MULTI-RESOLUTION LAPLACIAN DIFFUSION MODELS FOR KILOMETER-SCALE CLIMATE EMULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models have shown promise in emulating kilometer-scale climate simulations at a small fraction of their cost. A common design pairs a coarse-resolution global generator with a diffusion-based super-resolution stage that generates fine-scale fields conditioned on the coarse output. The two stages are trained independently across a single resolution gap and share no multiresolution structure, so they inhabit different representations and nothing constrains them to compose exactly. Weather data make this harder. Each channel distributes its variance across scales differently, with heavy tails spanning orders of magnitude, and any multiresolution construction must be realized on a spherical grid. We introduce **SPyD** (Spherical Pyramid Diffusion), which formulates coarse generation and super-resolution as a single diffusion over a Laplacian pyramid. The state splits into a coarse band and orthogonal finer bands, each with its own noise schedule, and the bands evolve under a system of probability-flow ODEs coupled through a shared denoiser. In this formulation, the conditioning is an exact projection of the diffusing state, so the super-resolution stage trains directly on this projected state and the coarse and fine bands compose exactly under the pyramid operators. We extend the EDM framework to channel-dependent noise schedules, calibrated per band from measured statistics, and we realize the pyramid on the sphere with global wavelet operators on HEALPix. On ICON simulation output at 5 km, a single SPyD model spanning the full $16\times$ gap in one stage (HPX64 to HPX1024) reproduces the spherical power spectra and pixel distributions of temperature and precipitation and extreme-precipitation percentiles through the 99.999th, improves on the deployed state of the art for global kilometer-scale emulation on every channel in CRPS at equal conditioning and equal cost, and produces well-calibrated conditional ensembles as measured by rank histograms. SPyD composes directly with pretrained coarse-resolution generators, supporting end-to-end sampling and reanalysis downscaling.

1 INTRODUCTION

Kilometer-scale atmospheric simulation resolves the processes that matter most for climate impacts. At 1–5 km grid spacing, deep convection, tropical cyclone inner cores, frontal structure, and orographic precipitation are represented explicitly rather than parameterized (Stevens et al., 2019; Hohenegger et al., 2022), and it is precisely these mesoscale phenomena that control local precipitation extremes, the quantity on which flood risk, infrastructure planning, and adaptation decisions turn. The obstacle is cost: global storm-resolving simulations are so expensive to run and to store that the large, multi-decadal ensembles needed for robust statistics of rare events remain out of reach, and their outputs accumulate at petabyte scale (Stevens et al., 2019).

Data-driven emulation offers one response to that cost. Global deterministic forecast models learned from reanalysis now rival operational numerical weather prediction at a fraction of the inference cost (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023), hybrid models embed learned components in a differentiable dynamical core (Kochkov et al., 2024), and learned atmospheric models have been pushed from forecasting to multi-year climate emulation (Watt-Meyer et al., 2023). Deterministic training, however, optimizes a conditional mean, and the fields it produces are correspondingly smooth, which suppresses precipitation extremes. Diffusion models (Ho et al., 2020; Song et al., 2020; Karras et al., 2022) avoid this by sampling the conditional distribution instead of averaging it,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

which is why probabilistic weather models built on them produce sharp, calibrated ensembles (Price et al., 2025), and why they are used to generate small-scale atmospheric structure.

Two relevant systems generate kilometer-scale global fields. Climate-in-a-Bottle (cBottle) (Brenowitz et al., 2025) demonstrated a two-stage pipeline in which a coarse ($\sim$100 km, HPX64) diffusion model generates the coarse global atmospheric state and a patch-based super-resolution model (cBottle-SR) downscales it to $\sim$5 km (HPX1024, 12.6 million pixels per channel) with an overlapping-patch multi-diffusion sampler; HiRO-ACE (Perkins et al., 2025) realizes the same vision on a 3 km training simulation, generating near-global precipitation two orders of magnitude faster than the simulator it emulates. These systems combine coarse generation with statistical downscaling (Wilby et al., 1998).

