GENERATE THE FOREST BEFORE THE TREES - A HIERARCHICAL DIFFUSION MODEL FOR CLIMATE DOWNSCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Downscaling is essential for generating the high-resolution climate data needed for local planning, but traditional methods remain computationally demanding. Recent years have seen impressive results from AI downscaling models, particularly diffusion models, which have attracted attention due to their ability to generate ensembles and overcome the smoothing problem common in other AI methods. However, these models typically remain computationally intensive. We introduce a Hierarchical Diffusion Downscaling (HDD) model, which introduces an easily-extensible hierarchical sampling process to the diffusion framework. A coarse-to-fine hierarchy is imposed via a simple downsampling scheme. HDD achieves competitive accuracy on the ERA5 reanalysis dataset and CMIP5 models, significantly reducing computational load by running on up to half as many pixels with competitive results. Additionally, a single model trained at 0.25° resolution transfers seamlessly across multiple CMIP5 models with much coarser resolution. HDD thus offers a lightweight alternative for probabilistic climate downscaling, facilitating affordable large-ensemble high-resolution climate projections; with a single model that can be applied across GCMs of varying input sizes. See a full code implementation at: https://github.com/HDD/HDD-Hierarchical-Diffusion-Downscaling.

1 INTRODUCTION

High resolution earth system data is critical for understanding and mitigating the impacts of anthropogenic climate change; however, generating it with traditional methodologies on a large scale is computationally prohibitive (Curran et al., 2024b) (Rampal et al., 2024). For example, general circulation models (GCMs) exhibit at least $\mathcal{O}(n^4)$ complexity with respect to resolution of the climate model due to the processing of variables in four dimensions (Balaji et al., 2022) (201, 2012) (Curran et al., 2024b). As a result, important historical and future climate datasets are often unavailable for use in local planning or only a limited subset of models is used, limiting the coverage of potential future climate outcomes(Riahi et al., 2017) (Rampal et al., 2024) (Curran et al., 2024b). In particular, for Australia, only two of the five Shared Socioeconomic Pathways (SSPs) from the IPCC's 6th Assessment Report have been downscaled to high resolutions, leaving large parts of the possible future scenario space unrepresented (Riahi et al., 2017).

In recent years, earth system modelling has undergone vast improvements owing to the proliferation of AI and advancements in computer science (Bi et al., 2022)(Lam et al., 2022)(Kochkov et al., 2024)(Curran et al., 2024a). Results from machine learning models in various earth-science tasks[1] are often competitive or superior on several metrics, but at a fraction of the inference cost (Hobeichi et al., 2023) (Bi et al., 2022)(Curran et al., 2024a) (de Burgh-Day and Leeuwenburg, 2023) . Downscaling, particularly, has seen promising advancements (Mardani et al., 2023), yet AI approaches are rarely compared against results from existing dynamical models and established climate metrics, which limits the trust that climate scientists can place in them.

[1]Weather prediction, Downscaling, Climate Emulation and many more

Additionally, recent advances in computer vision have led to significant progress in auto-regressive image generation, with models such as VAR and GPT-4o demonstrating state-of-the-art performance (Tian et al., 2024) (Chen et al., 2025). These models exhibit favorable scaling properties and excel in generating high-fidelity images by modeling pixel or patch sequences auto-regressively. Despite their success, such approaches remain largely unexplored in weather and climate image generation, primarily due to the computational complexity and the challenges of capturing spatiotemporal consistency and physical realism required in geoscientific applications.

A small but growing subset of the literature has examined the concept of dimension destruction in diffusion models (Jin et al., 2024) (Zhang et al., 2022) (Campbell et al., 2023). In this approach, in addition to corrupting the training images with Gaussian noise, the dimension or size of the image is destroyed gradually. We present an easily extensible addition to the noise process which can be incorporated in existing diffusion models. By encouraging the model to learn at multiple resolutions, we construct a hierarchical schedule that downscales autoregressively from coarse to fine. We show that this framework can be easily applied to most existing diffusion model setups with some minor adjustments[2]. We also show that these models produce results competitive with, and in some cases surpassing, traditional dynamically downscaled models that nest Regional Climate Models (RCMs) within a GCM. These results are achieved over the Australian domain at a fraction of the inference and training cost.

1. We propose HDD (Hierarchical Diffusion Downscaling), a model that learns multi-scale representations via a hierarchical diffusion process. It applies dimension destruction with noise injection to enable robust feature learning across resolutions, followed by a coarse-to-fine reverse generation during inference. HDD is trained on varying spatial shapes to enforce scale consistency and improve high-resolution reconstruction.
2. Our proposed methodology is architecture-agnostic, allowing integration with any existing diffusion model by augmenting it with shape-conditioning resulting in resolution-aware conditioning mechanism. The framework supports plug-and-play usage within standard diffusion pipelines, making it broadly applicable to weather and climate models without architectural re-design.
3. We train HDD on ERA5 over the Australian domain and provide a comprehensive evaluation benchmark with climate metrics. The model passes all evaluation metrics and is competitive with other AI downscaling models and traditional dynamical RCMs, while requiring significantly less computational cost.

2 Related Work

Diffusion models implicitly generate images in a coarse-to-fine manner (Dielman, 2024) (Rissanen et al., 2022). This is consistent with the way humans process images, where coarse features are recognised first and finer details later; i.e. we 'see' the forest before the trees (Ho et al., 2020; Oliva and Torralba, 2006; Navon, 1977; Kauffmann et al., 2014).

Many atmospheric variables also exhibit these same power laws Willeit et al. (2014), analogous to the $1/f$ fractal spectra of natural images (van der Schaaf and van Hateren, 1996) (Hyvärinen et al., 2009). This indicates variability across scales, with dominant energy at larger scales but a continuous scale-invariant distribution down to finer scales. For example, the Nastrom-Gage spectrum, derived from global aircraft data, demonstrates a robust kinetic energy scaling from approximately 3000 km to several kilometers, following k^{-3} at larger scales transitioning to $k^{-5/3}$ at smaller scales (Gage and Nastrom, 1986).

Regional analyses, including high-resolution reanalyses (e.g., ERA5 at 0.25°) and radar observations, also confirm this 1/f-like spectral behavior in atmospheric variables, such as precipitation intensity. These observations reveal continuous cascades from large storm systems down to small-scale showers without distinct breaks in scaling, consistent with findings from convection-permitting model forecasts in the U.S. (Gkioulekas and Tung, 2006).

The presence of self-similar, power-law spectra strongly motivates the application of coarse-to-fine multi-scale methods in atmospheric modeling and downscaling. Techniques like RainFARM exploit

[2]See methodology in section 3 for more details

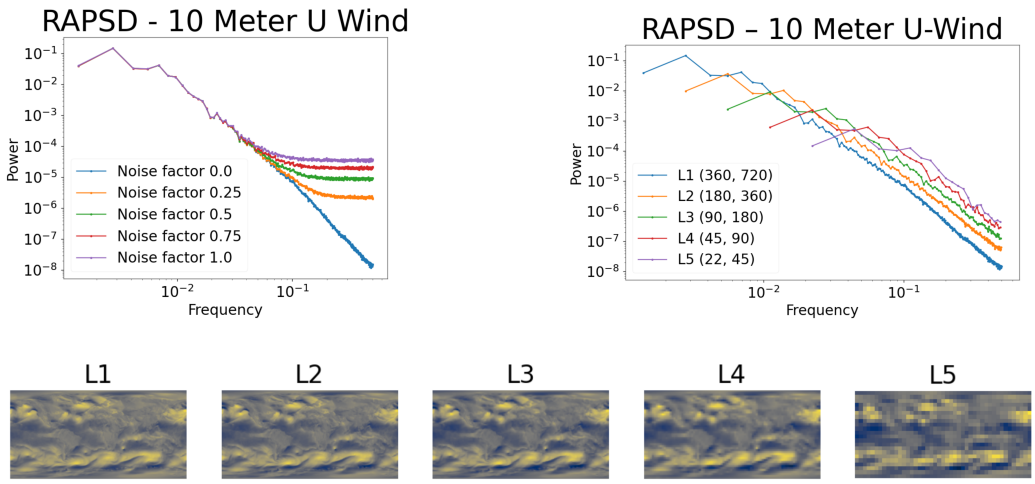

Figure 1: Radially Averaged Power Spectral Density (RAPSD) of the 10 metre U wind component. Left: High frequency finer features are the first to be corrupted by random noise. Conversely, the reverse diffusion process generates in a coarse-to-fine manner, contributing to the strong image generation capabilities of diffusion models. Enforcing this coarse-to-fine relationship explicitly can further improve results. Right: Weather data exhibits a clear power law. Coarser scales are much less information dense and closer to random noise. It is therefore, easier to model the coarser scales first, and progressively add higher frequency details. See appendix B for further details. Bottom: Progressively coarser wind data over grids of size 0.5 °(L1), 1.0 °(L2), 2.0 °(L3), 4.0 °(L4), 8 °(L5)

these scaling laws, generating fine-scale structures consistent with prescribed spectral properties, thereby reinforcing the conceptual and practical alignment between diffusion models in machine learning and traditional fractal-based atmospheric downscaling approaches (Rebora et al., 2006) (D'Onofrio et al., 2014).

3 Methodology

We impose an **explicit coarse-to-fine hierarchy** on the outputs of a noise-conditioned diffusion model by coupling the usual Gaussian noising process with a progressive *down-sampling / up-sampling* schedule that is sampled at every training step per the figure 4.

3.1 Baseline Elucidated Diffusion Model (EDM) formulation

We adopt a baseline implementation of EDM from (Watt and Mansfield, 2024) following Karras et al. (2022), which formalises and consolidates terminology from different diffusion methodologies. We refer the reader to the full paper (Karras et al., 2022) for further details. The basic setup is as follows:

Let $x_0 \in \mathbb{R}^{H\times W\times C}$ be a clean image (H: height, W: width, C: channels), and let $\{\sigma_t\}_{t=1}^{T}$ be a monotonically *increasing* noise schedule with $\sigma_0 = 0$ where $t = 1, ..., T$ represents the denoising timestep.

