GCD: GCM-consistent Diffusion for Zero-shot Downscaling across Heterogeneous GCMs

Ruian Tie[1,2,3,*] **Wenbo Xiong**[4,2,3,*] **Zhengyu Shi**[1,3,*] **Xinyu Su**[1,3]
Chenyu Jiang[4,3] **Libo Wu**[5,2,3,†] **Hao Li**[1,2,3,†]
[1] Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, China
[2] Shanghai Innovation Institute, Shanghai, China
[3] Shanghai Academy of AI for Science, Shanghai, China
[4] School of Data Science, Fudan University, Shanghai, China
[5] Institution for Big Data, Fudan University, Shanghai, China
{ratie25, wbxiong24, cyjiang25}@m.fudan.edu.cn,
{zhengyushi, xysu, wulibo, lihao_lh}@fudan.edu.cn

Abstract

Climate downscaling of General Circulation Models (GCMs) is an ill-posed inverse problem, as GCM outputs and reanalysis fields arise from different dynamical systems and lack fine-scale temporal pairing. Supervised methods thus rely on synthetic degradations or model-specific adaptation, limiting generalization across heterogeneous GCMs. Zero-shot diffusion posterior sampling (DPS) avoids paired supervision by combining a high-resolution climate prior with inference-time constraints from coarse GCM inputs, but faces consistency and efficiency challenges. We propose GCM-consistent Diffusion for Zero-shot Downscaling (GCD), which addresses these challenges with geographic conditioning, spectrally filtered measurement guidance, and distillation-accelerated Conjugate Gradient guidance. GCD reduces inference from 1,000 diffusion steps to 6 while maintaining consistency with the coarse GCM inputs. Across five heterogeneous GCMs, GCD achieves competitive 99th-percentile extreme-event accuracy and reconstructs plausible high-frequency tropical cyclone structures without model-specific retraining. Github: https://github.com/shawnvonbohr/GCD-Climate-Downscaling.git

1 Introduction

Climate downscaling is a fundamental task [1] that seeks to reconstruct plausible high-resolution atmospheric states from coarse-resolution (e.g., hundreds of kilometers) General Circulation Model (GCM) [2] outputs, supporting downstream applications such as climate impact assessment [3].

Despite its importance, climate downscaling remains fundamentally challenging due to the mismatch between GCM simulations and reanalysis data. Specifically, GCMs are free-running physical simulations whose generated weather trajectories do not chronologically align with reanalysis datasets such as ERA5. As a result, treating reanalysis data at the same timestamp as ground-truth labels introduces intrinsic supervision noise, making direct end-to-end training infeasible [4–6].

Previous studies address the mismatch between GCMs and reanalysis data by training downscaling models on proxy low-resolution inputs constructed from reanalysis data. Specifically, high-resolution reanalysis fields are artificially coarsened to generate synthetic low-resolution inputs, while the original reanalysis fields are used as high-resolution targets. The learned mapping is then applied directly to GCM outputs at inference time [7, 8], as shown in Fig. 1(a). However, the learned mapping

[*] Equal contribution. [†] Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

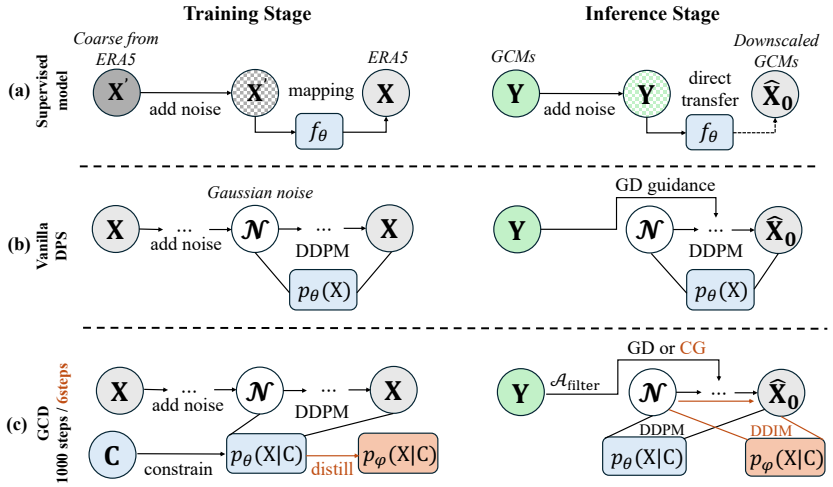


Figure 1: Schematic comparison of different downscaling paradigms during training and inference.

implicitly depends on the specific degradation process used to generate the synthetic inputs, which does not faithfully reflect the diverse error characteristics of real GCMs. As a result, a strong coupling exists between the trained model and a particular GCM, limiting its ability to generalize across heterogeneous GCMs with different spectral biases. This limitation becomes more pronounced in practical climate impact assessment, which typically relies on ensemble analysis over multiple heterogeneous GCMs, often ranging from 10 to 30 models [9–11]. Applying such approaches in this setting would require repeated retraining or adaptation for each GCM, resulting in poor scalability and substantial computational overhead.

Recent studies formulate GCM downscaling as an inverse problem and utilize diffusion posterior sampling (DPS) [12, 6] to overcome this limitation. By learning an unconditional prior from higher-resolution reanalysis data and enforcing GCM consistency at inference time, DPS improves generalization across heterogeneous GCMs without model-specific retraining, as shown in Fig. 1(b). However, directly applying vanilla DPS to global climate downscaling is bottlenecked by two critical challenges: consistency and efficiency. For consistency, the unconditional prior lacks geographic constraints and can produce physically implausible spatial patterns; meanwhile, standard measurement operators ignore spectral discrepancies between GCMs and the prior, degrading cross-model performance. For efficiency, the reverse diffusion process requires hundreds to thousands of function evaluations (NFEs), resulting in prohibitive inference costs. Beyond its application value, this setting poses a general machine learning challenge: zero-shot posterior sampling under domain shift, where a single pretrained diffusion prior must serve many heterogeneous, out-of-distribution measurement sources without retraining.

Motivated by the above challenges, we propose **GC**M-consistent Diffusion for Zero-shot **D**ownscaling across Heterogeneous GCMs (GCD), a zero-shot statistical downscaling framework for global GCMs. By eliminating the need for laborious model-specific retraining, GCD can be directly applied to diverse heterogeneous GCMs with substantially reduced computational costs. GCD targets cross-model consistency and efficient inference through three components: (1) *a geographically-constrained prior*, (2) *a filtering measurement operator*, and (3) *progressive distillation with conjugate gradient (CG) guidance*, as shown in Fig. 1(c). Our contributions are outlined as follows:

– A geographically-constrained climate prior to restore spatial consistency. We integrate static geographic boundaries ($\mathbf{C}$, e.g., topography) into the unconditional prior. Elevation and land–sea information has long been used in statistical downscaling [7]; we inject it into the score network of a diffusion prior. This explicit conditioning reduces physically implausible patterns and anchors the generated structures to local geography.

– A filtering measurement operator to resolve spectral inconsistency. We replace the asymmetric downsampling operator in vanilla DPS with a filtering measurement operator that acts as a symmetric low-pass filter. By extracting and aligning only the reliable large-scale signals (e.g., 5° resolution)

from both the estimated $\hat{\mathbf{X}}_0$ and GCM inputs, this operator substantially reduces the high-frequency spectral discrepancies between the ERA5 prior and heterogeneous models. Guidance thus constrains only scales at or above 5°; finer-scale structure is supplied by the geographically conditioned prior.

– A joint acceleration scheme for efficient posterior sampling. We combine progressive distillation with CG guidance to accelerate both the prior and guidance processes. For the prior, we first identify an effective NFE for the DDIM sampler (25 steps in our ablation) that preserves generation quality relative to the 1,000-step baseline, and then distill a compact prior that operates at this reduced step budget. For guidance, we replace vanilla gradient descent with the conjugate gradient (CG) method, which more effectively solves the underlying inverse problem and compensates for the truncation errors introduced by few-step sampling. Crucially, these two components are designed to be complementary: the distilled prior reduces the number of required sampling steps, while CG guidance improves convergence under limited iterations. Together, they compress inference from 1,000 steps to 6 while maintaining GCM consistency and competitive generation quality on most GCMs.

– Empirical evidence. Experiments on five CMIP6 GCMs show that GCD (1,000 NFE) is competitive with or better than supervised baselines in 99th-percentile errors, and the 6-step variant remains competitive on four of five GCMs, with a degradation on IPSL, whose raw fields deviate strongly from the ERA5 prior. GCD also reconstructs plausible high-frequency structures of tropical cyclones (TCs) with spectra consistent with ERA5. Because GCMs are free-running, we evaluate distributional fidelity rather than event-wise correctness.

2 Related Work

Unpaired Learning for Climate Downscaling. Supervised statistical downscaling trains on artificially coarsened reanalysis [7, 8], but the learned mapping is tied to the degradation used in training. Since GCM simulations and real-world observations lack paired data, researchers have explored unpaired and zero-shot downscaling. Early approaches aligned domain distributions via Normalizing Flows [13] or CycleGANs [4]. Recently, Schmidt et al. [6] formulated downscaling as a Bayesian inverse problem using vanilla Diffusion Posterior Sampling (DPS). However, applying vanilla DPS to global GCMs remains bottlenecked by physically implausible patterns from unconditional priors, spectral biases introduced by standard measurement operators, and prohibitive sampling costs. Our proposed GCD framework is designed to address these three limitations.

Solving Inverse Problems via Diffusion Priors. Leveraging diffusion models as priors for inverse problems has evolved from exploiting operator linearity to handling general non-linear degradations. For linear tasks, Kawar et al. [14] and Wang et al. [15] proposed DDRM and DDNM, respectively, which utilize SVD or null-space decomposition to analytically enforce data consistency. To extend applicability to general operators where such decomposition is infeasible, building upon the gradient guidance strategy in Chung et al. [16], Chung et al. [12] introduced DPS, which approximates the likelihood score to guide the reverse process. Addressing the high computational cost of such gradient-based guidance, Zhu et al. [17] focused on acceleration strategies, while Mardani et al. [18] proposed variational formulations that optimize the evidence lower bound directly. Closest to our guidance design, DDS [19] uses conjugate gradient (CG) to accelerate diffusion-based solvers for large-scale inverse problems. We build on the same idea but differ in setting and design: we couple CG with a spectral low-pass measurement operator and a distilled few-step student, targeting zero-shot downscaling across heterogeneous GCMs. WIND [20] is a concurrent diffusion-based framework for zero-shot atmospheric inverse problems; we discuss it as related work and do not compare against it.

Few-Step Generation. Progressive distillation [21] and consistency models [22] reduce sampling cost by learning large-stride transitions. We adopt progressive distillation rather than training a few-step model directly, for three reasons: (i) the student retains a discrete chain of intermediate noisy states, at which measurement guidance can be injected at inference time; (ii) teacher outputs allow explicit supervision with gradient and frequency losses to preserve high-frequency structure; and (iii) no per-GCM retraining is needed. Consistency models (CM, CM+noise) are included as baselines in Sec. 4.

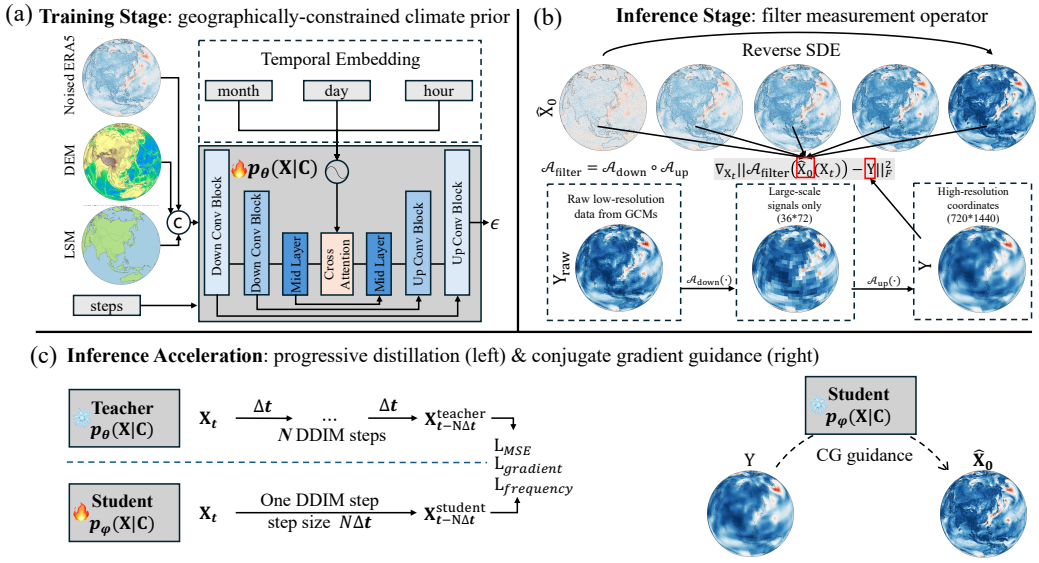


Figure 2: Overview of the GCD framework. (a) Training Stage: a geographically-constrained climate diffusion prior conditioned on static terrain and temporal embeddings. (b) Inference Stage: a filter measurement operator is applied during reverse diffusion sampling to reduce the spectral bias between the prior and GCMs. (c) Inference Acceleration: we use distillation and conjugate gradient guidance to compress the teacher into a fast student model.