These pipelines make one fixed jump in resolution: a super-resolution model is trained for a single coarse-to-fine ratio, with no intermediate resolutions and no way to reuse a stage at another scale. Climate simulation output is not released that way. The same run is published at a hierarchy of resolutions whose volumes differ by orders of magnitude; the ICON simulation used here is distributed as a separate store per HEALPix level (Segura et al., 2025). Coarse fields are cheap to store and widely available, while kilometer-scale fields are expensive and comparatively scarce. This motivates a framework for generating separate resolution bands. Laplacian-pyramid decompositions offer a natural framework for such pipelines: represent the state as orthogonal multiresolution bands, generate coarse bands first, and condition finer bands on them. The idea has a long lineage in image generation, from cascaded pipelines that generate at low resolution and successively super-resolve (Ho et al., 2022; Saharia et al., 2022) to Laplacian-band formulations that noise frequency bands at band-dependent rates (Atzmon et al., 2024). These arrive as collections of design choices (band definitions, per-band schedules, transition times) without a prescription for setting them, and they do not address what changes when the data are physical fields on a sphere with channels whose statistics differ by orders of magnitude.

The gap is sharpest at the stage handoff. A typical cascade trains its fine stage on simulation-derived conditioning and then deploys it on generated fields, often using noise-conditioning augmentation (Ho et al., 2022): the model is taught to tolerate a synthetic corruption whose statistics need not match those of generated fields. Nothing in the construction says what noise the conditioning ought to carry, because nothing ties the stages to a common clock. A shared diffusion clock specifies the noise carried by each band at the handoff. This paper develops the missing framework and its weather instantiation. We call it **SPyD** (Spherical Pyramid Diffusion). Our contributions are:

1. **Laplacian-pyramid diffusion for global weather.** We develop Laplacian-pyramid diffusion for weather fields: the state is split into a coarse band and orthogonal detail bands, each carrying its own noise schedule $\sigma_k(t) = A_k t$. The bands evolve as a single system of coupled probability-flow ODEs, providing a joint formulation for coarse generation and super-resolution that factorizes into independently trainable stages sharing one noise clock. The EDM losses and samplers (Karras et al., 2022) carry over, extended to support channel-dependent noise schedules (Section 3.1).
2. **Channel-dependent noise schedules.** Weather channels differ in their spectral properties by orders of magnitude, so one noise rate cannot serve them all. We measure band statistics per channel and set the rates from those measurements, without per-channel hand-tuning. To realize the bands on the sphere we develop global spherical wavelet operators on HEALPix (Gorski et al., 2005). Together these define models over bands at different resolutions (Sections 3.2 to 3.3).
3. **Native noisy conditioning at the stage handoff.** At the handoff from the coarse stage to the fine stage, a standard cascade (Ho et al., 2022) delivers generated fields to a model trained only on clean simulation fields. Because every band here is noised on one clock, the fine stage trains on noisy simulation-derived conditioning; Section 4.5 tests the distinct case of generated conditioning. Under generated conditioning the single-stage model's precipitation spectrum stays within 3% of the reference through the full pipeline, while cBottle-SR generates 36% too much fine-scale precipitation power (Sections 3.4 and 4.5).
4. **Higher fidelity than state-of-the-art 5 km global emulation.** On the ICON $\sim$5 km simulation, a single-stage instantiation matches pooled extreme-precipitation percentile estimates to within 1.5–4% through the 99.999th. It attains lower CRPS than the deployed cBottle-SR baseline on all twelve channels at equal conditioning and cost (Section 4).

