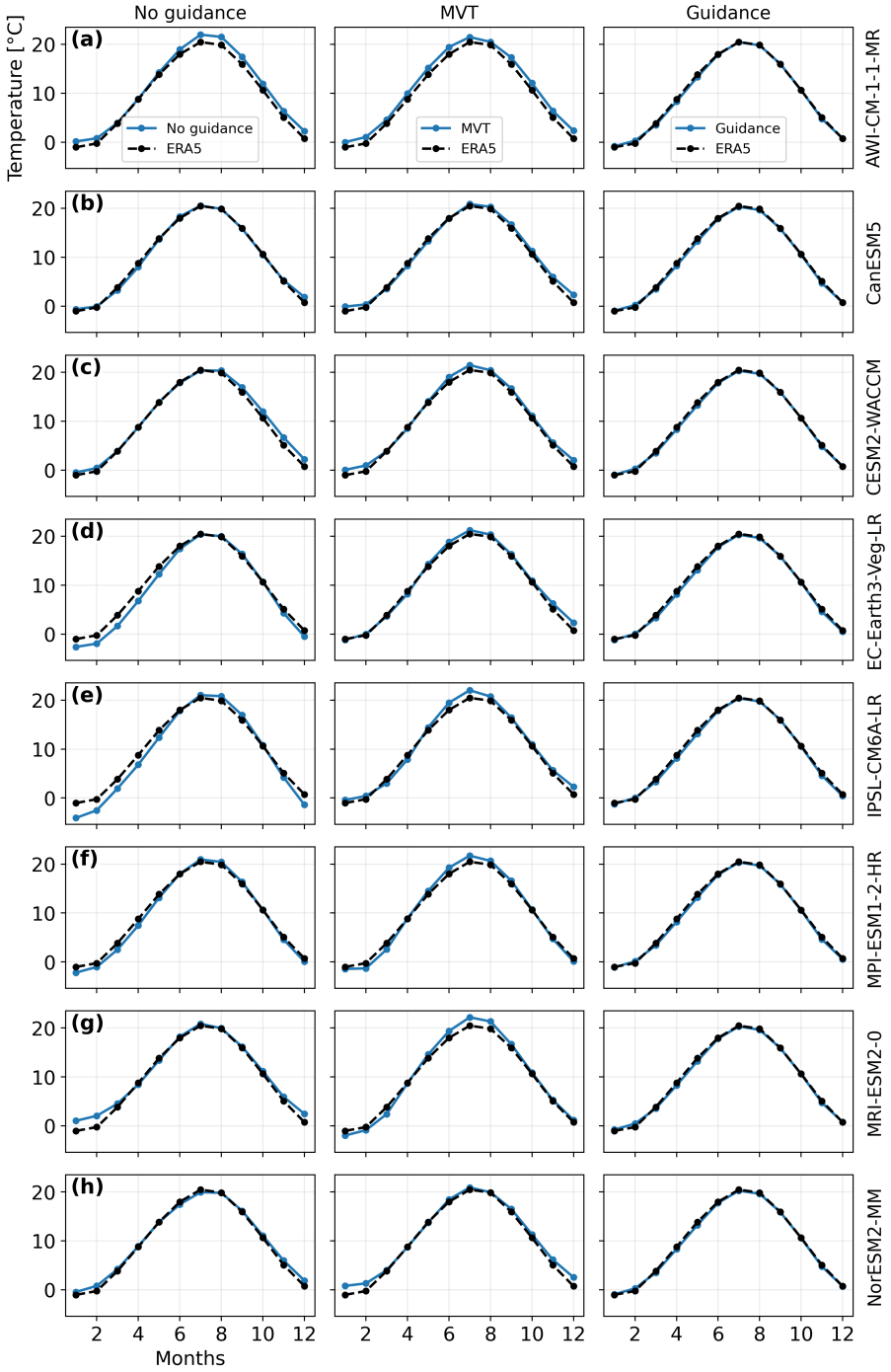


Figure F.4. Regional-mean monthly seasonal cycle for the CMIP6 models considered in this study (except MIROC6) over the 2001–2016 test period, compared with ERA5 (dashed black). Each row (panels (a)–(h)) is relative to one CMIP6 model; columns are uncorrected diffusion samples ($\gamma = 0$, left), MVT-corrected (center), and guidance-corrected at $\gamma^\star = 0.9$ (right). MIROC6 is shown in Figure F.1.