3 Methodology

We present the GCD framework, with its overall training and inference pipelines illustrated in Fig. 2. We first formulate the downscaling task as an inverse problem (Section 3.1). Next, we detail the geographically-constrained climate prior and the filter measurement operator used to guide the generation process (Section 3.2). Finally, we introduce a joint acceleration scheme combining distillation and conjugate gradient guidance to improve inference efficiency (Section 3.3).

3.1 Problem Formulation

Climate Downscaling as Inverse Problem. Let $\mathbf{Y} \in \mathbb{R}^{h\times w\times C}$ denote a coarse-resolution GCM output, where h and w are the spatial grid dimensions (latitude and longitude; they vary across GCMs, see Appendix C.1), and C is the number of physical variables (wind components, pressure, and geopotential; $C = 5$ in our experiments). The corresponding high-resolution state is $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ on a finer grid, with $H = 720$, $W = 1440$ (ERA5 at $0.25°$), $h \ll H$ and $w \ll W$. Climate downscaling aims to reconstruct plausible fine-scale states from coarse-resolution simulations. Although $\mathbf{X}$ and $\mathbf{Y}$ are three-dimensional tensors, we represent them as two-dimensional matrices for notational simplicity in the following text, without loss of generality. The task can be formulated as an ill-posed inverse problem, where the coarse-resolution field $\mathbf{Y}$ is degraded from the high-resolution state $\mathbf{X}$ through a non-invertible measurement operator:

$$\mathbf{Y} = \mathcal{A}(\mathbf{X}) + \mathbf{N}, \tag{1}$$

where $\mathcal{A}(\cdot) : \mathbb{R}^{H\times W} \mapsto \mathbb{R}^{h\times w}$ models spatial aggregation and resolution transform, $\mathbf{N} \sim \mathcal{MN}_{h\times w}(\mathbf{0}, \mathbf{\Sigma}_h, \mathbf{\Sigma}_w)$ is measurement noise following a matrix variate Gaussian distribution, and $\mathbf{\Sigma}_h$ and $\mathbf{\Sigma}_w$ are the row (latitudinal) and column (longitudinal) covariance matrices, respectively.

The objective is to learn a model f_θ that, given $\mathbf{Y}$, produces a high-resolution reconstruction $\hat{\mathbf{X}}$ consistent with the large-scale content of $\mathbf{Y}$ and statistically plausible at fine scales, rather than merely increasing resolution through interpolation.

Vanilla Diffusion Posterior Sampling in Climate Downscaling. We seek to reconstruct a plausible high-resolution state $\mathbf{X}$ from a coarse-resolution GCM output $\mathbf{Y}$ without access to paired coarse- and high-resolution samples. The ill-posed nature of the problem motivates a probabilistic formulation,

in which the solution is represented by the posterior distribution $p(\mathbf{X}|\mathbf{Y})$ rather than a point estimate. By Bayes' rule,

$$p(\mathbf{X}|\mathbf{Y}) \propto \underbrace{p(\mathbf{Y}|\mathbf{X})}_{\text{Data Consistency}} \cdot \underbrace{p(\mathbf{X})}_{\text{Climate Prior}}, \tag{2}$$

where $p(\mathbf{X})$ is a prior over physically and statistically plausible high-resolution states, and the likelihood $p(\mathbf{Y}|\mathbf{X})$ encodes the consistency between the reconstructed field and the observed coarse-resolution GCM output. In this zero-shot setting, the prior is trained on ERA5 alone, and the GCM output enters only at inference time, without paired supervision.

3.2 Geographically-Constrained Climate Prior & Filter Measurement Operator

Unlike vanilla DPS formulations that employ an unconditional diffusion prior, we parameterize the prior using a conditional DDPM [23]. To incorporate geographical information, we condition the reverse process on a context $\mathbf{C}$. This context integrates static spatial boundaries, specifically the land–sea mask (LSM) and digital elevation model (DEM), via channel-wise concatenation, in line with the use of elevation information in statistical downscaling [7]. It also includes cyclic temporal information (month, day, and hour) encoded via sine–cosine embeddings and cross-attention (see Fig. 2(a)). Given high-resolution reanalysis data $\mathbf{X}_0$, the network $\boldsymbol{\epsilon}_\theta(\mathbf{X}_t, t, \mathbf{C})$ is trained to approximate the added Gaussian noise (details in Appendix B.1).

During inference, vanilla DPS relies on a downsampling operator $\mathcal{A}_{\text{down}}$ to align the estimated state $\hat{\mathbf{X}}_0$ with coarse GCM outputs [6]. Inspired by Hess et al. [8], we observe that this operator overlooks the spectral bias between the ERA5-trained prior and heterogeneous GCMs. These discrepancies in high-frequency spatial signals degrade generation performance. Frequency-domain filtering, such as the Fast Fourier Transform (FFT), is a standard way to remove high frequencies, but hard frequency cutoffs on meteorological fields introduce ringing artifacts (the Gibbs phenomenon).

To avoid frequency-domain filtering, we propose a spatial-domain Filter Measurement Operator, $\mathcal{A}_{\text{filter}} = \mathcal{A}_{\text{up}} \circ \mathcal{A}_{\text{down}}$. By applying area-weighted downsampling to a 5° resolution (36×72 grid) before upsampling, $\mathcal{A}_{\text{filter}}$ acts as a spatial low-pass filter. It reduces high-frequency discrepancies and avoids the ringing of hard spectral cutoffs, providing large-scale guidance for the generative process. The same operator is applied to the generated field and to the GCM input (after regridding to the common 720×1440 grid), so both are compared at the same 5° scale. The inference details and algorithms of GCD are provided in Appendix B.2.

What is constrained. Guidance constrains only scales at or above 5°. Structure below 5° is supplied by the ERA5-trained, geographically conditioned prior. We therefore evaluate generated fields distributionally and do not claim event-level correctness (Sec. 4).

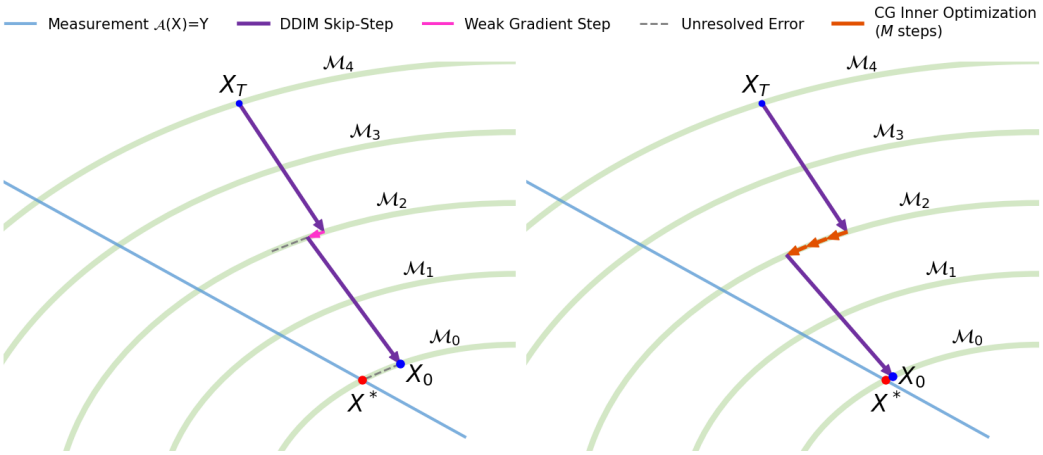
3.3 Distillation & Conjugate Gradient Guidance

To accelerate sampling, we distill a fast student model from a DDIM teacher. To limit distillation cost, we first identify an effective DDIM step budget that balances speed and generation quality (25 steps in our ablation, Sec. 4). The model at this "sweet spot" serves as the teacher. The student is trained in a single distillation stage to reproduce the teacher's cumulative transition over N consecutive teacher steps (stride Δt) in one step of stride $N\Delta t$, which compresses the 25-step teacher into a 6-step student on a non-uniform time schedule. Since distillation with a plain mean squared error ($\mathcal{L}_{\text{MSE}}$) on the intermediate denoised states $\mathbf{X}_t$ tends to yield over-smoothed results, we use a joint loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{grad}}\mathcal{L}_{\text{grad}} + \lambda_{\text{freq}}\mathcal{L}_{\text{freq}}. \tag{3}$$

The term $\mathcal{L}_{\text{MSE}}$ encourages global pixel consistency between the student states $\mathbf{X}^{\text{student}}_{t-N\Delta t}$ and the teacher states $\mathbf{X}^{\text{teacher}}_{t-N\Delta t}$:

$$\mathcal{L}_{\text{MSE}} = \|\mathbf{X}^{\text{student}}_{t-N\Delta t} - \mathbf{X}^{\text{teacher}}_{t-N\Delta t}\|_F^2. \tag{4}$$



(a) Geometry of DDIM (b) Geometry of DDIM with CG Guidance

Figure 3: Schematic illustration of the geometries of the two guidance processes (not measured data).

To preserve edge structures, the spatial gradient loss $\mathcal{L}_{\text{grad}}$ penalizes differences in local spatial variations of the predicted clean state $\hat{\mathbf{X}}_0$:

$$\mathcal{L}_{\text{grad}} = \|\nabla\hat{\mathbf{X}}_0(\mathbf{X}^{\text{student}}_{t-N\Delta t}) - \nabla\hat{\mathbf{X}}_0(\mathbf{X}^{\text{teacher}}_{t-N\Delta t})\|_F^2, \tag{5}$$

where ∇ denotes the spatial gradient operator (e.g., Sobel filter). The frequency loss $\mathcal{L}_{\text{freq}}$ is computed in the Fourier domain to counter the loss of high-frequency texture:

$$\mathcal{L}_{\text{freq}} = \|\mathcal{F}(\hat{\mathbf{X}}_0(\mathbf{X}^{\text{student}}_{t-N\Delta t})) - \mathcal{F}(\hat{\mathbf{X}}_0(\mathbf{X}^{\text{teacher}}_{t-N\Delta t}))\|_F^2, \tag{6}$$

where $\mathcal{F}$ denotes the FFT. Ablations (Sec. 4) show that adding $\mathcal{L}_{\text{grad}}$ and $\mathcal{L}_{\text{freq}}$ lowers the 6-step error compared with $\mathcal{L}_{\text{MSE}}$ alone.

While distillation reduces the number of denoising steps, consistency between the generated results and the observation must still be enforced. For notational simplicity, let $\mathbf{X}$ denote the intermediate denoised state at the current step and $\hat{\mathbf{X}}_0(\mathbf{X})$ the associated predicted clean state. Vanilla DPS updates $\mathbf{X}$ along the negative gradient of $\mathcal{J}(\mathbf{X}) = \|\mathcal{A}(\hat{\mathbf{X}}_0(\mathbf{X})) - \mathbf{Y}\|_F^2$ with gradient descent (GD). In the few-step regime, a single GD update per step corrects the deviations of large-stride prior transitions less accurately, and the residual errors accumulate across steps (Fig. 3(a); ablation in Sec. 4).

We therefore replace GD with the conjugate gradient (CG) method, applied to a regularized least-squares problem. Rather than relying on a local gradient step ($\mathbf{X} \leftarrow \mathbf{X} - \eta\nabla\mathcal{J}(\mathbf{X})$), we solve an inner-loop problem that strengthens observation consistency while staying close to the prior estimate:

$$\min_{\mathbf{X}} \|\mathcal{A}(\mathbf{X}) - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{X} - \hat{\mathbf{X}}_0\|_F^2, \tag{7}$$

where $\lambda > 0$ controls how strongly the solution is anchored to $\hat{\mathbf{X}}_0$. This is a Tikhonov regularization centered on $\hat{\mathbf{X}}_0$. Setting the gradient to zero gives the normal equation

$$(\mathcal{A}^T\mathcal{A} + \lambda\mathbf{I})\mathbf{X} = \mathcal{A}^T\mathbf{Y} + \lambda\hat{\mathbf{X}}_0. \tag{8}$$

For $\lambda > 0$ the matrix $\mathcal{A}^T\mathcal{A} + \lambda\mathbf{I}$ is symmetric positive definite, so CG applies, and its condition number is bounded by $(\|\mathcal{A}\|_2^2 + \lambda)/\lambda$. We run $M = 5$ CG iterations with $\lambda = 0.1$. Each iteration requires one application of $\mathcal{A}$ and one of $\mathcal{A}^T$; since $\mathcal{A}_{\text{filter}}$ is a sparse, local linear operator on $\mathbf{X} \in \mathbb{R}^{720\times1440\times5}$, the cost per iteration is $O(HWC)$.