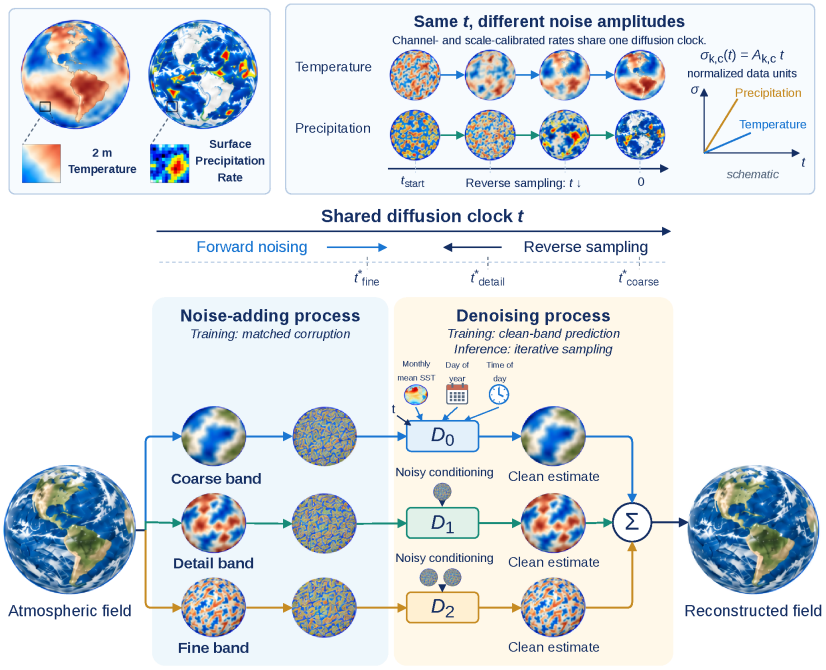

Figure 1: **SPyD overview.** A spherical field is split into coarse and finer bands, noised on a shared clock with channel- and band-dependent rates. Denoisers predict clean bands; finer bands receive noisy coarser-band conditioning. The predicted bands reconstruct the field. Globe textures and temperature/precipitation trajectories are schematic.

2 Related Work

Cascaded and multi-scale diffusion. Cascaded diffusion uses noisy conditioning augmentation at stage handoffs (Ho et al., 2022), while Edify applies band-dependent Laplacian diffusion to image synthesis (Atzmon et al., 2024). Other multiscale approaches include likelihood-oriented cascades, local conditional models, wavelet-domain diffusion models, and wavelet neural operators (Li et al., 2024; Kadkhodaie et al., 2023; Guth et al., 2022; Phung et al., 2023; Hu et al., 2025). SPyD defines the bands on a spherical atmospheric grid, calibrates their noise rates from each channel's measured band statistics rather than setting them by hand, and corrupts the coarse conditioning on the same clock as the generated band, so each stage trains on the noisy input it receives at the handoff.

Generative climate downscaling. CorrDiff (Mardani et al., 2025) and cBottle (Brenowitz et al., 2025) apply conditional diffusion to kilometer-scale regional and global downscaling, and diffusion models have also been applied to regional precipitation and climate downscaling (Srivastava et al., 2024; Watt & Mansfield, 2024). HiRO-ACE (Perkins et al., 2025) couples a coarse stochastic emulator to a 32× precipitation downscaler and separates coarse-generator error from downscaling error. HDD (Curran et al., 2025) imposes a coarse-to-fine hierarchy by changing spatial resolution during diffusion; our bands are instead explicit coarse and detail subspaces with channel-calibrated noise rates.

Weather and climate forecasting. Probabilistic forecasters such as GenCast (Price et al., 2025) model temporal evolution. Spherical DYffusion emulates a coarse global climate model (Rühling Cachay et al., 2024), while ERDM produces probabilistic ERA5 forecasts (Cachay et al., 2025). We address a different task, generating spatial detail conditioned on a coarse atmospheric state.

3 METHOD

3.1 SPYD: MULTI-BAND VARIANCE-EXPLODING DIFFUSION

Figure 1 summarizes the bandwise construction and its shared diffusion clock. Let $x \in \mathbb{R}^n$ represent the full C-channel atmospheric state on a HEALPix grid of order $l_{\max}$. Each channel has $N_{\text{pix}} = 12 \cdot 4^{l_{\max}}$ pixels, so $n = C\,N_{\text{pix}}$. We use $C = 12$ and $l_{\max} = 10$, corresponding to $12\,582\,912$ pixels per channel. Let $\{P_k\}_{k=0}^{K}$ be orthogonal projectors onto resolution bands ordered from coarse ($k = 0$) to fine ($k = K$), satisfying $P_k P_k = P_k$, $P_k P_m = 0$ for $k \neq m$, and $\sum_k P_k = I$, so that $x = \sum_k x_k$ with $x_k = P_k x$ and $\|x\|^2 = \sum_k \|x_k\|^2$. We define the forward process in the variance-exploding (EDM) convention (Karras et al., 2022) with band-specific noise rates,