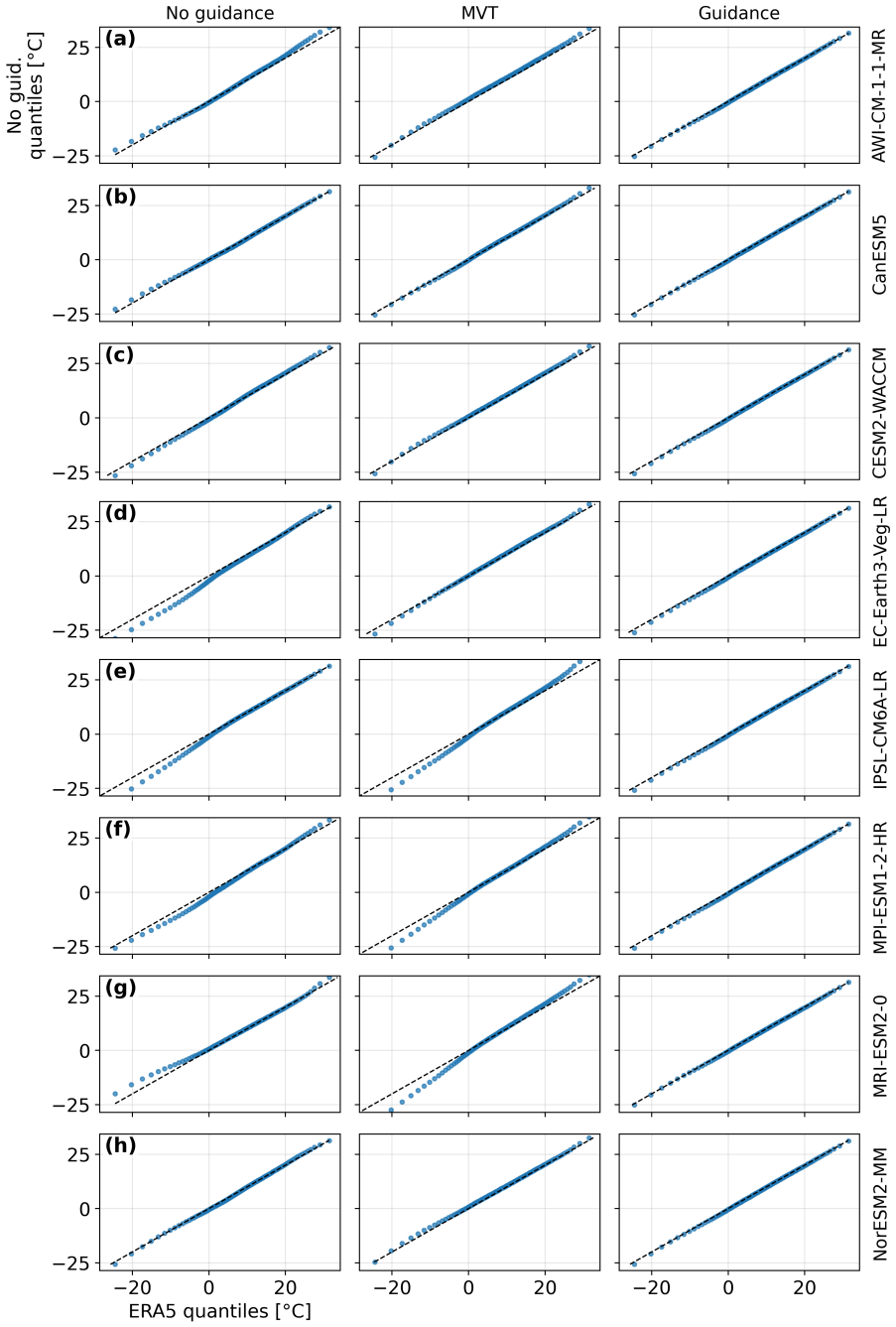


Figure F.5. Q-Q plots versus ERA5 for the CMIP6 models considered in this study (except MIROC6) over the 2001–2016 test period, pooled across grid cells, days, and the 20 daily samples; the dashed line is the 1:1 reference. Each row (panels (a)–(h)) is relative to one CMIP6 model; columns are uncorrected diffusion samples ($\gamma = 0$, left), MVT-corrected (center), and guidance-corrected at $\gamma^\star = 0.9$ (right). MIROC6 is shown in Figure 4b.