The CG solver relies on the linearity of $\mathcal{A}$; nonlinear operators would require other solvers or standard DPS-style gradient guidance. The overall procedure is summarized in Fig. 2(c), with details in Appendix B.3.

Table 1: Quantitative comparison of downscaling methods on different GCMs. The first five columns report the spatial MAE/RMSE averaged over the test period (2001-2009) for each respective GCM. Mean Inf. Time denotes the average inference time per sample across all five GCMs. Per GPU Hours reports the computational cost on a single GPU in training stage. Notably, the 'X' multiplier indicates that the cost must be multiplied by the number of target GCMs (e.g., requiring separate training for each specific GCM). Best results are highlighted in **bold** and underline denotes the second-best.

METHOD	IPSL	MIROC6	AWI	MPI-LR	MPI-HR	MEAN INF. TIME PER SAMPLE (S)	PER GPU HOURS (H)
BILINEAR	1.06 / 1.78	1.84 / 2.43	1.50 / 2.19	1.48 / 2.12	1.32 / 1.90	$\mathbf{1.80 \times 10^{-5}}$	–
CUBIC	1.03 / 1.79	1.82 / 2.42	1.50 / 2.18	1.47 / 2.12	1.30 / 1.87	$\underline{1.37 \times 10^{-4}}$	–
BCSD [30]	0.98 / 1.67	1.72 / 2.33	1.36 / 2.07	1.35 / 1.98	1.24 / 1.76	0.27	–
SWINIR [31]	1.01 / 1.67	1.52 / 2.05	1.37 / 2.02	1.35 / 1.95	1.18 / 1.68	0.25	11X
ESRGAN [32]	1.00 / 1.66	1.50 / 2.03	1.36 / 2.00	1.34 / 1.93	1.17 / 1.67	0.04	22X
CORRDIFF [33]	0.98 / 1.62	1.45 / 1.95	1.33 / 1.96	1.31 / 1.89	1.14 / 1.62	2.33	35X
CM [34]	0.97 / 1.61	1.46 / 1.96	1.34 / 1.97	1.30 / 1.88	1.15 / 1.63	0.16	33X
CM+NOISE [8]	<u>0.93 / 1.52</u>	1.32 / 1.78	1.30 / 1.92	1.28 / 1.85	1.10 / <u>1.56</u>	0.16	33
DDRM [14]	1.04 / 1.69	1.84 / 2.45	1.53 / 2.20	1.43 / 2.10	1.33 / 2.02	3.92	**12.5**
DPS [12]	1.03 / 1.67	1.84 / 2.40	1.48 / 2.13	1.40 / 2.08	1.32 / 1.89	588.38	**12.5**
GCD (1000 NFE)	**0.87 / 1.32**	<u>1.08</u> / **1.42**	<u>1.28 / 1.89</u>	<u>1.24 / 1.81</u>	**1.05 / 1.49**	589.62	**12.5**
GCD (6 NFE)	1.49 / 1.95	**1.05** / <u>1.50</u>	**1.14 / 1.71**	**1.07 / 1.63**	<u>1.08</u> / 1.59	4.38	<u>24.5</u>

4 Experiments

4.1 Experimental Setup

Datasets. *<u>GCMs Simulations.</u>* To evaluate the zero-shot generalization capability of our framework, we use 6-hourly outputs from the historical experiment of five different GCMs models in Coupled Model Intercomparison Project Phase 6 (CMIP6) [24], including IPSL-CM6A-LR [25], AWI-ESM-1-1-LR [26], MIROC6 [27], MPI-ESM1-2-LR and MPI-ESM1-2-HR [28]. *<u>ERA5.</u>* We utilize the ERA5 reanalysis dataset [29] as our high-resolution ground truth, using its native 0.25° spatial resolution (720×1440 grid points) and subsampling it to 6-hour intervals. The prior trained on ERA5 (1980–1999), validate on 2000, and test on 2001–2009 using both ERA5 and GCMs. GCMs data are not used in training or validation. More details about variables are in Appendix C.1

Baselines. We benchmark GCD against three distinct paradigms. First, traditional interpolations (Bilinear, Cubic). Second, supervised methods, which we evaluate in a direct-transfer setting to assess their cross-GCM generalization. This includes standard models (SwinIR, ESRGAN, CorrDiff, CM) that typically require model-specific retraining, as well as a domain-adapted variant (CM+noise) that injects artificial noise during training to mitigate domain gaps. Third, zero-shot solvers (DDRM, vanilla DPS) that bypass paired training entirely. Details are provided in Appendix C.2.

Implementation Details. Our prior is parameterized by a time- and condition-aware 2D U-Net. All models are trained using the AdamW optimizer. Comprehensive architectural details and training hyperparameters are detailed in Appendix C.3.

Evaluation Methods. Following recent established practices for evaluating unaligned climate fields [6], we evaluate distributional consistency and extreme-state characteristics by computing the latitude-weighted Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of the 99th percentile for each variable (formulas in Appendix B.4). To reduce the impact of randomness, we calculate the average over 5 ensemble members.

4.2 Main Results

We evaluate GCD's zero-shot capabilities directly across five heterogeneous GCMs through two critical dimensions. First, we benchmark overall performance using 99th percentile errors to assess extreme climate simulations, alongside a comparative analysis of inference efficiency and training costs. Second, we validate physical fidelity across scales by visualizing localized TCs structures and analyzing the global annual mean Power Spectral Density (PSD) to confirm the restoration of high-frequency atmospheric energy.

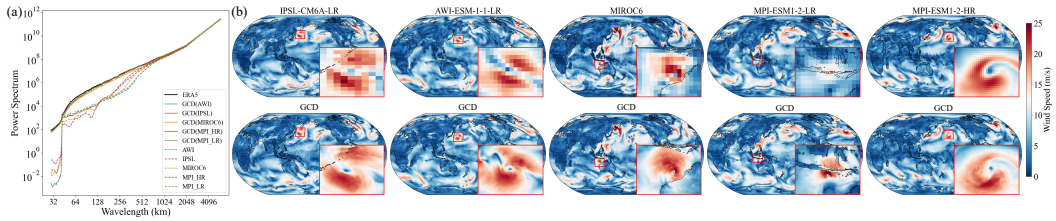


Figure 4: Spectral and spatial demonstration of GCD (1000 NFE) capabilities. **(a)** Annual mean power spectra of 10m Wind Speed (WS10m). **(b)** Visual comparison of WS10m and TCs. The **top row** displays raw outputs from various GCMs. The **bottom row** shows the corresponding high-resolution fields reconstructed by our GCD framework, revealing detailed vortex structures. Columns correspond to samples from (left to right): 2001-02-24 00:00, 2002-01-26 00:00, 2001-01-01 00:00, 2001-01-03 00:00, and 2001-03-16 00:00.

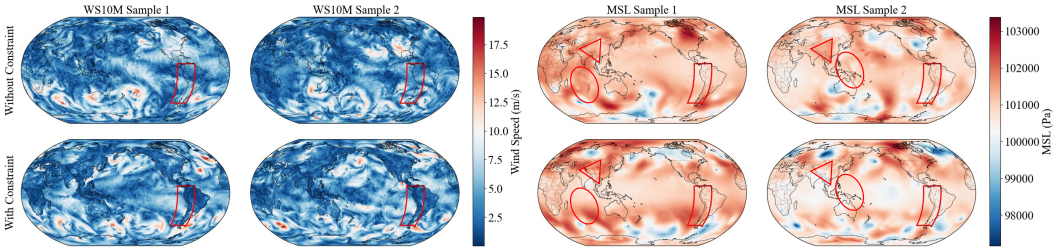


Figure 5: Impact of geographical constraints on generated climate states. The Top Row (without constraints) fails to capture topography-driven phenomena, lacking both the terrain blocking effect over the Andes (red box) and the expected high-pressure system over the Tibetan Plateau (red triangle), while also exhibiting spurious high-pressure artifacts (red ellipses). In contrast, the Bottom Row (with constraints) produces more physically plausible states, showing correct topography-induced wind deceleration, restoring missing pressure features, and eliminating unphysical anomalies.

Overall Performance. We evaluate GCD across five heterogeneous GCMs (Tab. 1), focusing on the trade-offs among accuracy, training costs, and inference efficiency. First, in terms of accuracy, the full-trajectory GCD (1000 NFE) demonstrates robust zero-shot generalization. It not only outperforms traditional zero-shot baselines like Vanilla DPS but also matches or exceeds supervised methods across most metrics. For instance, on the MIROC6 dataset, GCD (1000 NFE) achieves an MAE of 1.08, which is lower than the supervised models CorrDiff (1.45) and CM+NOISE (1.32).

Second, regarding scalability, GCD avoids the model-specific retraining limitations common in supervised approaches. Because GCMs and ERA5 lack chronologically paired data, standard supervised models (e.g., CorrDiff and SwinIR) typically learn a mapping from artificially degraded ERA5 to high-resolution ERA5. To ensure reliable inference, this artificial degradation must align with the specific characteristics of the target GCM. Consequently, the significant domain shifts between different GCMs often necessitate retraining for each new model. As shown in Tab. 1, this results in a linear increase in computational cost (denoted by the 'X' multiplier; e.g., CorrDiff requires 35 GPU hours per new GCM). While methods like CM+NOISE attempt to bypass retraining by injecting random noise to simulate domain shifts, simple statistical noise struggles to encapsulate complex physical discrepancies, which limits their generalization accuracy. In contrast, GCD requires only a single training phase on ERA5 data (12.5 h for the full-trajectory base model, and 24.5 h in total for the distilled 6-NFE version), and can be directly applied to any GCM in a zero-shot manner.

Finally, regarding inference efficiency, GCD (6 NFE) provides a practical solution that balances speed and accuracy. Standard zero-shot diffusion methods (e.g., Vanilla DPS) require nearly 588 seconds per sample due to heavy gradient computations. Through our joint acceleration strategy, GCD (6 NFE) reduces this inference time to 4.38 seconds. While this extreme step compression introduces higher variance on certain distributions, such as the IPSL dataset where the MAE increases to 1.49, GCD (6 NFE) maintains stable zero-shot accuracy on the remaining four GCMs (MIROC6, AWI, MPI-LR, and MPI-HR). In some cases, it even outperforms supervised baselines (e.g., achieving an MAE of 1.05 on MIROC6). Overall, although slightly slower than the deterministic CorrDiff (2.33 seconds), GCD (6 NFE) strikes a reasonable balance between eliminating retraining costs, improving inference efficiency, and maintaining generation quality.

TC Simulation. Fig. 4 evaluates the zero-shot capabilities of GCD on TCs. As mesoscale weather systems, TCs exhibit significant energy concentrated at finer spatial scales, at scales of tens of kilometers. Fig. 4(a) shows that raw GCMs experience substantial energy loss at these high-frequency bands. GCD effectively restores this missing energy, producing PSD curves that closely match the ERA5 reference at these critical scales. This spectral recovery directly corresponds to the spatial visualizations in Fig. 4(b). By restoring the high-frequency energy, GCD reconstructs coherent vortex structures and detailed textures that are otherwise absent in the diffuse coarse inputs. Additional variables are provided in Appendix D.1.

4.3 Ablation Study

Impact of Boundary Conditions on the Prior. To validate the design of our geographically-constrained prior, we conduct an ablation study focusing on the boundary conditions. Fig. 5 compares random samples generated by the prior with and without geographical conditioning. Without these conditions, the generated wind fields are overly smooth and lack spatial variations associated with terrain features. Specifically, the flow patterns fail to reflect sharp transitions at coastlines or topographically induced constraints.

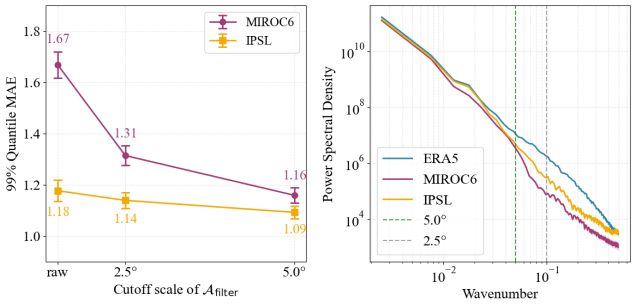


Figure 6: Impact of cutoff scale of the filter measurement operator ($\mathcal{A}_{\text{filter}}$). All evaluations are based on data from the test year 2001. (a) The 99% quantile MAE evaluated at different cutoff scales: raw (representing the native resolution of the GCMs), 2.5°, and 5.0°. (b) PSD comparison among ERA5, MIROC6, and IPSL. The vertical dashed lines indicate the wavenumbers corresponding to the 5.0° and 2.5° scales.