1. a *forward* (noising) kernel

$$q\big(x_t \mid x_{t-1}\big) = \mathcal{N}\big(x_t\,;\, x_{t-1}, \sigma_t^2 I\big), \qquad t = 1, \ldots, T,$$

2. and a learnable *reverse* kernel

$$p_\theta\big(x_{t-1} \mid x_t, \sigma_t\big) = \mathcal{N}\big(x_{t-1}\,;\, \mu_\theta(x_t, \sigma_t),\, \sigma_t^2 I\big),$$

where μ_θ is the output of a score network trained to minimise the EDM loss.

3.2 ADDING SPATIAL HIERARCHIES

Denote by $\mathbf{s}_t = (h_t, w_t)$ the *target resolution* at step t, with $(h_0, w_0) = (H, W)$ and $(h_T, w_T) \approx (1, 1)$. For each resolution we define

$$D_{\mathbf{s}_t} : \mathbb{R}^{h_{t-1} \times w_{t-1} \times C} \longrightarrow \mathbb{R}^{h_t \times w_t \times C}, \qquad U_{\mathbf{s}_t} : \mathbb{R}^{h_t \times w_t \times C} \longrightarrow \mathbb{R}^{h_{t-1} \times w_{t-1} \times C},$$

as bilinear *down-sampling* and matching *up-sampling* operators, respectively.

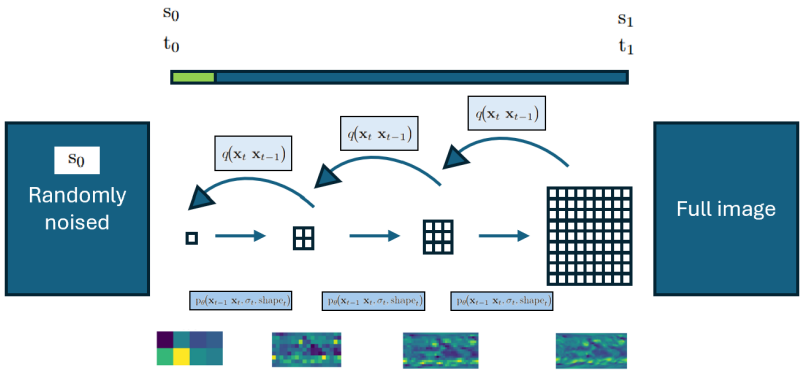

Figure 2: During training, the destruction kernel $q()$ progressively adds noise to the image and gradually destroys its dimensions according to the shape of the scheduler. At inference, the UNET $p()$ learns the inverse of this process, producing probabilistic downscaled outputs consistent with the original distribution. Note that although a UNET is used for $p()$ here, any function approximator could be applied. This hierarchical process is intuitive for image data, and in the climate setting we also show that competitive results can be achieved while processing up to half the pixels at inference.

Hierarchical reverse process. The network is trained to *simultaneously denoise and un-coarsen*. Conditioning on both σ_t and the shape $\mathbf{s}_t$ we write

$$q\big(x_t \mid x_{t-1}\big) = \mathcal{N}\big(x_t\,;\, x_{t-1}; \mathbf{s}_t, \sigma_t^2 I\big) \tag{1}$$

$$p_\theta\big(x_{t-1} \mid x_t, \sigma_t, \mathbf{s}_t\big) = \mathcal{N}\big(x_{t-1}\,;\, \mu_\theta(x_t, \sigma_t, \mathbf{s}_t),\, \sigma_t^2 I\big), \tag{2}$$

Embedding every latent back to full resolution with $\tilde{x}_t = U_{\mathbf{s}_t}(x_t)$, the hierarchical EDM loss (HDD) is

$$\mathcal{L}_{\text{HDD}}(\theta) = \mathbb{E}_{t,\, x_0,\, \epsilon}\Big[w(\sigma_t)\, \big\|\epsilon - f_\theta\big(\tilde{x}_t + \sigma_t \epsilon,\, \sigma_t,\, \mathbf{s}_t\big)\big\|_2^2\Big],$$

where f_θ is the score network and $w(\sigma_t)$ is the usual EDM weighting term.

Because each (h_t, w_t) is sampled once per example, the model learns a coarse-to-fine mapping with no additional passes through the data. Equations 1–2 reduce to the standard EDM when $D_{\mathbf{s}_t}$ and $U_{\mathbf{s}_t}$ are the identity[3], ensuring drop-in compatibility with existing code. The formulation aligns with the visual narrative in : blue arrows depict the noise and dimension-destruction kernels $q(x_t \,|\, x_{t-1})$, while straight arrows illustrate the learned reverse kernels $p_\theta(x_{t-1} \,|\, x_t, \sigma_t, \mathbf{s}_t)$.

The algorithm in Table1 outlines the sampling procedure for training and inference. Steps highlighted in blue are unique to the hierarchical model, which enforces a coarse-to-fine generation while exactly mirroring the usual practice of sampling the noise level σ_t uniformly in log-space. Hence, the network encounters the full spectrum of noise levels *and* spatial resolutions while seeing each training sample only once.

[3]In this context, the identity would be if a shape schedule of the full final resolution is used throughout the whole schedule process: $s_t = (x_T, y_T)$ where $t = 1...T$ for each intermediate step

Algorithm 1 Training (Hierarchical Forward Process)

Require: dataset $q(x_0)$, noise schedule $\{\sigma_t\}_{t=1}^{T}$, shape schedule $\{(h_t, w_t)\}_{t=1}^{T}$, network f_θ

repeat
  $x_0 \sim q(x_0)$
  $t \sim \text{Uniform}(\{1, \dots, T\})$
  $\epsilon \sim \mathcal{N}(0, I)$
  $x_t \leftarrow D_t(x_0)$ ▷ down-sample to (h_t, w_t)
  $z \leftarrow \sqrt{\bar{\alpha}_t}\, x_t + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$
  gradient step on $\nabla_\theta \left\| \epsilon - f_\theta\big(U_t(z),\, t,\, (h_t, w_t)\big) \right\|_2^2$
until converged

Algorithm 2 Sampling (Hierarchical Reverse Process)

Require: f_θ, noise $\{\sigma_t\}_{t=1}^{T}$, shape schedule $\{(h_t, w_t)\}_{t=1}^{T}$

$x_T \sim \mathcal{N}(0, I_{h_T \times w_T})$
for $t = T, \dots, 1$ **do**
  $\tilde{x}_t \leftarrow U_t(x_t)$ ▷ upsample to full res
  $\epsilon_t \leftarrow f_\theta\big(\tilde{x}_t, t, (h_t, w_t)\big)$
  $x_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\big(\tilde{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_t\big)$
  if $t > 1$ **then**
    $z \sim \mathcal{N}(0, I)$
    $x_{t-1} \leftarrow x_{t-1} + \sigma_t z$
  $x_{t-1} \leftarrow D_{t-1}(x_{t-1})$ ▷ project to next latent size
return x_0

Hence, the model first learns to minimise the coarse EDM loss and then progresively unlocks finer scales. We justify this two-phase optimisation strategy mathematically in Appendix I.

4 What is the Theoretical Speedup?

We seek to define the theoretical improvement in processing speed for different shape schedules at training and inference. Note that as a standard UNET is being used as the underlying architecture, processing time scales linearly with the number of pixels; this would not be the case for an attention-based architecture like ViT, where processing time scales quadratically.

Note that this section represents the theoretical upper bound for performance improvement as this varies slightly with the size of image, the number timesteps T and ignores any overhead operations.

Let a clean image be $x_0 \in \mathbb{R}^{H \times W \times C}$ with full area $A = H\,W$. During training we draw a *single* noise–shape index $t \sim \text{Uniform}\{1, \dots, T\}$ per minibatch and replace the Gaussian–only corruption of EDM with the composite operator $x_{t-1} \mapsto D_{\mathbf{s}_t}(x_{t-1} + \sigma_t \varepsilon)$ that *downsamples first, denoises later*. Write $A_t = h_t w_t$ for the area at step t and define the *normalised mean area*

$$\alpha \;=\; \frac{1}{T\,A} \sum_{t=1}^{T} A_t, \qquad \alpha \in (0, 1]. \tag{1}$$

α is the fraction of pixels—relative to baseline EDM—consumed *on average* by one network call; its reciprocal is therefore the ideal pixel/FLOP speed-up:

$$S_{\text{train}} \;=\; S_{\text{infer}} \;=\; \frac{1}{\alpha}. \tag{2}$$

We refer the reader to appendix H for the full proof but note that the hierarchical shape schedules have the following speedups:

Shape scheduler	α	Speed-up $S = 1/\alpha$
Linear shrink $(h_t, w_t) \propto 1 - \frac{t-1}{T-1}$	$\frac{1}{3}$	$3\times$
Unit-shrink $\big(h_t, w_t\big) = \big(H - (t-1),\, W - (t-1)\big)$	see Eq. (3)	$\approx 1.32\times$ (50 steps)

All schedules are *drop-in*: when $D_{\mathbf{s}_t} = U_{\mathbf{s}_t} = I$ they revert to vanilla EDM. Eq. equation 2 therefore gives an upper-bound on pixel, FLOP and memory savings obtainable with the HDD framework. We note that this is a higher speed up than comparable image-based approaches due to the choice of shape scheduler (Zhang et al., 2022).

5 EXPERIMENTS

5.1 ERA5 EXPERIMENT

We trained a model on 30 years of ERA5 reanalysis data across temperature, u/v component of wind and precipitation from 1990 to 2019 over the Australian domain[4]. The task was to downscale the resolution from 1.5° to 0.25°. This was trained for 144 hours on two A100 GPUs for 360 epochs. Training time/resources were mimicked for Earth-ViT and the base EDM and all models achieved convergence. We then evaluate each of these models on the same task for five years of ERA5 data from January 2020 to December 2024 [5]. Note that Earth-ViT is based on the popular weather forecasting model panguweather with several slight modifications for the downscaling setting. See (Curran et al., 2024a) for further information on Earth-ViT and Appendix E for information on the training procedure for this setting.