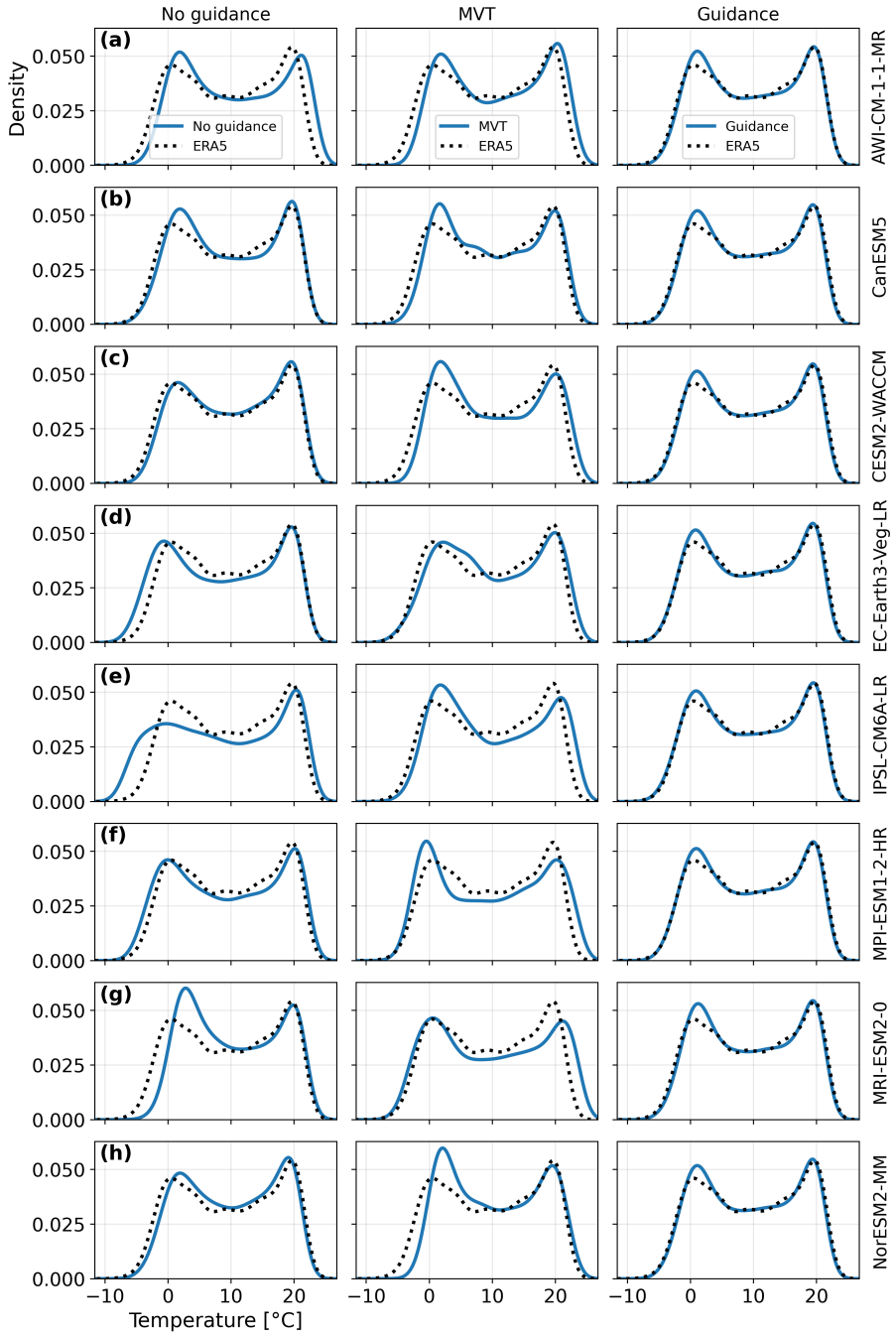


Figure F.6. KDE of regional-mean daily temperature for the CMIP6 models considered in this study (except MIROC6) over the 2001–2016 test period; method shown in solid yellow, ERA5 in dotted red. Each row (panels (a)–(h)) is relative to one CMIP6 model; columns are uncorrected diffusion samples ($\gamma = 0$, left), MVT-corrected (center), and guidance-corrected at $\gamma^\star = 0.9$ (right). MIROC6 is shown in Figure 4a.