By contrast, conditioning on the Digital Elevation Model (DEM) and Land Sea Mask (LSM) yields structures that closely align with local geography. For example, along the west coast of South America, the conditioned model captures sharp gradients at the coastal boundaries and reflects the topographical influence of the Andes. These observations indicate that geographical conditioning is essential for preventing overly smoothed outputs and generating terrain-aware flow structures. A detailed breakdown of the individual contributions of DEM and LSM is provided in Appendix D.2.

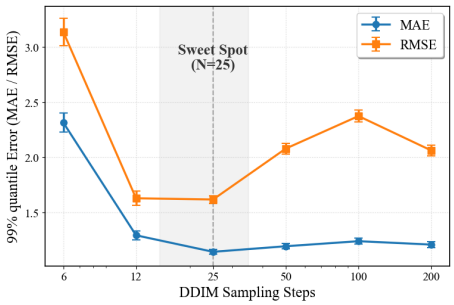
Impact of Filter Measurment Operator. Increasing the cutoff scale of the filter measurement operator ($\mathcal{A}_{\text{filter}}$) effectively reduces the downscaling error. We evaluate this relationship using the 99% quantile MAE on the 2001 test set. Fig. 6(a) compares the performance across three scales: the native GCM resolution (raw), 2.5°, and 5.0°. For both MIROC6 and IPSL, the error consistently decreases as the cutoff scale becomes larger. The lowest MAE is achieved at the 5.0° scale. This trend indicates that removing small-scale spatial information from the input variables improves the final prediction accuracy.

This performance improvement stems from mitigating the high-frequency spectral discrepancies between the GCMs and the ERA5 reference. Fig. 6(b) illustrates the PSD of the datasets. At higher wavenumbers corresponding to the raw resolutions, the GCM spectra deviate significantly from the ERA5 distribution. Applying a 5.0° low-pass filter successfully removes these mismatched high-frequency components. Consequently, the aligned large-scale features serve solely as inference-time guidance, preventing the downscaling process from being misled by inaccurate local inputs.

Additionally, unlike spatial downsampling measurement operator that suffers from vanishing gradients at the 5.0° scale, this filter operator maintains stable guidance gradients throughout the generative process (see Appendix D.3 for detailed gradient analysis).

Impact of Distillation & Conjugate Gradient Guidance. To balance downscaling accuracy and inference efficiency, we first establish the performance upper bound of the teacher model by evaluating standard DDIM sampling on the MIROC6 dataset. As shown in Fig. 7 (left), the 99% quantile errors exhibit a clear U-shaped trend with respect to the NFE. The errors reach their minimum at NFE = 25 (MAE: 1.1434, RMSE: 1.6177), representing an optimal trade-off between detail generation and

denoising stability. Notably, increasing sampling steps beyond 25 (e.g., to 50 or 200) yields no further improvements and even slightly degrades the performance. Consequently, we identify the NFE = 25 trajectory as the optimal trade-off to serve as the baseline for subsequent distillation.



Distillation Loss	Guidance	MAE	RMSE
$\mathcal{L}_{\text{MSE}}$	GD	1.81	2.40
$\mathcal{L}_{\text{MSE}}$ & $\mathcal{L}_{\text{freq}}$	GD	1.79	2.36
$\mathcal{L}_{\text{MSE}}$ & $\mathcal{L}_{\text{grad}}$	GD	1.71	2.25
$\mathcal{L}_{\text{MSE}}$ & $\mathcal{L}_{\text{grad}}$ & $\mathcal{L}_{\text{freq}}$	GD	1.69	2.23
$\mathcal{L}_{\text{MSE}}$	CG	1.25	1.71
$\mathcal{L}_{\text{MSE}}$ & $\mathcal{L}_{\text{grad}}$ & $\mathcal{L}_{\text{freq}}$	CG	1.16	1.59

Figure 7: Performance evaluation on the MIROC6 (2001) dataset. The left panel evaluates the impact of DDIM NFE to identify the optimal sweet spot (NFE=25). Based on this optimal trajectory, the right table presents an ablation study of student models distilled down to 6 steps, comparing different distillation losses and guidance strategies (GD vs. CG).

Using the 25-step model as our teacher, we distill the generation process down to 6 steps for ablation analysis (Fig. 7, right). Results show that standard distillation at such a low-step regime yields limited accuracy. By explicitly introducing $\mathcal{L}_{\text{grad}}$ and $\mathcal{L}_{\text{freq}}$ into the training phase, we force the student to accurately mimic the teacher's structural and spectral variations, significantly reducing both MAE and RMSE. Finally, replacing the standard GD solver with CG provides a more precise and stable optimization, further minimizing the overall error. Note that we opt for a direct one-step distillation rather than standard progressive halving (e.g., $25 \to 12 \to 6$); detailed justifications and empirical results demonstrating their comparable performance are provided in Appendix D.4.

Overall, our two-stage evaluation paradigm demonstrates that blindly increasing sampling steps is computationally inefficient. By anchoring the distillation process to the empirically verified sweet spot (NFE = 25) and applying the enhanced loss alongside CG, our method achieves competitive downscaling accuracy in only 6 inference steps, offering a practical solution for efficient high-resolution meteorological data generation.

5 Conclusion

We introduced GCD, a GCM-consistent diffusion framework for zero-shot downscaling that is trained once on ERA5 and applied to heterogeneous CMIP6 GCMs without model-specific retraining. GCD conditions the diffusion prior on static boundaries (land–sea mask and elevation), which yields more terrain-consistent samples in our qualitative comparison. A spatial-domain filter measurement operator restricts guidance to scales of 5° and above, which reduces the high-frequency spectral mismatch between GCM inputs and the ERA5-trained prior. To lower inference cost, we combine single-stage distillation ($25 \to 6$ steps) with conjugate gradient guidance, which reduced the 6-step error relative to gradient-descent guidance in our ablations. Across five GCMs, the full 1000-step GCD attains lower 99th-percentile errors than all compared baselines. The 6-step version does so on three of the five GCMs while running over $130\times$ faster than vanilla DPS, but it degrades on IPSL. The generated fields also show vortex structures consistent with tropical cyclones, which we assess distributionally rather than event by event. The assumptions and limits of these results are discussed in Section A.

6 Acknowledgements

This work was supported by the Key Program of the National Natural Science Foundation of China (NSFC) (Grant No. 72234002), the National Natural Science Foundation of China (NSFC) (Grant No. 72504061), and the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (No. 2025-GZL-RGZN-BTBX-02031).

References

[1] Bruce C Hewitson and Robert George Crane. Climate downscaling: techniques and application. *Climate Research*, 7(2):85–95, 1996.

[2] Gerald A Meehl, George J Boer, Curt Covey, Mojib Latif, and Ronald J Stouffer. The coupled model intercomparison project (cmip). *Bulletin of the American Meteorological Society*, 81(2): 313–318, 2000.

[3] Robert William Kates, Jesse H Ausubel, and Mimi Berberian. Climate impact assessment: Studies of the interaction of climate and society. 1987.

[4] Haijie Li, Ya Wang, Gang Huang, Weichen Tao, and Pengfei Lin. Generative Downscaling and Bias Correction of Multivariable Earth System Model Simulations. *Geophysical Research Letters*, 52(18):e2025GL117397, September 2025. ISSN 0094-8276, 1944-8007.

[5] Ruian Tie, Xiaohui Zhong, Zhengyu Shi, Hao Li, Bin Chen, Jun Liu, and Wu Libo. Generative artificial intelligence improves projections of climate extremes. *arXiv preprint arXiv:2508.16396*, 2025.

[6] Jonathan Schmidt, Luca Schmidt, Felix M Strnad, Nicole Ludwig, and Philipp Hennig. A generative framework for probabilistic, spatiotemporally coherent downscaling of climate simulation. *npj Climate and Atmospheric Science*, 8(1):270, 2025.

[7] Thomas Vandal, Evan Kodra, Sangram Ganguly, Andrew Michaelis, Ramakrishna Nemani, and Auroop R Ganguly. Deepsd: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 1663–1672, 2017.

[8] Philipp Hess, Michael Aich, Baoxiang Pan, and Niklas Boers. Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning. *Nature Machine Intelligence*, pages 1–11, 2025.

[9] Christopher H Trisos, Cory Merow, and Alex L Pigot. The projected timing of abrupt ecological disruption from climate change. *Nature*, 580(7804):496–501, 2020.

[10] Maximilian Kotz, Anders Levermann, and Leonie Wenz. Retracted article: The economic commitment of climate change. *Nature*, 628(8008):551–557, 2024.

[11] Camilo Mora, Bénédicte Dousset, Iain R Caldwell, Farrah E Powell, Rollan C Geronimo, Coral R Bielecki, Chelsie WW Counsell, Bonnie S Dietrich, Emily T Johnston, Leo V Louis, et al. Global risk of deadly heat. *Nature climate change*, 7(7):501–506, 2017.

[12] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[13] Brian Groenke, Luke Madaus, and Claire Monteleoni. Climalign: Unsupervised statistical downscaling of climate variables via normalizing flows. In *Proceedings of the 10th International Conference on Climate Informatics*, pages 60–66, 2020.

[14] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

[15] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

[16] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

[17] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1219–1229, 2023.

[18] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023.

[19] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *International conference on learning representations*, volume 2024, pages 38922–38949, 2024.

[20] Michael Aich, Andreas Fürst, Florian Sestak, Carlos Ruiz-Gonzalez, Niklas Boers, and Johannes Brandstetter. Wind: Weather inverse diffusion for zero-shot atmospheric modeling. *arXiv preprint arXiv:2602.03924*, 2026.

[21] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[22] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[24] Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5): 1937–1958, 2016.

[25] Olivier Boucher, Jérôme Servonnat, Anna Lea Albright, Olivier Aumont, Yves Balkanski, Vladislav Bastrikov, Slimane Bekki, Rémy Bonnet, Sandrine Bony, Laurent Bopp, et al. Presentation and evaluation of the ipsl-cm6a-lr climate model. *Journal of Advances in Modeling Earth Systems*, 12(7):e2019MS002010, 2020.

[26] Tido Semmler, Sergey Danilov, Paul Gierz, Helge F Goessling, Jan Hegewald, Claudia Hinrichs, Nikolay Koldunov, Narges Khosravi, Longjiang Mu, Thomas Rackow, et al. Simulations for cmip6 with the awi climate model awi-cm-1-1. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002009, 2020.

[27] Hiroaki Tatebe, Tomoo Ogura, Tomoko Nitta, Yoshiki Komuro, Koji Ogochi, Toshihiko Takemura, Kengo Sudo, Miho Sekiguchi, Manabu Abe, Fuyuki Saito, et al. Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in miroc6. *Geoscientific Model Development*, 12(7):2727–2765, 2019.

[28] Thorsten Mauritsen, Jürgen Bader, Tobias Becker, Jörg Behrens, Matthias Bittner, Renate Brokopf, Victor Brovkin, Martin Claussen, Traute Crueger, Monika Esch, et al. Developments in the mpi-m earth system model version 1.2 (mpi-esm1. 2) and its response to increasing co2. *Journal of Advances in Modeling Earth Systems*, 11(4):998–1038, 2019.

[29] Hans Hersbach, Bill Bell, Paul Berrisford, Gionata Biavati, András Horányi, Joaquín Muñoz Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Iryna Rozum, et al. Era5 hourly data on pressure levels from 1940 to present. *Copernicus climate change service (c3s) climate data store (cds)*, 10:24381, 2023.

[30] Andrew W Wood, Edwin P Maurer, Arun Kumar, and Dennis P Lettenmaier. Long-range experimental hydrologic forecasting for the eastern united states. *Journal of Geophysical Research: Atmospheres*, 107(D20):ACL–6, 2002.

[31] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[32] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[33] Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, et al. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6(1):124, 2025.

[34] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023. URL `https://arxiv.org/abs/2303.01469`.

[35] Kwanyoung Kim and Jong Chul Ye. Noise2score: Tweedie's approach to self-supervised image denoising without clean images. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 864–874. Curran Associates, Inc., 2021.

[36] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[37] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[38] Xiang Gao, Meera Sitharam, and Adrian E. Roitberg. Bounds on the jensen gap, and implications for mean-concentrated distributions. *arXiv preprint arXiv:1712.05267*, 2017.

[39] Slavko Simic. On a global upper bound for jensen's inequality. *Journal of Mathematical Analysis and Applications*, 343(1):414–419, 2008. ISSN 0022-247X.