Table 1: Performance of models trained on the ERA5 downscaling task over Australia, evaluated using Root Mean Squared Error (RMSE), Peak Signal-to-Noise Ratio (PSNR) and Continuous Ranked Probability Score (CRPS) . Best performance for each metric shown in bold. The 'Base EDM' model corresponds to the backbone for popular downscaling model Corrdiff from NVIDIA (Mardani et al., 2023; Karras et al., 2022). The inclusion of different shape schedules is further examined in the ablation analysis in Section 5.2

Model	**RMSE**	**PSNR**	**CRPS**
Bilinear Interpolation	0.362755	9.54	–
Earth-ViT (Finetuned PanguWeather (Curran et al., 2024a))	0.317410	10.63	–
Base EDM - 500 steps	0.000143	31.45	0.0002319
HDD **(Ours)**	**0.000128**	**33.34**	**0.0002308**

5.2 ABLATION RESULTS - HIERARCHICAL SCHEDULER

We now take the trained HDD model and examine the effect of different shape schedules to assess at what frequency dimensions should be increased versus just denoising. We take the best model on the ERA5 task (HDD - Hierarchical - 500 steps - 3 denoise steps per shape step) and reapply it with different shape schedules. As expected, there is a tradeoff between the total intermediate number of pixels processed at inference and RMSE. Interestingly, even the most extreme model, with equally spaced dimension jumps between (1,1) and (144,272), achieves very similar results to the base case. This suggests that passing additional shape information to the model enables it to capture coarse details similarly to a standard diffusion model, but at a fraction of the inference cost.

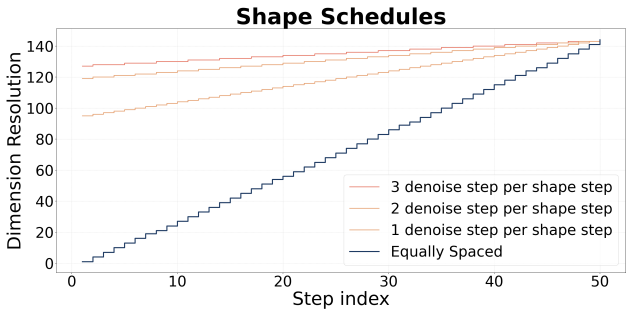

Figure 3: The dimensions of the downscaled image as it evolves with the number of steps. Note that in the extreme case where we equally our dimension jumps over the steps, we only process one third as many pixels at inference - see section four for further breakdown on this.

Interestingly, enforcing a slight coarse-to-fine generation with minimal overall pixel reduction appears to yield the best results. However, Different shape schedules show that increasing the number of denoising steps per shape (visualised in 3 as the flattening slope) leads to an initial improvement before leveling out. The final ablation '500 denoise steps per shape step' is equivalent to having no

[4]Over a latitude/longitude bounding box of (-7.75,109) to (-43.5,176.75)

[5]See appendix E for further information on sampling procedures and ablations on metrics

Table 2: Same task as in section 5.1 but with differing shape schedules ablated

Model	RMSE	PSNR
Equally Spaced Steps	0.000131	33.262
Move in tandem – 1 denoise step per shape step	0.000129	**33.40**
2 denoise steps per shape step	0.000129	33.36
3 denoise steps per shape step **(Ours)**	**0.000128**	33.34
50 denoise steps per shape step	0.000131	33.21
500 denoise steps per shape step	0.000133	33.11

dimension-reducing steps, yet still performs much better than the base model in section 5.1. We hypothesise that by taking advantage of the hierachical nature of weather/climate data, the model is able to more efficiently model the coarser features in earlier diffusion steps. Additionally, the score function is able to more easily approximate the earlier coarsened steps, allowing the model to discretise the generation problem further. We expand on this reasoning and discuss several other hypotheses for this improved performance in Appendix A / B / D and Section 2.

5.3 GCM Application and comparison to RCM simulations

We then applied the models trained in section 5.1 to precipitation simulations from multiple GCMs. GCM data is available at coarse resolution of approximately 1.5°, though finer and coarser resolutions also exist depending on the model.[6] grids due to the aforementioned computational constraints with generating this data[7]. We downscaled historical daily precipitation simulations from multiple GCMs: MIROC 5, CNRM-CM5, HadGEM and GFDL-ESM2M to 0.5° resolution over Australia and evaluated against precipitation observations from the Australian Gridded Climate Dataset (AGCD) (Jones, 1999).

The evaluation employs a set of minimum standard metrics focused on fundamental rainfall characteristics: total precipitation, spatial distribution, and seasonal cycle. We compute four metrics benchmarked against observations, with acceptance thresholds defined following (Isphording et al., 2024). The four metrics are detailed in appendix D alongside further details regarding this experiment..

Performance scores are displayed in table 4. Results show that both HDD and baseline EDM models achieve a pass across all four precipitation evaluation benchmarks. In comparison 20 out of 24 RCM simulations meet the benchmark criteria (Isphording et al., 2024).

Table 3: Relevant climate metrics SCorr, MAPE from (Isphording et al., 2024) and Computational Efficiency for MIROC5–driven downscaling runs

Model	**NRMSE**	**MAD**	**SCorr**	**MAPE**	**kgCO_2 Emitted**
CCAM-1704	0.78	0.97	0.87	0.38	$\sim$1032kg
Earth-ViT	0.33	1.30	0.82	0.44	$\sim$51kg Training + $\sim$2kg inference
Base EDM	0.54	1.30	0.81	0.55	$\sim$105kg Training + $\sim$5kg inference
HDD	0.45	1.09	0.85	1.18	$\sim$50kg Training + $\sim$2kg[8] inference

5.4 Resolution-agnostic capabilities of HDD

A practical requirement for downscaling is that a single model should transfer across General Circulation Models (GCMs) with *different native grid spacings*. We therefore evaluate HDD zero-shot across four CMIP5 GCMs spanning input resolutions from $\sim 1.4°$ to $2.5° \times 2.0°$ and compare against the same backbone without hierarchies (EDM). Results below collate the precipitation metrics from (Isphording et al., 2024) (NRMSE, MAD in months, MAPE, spatial correlation).

[6] This equates to grids of approximately 167km x 167km

[7] See appendix D for further information

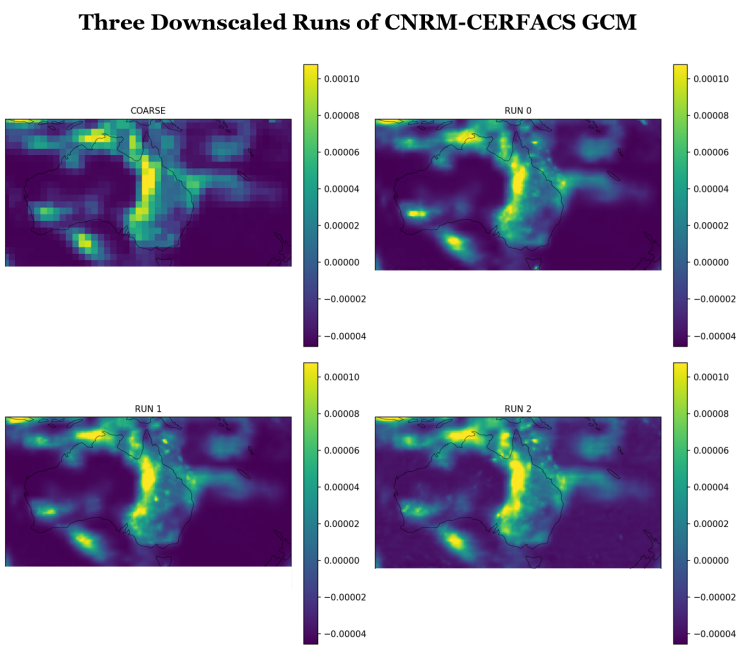

Figure 4: Model applied over the Australian extent of the CNRM-CERFACS General Circulation model

Table 4: Performance across MIROC5, CNRM-CM5, HadGEM, and GFDL-ESM2M. Best values for the downscaled benchmarks within each GCM group are highlighted in bold and second best are italicised.

GCM	Resolution	Model	NRMSE	MAD	MAPE	SpCor
MIROC5	1.41°×1.41°	CCAM-2008	0.62	**0.85**	0.49	**0.88**
	1.41°×1.41°	CCAM-1704	0.78	*0.97*	0.38	*0.87*
	1.41°×1.41°	Earth-ViT	**0.33**	1.30	**0.44**	0.82
	1.41°×1.41°	EDM	0.54	1.30	*0.55*	0.81
	1.41°×1.41°	HDD	*0.45*	1.09	1.18	0.85
CNRM-CM5	1.406°×1.401°	CCAM-1704	0.68	**0.87**	0.54	**0.82**
	1.406°×1.401°	Earth-ViT	**0.50**	*1.07*	*0.30*	*0.79*
	1.406°×1.401°	EDM	*0.56*	1.27	**0.25**	0.78
	1.406°×1.401°	HDD	0.62	1.39	0.57	*0.79*
HadGEM	1.875°×1.25°	CCAM-1704	*0.78*	**0.95**	**0.27**	**0.89**
	1.875°×1.25°	EDM	0.94	1.90	*0.54*	0.76
	1.875°×1.25°	HDD	**0.49**	*1.26*	1.14	*0.85*
GFDL-ESM2M	2.5°×2.0°	CCAM-2008	**0.49**	**0.92**	*0.31*	**0.92**
	2.5°×2.0°	CCAM-1704	0.69	**0.92**	*0.31*	*0.90*
	2.5°×2.0°	Earth-ViT	0.61	1.15	2.51	-0.13
	2.5°×2.0°	EDM	0.73	1.35	0.35	0.50
	2.5°×2.0°	HDD	*0.51*	*0.96*	**0.27**	0.71

Analysis. **(i) Spatial structure holds up as inputs get coarser.** HDD consistently *improves* spatial correlation over EDM across all GCMs (+1.3 to +20.3%), with the largest gain on the coarsest input (GFDL-ESM2M, $+20.3\%$ SpCor), indicating that hierarchical shape-conditioning helps reconstruct coherent fine-scale patterns even when the upstream grid is very coarse (Table 5).