$$x(t) \;=\; x(0) + \sum_k \sigma_k(t)\, P_k \epsilon \;=:\; \psi_t^{\epsilon}\big(x(0)\big), \qquad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

so under the forward process each band evolves as an independent EDM process $x_k(t) = x_k(0) + \sigma_k(t) P_k \epsilon$; the bands are coupled only through the denoiser, and the losses, preconditioning, and samplers of EDM apply per band. Because the forward process factorizes over bands, the stages can be trained independently. We use linear schedules $\sigma_k(t) = A_k t$, where A_k acts channelwise with entries $A_{k,c}$; hence $\sigma_{k,c}(t) = A_{k,c} t$. For this joint formulation, let $\mathcal{D}(x, t; \theta)$ be a full-state denoiser. Its band-k prediction is $P_k \mathcal{D}(x, t; \theta)$. In each band we use the EDM loss

$$L_k(t, \theta) = \mathbb{E}_{\epsilon, x} \sum_{c=1}^{C} \lambda_{k,c}(t) \big\| [P_k \mathcal{D}(\psi_t^{\epsilon}(x(0)), t; \theta) - x_k(0)]_c \big\|_2^2, \qquad L = \mathbb{E}_{t \sim p_t} \sum_k L_k(t, \theta). \tag{2}$$

Here $\lambda_{k,c}(t)$ weights the squared denoising error for channel c in band k. The channelwise EDM weights and preconditioning coefficients (c_{in}, c_{out}, and c_{skip}) are given in Appendix B. The reverse process is a *system* of probability-flow ODEs, one per band, coupled through a shared denoiser:

$$\dot{x}_k \;=\; \frac{d \log \sigma_k}{dt} \big(x_k - P_k \mathcal{D}(x, t)\big), \qquad k = 0, \ldots, K. \tag{3}$$

With the linear schedules used here, this system reduces to the standard EDM ODE on the full state (Appendix B); the bands differ in their noise rates, which set when each band becomes informative, rather than in their reverse dynamics. Under the calibrated rates of Section 3.3, fine bands become noise-dominated before coarse bands as noise is added. Reverse sampling runs in the opposite order: coarse structure emerges first, then progressively finer detail. Handoff times make this ordering precise.

Handoff times. A handoff $t_k^{\star}$ is a reference time at which band k is still noise-dominated. We place the next finer band's handoff after the coarser bands have become informative in reverse sampling. Section 3.3 gives the recursive calibration rule.

3.2 SPHERICAL BAND CONSTRUCTION: GLOBAL WAVELET OPERATORS

The remaining challenge is to realize the projectors on the spherical HEALPix grid while retaining a consistent coarse-to-fine hierarchy. We build them from an orthonormal discrete wavelet transform W (Daubechies, 1992; Mallat & Mallat, 1999). The level-j coarsening

$$\mathcal{C}_j \;=\; W^{\top} \Pi_j W, \qquad \mathcal{C}_j \mathcal{C}_{j'} \;=\; \mathcal{C}_{\max(j,j')}, \tag{4}$$

with Π_j the diagonal $0/1$ selector retaining the level-j approximation coefficients, is an orthogonal projector, and the nesting of the approximation spaces orders the family. A pyramid is specified by the grid orders it retains, $a_0 < a_1 < \cdots < a_K = l_{\max}$. Writing P_a^b for the band carrying the content between orders a and b,

$$P_a^b \;:=\; \mathcal{C}_{l_{\max}-b} - \mathcal{C}_{l_{\max}-a}, \qquad P_0 \;=\; \mathcal{C}_{l_{\max}-a_0}, \qquad P_k \;=\; P_{a_{k-1}}^{a_k}, \quad k = 1, \ldots, K, \tag{5}$$

with $\mathcal{C}_0 := I$, so band index and spatial scale run together: P_0 is the coarsest band and P_K the finest. The axioms of Section 3.1 follow from $W^{\top} W = I$ and $\Pi_j \Pi_{j'} = \Pi_{\max(j,j')}$: each band is idempotent and symmetric, the bands are mutually orthogonal, and they sum to I.