[40] AK Gupta and DK Nagar. *Matrix Variate Distributions*, volume 104. CRC Press, 1999.

A Limitations

GCD has several limitations.

Static boundary conditions. The geographically-constrained prior uses only static boundaries (land–sea mask and elevation). Dynamic physical constraints (e.g., time-varying ocean states) could further improve physical consistency and are left for future work.

Inference cost. The 6-step scheme reduces inference time substantially, but diffusion sampling remains more expensive than deterministic regression: GCD (6 NFE) takes 4.38 s per sample, versus 0.04 s for ESRGAN, 0.25 s for SwinIR, and 2.33 s for CorrDiff. This is a challenge for large ensemble generation. Distillation also adds one-time training (24.5 GPU hours in total).

Linear measurement operators. The CG solver relies on the linearity of $\mathcal{A}$. Nonlinear operators would require other solvers or standard DPS-style gradient guidance.

Scales below 5°. Guidance constrains only scales at or above 5°. Finer structure comes from the ERA5-trained prior and is not constrained by the GCM.

No event-level validation. GCMs are free-running and not temporally aligned with ERA5, so we evaluate 99th-percentile statistics and spectra and show tropical cyclones as qualitative cases. Our results do not establish that individual generated events are correct. Evaluation on nudged or reanalysis-driven simulations is a natural next step.

Degradation on IPSL. GCD (6 NFE) is worse than bilinear interpolation on IPSL (MAE 1.49 versus 1.06). The IPSL spectrum deviates strongly from ERA5 (Fig. 6(b)). We hypothesize that the 6-step student is less tolerant of this shift. The λ and NFE sweeps in Appendix D.8 are consistent with this hypothesis, but we did not directly diagnose the input error by wavenumber band.

Scope of experiments. We evaluate five CMIP6 GCMs in the historical experiment and five variables. The teacher step budget (NFE=25) was selected on MIROC6 only. We have not tested future scenarios, other variables, or other GCMs. Appendix D.9 reports NFE and λ sensitivity on all five GCMs.

B Supplementary Theory and Methods

B.1 Supplementary Theory in Sec. 3.2

The forward process diffuses the data distribution into Gaussian noise, while the reverse process learns to recover the data under the guidance of $\mathbf{C}$. The forward marginal distribution $q(\mathbf{X}_t|\mathbf{X}_0)$ and the conditional reverse transition $p_\theta(\mathbf{X}_{t-1}|\mathbf{X}_t, \mathbf{C})$ are parameterized as

$$\begin{aligned} q(\mathbf{X}_t|\mathbf{X}_0) &= \mathcal{N}(\mathbf{X}_t; \sqrt{\bar{\alpha}_t}\mathbf{X}_0, (1-\bar{\alpha}_t)\mathbf{I}), \\ p_\theta(\mathbf{X}_{t-1}|\mathbf{X}_t, \mathbf{C}) &= \mathcal{N}(\mathbf{X}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{X}_t, t, \mathbf{C}), \beta_t\mathbf{I}), \end{aligned} \tag{9}$$

where β_t represents the pre-defined noise variance schedule, $\alpha_t \triangleq 1-\beta_t$, $\bar{\alpha}_t \triangleq \prod_{s=1}^{t} \alpha_s$. $\boldsymbol{\mu}_\theta$ denotes the reverse process mean function approximator and $\boldsymbol{\mu}_\theta(\mathbf{X}_t, t, \mathbf{C}) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{X}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{X}_t, t, \mathbf{C}))$. The network $\boldsymbol{\epsilon}_\theta$ is trained to approximate the Gaussian noise $\boldsymbol{\epsilon}$ added during the forward process by minimizing the variational lower bound:

$$\mathcal{L} = \mathbb{E}_{\mathbf{X}_0, \boldsymbol{\epsilon}, t, \mathbf{C}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{X}_t, t, \mathbf{C})\|_F^2\right]. \tag{10}$$

Let $\mathbf{X}_0 \sim q(\mathbf{X}_0)$ denote the high-fidelity atmospheric state variables sampled from the reanalysis data distribution. We formulate the prior using a conditional Denoising Diffusion Probabilistic Model (DDPM).

B.2 Inference Algorithm of GCD

From the noisy state $\mathbf{X}_t$ using Tweedie's formula [35, 36], we have the posterior mean:

$$\hat{X}_0(\mathbf{X}_t, \mathbf{C}) := \mathbb{E}[\mathbf{X}_0|\mathbf{X}_t, \mathbf{C}] = \frac{\mathbf{X}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{X}_t, t, \mathbf{C})}{\sqrt{\bar{\alpha}_t}}. \tag{11}$$

However, using only Eq. (11) for climate downscaling will result in inconsistencies with low-resolution observations and other issues. Therefore, posterior sampling based on Bayes' rule is required. We have:

$$\log p(\mathbf{X}_t|\mathbf{Y},\mathbf{C}) \propto \log p(\mathbf{X}_t|\mathbf{C}) + \log p(\mathbf{Y}|\mathbf{X}_t,\mathbf{C}). \tag{12}$$

Taking the gradient with respect to $\mathbf{X}_t$ to obtain the conditional score function:

$$\nabla_{\mathbf{X}_t} \log p(\mathbf{X}_t|\mathbf{Y},\mathbf{C}) = \nabla_{\mathbf{X}_t} \log p(\mathbf{X}_t|\mathbf{C}) + \nabla_{\mathbf{X}_t} \log p(\mathbf{Y}|\mathbf{X}_t,\mathbf{C}). \tag{13}$$

In the framework of diffusion models, the noise-prediction network $\boldsymbol{\epsilon}_\theta(\mathbf{X}_t, t)$ is essentially equivalent to the time-dependent score function $\nabla_{\mathbf{X}_t} \log p(\mathbf{X}_t)$ [37]. Therefore, once $\boldsymbol{\epsilon}_{\theta^*}$ is acquired by minimizing the variational lower bound in Eq. (10), we can use the approximation $\nabla_{\mathbf{X}_t} \log p(\mathbf{X}_t|\mathbf{C}) \simeq -\boldsymbol{\epsilon}_{\theta^*}(\mathbf{X}_t, t, \mathbf{C})/\sqrt{1-\bar{\alpha}_t}$, and the approximation error arises from an optimization error with respect to Eq. (10).

The second term $\nabla_{\mathbf{X}_t} \log p(\mathbf{Y}|\mathbf{X}_t,\mathbf{C})$ in Eq. (13) is intractable, but according to $p(\mathbf{Y}|\mathbf{X}_t,\mathbf{C}) = \mathbb{E}_{\mathbf{X}_0\sim p(\mathbf{X}_0|\mathbf{X}_t,\mathbf{C})}[p(\mathbf{Y}|\mathbf{X}_0)]$ and $\hat{\mathbf{X}}_0 := \mathbb{E}[\mathbf{X}_0|\mathbf{X}_t,\mathbf{C}] = \mathbb{E}_{\mathbf{X}_0\sim p(\mathbf{X}_0|\mathbf{X}_t,\mathbf{C})}[\mathbf{X}_0]$, it can be approximated by conditioning on the estimated clean image $\hat{\mathbf{X}}_0$:

$$p(\mathbf{Y}|\mathbf{X}_t,\mathbf{C}) \simeq p(\mathbf{Y}|\hat{\mathbf{X}}_0(\mathbf{X}_t,\mathbf{C})), \tag{14}$$

implying that the outer expectation of $p(\mathbf{Y}|\mathbf{X}_0)$ over the posterior distribution is replaced with the inner expectation of $\mathbf{X}_0$. Thus, the approximation error in Eq. (14) can be quantified with the Jensen gap [38, 39]. Furthermore, this Jensen's gap has an exact upper bound under our inverse problem subject to Gaussian noise [12].

To explicitly formulate this likelihood for our downscaling task, we consider the common measurement model stated in Eq. (1). In our framework, $\boldsymbol{\Sigma}_W$ is set to $\mathbf{I}$ by assuming longitudinal homoscedasticity, as the longitudinal span of grid cells remains invariant across the domain. Therefore, the log-likelihood is instantiated as follows:

$$\log p(\mathbf{Y}|\hat{\mathbf{X}}_0(\mathbf{X}_t,\mathbf{C})) \simeq -\|\mathbf{Y} - \mathcal{A}_{\text{filter}}(\hat{\mathbf{X}}_0(\mathbf{X}_t,\mathbf{C}))\|^2_{\boldsymbol{\Lambda},\mathbf{I}}, \tag{15}$$

where $\boldsymbol{\Lambda} \triangleq \boldsymbol{\Sigma}_H^{-1}$, $\|\mathbf{X}\|^2_{\boldsymbol{\Lambda},\mathbf{I}} \triangleq \text{Tr}(\boldsymbol{\Lambda}\mathbf{X}\mathbf{X}^\top)$. The formulation of the above is derived from the density function for the Matrix Variate Gaussian (MVG) distribution [40].

Then, we rewrite the log-likelihood in Eq. (15) to quantify the distance between $\mathcal{A}_{\textbf{filter}}(\hat{\mathbf{X}}_0)$ and $\mathbf{Y}$. We also incorporate latitude weighting into this distance calculation:

$$\mathcal{E}(\mathbf{X}_t,\mathbf{Y}) = \|\mathbf{W}(\mathcal{A}_{\textbf{filter}}(\hat{\mathbf{X}}_0(\mathbf{X}_t,\mathbf{C})) - \mathbf{Y})\|^2_F, \tag{16}$$

where $\mathbf{W}^\top\mathbf{W} = \boldsymbol{\Sigma}_H^{-1}$.

Notably, the distance in Eq. (16) differs from the log-likelihood in Eq. (15) by only one negative sign. The weighting matrix $\mathbf{W}$ combines the spherical geometry of the Earth and the measurement noise at different latitudes, with weights expressed as $\mathbf{W}_{ii} = \sqrt{\boldsymbol{\Sigma}^{-1}_{H_{ii}}} \propto \cos(\phi_i)$, where ϕ_i is the latitude of the i-th grid row.

Finally, we use the gradient $\nabla_{\mathbf{X}_t}\mathcal{E}$ to guide the generation process. This ensures that the generated high-resolution data is consistent with the low-resolution input. Specifically, when downsampled back to the coarse scale, the generated output matches the original GCMs projection while preserving fine-scale variability. This approach preserves the fidelity of large-scale structures while augmenting realistic local details. The algorithm of inference refer to Algorithm 1.

B.3 Accelerated Inference of GCD

Algorithm 2 outlines the accelerated inference scheme of the GCD framework, which effectively compresses the generation trajectory by integrating progressive distillation with Conjugate Gradient (CG) guidance. The algorithm requires several specific inputs for the few-step regime: the reduced

Algorithm 1 GCD inference

Require: $T, \mathbf{Y}_{\text{raw}}, \mathbf{C}, \zeta, \mathbf{W}$
Init: $\mathbf{X}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{Y} \leftarrow \mathcal{A}_{\text{filter}}(\mathbf{Y}_{\text{raw}})$
for $t = T - 1$ **to** 0 **do**
  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\boldsymbol{\epsilon} \leftarrow \boldsymbol{\epsilon}_\theta(\mathbf{X}_t, t, \mathbf{C})$
  $\hat{\mathbf{X}}_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{X}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon})$
  $\mathbf{G} \leftarrow \nabla_{\mathbf{X}_t} \|\mathbf{W}(\mathcal{A}_{\text{filter}}(\hat{\mathbf{X}}_0) - \mathbf{Y})\|_F^2$
  $\boldsymbol{\mu}_t \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{X}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)$
  $\mathbf{X}_{t-1} \leftarrow \boldsymbol{\mu}_t + \sqrt{\beta_t}\mathbf{Z} - \zeta\mathbf{G}$
end for
return $\hat{\mathbf{X}}_0$

Algorithm 2 GCD inference acceleration

Require: $K, \mathbf{Y}_{\text{raw}}, \mathbf{C}, \{t_k\}_{k=1}^K, \tau, \lambda, M$
Init: $\mathbf{X}_{t_K} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{Y} \leftarrow \mathcal{A}_{\text{filter}}(\mathbf{Y}_{\text{raw}})$
for $k = K - 1$ **to** 0 **do**
  $\boldsymbol{\epsilon} \leftarrow \boldsymbol{\epsilon}_\psi(\mathbf{X}_{t_k}, t_k, \mathbf{C})$
  $\hat{\mathbf{X}}_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_{t_k}}}\left(\mathbf{X}_{t_k} - \sqrt{1-\bar{\alpha}_{t_k}}\boldsymbol{\epsilon}\right)$
  $\mathbf{X}_{t_{k-1}} \leftarrow \sqrt{\bar{\alpha}_{t_{k-1}}}\hat{\mathbf{X}}_0 + \sqrt{1-\bar{\alpha}_{t_{k-1}}}\boldsymbol{\epsilon}$
  $\mathbf{X}_0^{(0)} \leftarrow \hat{\mathbf{X}}_0$ *# Initialize CG*
  for $m = 1$ **to** M **do**
    $\mathbf{X}_0^{(m)} \leftarrow \text{CG_Step}(\mathbf{X}_0^{(m-1)}, \mathbf{Y}, \lambda, \tau)$
  end for
  $\mathbf{X}_0^\star \leftarrow \mathbf{X}_0^{(M)}$
  $\mathbf{X}_{t_{k-1}} \leftarrow \mathbf{X}_{t_{k-1}} + \sqrt{\bar{\alpha}_{t_{k-1}}}(\mathbf{X}_0^\star - \hat{\mathbf{X}}_0)$
end for
return $\mathbf{X}_{t_0}$

total number of sampling steps K (e.g., $K = 6$), the raw coarse observation $\mathbf{Y}_{\text{raw}}$, the physical conditioning context $\mathbf{C}$, and the discrete time schedule $\{t_k\}_{k=1}^K$ tailored for the distilled model. Furthermore, it takes hyperparameters for the inner CG solver, including the number of inner iterations M, the regularization weight λ, and the solver tolerance τ.