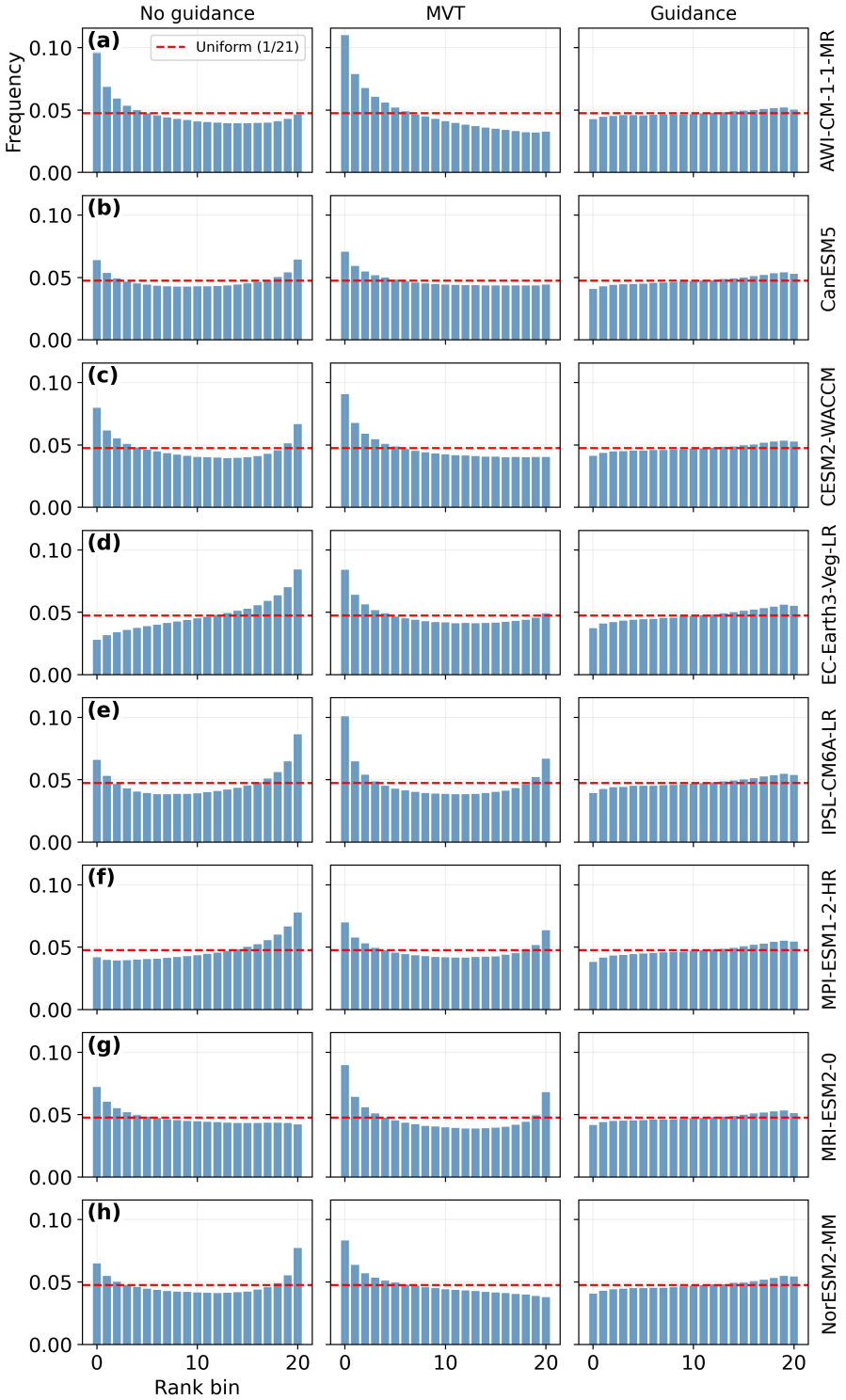


Figure F.7. Rank histograms versus ERA5 for the CMIP6 models considered in this study (except MIROC6) over the 2001–2016 test period; $N_s = 20$ samples per day. The dashed red line at $1/(N_s + 1)$ is the uniform calibration reference. Each row (panels (a)–(h)) is relative one CMIP6 model; columns are uncorrected diffusion samples ($\gamma = 0$, left), MVT-corrected (center), and guidance-corrected at $\gamma^\star = 0.9$ (right). Across all eight simulators, guidance produces visibly flatter histograms than either uncorrected or MVT. MIROC6 is shown in Figure F.2.