(ii) Error tradeoffs depend on the GCM and native resolution. Relative to EDM, HDD often lowers distributional errors (MAPE: -114% to -128% on MIROC5/CNRM; -111% on HadGEM3),

Table 5: Paired comparison (HDD−EDM): percentage change by GCM/input resolution. All values are taken as a percentage with improvement with positive values being better across all metrics.

GCM (input resolution)	ΔNRMSE (%)	ΔMAD (%)	ΔMAPE (%)	ΔSpCor (%)
MIROC5 ($1.41°\times1.41°$)	$+16.7$	$+16.2$	-114.5	$+4.9$
CNRM-CM5 ($1.406°\times1.401°$)	-10.7	-9.5	-128.0	$+1.3$
HadGEM3 ($1.875°\times1.25°$)	$+47.9$	$+33.7$	-111.1	$+11.8$
GFDL-ESM2M ($2.5°\times2.0°$)	$+30.1$	$+28.9$	$+22.9$	$+20.3$

but can raise NRMSE/MAD for some models (e.g., HadGEM3 NRMSE $+47.9\%$). This suggests HDD locks in where/when rainfall occurs (timing/structure) more reliably than exact amplitude for certain GCMs, which aligns with its coarse-to-fine inductive bias..

(iii) Resolution-agnostic generalisation emerges from the hierarchy. HDD was trained at a $1.5°$ input resolution but transfers to GFDL-ESM2M at $2.5°\times2.0°$ with markedly higher SpCor than EDM (+20.3%) and competitive errors (Table 4). We hypothesise that the randomised shape schedules during training have taught the network to interpolate missing intermediate scales at inference, effectively acting as a learned multi-resolution prior; this makes it particularly suited to the downscaling problem where GCMs come in varying resolutions. We expand on this reasoning and discuss several other hypotheses for this improved performance in appendix A / B.

(iv) Positioning vs. non-probabilistic baselines. Where available, Earth-ViT attains strong NRMSE on some GCMs (e.g., MIROC5), but (a) lacks native ensemble capability and (b) sometimes degrades spatial skill on very coarse inputs (GFDL, negative SpCor), whereas HDD keeps positive spatial skill and remains probabilistic for ensemble generation. This complements Section 5.1 where HDD matched or exceeded EDM at lower pixel budgets.

Takeaways. HDD preserves spatial coherence better than EDM as inputs get coarser; amplitude/timing errors show GCM-dependent tradeoffs that can be tuned by the shape schedule; the hierarchical conditioning confers practical resolution-agnostic behaviour, enabling a *single* trained model to service multiple GCMs of varying native grid size without re-architecture; and (iv) HDD retains the probabilistic benefits of diffusion (ensembles, uncertainty) while operating on fewer pixels (cf. Section 4).

6 Conclusion

Overall, HDD demonstrates that a coarse-to-fine ladder inside a diffusion model reduces the pixel budget - and therefore the FLOPs and CO_2 - by roughly two-thirds, while preserving or even improving performance when tested on ERA5 and historical simulations from CMIP5 GCMs. Regional climate fields are produced at a fraction of the computational and monetary cost of both dynamical downscaling (RCM simulations) and standard diffusion models, with performance comparable to dynamical downscaling. The method is architecture-agnostic and applicable to any diffusion framework, although retraining is needed to incorporate two additional scalar inputs[9]. HDD can also be extended to a multi-model ensemble, enabling the generation of large ensembles of regional climate projections and providing a pathway to assess the full spectrum of possible future climate outcomes, which is currently underrepresented. Although we note that the model is not tested on future GCM data, this could be an interesting next direction for this work. [10] Finally, HDD and other AI-based downscaling methods can viewed as an alternative that complements dynamically downscaled results, expanding the range of feasible ensemble sizes and scenario coverage while reducing computational cost.

[9] One scalar for each of the (h_t, w_t) shapes in s_t

[10] Lack of ground truth makes this difficult in addition to shifting climate distribution

REFERENCES

A National Strategy for Advancing Climate Modeling. National Academies Press, December 2012. ISBN 9780309259774. doi: 10.17226/13430. URL http://dx.doi.org/10.17226/13430.

V. Balaji, Fleur Couvreux, Julie Deshayes, Jacques Gautrais, Frédéric Hourdin, and Catherine Rio. Are general circulation models obsolete? *Proceedings of the National Academy of Sciences*, 119(47), November 2022. ISSN 1091-6490. doi: 10.1073/pnas.2202075119. URL http://dx.doi.org/10.1073/pnas.2202075119.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast, 2022. URL https://arxiv.org/abs/2211.02556.

Andrew Campbell, William Harvey, Christian Weilbach, Valentin De Bortoli, Tom Rainforth, and Arnaud Doucet. Trans-dimensional generative modeling via jump diffusion models, 2023. URL https://arxiv.org/abs/2305.16261.

Sixiang Chen, Jinbin Bai, Zhuoran Zhao, Tian Ye, Qingyu Shi, Donghao Zhou, Wenhao Chai, Xin Lin, Jianzong Wu, Chao Tang, Shilin Xu, Tao Zhang, Haobo Yuan, Yikang Zhou, Wei Chow, Linfeng Li, Xiangtai Li, Lei Zhu, and Lu Qi. An empirical study of gpt-4o image generation capabilities, 2025. URL https://arxiv.org/abs/2504.05979.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, April 2005. ISBN 9780471748823. doi: 10.1002/047174882x. URL http://dx.doi.org/10.1002/047174882X.

Declan Curran, Hira Saleem, Sanaa Hobeichi, and Flora Salim. Resolution-agnostic transformer-based climate downscaling, 2024a. URL https://arxiv.org/abs/2411.14774.

Declan Curran, Hira Saleem, and Flora Salim. Identifying high resolution benchmark data needs and novel data-driven methodologies for climate downscaling, 2024b. URL https://arxiv.org/abs/2405.20346.

Catherine O. de Burgh-Day and Tennessee Leeuwenburg. Machine learning for numerical weather and climate modelling: a review. *Geoscientific Model Development*, 16(22):6433–6477, November 2023. ISSN 1991-9603. doi: 10.5194/gmd-16-6433-2023. URL http://dx.doi.org/10.5194/gmd-16-6433-2023.

Sandar Dielman. Diffusion is spectral autoregression. https://sander.ai/2024/09/02/spectral-autoregression.html, 2024.

D. D'Onofrio, E. Palazzi, J. von Hardenberg, A. Provenzale, and S. Calmanti. Stochastic rainfall downscaling of climate models. *Journal of Hydrometeorology*, 15(2):830–843, April 2014. ISSN 1525-7541. doi: 10.1175/jhm-d-13-096.1. URL http://dx.doi.org/10.1175/JHM-D-13-096.1.

K. S. Gage and G. D. Nastrom. Theoretical interpretation of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft during gasp. *Journal of the Atmospheric Sciences*, 43(7):729–740, April 1986. ISSN 1520-0469. doi: 10.1175/1520-0469(1986)043<0729:tioaws>2.0.co;2. URL http://dx.doi.org/10.1175/1520-0469(1986)043<0729:TIOAWS>2.0.CO;2.

Eleftherios Gkioulekas and Ka-Kit Tung. Recent developments in understanding two-dimensional turbulence and the nastrom–gage spectrum. *Journal of Low Temperature Physics*, 145(1–4):25–57, December 2006. ISSN 1573-7357. doi: 10.1007/s10909-006-9239-z. URL http://dx.doi.org/10.1007/s10909-006-9239-z.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Sanaa Hobeichi, Nidhi Nishant, Yawen Shao, Gab Abramowitz, Andy Pitman, Steve Sherwood, Craig Bishop, and Samuel Green. Using machine learning to cut the cost of dynamical downscaling. *Earth's Future*, 11(3), March 2023. ISSN 2328-4277. doi: 10.1029/2022ef003291. URL http://dx.doi.org/10.1029/2022EF003291.

Aapo Hyvärinen, Jarmo Hurri, and Patrik O. Hoyer. *Natural Image Statistics*. Springer London, 2009. ISBN 9781848824911. doi: 10.1007/978-1-84882-491-1. URL http://dx.doi.org/10.1007/978-1-84882-491-1.

Rachael N. Isphording, Lisa V. Alexander, Margot Bador, Donna Green, Jason P. Evans, and Scott Wales. A standardized benchmarking framework to assess downscaled precipitation simulations. *Journal of Climate*, 37(4):1089–1110, February 2024. ISSN 1520-0442. doi: 10.1175/jcli-d-23-0317.1. URL http://dx.doi.org/10.1175/JCLI-D-23-0317.1.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling, 2024. URL https://arxiv.org/abs/2410.05954.

Philip W. Jones. First- and second-order conservative remapping schemes for grids in spherical coordinates. *Monthly Weather Review*, 127(9):2204–2210, September 1999. ISSN 1520-0493. doi: 10.1175/1520-0493(1999)127<2204:fasocr>2.0.co;2. URL http://dx.doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Louise Kauffmann, Stephen RamanoÃ«l, and Carole Peyrin. The neural bases of spatial frequency processing during scene perception. *Frontiers in Integrative Neuroscience*, 8, May 2014. ISSN 1662-5145. doi: 10.3389/fnint.2014.00037. URL http://dx.doi.org/10.3389/fnint.2014.00037.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Berlin Heidelberg, 1992. ISBN 9783662126165. doi: 10.1007/978-3-662-12616-5. URL http://dx.doi.org/10.1007/978-3-662-12616-5.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, July 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07744-y. URL http://dx.doi.org/10.1038/s41586-024-07744-y.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729694. URL http://dx.doi.org/10.1214/AOMS/1177729694.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Graphcast: Learning skillful medium-range global weather forecasting, 2022. URL https://arxiv.org/abs/2212.12794.

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling, 2023. URL https://arxiv.org/abs/2309.15214.