Similar to the standard procedure, the process initializes with a pure noise state $\mathbf{X}_{t_K} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and extracts the smoothed observation $\mathbf{Y} = \mathcal{A}_{\text{filter}}(\mathbf{Y}_{\text{raw}})$. The reverse generative process iterates over the compressed schedule from $k = K - 1$ down to 0. At each step, the distilled student model $\boldsymbol{\epsilon}_\psi$ rapidly predicts the noise component, enabling the estimation of the clean high-resolution state $\hat{\mathbf{X}}_0$ across a large time stride. Based on this uncorrected estimate, a preliminary next state $\mathbf{X}_{t_{k-1}}$ is computed.

However, because aggressive large-step sampling inevitably introduces significant truncation errors and deviates from the data manifold, an inner optimization loop is introduced to improve data consistency. Starting with the initial guess $\mathbf{X}_0^{(0)} = \hat{\mathbf{X}}_0$, the algorithm executes M iterations of the 'CG_Step'. This inner solver efficiently traverses the current data manifold to solve the optimization problem associated with the inverse problem (Eq. (7)). This iterative refinement yields an optimized state $\mathbf{X}_0^\star$ that is consistent with the coarse input. Finally, to correct the diffusion trajectory, the scaled difference between the optimized state and the original estimate, $\sqrt{\bar{\alpha}_{t_{k-1}}}(\mathbf{X}_0^\star - \hat{\mathbf{X}}_0)$, is injected back into the preliminary state $\mathbf{X}_{t_{k-1}}$. This dual-stage mechanism ensures that the ultra-fast prior transitions provided by distillation are precisely calibrated by the CG guidance, resulting in high-fidelity downscaling in minimal steps.

B.4 Supplementary Evaluation Methods

Climate downscaling differs from short-term weather prediction in evaluation method. Although GCMs and ERA5 reanalysis share nominal timestamps in the historical period, their internal weather realizations are not phase-aligned, rendering pointwise time-synchronous comparison statistically inappropriate. Consequently, evaluation focuses on distributional consistency and extreme-state characteristics under comparable climatological conditions. To evaluate the model's ability to reproduce extreme climate states, we analyze the spatial statistics of upper-tail behavior.

We define the MAE and RMSE at the 99^{th} percentile as follows:

$$\text{MAE} = \frac{1}{W\sum_{j=1}^{H} w_j} \sum_{i=1}^{W}\sum_{j=1}^{H} w_j \left|Q_{\text{model},i,j} - Q_{\text{ERA5},i,j}\right|, \tag{17}$$

$$\text{RMSE} = \sqrt{\frac{1}{W\sum_{j=1}^{H} w_j} \sum_{i=1}^{W}\sum_{j=1}^{H} w_j \left(Q_{\text{model},i,j} - Q_{\text{ERA5},i,j}\right)^2}, \tag{18}$$

where $W = 1440$ and $H = 720$ denote the number of grid points in the longitude and latitude directions, respectively. $Q_{\text{model},i,j}$ and $Q_{\text{ERA5},i,j}$ represent the 99th percentile values at grid point (i, j). Crucially, $w_j = \cos(\phi_j)$ represents the latitude weight, which corrects for the area distortion inherent in the spherical geometry of the Earth.

C Supplementary Experiments Details

C.1 Details of Datasets

As shown in Table 2, these models cover a range of resolutions from about 100 km to 250 km. This diversity enables an assessment of whether our method generalizes across different grid resolutions without model-specific retraining.

Table 2: Summary of GCMs used in this study.

MODEL NAME	INSTITUTION	NOM. RES.	GRID SIZE
IPSL-CM6A-LR	IPSL	∼250 KM	144 × 143
AWI-ESM-1-1-LR	AWI	∼250 KM	192 × 96
MPI-ESM1-2-LR	MPI-M	∼250 KM	192 × 96
MIROC6	JAMSTEC	∼140 KM	256 × 128
MPI-ESM1-2-HR	MPI-M	∼100 KM	384 × 192

To provide a clear and challenging evaluation of downscaling performance, we focus on TCs–related atmospheric fields as a representative case. Specifically, we select five variables: Mean Sea Level Pressure (MSL) and 10 m Wind Components (U10M, V10M) to characterize near-surface intensity, and Geopotential Height at 500hPa and 250hPa (Z500, Z250) to represent upper-level circulation patterns.

We select the same variables as in the ERA5 dataset. In CMIP6, these are named 'psl' (MSL), 'uas' and 'vas' (10m wind), and 'zg' (geopotential height). Specifically, we multiply the 'zg' fields at 500hPa and 250hPa by the gravitational acceleration g to convert them into geopotential (Z), ensuring consistency with the ERA5 format.

C.2 Settings Details of Baselines

In our evaluation, the standard supervised baselines (SwinIR, ESRGAN, CorrDiff, CM) are trained on artificially degraded ERA5 data. Since their inputs must strictly align with the spatial characteristics of the training distribution, applying them directly to unseen heterogeneous GCMs constitutes a direct-transfer task. We use this setting to explicitly demonstrate the sub-optimal cross-domain performance caused by domain shifts when model-specific retraining is omitted. For CM+noise, we follow Hess et al. [8] by injecting statistical noise during training to mitigate these domain gaps.

SwinIR. The network was configured with a window size of 8×8, 4 Residual Swin Transformer Blocks (RSTB), 4 Swin Transformer Layers (STL) per block, and 128 channels. Training was conducted using the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with an initial learning rate of 2×10^{-4} and an $\mathcal{L}_1$ pixel loss.

ESRGAN. The generator utilizes 12 Residual-in-Residual Dense Blocks (RRDB) with 64 channels. The model was optimized using Adam with a learning rate of 1×10^{-4}. The total loss function was formulated as a weighted sum of the $\mathcal{L}_1$ pixel loss ($\lambda = 1 \times 10^{-2}$), a VGG-based perceptual loss ($\lambda = 1.0$), and a relativistic adversarial loss ($\lambda = 5 \times 10^{-3}$).

CorrDiff. The architecture comprises two distinct networks: a deterministic regression UNet and a generative diffusion UNet. Both networks share an identical backbone configuration with channel dimensions $(64, 128, 256, 512)$, though they are optimized independently. The regression UNet is first trained using an $\mathcal{L}_2$ loss to predict the smoothed high-resolution mean directly from the coarse input. Subsequently, the diffusion UNet is trained to generate the high-frequency residual, conditioned on the coarse input. The diffusion model was trained with a linear noise schedule over $T = 1000$ steps. During inference, we utilized the DDIM sampler for 12 NFEs to balance computational efficiency and sample quality.

CM and CM+NOISE. The baseline Consistency Model (CM) was trained using the continuous-time formulation with a maximum time boundary of $T = 80$ and an EMA decay rate of 0.999. The optimization objective combined an $\mathcal{L}_2$ pixel loss with an LPIPS perceptual metric. To maximize inference efficiency, sampling was executed using the single-step consistency sampler (1 NFE). For the CM+NOISE variant, we followed the exact same architecture but introduced a data augmentation strategy to simulate domain shifts. Specifically, during training, artificial noise was injected into the

coarse-resolution ERA5 inputs by applying the standard DDPM forward noise scheduler for $t = 50$ steps. Correspondingly, during the inference phase, this identical 50-step DDPM noise degradation was applied to the raw GCM inputs prior to single-step generation.

DDRM. Following the original implementation, all experiments were performed with the default settings of $\eta_B = 1.0$ and $\eta = 0.85$. We leveraged the DDIM sampler for 20 NFEs. The forward measurement operator was implemented using 2D average pooling to construct the singular value decomposition (SVD) matrices efficiently.

Vanilla DPS. We utilized the same backbone without the geographic and temporal conditioning. Inference was performed using the standard DDPM sampler for the full 1000 NFEs. The gradient guidance scale was empirically set to $\zeta = 0.5$, and the measurement operator was defined as standard spatial average pooling without any frequency-domain filtering.

C.3 Settings Details of GCD

The backbone of our pre-trained prior is a 2D U-Net architecture with a four-level encoder–decoder structure and skip connections to preserve multi-scale spatial information. The channel dimensions for the downsampling and upsampling paths are set to $\{64, 128, 256, 512\}$, respectively. Crucially, the bottleneck stage incorporates cross-attention mechanisms to explicitly integrate temporal embeddings, conditioning the spatial features on time information. Each resolution level comprises two stacked residual blocks to enhance feature extraction capabilities.

The model was pre-trained and inferred on a cluster of 8 NVIDIA H200 GPUs (141 GB memory per GPU). We used the AdamW optimizer with an initial learning rate of 10^{-4}. A StepLR scheduler with decays factor $\gamma = 0.9$ was applied every 50 epochs. Training was conducted for 75 epochs. Subsequently, the distillation model was initialized from this pre-trained model and trained for an additional 30 epochs, maintaining the exact same learning rate and scheduler settings. We empirically set the balancing weights to $\lambda_{\text{grad}} = 0.1$ and $\lambda_{\text{freq}} = 0.01$ in distill loss function.

D Supplementary Results

D.1 Supplementary Results in Main Results

This section complements Fig. 4 by further illustrating the downscaling results for pressure-related variables. Since upper-level fields are generally smooth, the visual differences before and after downscaling are not significant for Z250 and Z500 (see Fig. 9 and Fig. 10). In contrast, MSL shows clear improvements in detail over regions like the Tibetan Plateau and the Andes (see Fig. 8). To capture nuances beyond visual inspection, we further analyzed the spectral characteristics (Fig. 11). In the frequency domain, distinct behaviors emerge at small scales: for MSL, GCMs tend to overestimate high-frequency energy, whereas our results align consistently with ERA5. Conversely, for Z250 and Z500, where the small-scale energy in GCMs nearly vanishes, our method effectively recovers these high-frequency components, exhibiting slightly higher energy levels than ERA5. Despite the limited visual change, it is important to note that Z250 and Z500 remain crucial for the accurate localization and identification of TCs.

D.2 Further Ablation Study About Prior

In the main text, we demonstrated that geographical constraints are essential for correcting unphysical artifacts. To isolate their individual contributions, we further evaluate the model conditioned on only the Digital Elevation Model (DEM) or only the Land Sea Mask (LSM).

As shown in Figure 12, both individual constraints successfully reproduce the terrain blocking effect over the Andes (red boxes in WS10M). Naturally, the DEM constraint works by providing explicit topographic heights to guide wind deceleration. Interestingly, the LSM constraint alone achieves a similar blocking effect. Because the Andes mountain range tightly follows the South American coastline, this sharp land-sea boundary acts as an effective spatial proxy for the mountain barrier.

Similarly, for the Mean Sea Level pressure (MSL), both individual constraints independently suppress the spurious high-pressure anomalies observed in the unconstrained baseline. These results confirm

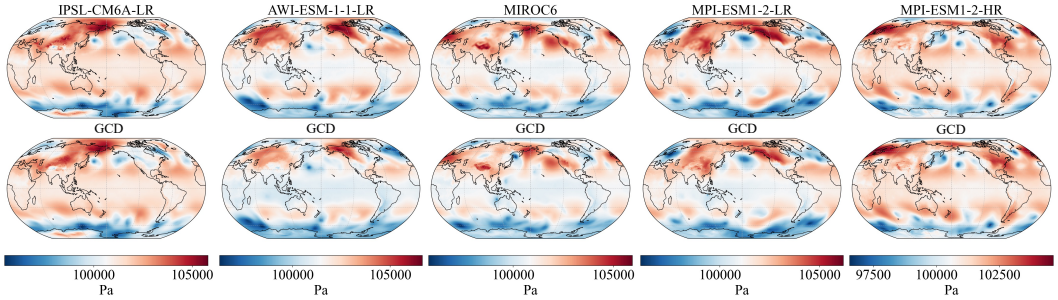


Figure 8: Visualization of MSL fields across different GCMs before and after downscaling. The time period is consistent with Fig. 4. Note that unlike surface pressure, which mirrors rugged topography, MSL is a derived variable designed to filter out orographic effects, resulting in a naturally smoother field at synoptic scales, although major features like the Tibetan Plateau and the Andes remain visually discernible. To quantify the variations that are less apparent, we provide a spectral analysis in Fig. 11 to illustrate the differences in small-scale energy.