On the order-10 grid, our single-stage models take $K = 1$: the base $P_0 = \mathcal{C}_4$ carries the order-6 content used as conditioning, and the generated band is $P_1 = P_6^{10}$. The multi-stage model of Section 3.4 takes $K = 2$ over the same grid, with bands $\mathcal{C}_4$, P_6^8, and P_8^{10}. The wavelet family chooses W and thereby the character of each band. Under Haar ($\mathrm{db1}$, average pooling), a band inherits blocky aliasing at every coarse-pixel boundary, structure the denoiser must spend capacity reproducing. Daubechies-4 ($\mathrm{db4}$) reduces this aliasing while keeping compact support; we use it throughout.

The HEALPix grid (Gorski et al., 2005) makes Eq. (4) practical on the sphere. It is a quadtree over twelve base faces, and each face refines into a $2^l \times 2^l$ array whose levels align with wavelet levels. Nested ordering encodes this quadtree as a Morton (bit-interleaved) index rather than a raster, so we reindex each face to a row-major (x, y) layout and apply a separable planar multilevel DWT per face. We first pad each face with a halo of true neighbor-face content (Karlbauer et al., 2024), which keeps the length-8 filters continuous across face seams. We decompose the data and project the diffusion noise on the full sphere before extracting training patches, so training and sampling apply the same operators. Because the DWT is not shift-invariant, we also phase-align the decimating analysis/synthesis pair used at the cross-resolution handoff with $\mathcal{C}_j$, so the handoff represents the same coarse content on either grid up to measured face-seam residuals. Appendix A details the transform, its interplay with the HEALPix hierarchy, and these residuals.

3.3 Recursive calibration of band noise and handoffs

One noise rate cannot serve channels whose band statistics differ by orders of magnitude. We therefore set each band's per-channel rate from measured statistics and place handoffs recursively from coarse to fine. For band k and channel c, let $d_k = \mathrm{rank}(P_k)/C$ be its number of retained modes and define the per-mode signal amplitude and dimensionless noise-to-signal amplitude ratio as

$$\ell_{k,c} = \sqrt{\frac{\mathbb{E}\|[P_k x]_c\|_2^2}{d_k}}, \qquad r_{k,c}(t) = \frac{A_{k,c}\, t}{\ell_{k,c}}. \tag{6}$$

Thus $r \ll 1$ is signal-dominated, $r \approx 1$ has comparable signal and noise, and $r \gg 1$ is noise-dominated. We measure $\ell_{k,c}$ on the training split, together with a tail factor $\beta_{k,c} = \max\big(1, q_{0.999}(|[P_k x]_c|)/(z_{0.999}\sigma_{k,c}^{\mathrm{data}})\big)$. Here $\sigma_{k,c}^{\mathrm{data}}$ is the band's per-pixel standard deviation and $z_{0.999} \approx 3.29$, so $\beta_{k,c} > 1$ flags a band heavier-tailed than a Gaussian. At the handoff, we require the new band to be noise-dominated by a margin:

$$r_{k,c}(t_k^\star) = m b_{k,c}, \qquad A_{k,c} = m b_{k,c} \frac{\ell_{k,c}}{t_k^\star}, \qquad b_{k,c} = \min(\beta_{k,c}, 2), \quad m = 10. \tag{7}$$

We calibrate handoffs recursively from coarse to fine. Before introducing band $k+1$, the already-generated bands must be signal-bearing, giving the rule $t_{k+1}^\star \le \min_{j \le k,\, c} \frac{\ell_{j,c}}{A_{j,c}}$. In particular, at the first detail-band handoff the coarse condition must remain signal-bearing, $r_{0,c}(t_1^\star) \lesssim 1$, while the new band is noise-dominated, $r_{1,c}(t_1^\star) = m b_{1,c} \gg 1$. The same condition applies to every retained coarser band at later handoffs. Once $t_{k+1}^\star$ is chosen, we measure the new band's statistics and set its rates by the same margin rule: $t_k^\star \to \{A_{k,c}\}_c \to \min_{j \le k,c}(\ell_{j,c}/A_{j,c}) \to t_{k+1}^\star \to \{A_{k+1,c}\}_c$. We choose $t_{\min,k}$ and $t_{\max,k}$ so $r_{k,c}(t_{\min,k}) \ll 1$ and $r_{k,c}(t_{\max,k}) \gg 1$ for every channel. Appendix B gives the practical diagnostics and the evaluated times; our experiments use $t^\star = 10^{-3}$ for the $6 \to 10$ and $6 \to 8$ bands and $t^\star = 10^{-5}$ for the $8 \to 10$ band.