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, July 1977. ISSN 0010-0285. doi: 10.1016/0010-0285(77)90012-3. URL http://dx.doi.org/10.1016/0010-0285(77)90012-3.

Aude Oliva and Antonio Torralba. *Chapter 2 Building the gist of a scene: the role of global image features in recognition*, page 23–36. Elsevier, 2006. doi: 10.1016/s0079-6123(06)55002-2. URL http://dx.doi.org/10.1016/S0079-6123(06)55002-2.

Neelesh Rampal, Sanaa Hobeichi, Peter B. Gibson, Jorge Baño-Medina, Gab Abramowitz, Tom Beucler, Jose González-Abad, William Chapman, Paula Harder, and José Manuel Gutiérrez. Enhancing regional climate downscaling through advances in machine learning. *Artificial Intelligence for the Earth Systems*, 3(2), April 2024. ISSN 2769-7525. doi: 10.1175/aies-d-23-0066.1. URL http://dx.doi.org/10.1175/AIES-D-23-0066.1.

Nicola Rebora, Luca Ferraris, Jost von Hardenberg, and Antonello Provenzale. Rainfarm: Rainfall downscaling by a filtered autoregressive model. *Journal of Hydrometeorology*, 7(4):724–738, August 2006. ISSN 1525-755X. doi: 10.1175/jhm517.1. URL http://dx.doi.org/10.1175/JHM517.1.

Keywan Riahi, Detlef P. van Vuuren, Elmar Kriegler, Jae Edmonds, Brian C. O'Neill, Shinichiro Fujimori, Nico Bauer, Katherine Calvin, Rob Dellink, Oliver Fricko, Wolfgang Lutz, Alexander Popp, Jesus Crespo Cuaresma, Samir KC, Marian Leimbach, Leiwen Jiang, Tom Kram, Shilpa Rao, Johannes Emmerling, Kristie Ebi, Tomoko Hasegawa, Petr Havlik, Florian Humpenöder, Lara Aleluia Da Silva, Steve Smith, Elke Stehfest, Valentina Bosetti, Jiyong Eom, David Gernaat, Toshihiko Masui, Joeri Rogelj, Jessica Strefler, Laurent Drouet, Volker Krey, Gunnar Luderer, Mathijs Harmsen, Kiyoshi Takahashi, Lavinia Baumstark, Jonathan C. Doelman, Mikiko Kainuma, Zbigniew Klimont, Giacomo Marangoni, Hermann Lotze-Campen, Michael Obersteiner, Andrzej Tabeau, and Massimo Tavoni. The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, 42: 153–168, January 2017. ISSN 0959-3780. doi: 10.1016/j.gloenvcha.2016.05.009. URL http://dx.doi.org/10.1016/j.gloenvcha.2016.05.009.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation, 2022. URL https://arxiv.org/abs/2206.13397.

H.L. Royden and P. Fitzpatrick. *Real Analysis*. Prentice Hall, 2010. ISBN 9780135113554. URL https://books.google.com.au/books?id=H65bQgAACAAJ.

Walter Rudin. *Real and complex analysis, 3rd ed.* McGraw-Hill, Inc., USA, 1987. ISBN 0070542341.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2020. URL https://arxiv.org/abs/2011.13456.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. 2024.

A. van der Schaaf and J.H. van Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759–2770, September 1996. ISSN 0042-6989. doi: 10.1016/0042-6989(96)00002-8. URL http://dx.doi.org/10.1016/0042-6989(96)00002-8.

Zhong Yi Wan, Ricardo Baptista, Yi-fan Chen, John Anderson, Anudhyan Boral, Fei Sha, and Leonardo Zepeda-Núñez. Debias coarsely, sample conditionally: Statistical downscaling through optimal transport and probabilistic diffusion models, 2023. URL https://arxiv.org/abs/2305.15618.

Robbie A. Watt and Laura A. Mansfield. Generative diffusion-based downscaling for climate, 2024. URL https://arxiv.org/abs/2404.17752.

Matteo Willeit, Roberta Amorati, Andrea Montani, Valentina Pavan, and Maria Stefania Tesini. Comparison of spectral characteristics of precipitation from radar estimates and cosmo-model predicted fields. *Meteorology and Atmospheric Physics*, 127(2):191–203, November 2014. ISSN 1436-5065. doi: 10.1007/s00703-014-0359-8. URL http://dx.doi.org/10.1007/s00703-014-0359-8.

Han Zhang, Ruili Feng, Zhantao Yang, Lianghua Huang, Yu Liu, Yifei Zhang, Yujun Shen, Deli Zhao, Jingren Zhou, and Fan Cheng. Dimensionality-varying diffusion process, 2022. URL https://arxiv.org/abs/2211.16032.

A RADIALLY-AVERAGED POWER-SPECTRAL DENSITY (RAPSD)

Figures 1 and 2 in the main text visualise how power is redistributed when an ERA5 field is either (i) coarsened by repeated $2\times$ down-sampling or (ii) contaminated with Gaussian noise of increasing variance. This appendix formalises the metrics that underlie those plots and derives the theoretical curves that explain their shapes.

A.1 FROM A 2-D FIELD TO A 1-D SPECTRUM

Let the clean field be a real-valued array $x \in \mathbb{R}^{N_y \times N_x}$, indexed by pixel coordinates $\mathbf{r} = (n, m)$. Its discrete Fourier transform (DFT) is

$$X(\mathbf{k}) \;=\; \mathcal{F}\{x\}(\mathbf{k}) \;=\; \sum_{n=0}^{N_y-1} \sum_{m=0}^{N_x-1} x_{n,m}\, e^{-2\pi i \left(k_y n/N_y + k_x m/N_x\right)}, \quad \mathbf{k} = (k_x, k_y). \tag{3}$$

The *power-spectral density* (PSD) is the squared magnitude $P(\mathbf{k}) = |X(\mathbf{k})|^2$. For many geophysical or photographic images the statistics are approximately isotropic, so it is convenient to collapse the 2-D PSD into a 1-D function of radius $f = \|\mathbf{k}\|$ (spatial frequency):

$$\mathrm{RAPSD}(f_j) \;=\; \frac{1}{|\mathcal{A}_j|} \sum_{\mathbf{k} \in \mathcal{A}_j} P(\mathbf{k}), \quad \mathcal{A}_j = \big\{\mathbf{k} :\; f_j \le \|\mathbf{k}\| < f_{j+1}\big\}, \tag{4}$$

where the annuli $\{\mathcal{A}_j\}_{j=0}^{J-1}$ partition the Fourier plane into logarithmically-spaced bins ($f_{j+1}/f_j = \mathrm{const}$). Log–log plots of $f \mapsto \mathrm{RAPSD}(f)$ often reveal the empirical power law $\mathrm{RAPSD}(f) \propto f^{-\alpha}$ with $\alpha \approx 2$ for natural images and for many meteorological fields (e.g. the Nastrom–Gage kinetic-energy spectrum).

A.2 EFFECT OF 2 X SPATIAL DOWN-SAMPLING

Down-sampling by an integer factor s reduces the Nyquist frequency from $f_{\max} = 1/2$ (in pixel units) to $f_{\max}/s$. Assuming ideal low-pass pre-filtering, all energy above $f_{\max}/s$ is discarded; below that limit the PSD is merely scaled by s^2 to conserve total variance. For five binary-decade reductions $s \in \{2, 4, 8, 16, 32\}$ (as in **Fig. 1**) one expects

$$\mathrm{RAPSD}_s(f) \;=\; \begin{cases} s^2\, \mathrm{RAPSD}_{\text{orig}}(f), & f < \frac{1}{2s}, \\ 0, & f \ge \frac{1}{2s}. \end{cases}$$

Hence the curves in Fig. 1 coincide at low f and peel off successively at their (progressively smaller) Nyquist cut-offs—exactly the trend observed.

A.3 EFFECT OF ADDITIVE WHITE NOISE

Let $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ be i.i.d. pixel noise. Because the DFT is linear and because white noise is spectrally flat,

$$Y = X + \varepsilon, \tag{5}$$

$$\mathbb{E}\big[P_Y(\mathbf{k})\big] = P_X(\mathbf{k}) + \sigma_n^2, \tag{6}$$

so every RAPSD curve is translated upward by the *same* constant σ_n^2. Define the *hinge frequency* $f^\star(\sigma_n) = \min\{f : \mathrm{RAPSD}_X(f) \le \sigma_n^2\}$. For $f < f^\star$ the spectrum is signal-dominated and remains unchanged; for $f > f^\star$ the spectrum is noise-dominated and becomes flat at the level σ_n^2. Because the notebook sets $\sigma_n = \text{noise\_factor} \times \sigma_X$, the plateau height scales quadratically with the chosen `noise_factor`. The five coloured curves of **Fig. 2** therefore realise

$$\mathrm{RAPSD}_{\text{noise}}(f \mid \lambda) \;=\; \mathrm{RAPSD}_X(f) \;+\; \lambda^2\, \sigma_X^2, \quad \lambda \in \{0, 0.25, 0.5, 0.75, 1.0\}.$$

Each ten-fold rise in λ^2 shifts the horizontal tail up by the same factor and pushes $f^\star$ leftward, reducing the bandwidth in which the underlying flow field is recoverable.

A.4 Relevance for coarse-to-fine diffusion scheduling

In a Gaussian diffusion process the state at noise level σ_t is $x_t = \alpha_t\, x_0 + \sigma_t\, \varepsilon$ (with $\varepsilon \sim \mathcal{N}(0,1)$). Because RAPSD adds linearly, low frequencies re-emerge first as σ_t decays, while high-frequency detail appears only when $\sigma_t^2 \lesssim \mathrm{RAPSD}_X(f)$. Thus the reverse-time sampler automatically follows a coarse-to-fine trajectory in spectral space. The patterns in Figures 1 and 2 therefore provide the mathematical justification for conditioning a diffusion model either on a sequence of coarsened resolutions or on a carefully curated noise schedule when generating meteorological fields.