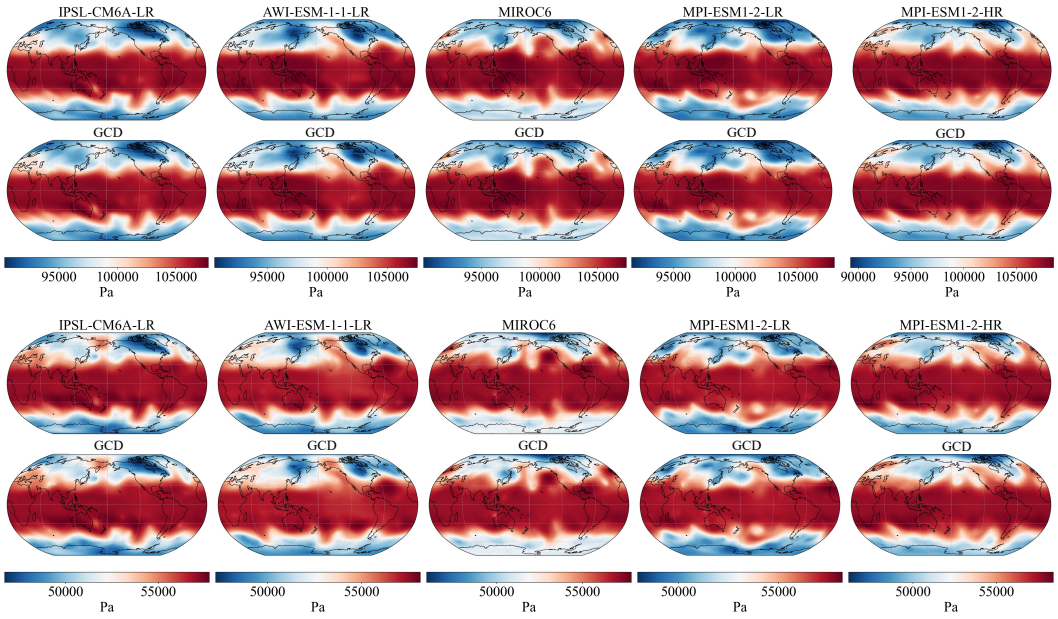


Figure 10: Visualization of Z500 fields across different GCMs before and after downscaling. The time period is consistent with Fig. 4. Since upper-level fields are generally smooth, making differences in small-scale details difficult to distinguish visually, we conducted a spectral analysis in Fig. 11.

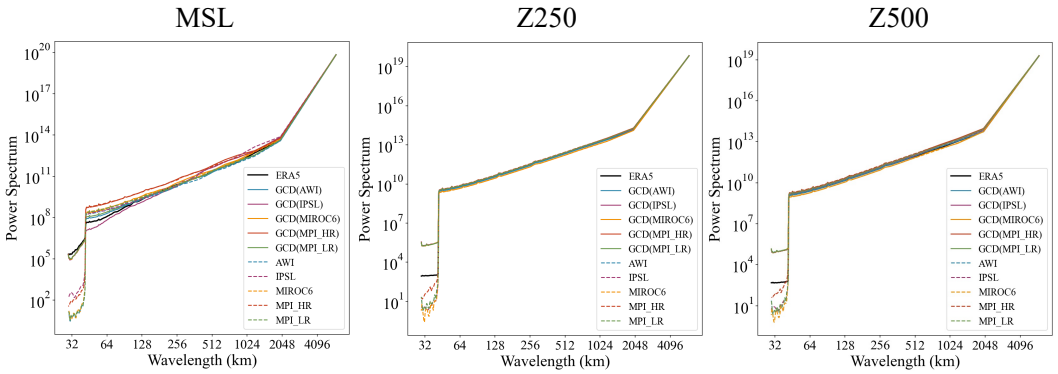


Figure 11: PSD of MSL, Z250, Z500 fields across different GCMs before and after downscaling. The time period is consistent with Fig. 4.

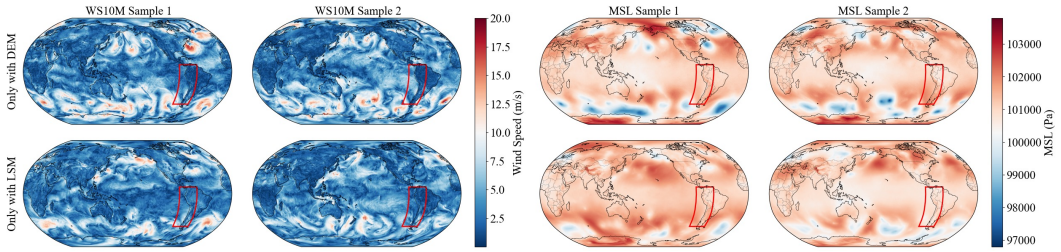


Figure 12: Ablation study of individual geographical constraints. Both the "Only with DEM" and "Only with LSM" configurations successfully capture the terrain blocking effect over the Andes (red boxes in WS10M) and maintain stable MSL patterns.

that while DEM and LSM are individually robust geographical priors, combining them (as in our final model) ensures the most comprehensive physical fidelity.

D.3 Further Ablation Study About Measurement Operator

To investigate the advantages of $\mathcal{A}_{\text{filter}}$ over the vanilla downsampling operator ($\mathcal{A}_{\text{down}}$) in conditional guidance, we conducted an ablation study using 5° ERA5 data as the guidance signal. Under this setting, the core difference between the two operators lies in the presence or absence of the upsampling operator, $\mathcal{A}_{\text{up}}(\cdot)$.

As shown in Fig. 13(a), removing $\mathcal{A}_{\text{up}}(\cdot)$ causes the model to lose effective constraints on large-scale physical fields, resulting in massive spatial biases in key variables such as Z250 and Z500. Conversely, incorporating $\mathcal{A}_{\text{up}}(\cdot)$ significantly stabilizes the spatial performance, nearly eliminating the biases.

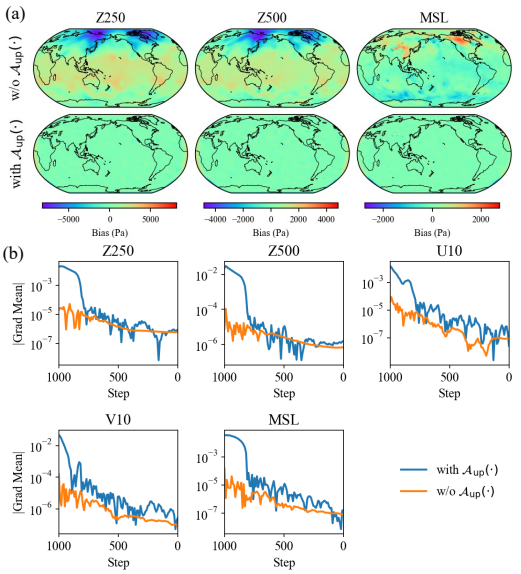


Figure 13: Impact of $\mathcal{A}_{\text{up}}(\cdot)$ in $\mathcal{A}_{\text{filter}}(\cdot)$. (a) Bias distributions. (b) Evolution of gradients during diffusion posterior sampling, illustrating the trajectory from timestep $t = 1000$ to 0 for both methods.

To uncover the underlying mechanism, we tracked the evolution of gradients during the diffusion posterior sampling (DPS) process. As illustrated in Fig. 13(b), during the early stages of denoising (i.e., timestep $t \approx 1000$, when the model constructs macroscopic structures), the operator with $\mathcal{A}_{\text{up}}(\cdot)$ provides a sufficiently large and stable gradient guidance signal. In contrast, the absence of $\mathcal{A}_{\text{up}}(\cdot)$ leads to extremely small and noisy gradients early on, depriving the model of effective conditional constraints during the critical structure-formation phase, which ultimately causes the severe spatial distortion seen in (a). This clearly demonstrates the crucial role of $\mathcal{A}_{\text{filter}}$ in maintaining gradient stability and ensuring physical consistency.

D.4 Further Ablation Study About Distillation

Standard diffusion distillation often halves the step count iteratively (e.g., $25 \to 12 \to 6$). We instead distill directly ($25 \to 6$). To check whether this simplification hurts accuracy, we compare the two trajectories on MIROC6 (test year 2001). To isolate the effect of the trajectory, both students are trained with the same objective, $\mathcal{L}_{\text{MSE}}$ only, and evaluated with CG guidance at $NFE = 6$. The absolute errors are therefore comparable to the $\mathcal{L}_{\text{MSE}}$ + CG row of Fig. 7, not to the full-loss row.

The two trajectories give nearly identical errors ($\Delta\text{MAE} = 0.0018$, $\Delta\text{RMSE} = 0.0002$), and the direct trajectory needs no intermediate model. A plausible explanation is that iterative halving mainly helps when compressing hundreds of steps, whereas our teacher already uses only 25; we

Table 3: 99th-percentile errors on MIROC6 (2001) for two distillation trajectories, both trained with $\mathcal{L}_{\text{MSE}}$ only and evaluated with CG guidance.

Distillation Trajectory	Intermediate Models	MAE	RMSE
$25 \to 12 \to 6$ (iterative halving)	1	1.2523	1.7133
$25 \to 6$ (direct, ours)	0	1.2541	1.7135

Table 4: MAE of the mean field, 50th and 90th percentile. Computed as for the 99th-percentile results in Table 1.

Method	MIROC6	IPSL	AWI	MPI-LR	MPI-HR
Mean field					
Bilinear	0.76	0.48	0.68	0.65	0.59
SwinIR	0.69	0.42	0.63	0.58	0.54
ESRGAN	0.67	0.41	0.61	0.56	0.52
CorrDiff	0.62	0.44	0.57	0.53	0.49
CM+NOISE	0.59	0.39	0.55	0.50	0.46
GCD (1000 NFE)	0.49	0.36	0.51	0.45	0.41
GCD (6 NFE)	0.48	0.68	0.47	0.42	0.44
50th percentile					
Bilinear	0.69	0.41	0.61	0.61	0.54
SwinIR	0.61	0.37	0.55	0.53	0.48
ESRGAN	0.58	0.36	0.54	0.51	0.47
CorrDiff	0.55	0.38	0.49	0.48	0.44
CM+NOISE	0.51	0.34	0.48	0.45	0.41
GCD (1000 NFE)	0.43	0.31	0.44	0.41	0.36
GCD (6 NFE)	0.42	0.59	0.41	0.38	0.39
90th percentile					
Bilinear	1.52	0.88	1.24	1.19	1.08
SwinIR	1.25	0.81	1.12	1.07	0.98
ESRGAN	1.21	0.79	1.09	1.04	0.95
CorrDiff	1.14	0.77	1.04	0.98	0.91
CM+NOISE	1.08	0.72	0.99	0.93	0.85
GCD (1000 NFE)	0.88	0.68	0.92	0.85	0.78
GCD (6 NFE)	0.85	1.21	0.86	0.76	0.82

have not tested this explanation directly. We did not run the halving trajectory with the full loss $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{freq}}$, so the comparison does not cover that setting. We use the direct trajectory because it is simpler and cheaper to train at no measurable accuracy cost in this experiment.

D.5 Full-distribution metrics

While the 99th-percentile MAE effectively captures the model's capability to reconstruct extreme weather events, evaluating the full distribution is essential to ensure that the generative process does not introduce systemic biases into normal climate states. Table 4 presents the Mean Absolute Error (MAE) for the mean field, as well as the 50th and 90th percentiles.

The results reveal the limitations of traditional interpolation methods in modeling tail distributions. For instance, on the MIROC6 dataset, while the Bilinear method achieves a reasonable mean field MAE of 0.76, its absolute error widens significantly to 1.52 at the 90th percentile. In contrast, GCD (1000 NFE) maintains high absolute accuracy across the distribution, with its 90th-percentile error (0.88) remaining highly competitive. Among all baselines, GCD (1000 NFE) has the lowest mean-field, P50 and P90 errors on all five GCMs.