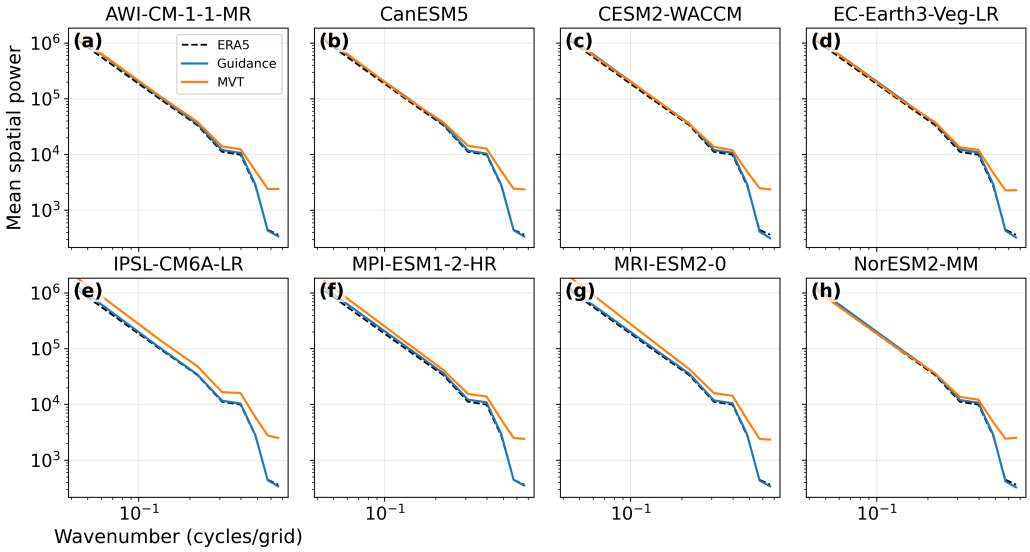


Figure F.8. Radially-averaged 2D spatial power spectra of ERA5 (dashed black), guidance-corrected samples at $\gamma^\star = 0.9$ (blue), and MVT-corrected samples (orange) for the CMIP6 models considered in this study (except MIROC6) over the 2001–2016 test period (panels (a)–(h)). For every CMIP6 model, guidance overlaps ERA5 across all wavenumbers, while MVT sits slightly above at intermediate wavenumbers and substantially exceeds ERA5 at the highest wavenumber bin (the same pixel-independence signature seen for MIROC6 in Figure 5).