B Coarse-to-Fine Diffusion Approximates the Score Function on Simpler Distributions

Recall that in standard diffusion models, the *score function* at time t is

$$\nabla_x \log p_t(x),$$

which describes the gradient of the log-density for a progressively noisier version of x_0. Learning the reverse process is equivalent to learning to denoise (or equivalently approximate this score) at various noise levels.

Coarse Distributions as Simpler Targets. When the image is downsampled to coarser resolutions (fewer pixels, fewer degrees of freedom), the induced data distribution

$$p_{\text{coarse}}(x) \;=\; \text{distribution of downsampled images}$$

is generally "simpler" to model. Intuitively, high-frequency details are removed, so spatial correlations are more tractable, and the manifold of coarse-resolution images is lower-dimensional. Consequently, predicting the score

$$\nabla_x \log p_{\text{coarse}}(x)$$

becomes easier: there are fewer fine-grained features to learn, and the model focuses on broad, low-frequency structure.

Hierarchical vs. Single-Scale Approach. By first learning to approximate the score at a coarse distribution,

$$s_t^{(\text{coarse})}(x) \;\approx\; \nabla_x \log p_{\text{coarse},t}(x),$$

the model handles an easier inverse problem. Then, as the resolution is gradually increased, each subsequent diffusion (and corresponding score function) refines the result:

$$s_t^{(\text{finer})}(x) \;\approx\; \nabla_x \log p_{\text{finer},t}(x).$$

Each finer scale deals with distributions that are "closer" to the full-resolution distribution but still easier than jumping directly from pure noise to a full-resolution image in one step.

Connection to Score-Based Diffusion. In the continuous SDE view, we can think of downsampling as reducing the dimensionality or bandwidth of the data, so at time t, the *score* $\nabla_x \log p_t(x)$ lives on a simpler manifold. Discretizing this idea across multiple resolutions amounts to learning a sequence of denoising (or score) functions:

$$f_\theta\big(\mathcal{U}(x_t),\, t\big) \;\approx\; \nabla_x \log p_{\text{downsampled},t}(x),$$

where $\mathcal{U}(x_t)$ is an upsampled version of x_t. Such a hierarchical approach effectively breaks a complex score-estimation problem into stages where each stage handles a simpler, coarser distribution; with each increase in resolution progressively conditioning on the previous result. The autoregressive coarse to fine nature of this generation is intuitive–as previously discussed, humans understand images in a coarse to fine manner with

Summary. Hence, this method approximates the score function on these coarser distributions—where the image space is significantly reduced in size and complexity—before progressively shifting to higher resolutions. This *coarse-to-fine* strategy stabilizes training and provides a more direct way for the network to focus first on large-scale structure and then on finer details, thus approximating the score function in a stepwise manner from simpler (coarse) to more complex (full-resolution) distributions.

C WHY INCREASING THE NUMBER OF DIMENSION-DESTRUCTION STEPS IMPROVES RESULTS

C.1 DISCRETIZING A CONTINUOUS DIFFUSION PROCESS IN TIME

One perspective introduced in (Song et al., 2020), is to view diffusion models as discretized solutions to a *continuous-time* Stochastic Differential Equation (SDE). For standard DDPM (Ho et al., 2020) (without dimension destruction), the forward (noising) process in continuous time can be written as:

$$d\mathbf{x} = f(\mathbf{x}, t)\, dt \;+\; g(t)\, d\mathbf{w}, \tag{7}$$

where $\mathbf{w}$ is a standard Wiener process (Brownian motion) and $t \in [0, 1]$. Conceptually, $f(\mathbf{x}, t)$ and $g(t)$ define the drift and diffusion coefficients that gradually corrupt data into noise.

Reverse-Time SDE. By reversing the time variable from $t = 1$ down to $t = 0$, one obtains the *reverse* SDE:

$$d\mathbf{x} \;=\; \Big[f(\mathbf{x}, t) - g(t)^2\, \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\Big]\, dt \;+\; g(t)\, d\overline{\mathbf{w}}, \tag{8}$$

where $p_t(\mathbf{x})$ is the (instantaneous) distribution of $\mathbf{x}$ at time t, and $\overline{\mathbf{w}}$ is a Brownian motion in reverse time. This reverse SDE perfectly "denoises" the corrupted data back to a clean sample as t goes from 1 down to 0.

Discrete Approximation via Euler–Maruyama. In practice, we discretize the interval $[0, 1]$ into N steps, $t_1 < t_2 < \cdots < t_N$, and approximate the update with an explicit numerical scheme (e.g., Euler–Maruyama):

$$\mathbf{x}_{t_{k-1}} \;\approx\; \mathbf{x}_{t_k} \;+\; \big[f(\mathbf{x}_{t_k}, t_k) - g(t_k)^2\, \nabla_{\mathbf{x}} \log p_{t_k}(\mathbf{x}_{t_k})\big]\, \Delta t \;+\; g(t_k)\, \sqrt{\Delta t}\, \boldsymbol{\eta}_k,$$

where $\Delta t = t_{k-1} - t_k$ is small, and $\boldsymbol{\eta}_k \sim \mathcal{N}(0, \mathbf{I})$. The fewer steps N we use, the *larger* each Δt—and hence the larger the local approximation error at each step.

Key Point (Time Discretization). As $N \to \infty$, $\Delta t \to 0$, and the discrete chain converges to the exact solution of the SDE. Therefore, *more steps* $\Longrightarrow$ *smaller local errors* $\Longrightarrow$ *better final reconstruction*. Empirically, one observes that generating samples with more reverse steps (e.g., 1000 steps vs. 50) yields sharper images, because smaller increments in each denoising step incur fewer approximation artifacts (Ho et al., 2020).

C.2 ACCUMULATION OF LOCAL ERRORS

Each discrete reverse step $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ can introduce some mismatch (e.g. KL divergence) compared to the true $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$. If a single step has a small per-step error ε, over N steps the total discrepancy might accumulate on the order of $O(N \cdot \varepsilon)$. However, when we *increase* N, the per-step error ε often *decreases* because each denoising increment is smaller.

This can be made rigorous by considering the continuum limit $N \to \infty$, where $\Delta t = 1/N$. Classical SDE analysis (see (Kloeden and Platen, 1992)) shows that under certain regularity conditions, the global approximation error converges to zero as $\Delta t \to 0$. Thus,

$$\lim_{N \to \infty} p_\theta(\mathbf{x}_0) \;=\; q(\mathbf{x}_0),$$

meaning the learned model recovers the true data distribution in the idealized infinite-step regime (assuming perfect training).

C.3 DIMENSION-DESTRUCTION VIEWPOINT (COARSE-TO-FINE)

In the hierarchical shape-conditioning setting, we introduce a *second axis* of discretization: not only do we add noise at each step, but we also *downsample* (i.e. reduce the spatial resolution). Let us denote the forward dimension-destruction step at time index k as

$$x_k \;=\; \mathcal{D}_k\Big(x_{k-1} + \epsilon_k\Big), \quad \epsilon_k \sim \mathcal{N}(0, \sigma_k^2 \mathbf{I}), \tag{9}$$

where $\mathcal{D}_k$ maps $(h_{k-1} \times w_{k-1})$ pixels to $(h_k \times w_k)$ pixels, typically $h_k < h_{k-1}$ and $w_k < w_{k-1}$. The reverse process then *upsamples* to full resolution before denoising. In effect, we are discretizing both (a) the *time* variable **and** (b) the *spatial dimension*.

Finer Discretisation in Dimension. When dimension changes are large (e.g. $64 \times 64 \rightarrow 4 \times 4$ in a single step), the model loses significant high-frequency content all at once, and the reverse step must hallucinate many details in one jump. This is akin to having a large Δt in the SDE sense; the local error can be large and difficult to reverse.

Conversely, if we *split* that dimension change into multiple steps

$$(64\times 64) \rightarrow (32\times 32) \rightarrow (16\times 16) \rightarrow (8\times 8) \rightarrow (4\times 4),$$

each step is a smaller "destruction" of high frequencies and structure, so the reverse upsampling + denoising is more accurate (less local error). By increasing the number of dimension-destruction steps, we make these changes more gradual, pushing the discrete approximation closer to the (hypothetical) continuous limit in scale space:

$$\Delta(\text{dimension}) \rightarrow 0.$$

Hence, *more dimension-destruction steps* is exactly parallel to *more time steps* in standard DDPM: it yields finer increments, lower local error, and a more faithful final reconstruction.

C.4 Conclusion: A Double Discretization Argument

Overall, we have two types of discretization:

- **Noise-Level Discretization:** Splitting the time interval $[0, 1]$ into small Δt's (as in standard diffusion).
- **Dimension Discretization:** Splitting the resolution reduction into many smaller downsampling increments.

Both can be justified under the same SDE-based argument: *smaller steps in each dimension of transformation* $\implies$ *lower approximation error per step* $\implies$ *better overall fidelity*. Empirical evidence (Tables in Section 5) corroborates that increasing either (or both) the number of time steps *and* dimension steps significantly improves generation quality metrics such as PSNR, SSIM, and FID

D Climate Benchmark Metrics

This appendix concisely defines the minimum-standard metrics proposed by (Isphording et al., 2024) and adopted in the present study to assess the ability of ML models to capture three fundamental characteristics of rainfall: *How much does it rain? Where does it rain?*, and *When does it rain?* The metrics used to quantify these characteristics are listed in Table 2 below. Each metric is accompanied by (*i*) its mathematical formulation as implemented in our analysis code, (*ii*) the numerical benchmark that constitutes a 'pass', and (*iii*) a brief explanation of what the metric tells us from a climate science perspective. Where relevant, n denotes the number of non-missing land grid-cells after the AGCD quality mask is applied and w_i the cosine-latitudinal area-weight for cell i. P_i and O_i are the predicted and observed climatological-mean annual rainfall total respectively. We also apply these metrics to evaluate three dynamically downscaled precipitation simulations produced CNRM-CM5 CCAM-1704, HadGEM2-ES RegCM4-7, and MIROC5 CCAM-1704. These datasets are part of the CORDEX-CMIP5 ensemble over the Australasian domain.