D.6 Probabilistic evaluation

To quantify the probabilistic skill of our generative downscaling framework, we evaluate the Continuous Ranked Probability Score (CRPS). CRPS measures the distance between the predicted cumulative

Table 5: Continuous Ranked Probability Score (CRPS) for 10 ensemble members, averaged over variables. ERA5 reanalysis is used as the ground truth reference. For deterministic methods, CRPS is mathematically equivalent to MAE.

Method	MIROC6	IPSL	AWI	MPI-LR	MPI-HR
Bilinear (= MAE)	0.76	0.48	0.68	0.65	0.59
SwinIR (= MAE)	0.69	0.42	0.63	0.58	0.54
ESRGAN (= MAE)	0.67	0.41	0.61	0.56	0.52
CorrDiff	0.58	0.42	0.54	0.50	0.46
CM+NOISE	0.54	0.37	0.51	0.47	0.43
GCD (1000 NFE)	0.43	0.34	0.46	0.41	0.37
GCD (6 NFE)	0.45	0.66	0.44	0.40	0.41

distribution function (CDF) and the empirical CDF of the observation. For an ensemble of predictions, it is defined as:

$$\text{CRPS} = \frac{1}{N}\sum_{i=1}^{N}|x_i - y| - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|$$

where x_i represents the generated ensemble members, y is the ERA5 ground truth, and $N = 10$ is the ensemble size. For deterministic baselines (e.g., Bilinear, SwinIR), the ensemble size is $N = 1$, and the CRPS mathematically reduces to the standard MAE.

Table 5 shows that the lowest CRPS on each GCM is obtained by one of the two GCD variants: GCD (1000 NFE) on MIROC6, IPSL and MPI-HR, and GCD (6 NFE) on AWI and MPI-LR. The exception is GCD (6 NFE) on IPSL, whose CRPS (0.66) is worse than every baseline (Appendix D.8).

The CRPS is computed on the time-mean field, so the deterministic rows coincide with the mean-field MAE in Table 4; the P99 results in Table 1 average over 5 members, whereas the CRPS uses 10. An ensemble CRPS is by construction no larger than the mean member error, so it favors ensemble methods over deterministic ones. For example, GCD (1000 NFE) on MIROC6 has a mean-field MAE of 0.49 and a CRPS of 0.43. The comparison therefore mixes the effect of a more accurate model with the benefit of ensembling.

D.7 Quantitative spectral error

To quantify the recovery of small-scale structure shown in the tropical cyclone analysis, we compute the Log-Spectral Distance (LSD) between the radially averaged one-dimensional power spectral density (PSD) of the generated fields and that of ERA5.

The filter measurement operator $\mathcal{A}_{\text{filter}}$ applies a 5° low-pass filter, which corresponds to a wavelength of roughly 550 km at the equator, or a global wavenumber of $k \approx 72$. Guidance therefore acts mainly at scales larger than this, and finer scales are supplied largely by the ERA5-trained prior. To evaluate these largely unguided scales, we compute the LSD only over the band $72 \le k \le k_{\max}$, where $k_{\max}$ is the highest resolved wavenumber:

$$\text{LSD} = \sqrt{\frac{1}{|K|}\sum_{k=72}^{k_{\max}}\left(\log_{10} P_{\text{model}}(k) - \log_{10} P_{\text{ERA5}}(k)\right)^2},$$

where $P(k)$ is the unnormalized power spectrum and $|K|$ is the number of wavenumbers in the band. Because the band starts at the filter cutoff, it is not strictly free of guidance: the spatial box filter is not an ideal low-pass filter and leaks some signal beyond $k \approx 72$. Wavelengths are quoted at the equator, and the same k corresponds to shorter physical wavelengths at higher latitudes.

Table 6 reports the results. Bilinear interpolation has the largest distances (2.75–3.22), which reflects the lack of small-scale energy in the interpolated fields. The supervised regression baselines SwinIR and ESRGAN remain far from ERA5 (1.62–1.92). The generative supervised baselines CorrDiff and CM+noise are closer (0.68–1.05), but still clearly worse than GCD. GCD (1000 NFE) obtains the

Table 6: Log-Spectral Distance (LSD) between the radially averaged power spectrum of generated fields and ERA5, computed over the largely unguided high-frequency band ($k \geq 72$, wavelengths $\lesssim 550$ km at the equator). Lower is better.

Method	MIROC6	IPSL	AWI	MPI-LR	MPI-HR
Bilinear	3.15	3.22	2.95	2.88	2.75
SwinIR	1.85	1.92	1.78	1.72	1.65
ESRGAN	1.82	1.88	1.75	1.68	1.62
CorrDiff	0.95	1.05	0.88	0.85	0.82
CM+NOISE	0.82	0.94	0.75	0.72	0.68
GCD (1000 NFE)	0.18	0.24	0.20	0.17	0.15
GCD (6 NFE)	0.26	2.15	0.28	0.24	0.22

Table 7: P99 error on IPSL for different configurations. The 15-step model is obtained by evaluating the continuous-time distilled model with a finer 15-step ODE solver schedule. Time is measured in seconds per sample.

Configuration	MAE	RMSE	Time (s)
Bilinear	1.06	1.78	—
GCD, 6 NFE ($\lambda = 0.1$)	1.49	1.95	4.38
GCD, 6 NFE ($\lambda = 0.5$)	0.98	1.48	4.38
GCD, 6 NFE ($\lambda = 1.0$)	0.92	1.38	4.38
GCD, 15 NFE ($\lambda = 0.1$)	0.89	1.35	10.95
GCD, 1000 NFE	0.87	1.32	589.62

lowest distance on every GCM (0.15–0.24). Since the LSD is computed on a $\log_{10}$ scale, a value of 0.2 corresponds to a typical spectral mismatch of about a factor of 1.6 in power, compared with a factor of roughly 10^3 for bilinear interpolation.

GCD (6 NFE) is close to the 1000-step model on four GCMs (0.22–0.28). On IPSL it reaches 2.15, which is consistent with the degradation of its 99th-percentile error on this GCM (Table 1). The LSD is unsigned, so this value does not show whether the generated IPSL fields contain too much or too little small-scale energy. We therefore do not draw conclusions about the mechanism from this metric alone.

This metric compares globally averaged, radially averaged spectra. It shows that the generated fields have realistic small-scale energy levels, but it does not show that individual small-scale structures are correct, and it is not a substitute for event-level validation (see Limitations).

D.8 Failure analysis on IPSL

As observed in Table 1, the 6-step distilled GCD experiences a noticeable performance degradation on the IPSL dataset (MAE 1.49), performing worse than the Bilinear baseline (1.06). To understand this failure mode, we conduct an ablation study on the inference configurations, as detailed in Table 7.

This degradation likely reflects the severe distribution shift of the raw IPSL fields. Under a compressed 6-step budget with weak regularization ($\lambda = 0.1$), the conjugate gradient solver may over-trust the shifted GCM input.

This issue can be mitigated through two distinct mechanisms. First, increasing the sampling steps to 15 (MAE 0.89) reduces the ODE truncation error, allowing the prior to naturally pull the trajectory back to the physically valid data manifold. Alternatively, keeping the 6-step budget but increasing the regularization weight to $\lambda = 1.0$ (MAE 0.92) anchors the solution more strongly to the prior estimate $\hat{\mathbf{X}}_0$. This is consistent with sensitivity to out-of-distribution inputs under extreme step compression.

D.9 Generalization of the step budget and regularization weight

To justify the default hyperparameters used in our distilled student model, Table 8 presents an ablation on the teacher's step budget (NFE) and the conjugate gradient regularization weight (λ).

Table 8: P99 MAE of the 6-step student for different teacher step budgets and regularization weights λ, on all five GCMs. Defaults are NFE=25 and $\lambda = 0.1$.

Setting	MIROC6	IPSL	AWI	MPI-LR	MPI-HR
Teacher NFE					
15	1.18	1.67	1.25	1.19	1.17
25 (default)	1.05	1.49	1.14	1.07	1.08
50	1.09	1.54	1.17	1.11	1.10
Regularization weight λ					
0.01	1.12	1.73	1.21	1.15	1.16
0.1 (default)	1.05	1.49	1.14	1.07	1.08
1.0	1.22	0.92	1.31	1.26	1.24

Teacher Capacity Bottleneck: We observe a U-shaped performance curve regarding the teacher NFE. While a 15-step teacher introduces excessive truncation errors, increasing the teacher steps to 50 also slightly degrades the student's performance (e.g., MIROC6 MAE increases from 1.05 to 1.09). This may reflect limited capacity of the 6-step student to mimic more complex ODE trajectories, although we did not verify this directly. Therefore, NFE=25 serves as the optimal distillation sweet spot.

The Regularization Trade-off: The parameter λ controls the weight placed on the prior estimate $\hat{\mathbf{X}}_0$ relative to the GCM measurement. A strong prior constraint ($\lambda = 1.0$) successfully rescues the model on the heavily shifted IPSL dataset (MAE improves from 1.49 to 0.92) by rejecting corrupted input signals. However, for well-behaved GCMs (e.g., AWI, MIROC6), over-trusting the prior causes the model to ignore valid extreme-weather signals present in the GCM inputs, leading to higher P99 errors (e.g., AWI MAE increases to 1.31). We use $\lambda = 0.1$ as the global default, since it gives the lowest error on four of the five GCMs. NFE=25 and $\lambda = 0.1$ were chosen on MIROC6 using 2001, which lies within the test period. Table 8 reports sensitivity on all five GCMs and was not used for per-GCM tuning.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims regarding the our framework—including the geographically-constrained prior, the filter measurement operator, and the 6-step acceleration scheme—are clearly stated in Section 1 and fully supported by the methodology in Section 3 and the comprehensive empirical results across five GCMs in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses its limitations in Appendix A, noting that the geographically-constrained prior currently relies only on static boundary conditions (e.g., topography), and that despite the 6-step acceleration, the inherent computational overhead of diffusion models remains higher than deterministic regression, posing challenges for massive ensemble generations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper focuses on methodological innovations and empirical evaluations for climate downscaling. While it formulates the problem using established diffusion models and inverse problem frameworks (detailed in Section 3 and Appendix B), it does not introduce new theoretical theorems or require mathematical proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides comprehensive details for reproducibility, including dataset specifics (ERA5 and CMIP6), baseline configurations, network architectures, and training hyperparameters, which are extensively documented in Section 4.1 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The datasets used in this study (ERA5 and CMIP6) are publicly available open-access datasets. The source code for the proposed GCD framework is currently withheld during the review process but will be fully open-sourced with comprehensive instructions upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All necessary experimental details, including chronological data splits (train/val/test years), optimizer types (e.g., AdamW), and specific hyperparameters for both the proposed method and baselines, are explicitly detailed in Section 4.1 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The statistical variability of our experiments is reported in key figures (e.g., Figures 6 and 7), capturing the variance across 5 ensemble members. For tabular results, we report the ensemble averages evaluated over an extensive 9-year test period to ensure statistical robustness, as detailed in Section 4.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper explicitly details the compute resources in Appendix C.3, specifying the use of a cluster of 8 NVIDIA H200 GPUs (141 GB memory per GPU). Additionally, Table 1 in Section 4.2 provides a comprehensive breakdown of training costs (GPU hours) and inference times for all evaluated methods.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research strictly conforms to the NeurIPS Code of Ethics. It utilizes publicly available geophysical and climate datasets for environmental science purposes and does not involve human subjects, sensitive personal data, or applications with foreseeable malicious use.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This research focuses on foundational algorithmic improvements for meteorological downscaling. While it inherently offers positive societal benefits by aiding climate impact assessments and extreme weather analysis (as noted in the Introduction), it does not pose direct negative societal risks (e.g., privacy, fairness, or security concerns) that typically necessitate a dedicated broader impact discussion.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The datasets utilized in this study, including ERA5 and CMIP6, are publicly available for scientific research under their respective open-use licenses (e.g., Copernicus open data license for ERA5). We have properly cited all data sources in the main text and detailed the data acquisition process in the Appendix.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The datasets utilized in our experiments (ERA5 and CMIP6) consist exclusively of geophysical and meteorological measurements (e.g., temperature, precipitation). Consequently, they do not contain any personally identifiable information (PII), human subjects data, or offensive content. Furthermore, this work does not involve collecting any new primary data.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This research relies exclusively on objective meteorological and climate datasets (ERA5 and CMIP6). It does not involve human subjects, crowdsourcing, or any behavioral experiments, making Institutional Review Board (IRB) approval or participant consent inapplicable.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This research focuses entirely on geophysical and climate data (ERA5 and CMIP6). It does not involve any crowdsourcing experiments, data annotation by workers, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: As this paper does not involve any research with human subjects, Institutional Review Board (IRB) approvals or equivalent ethical reviews for human research are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research (our framework for climate downscaling) is based entirely on diffusion models and does not involve Large Language Models (LLMs) as any component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.