3.4 Conditional stages and global sampling

We exploit this factorization and train three conditional band generators independently, each realizing the joint forward process on its own subspaces. For a stage from order a to order b, let $\mathcal{D}_a^b(u, v, t)$ denote its conditional denoiser, which predicts the clean band $P_a^b x(0)$ from a noisy band u and noisy coarse conditioning v, with $P_{\mathrm{band}} = P_a^b$ and $P_{\mathrm{coarse}} = \mathcal{C}_{l_{\max}-a}$. The single-stage model $\mathcal{D}_6^{10}$ generates the HPX64-to-HPX1024 band. The two-stage cascade, which we call multi-stage SPyD, applies $\mathcal{D}_6^8$ and then $\mathcal{D}_8^{10}$; its fine stage conditions on the HPX64 base plus the generated intermediate band, which we represent on HPX1024. Appendix B gives the time mapping for a pretrained coarse EDM model.

Conditioning with matched noise. The stage receives a noisy version u of the band it must generate and a noisy version v of the coarser field it conditions on. With $\epsilon \sim \mathcal{N}(0, I)$, these inputs are

$$\begin{aligned} u &= P_{\text{band}} x(0) + A_{\text{band}} t \, P_{\text{band}} \epsilon, \\ v &= P_{\text{coarse}} x(0) + A_{\text{lr}} t \, P_{\text{coarse}} \epsilon. \end{aligned} \tag{8}$$

Here A_{band} contains the calibrated per-channel rates. We train $\mathcal{D}_a^b(u, v, t)$ to recover $P_{\text{band}} x(0)$; at inference, v is formed from the supplied or previously generated coarse field. Both noise levels anneal on the same clock. Appendix B relates this stage-specific construction to Section 3.1.

Global sampling. Multi-diffusion maintains one noisy band y over the full sphere throughout sampling (Bar-Tal et al., 2023; Brenowitz et al., 2025). At each Heun evaluation, the local denoiser processes overlapping patches of y and the noisy coarse condition v. Its predictions are assembled into one global denoiser estimate:

$$\overline{\mathcal{D}}_a^b(y, v, t) = \frac{\sum_i G_i^* \mathcal{D}_a^b(G_i y, G_i v, t)}{\sum_i G_i^* \mathbf{1}}, \tag{9}$$

where G_i extracts patch i from the padded HEALPix field, G_i^* places its values back on the global grid, and division is pixelwise. The denominator counts the patches covering each pixel, so overlapping predictions are averaged before the global Heun update. After a stage finishes, its generated band is added to the retained coarse field, and that sum conditions the next stage. Algorithm 1 applies this operation at arbitrary pyramid resolutions; Appendix B gives patch geometry and sampling settings. Section 4 evaluates both samplers on coarsened ICON states, and Section 4.5 on fields from a pretrained coarse generator.

4 Empirical Evaluation

We first downscale HPX64 fields obtained by coarsening held-out ICON states, comparing single-stage SPyD ($6 \to 10$), multi-stage SPyD ($6 \to 8 \to 10$), and cBottle-SR, then condition the models on HPX64 samples from cBottle's coarse generator. The multi-stage model tests whether SPyD can compose intermediate bands.

4.1 Experimental setup

We use twelve fields from the five-year nextGEMS ICON Cycle 3 simulation (run ngc3028), sampled every 30 minutes and mapped to HPX1024 (Koldunov et al., 2023). This is the same simulation used by cBottle-SR (Brenowitz et al., 2025). We hold out the final 25% of timesteps for evaluation. Each conditional stage uses a SongUNet with rates calibrated as in Section 3.3; all models use global multi-diffusion Heun sampling. A single-stage global sample takes ≈6 minutes on one NVIDIA H100, comparable to cBottle-SR, which uses a backbone of the same size with the same sampler, step count, and patch tiling; multi-stage SPyD adds a further 8–17%. The channels follow cBottle-SR (Brenowitz et al., 2025) and are listed with their units in Table 4; training and sampling settings and evaluation protocols are in Appendix B, and Appendix D reports train/validation denoising-loss ratios for single-stage SPyD per noise level and checkpoint.