D.1 Mean Absolute Percentage Error (MAPE) - How much does it rain?

$$\text{MAPE} = \frac{1}{W}\sum_{i=1}^{n} w_i \Big|\frac{P_i - O_i}{O_i}\Big|, \qquad W := \sum_{i=1}^{n} w_i. \tag{10}$$

Benchmark: MAPE ≤ 0.75.

Climate meaning. MAPE gauges the proportional bias in climatological mean annual rainfall: values ≤ 0.75 require simulations to be, on average, within 75% of observations—a pragmatic trade-off between model realism and current skill levels.

Table 6: Downscaling performance metrics and estimated carbon footprint for machine-learning (ML) and dynamical-downscaling (DD) configurations over Australia (1976–2005). Results for both HDD and the base EDM were both performed using 50 inference steps and 3 denoise steps per shape step for the HDD models

Driving GCM	Model / RCM	NRMSE	MAD	SCorr	MAPE	$kgCO_2$ Train [11]
MIROC5	HDD	0.45	1.0865	0.85	1.1752	∼53 kg
CNRM-CM5	HDD	0.62	1.3897	0.79	0.5185	∼53 kg
MIROC5	Base EDM	0.54	1.2961	0.81	0.5505	∼105 kg
CNRM-CM5	Base EDM	0.56	1.2654	0.78	0.2481	∼105 kg
CNRM-CM5	Earth-ViT	0.50	1.1044	0.80	0.3498	∼51 kg
CNRM-CM5	CCAM-1704	0.68	0.8698	0.89	0.2600	1032 kg (total)
HadGEM2-ES	RegCM4-7	1.47	1.1586	0.90	1.0223	588 kg (total)
MIROC5	CCAM-1704	0.78	1.1083	0.88	0.3374	Not Available

Table 7: Minimum-standard rainfall metrics (adapted from (Isphording et al., 2024)). Metrics are computed from area-weighted average total rainfall over Australia using the AGCD observational dataset. Amplitude is the difference between maximum and mean monthly rainfall; phase is the month of maximum rainfall.

Fundamental rainfall characteristic	Quantifying metric	Benchmark threshold
How much does it rain?	Mean absolute percentage error (MAPE)	MAPE ≤ 0.75
Where does it rain?	Spatial correlation (SCor)	SCor ≥ 0.7
When does it rain?	*Amplitude*: normalised root mean squared error (NRMSE) *Phase*: mean absolute deviation (MAD; months) of the maximum-rainfall month	*Amplitude*: NRMSE ≤ 0.6 *Phase*: MAD ≤ 2

D.2 Spatial Correlation (SCor) - (Where does it rain?)

$$\text{SCor} = \frac{\sum_{i=1}^{n} w_i,(P_i-\hat{P})(O_i-\hat{O})}{\sqrt{\sum_{i=1}^{n} w_i,(P_i-\hat{P})^2}\sqrt{\sum_{i=1}^{n} w_i,(O_i-\hat{O})^2}}, \quad \hat{P} := \frac{1}{W}\sum_{i=1}^{n} w_i P_i, \quad \hat{O} := \frac{1}{W}\sum_{i=1}^{n} w_i O_i. \tag{11}$$

where $\hat{P}$ and $\hat{O}$ are the area-weighted spatial means of P_i and O_i.

Benchmark: SCor ≥ 0.7.

Climate meaning. SCor evaluates how well the model reproduces the *spatial pattern* of mean annual rainfall. correlations (≥ 0.7) imply that regional wet and dry zones are captured in the right places, even if the absolute totals differ.

D.3 Seasonal-cycle metrics (When does it rain?)

We define the grid-cell seasonal amplitude A_i, where $A_i = P_i^{\max} - \bar{P}_i$, and M_i as the phase (month index) of that maximum $P_i^{\max}$. The Normalised RMSE of amplitudes:

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{W}\sum_i w_i (A_i^{\text{mod}} - A_i^{\text{obs}})^2}}{\frac{1}{W}\sum_i w_i, {A_i^{\text{obs}}}^2} \tag{12}$$

where A_i^{mod} and A_i^{obs} are the amplitudes of the prediction and observation respectively.

Benchmark: NRMSE ≤ 0.60.

$$\text{MAD} = \frac{1}{W}\sum_{i=1}^{n} w_i \left|\Delta\phi(M_i^{\text{mod}}, M_i^{\text{obs}})\right| \tag{13}$$

where $\Delta\phi$ is the shortest circular distance on a 12-month circle. M_i^{mod} and M_i^{obs} are the predicted and observed phases respectively.

Benchmark: MAD ≤ 2.

E Calculations for Carbon Emitted into Atmosphere from Model Training/Inference

E.1 Key inputs and assumptions

- **Service-unit (SU) charging on HPC software.** CPU queues are charged at $2\ \text{SU}\,\text{core}^{-1}\,\text{h}^{-1}$, the V100 queue at 3 SU per "resource*hour", and the A100 queue at $4.5\ \text{SU}\,\text{resource}^{-1}\,\text{h}^{-1}$.
- **Hardware power draw.**
  - *CPU node* $P_{\text{system}} = 2.90$ MW , i.e. $P_{\text{core}} = 14.2$ W.
  - *DGX A100 node.* $P_{\text{node}} = 6.5$ kW for 8×A100 GPUs[12] $\Rightarrow P_{\text{GPU}} = 0.81$ kW.
  - *DGX-1 V100 node.* Thermal-design power $P_{\text{node}} = 3.2$ kW for 8×V100 GPUs $\Rightarrow P_{\text{GPU}} = 0.40$ kW.
- **Data-centre overhead.** Power-usage-effectiveness (PUE) assumed $\text{PUE} = 1.3$[13].
- **Grid-emission factor** Scope-2 factor $0.68\ \text{kg}\,\text{CO}_2\,\text{kWh}^{-1}$ and scope-3 factor $0.05\ \text{kg}\,\text{CO}_2\,\text{kWh}^{-1}$ from the
  Combined factor used below: $\gamma = 0.73\ \text{kg}\,\text{CO}_2\,\text{kWh}^{-1}$.

E.2 CPU workloads

Energy per core-hour $E_{\text{core}} = P_{\text{core}} \times \text{PUE} = 14.2\ \text{W} \times 1.3 = 18.5\ \text{W} = 0.0185\ \text{kWh}$.

SU–to–energy conversion CPU: 2 SU per core-hour, hence $E_{\text{SU}} = 0.0185\ \text{kWh}/2 = 9.25 \times 10^{-3}\ \text{kWh}$.

Carbon per kSU $1\ \text{kSU} = 1\,000\ \text{SU} \Rightarrow m_{\text{kSU}} = 1\,000\,E_{\text{SU}}\,\gamma = 1\,000 \times 9.25 \times 10^{-3} \times 0.73 \approx 6.8\ \text{kg}\,\text{CO}_2\,\text{e}$.

E.3 GPU workloads

A100

$$E_{\text{GPU h}} = P_{\text{GPU}} \times \text{PUE} = 0.8125\ \text{kW} \times 1.3 = 1.06\ \text{kWh},$$
$$m_{\text{GPU h}} = E_{\text{GPU h}}\,\gamma \approx 1.06 \times 0.73 = 0.77\ \text{kg}\,\text{CO}_2\,\text{e}.$$

Six-hour run on a single A100: $m = 6 \times 0.77 \approx 4.6\ \text{kg}\,\text{CO}_2\,\text{e}$.

V100

$$E_{\text{GPU h}} = 0.40\ \text{kW} \times 1.3 = 0.52\ \text{kWh},$$
$$m_{\text{GPU h}} = 0.52 \times 0.73 = 0.38\ \text{kg}\,\text{CO}_2\,\text{e}.$$

One hour per V100: $0.38\ \text{kg}\,\text{CO}_2\,\text{e}$.

[12] See A100 power draw per nvidia specifications: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-dgx-a100-datasheet.pdf

[13] Specific to local cluster

E.4 SUMMARY

Workload	Energy (kWh h^{-1})	CO_2 (kg h^{-1})
CPU	0.0185	0.014
A100 GPU	1.06	0.77
V100 GPU	0.52	0.38

These values were scaled for the various runtimes involved in each script.

F CHAIN-RULE PROOF OF THE KL–DECOMPOSITION

Note that the terminology and logic here is adapted from (Rudin, 1987) (Royden and Fitzpatrick, 2010) but has been adapted for the downscaling setting.

Throughout we write λ_d for the *d-dimensional Lebesgue measure*, (i.e. the ordinary notion of volume for a distribution in $\mathbb{R}^d$ - we only define this here to get some nice continuity guarantees for our marginals later)

Formally, a probability law q on $\mathbb{R}^d$ is said to be *absolutely continuous* with respect to λ_d, written $q \ll \lambda_d$, if there exists a non-negative integrable function $q(x)$—the *density*—such that $q(A) = \int_A q(x)\,dx$ for every measurable set A. Absolute continuity is what licenses the familiar integral form of the Kullback–Leibler divergence $KL(q\|p) = \int q(x) \log \frac{q(x)}{p(x)}\,dx$.

In the hierarchical-diffusion setting the fine and coarse image tensors live in Euclidean spaces

$$\mathcal{X} = \mathbb{R}^{h_{t-1}\times w_{t-1}\times C}, \qquad \mathcal{Y} = \mathbb{R}^{h_t \times w_t \times C},$$

so we denote their Lebesgue measures by $\lambda_{\mathcal{X}}$ and $\lambda_{\mathcal{Y}}$, respectively. Assuming $q_{t-1}, p_{t-1} \ll \lambda_{\mathcal{X}}$ and $q_t, p_t \ll \lambda_{\mathcal{Y}}$ simply states that all four distributions possess densities, allowing us to manipulate KL integrals rigorously.

The decomposition proved below is the *chain rule of relative entropy*, first written explicitly by Kullback & Leibler (Kullback and Leibler, 1951) and now standard in information theory (Cover and Thomas, 2005). We give the brief self-contained derivation here for completeness and to show notation for the downscaling/super-resolution setting.

Throughout we let $\mathcal{X} = \mathbb{R}^{h_{t-1}\times w_{t-1}\times C}$ and $\mathcal{Y} = \mathbb{R}^{h_t \times w_t \times C}$ denote the fine– and coarse–resolution spaces at step t, and we assume $q_{t-1}, p_{t-1} \ll \lambda_{\mathcal{X}}$ and $q_t, p_t \ll \lambda_{\mathcal{Y}}$ for the appropriate Lebesgue measures (absolute continuity guarantees the existence of densities).

Let the deterministic down-sampling operator be $D_t : \mathcal{X} \to \mathcal{Y}$ and write $Y = D_t(X)$. Because D_t is measurable and information–non-increasing, the push-forwards $q_t := D_t \# q_{t-1}$ and $p_t := D_t \# p_{t-1}$ exist and are again absolutely continuous.

[KL chain rule under a measurable map] For any pair of measures q_{t-1}, p_{t-1} on $\mathcal{X}$ and any measurable mapping $D_t : \mathcal{X} \to \mathcal{Y}$,

$$KL\big(q_{t-1} \,\|\, p_{t-1}\big) \;=\; KL\big(q_t \,\|\, p_t\big) \;+\; E_{y\sim q_t}\Big[KL\big(q_{t-1}(\cdot \mid Y{=}y) \,\|\, p_{t-1}(\cdot \mid Y{=}y)\big)\Big], \tag{14}$$

where the inner KL is taken between the regular conditional distributions of X given $Y = y$. Both terms on the right-hand side are non-negative, hence splitting the coarse-scale divergence from the fine-scale residual.

Proof. Let $p(x), q(x)$ be the densities of p_{t-1}, q_{t-1} w.r.t. $\lambda_{\mathcal{X}}$, and denote the joint law of (X, Y) under q by $q(x, y) = q(x)\,\delta\big(y - D_t(x)\big)$, with an analogous definition for p. Because Y is a deterministic function of X, we may factorise $q(x) = q(y)\,q(x \mid y)$ and $p(x) = p(y)\,p(x \mid y)$, where $q(y)$ and $p(y)$ are the coarse densities and $q(x \mid y)$, $p(x \mid y)$ are the conditional densities.

Using $KL(q\|p) = \int q(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x$ and substituting the factorisations,

$$KL(q\|p) = \int q(y)\, q(x \,|\, y) \Big[\log \tfrac{q(y)}{p(y)} + \log \tfrac{q(x|y)}{p(x|y)} \Big] \, \mathrm{d}x \, \mathrm{d}y$$
$$= \int q(y) \log \tfrac{q(y)}{p(y)} \, \mathrm{d}y \; + \; \int q(y) \Big[\int q(x \,|\, y) \log \tfrac{q(x|y)}{p(x|y)} \, \mathrm{d}x \Big] \mathrm{d}y.$$

The first integral is $KL(q_t\|p_t)$; the term in square brackets is $KL\big(q_{t-1} \,|\, Y{=}y \,\|\, p_{t-1} \,|\, Y{=}y\big)$. Taking the expectation over $y \sim q_t$ gives Eq. equation 14. □

Consequence for the HDD process. Setting $Y = D_t(X)$ and identifying q_{t-1}, p_{t-1} with the forward and reverse marginals at scale $h_{t-1} \times w_{t-1}$ yields exactly Eq. (14) of the main text:

$$KL\big(q_{t-1} \,\|\, p_{t-1}\big) \;=\; \underbrace{KL\big(D_t q_{t-1} \,\|\, D_t p_{t-1}\big)}_{\text{coarse term}} \;+\; E_{x_t \sim q_t} KL\big(q_{t-1} \,|\, x_t \,\|\, p_{t-1} \,|\, x_t\big).$$

Because the conditional KL is non-negative, matching the down-sampled marginals can only reduce the divergence at the fine scale, proving the monotone-improvement property stated in Theorem 3.1. □

G Additional Result/Hyperparameters and Disclosures

We also tested the model over 50 denoisng steps as an ablation and report these below. We note these

Table 8: Results extended to include CRPS metric. HDD reports values over a probabilistic distribution which are closer to the true underlying value

Model	RMSE	PSNR	CRPS
Base EDM – 50 Steps	0.000197	29.17	0.0002415
HDD – 50 steps – 3 denoise steps per shape step	**0.000157**	**31.40**	**0.0002402**

Below we summarise the key hyperparameters used for sampling and inference in our experiments:

Table 9: Hyperparameters used for sampling and inference. Note that these were generally kept the same as the original EDM implementation in (Karras et al., 2022)

Hyperparameter	Value
Inference sampling	
Number of steps	50 [14]
Noise schedule	
$\sigma_{\min}$	0.002
$\sigma_{\max}$	80
ρ	7
Stochasticity (churn)	
S_{churn}	1
$S_{\min}$	0
$S_{\max}$	$+\infty$
S_{noise}	1
Hierarchical scheduler	
Full resolution (H, W)	(144, 272)
Noise steps per split	1 - 50
Hardware & batch	
GPU	A100
Batch size	1

Per ICLR policy, we disclose that LLMs were used to aid/polish writing, for the retrieval of relevant related work, and in reviewing portions of the paper.

H SPEEDUP

H.1 DROP-IN SHAPE SCHEDULERS

(i) *Equally Spaced Shrink* At each diffusion step we follow a linear ramp from $(1,1) \to (H,W)$ in T equal increments:

$$h_t = H - \frac{t-1}{T-1}\,(H-1), \quad w_t = W - \frac{t-1}{T-1}\,(W-1).$$

The instantaneous area therefore decays quadratically, $A_t = h_t w_t = \big(1 - \frac{t-1}{T-1}\big)^2 A$. Averaging over the schedule gives the dimension-agnostic mean area

$$\alpha_{\text{lin}} = \frac{1}{T\,A}\sum_{t=1}^{T} A_t = \frac{1}{T}\sum_{k=0}^{T-1}\Big(1 - \tfrac{k}{T-1}\Big)^2 = \tfrac{1}{3}.$$

Via the general rule $S = 1/\alpha$ this implies a tight $\boxed{3\times}$ pixel– and FLOP–saving ceiling, drop-in for vanilla EDM.

(ii) *Unit-shrink per denoise step.* At every diffusion step *both* spatial dimensions drop by a single pixel until reaching one:

$$h_t = \max\big(1,\, H-(t-1)\big), \quad w_t = \max\big(1,\, W-(t-1)\big).$$

For $T \le \min(H,W)$ no clamping is active, giving

$$\sum_{t=1}^{T} A_t = T\,H\,W - \frac{(T-1)T}{2}\,(H+W) + \frac{(T-1)T(2T-1)}{6}.$$

Plugging this sum into equation 1–equation 2 yields the closed-form

$$\boxed{S_{\text{unit}} = \Big[1 - \frac{(T-1)}{2A}\,(H+W) + \frac{(T-1)(2T-1)}{6A}\Big]^{-1}}.$$

Example. $H{=}144,\ W{=}272,\ T{=}50$: $\alpha{\approx}0.760 \Rightarrow S{\approx}1.32\times$.

H.2 SUMMARY OF THEORETICAL PIXEL SAVINGS

Shape scheduler	α	Speed-up $S = 1/\alpha$
Linear shrink $(h_t, w_t) \propto 1 - \frac{t-1}{T-1}$	$\frac{1}{3}$	$3\times$
Unit-shrink $\big(h_t, w_t\big) = \big(H-(t-1),\, W-(t-1)\big)$	see Eq. (3)	$\approx 1.32\times$ (50 steps)

All schedules are *drop-in*: when $D_{\mathbf{s}_t} = U_{\mathbf{s}_t} = I$ they revert to vanilla EDM. Eq. equation 2 therefore gives an upper-bound on pixel, FLOP and memory savings obtainable with the HDD framework. We note that this is a higher speed up than comparable image-based approaches due to the choice of shape scheduler (Zhang et al., 2022).

I MONOTONE DECOMPOSITION OF KULLBACK–LEIBLER (KL) DIVERGENCE ACROSS SCALES

At its core, downscaling can be framed as an optimal transport problem between the low- and high- spatial-resolution weather distributions (Wan et al., 2023). We seek to determine the optimal transformation

Write q_t and p_t for the true and model marginals of x_t in 1–2. Because D_t is information-non-increasing and U_t is a right-inverse in expectation, the *chain rule of relative entropy* yields[15]

$$KL\big(q_{t-1} \,\|\, p_{t-1}\big) \;=\; \underbrace{KL\big(D_t q_{t-1} \,\|\, D_t p_{t-1}\big)}_{\text{coarse divergence}} + E_{x_t \sim q_t} KL\big(q_{t-1}|\, x_t \,\|\, p_{t-1}|\, x_t\big), \qquad (15)$$

[15] A proof appears in Appendix E.

the second term being always non-negative. 15 shows that *matching the down-sampled marginals can only* decrease *the fine-scale KL*. Summation over t telescopes:

For the HDD forward–reverse pair 1–2,

$$KL\big(q_0 \,\|\, p_0\big) \;=\; \sum_{t=1}^{T} \Big[KL\big(D_t q_{t-1} \,\|\, D_t p_{t-1}\big) - KL\big(D_t q_t \,\|\, D_t p_t\big) \Big] \;\geq\; 0,$$

and the summand is non-negative for every t. Consequently the coarse-to-fine procedure is *monotonically improving*: each successful fit at scale t tightens an upper bound on the ultimate divergence at full resolution.

Proof. Apply 15 at steps t and $t+1$, subtract, and note that $KL(D_t q_t \,\|\, D_t p_t) = KL(q_t \,\|\, p_t)$ because D_t is the identity on $\mathbb{R}^{h_t \times w_t \times C}$. Summation over t finishes the argument. □

Implications. Section I justifies a two-phase optimisation strategy: (i) minimise the *coarse* EDM loss (large σ_t, small shape) until $KL(D_t q_{t-1} \,\|\, D_t p_{t-1})$ plateaus; (ii) progressively unlock finer scales. Empirically this drastically improves inference time for the similiar RMSE.