4.2 SPyD reproduces weather structure and precipitation variability

A downscaler must recover weather structures that its coarse input cannot resolve. Figure 2 compares ICON, cBottle-SR, and single-stage SPyD under the same coarse conditioning for one held-out global state. Precipitation and outgoing longwave radiation (OLR) show a cyclone, fronts, tropical convection, and the intertropical convergence zone at global and local scales. In the cyclone zoom, SPyD preserves a clear eye inside the rain-filled eyewall, whereas cBottle-SR fills much of the core with rain. Storm-centered comparisons of the four deepest storms in the simulation (Figure 8) show the same pattern: both SPyD models form a closed eyewall around an open eye, while cBottle-SR produces a broader, less distinct precipitation core. The other zooms show localized convection and frontal bands across different weather regimes, see Appendix C.

However, visual detail alone cannot show whether the models generate the right amount of variability. Figure 3 compares pixel-value distributions and power spectra over 32 held-out states. Spectral power

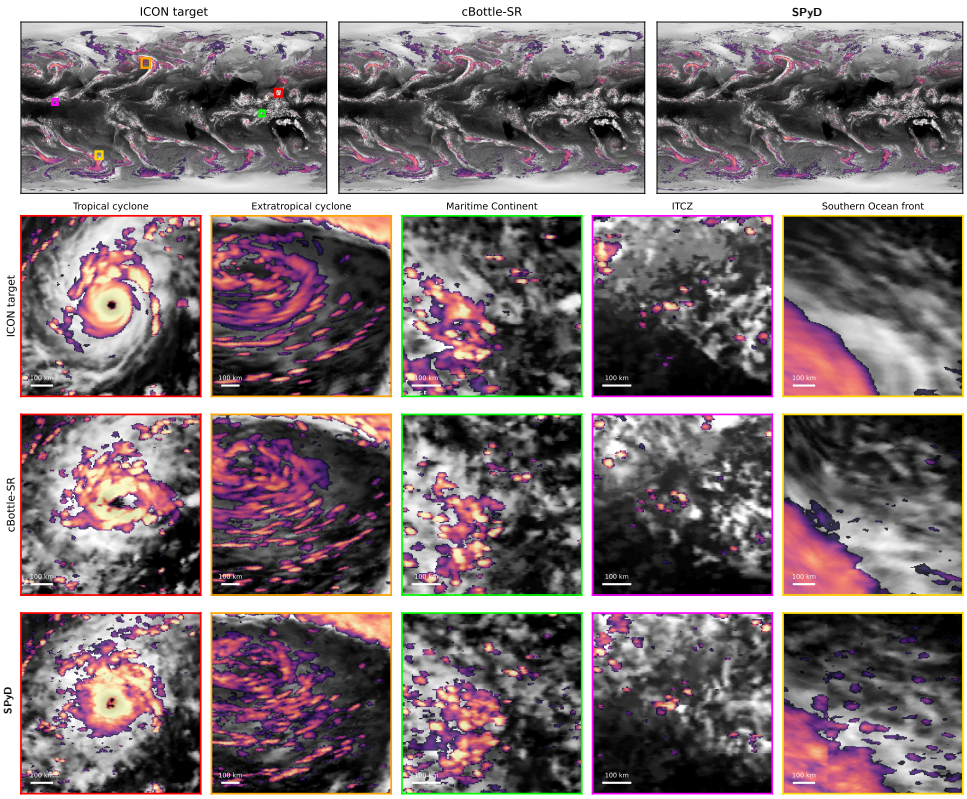

Figure 2: Global two-layer view (outgoing longwave radiation, grayscale; precipitation overlay, logarithmic scale) with regime zooms at native resolution, for the held-out sample at 2024-12-03T09:00:00: ICON (left), cBottle-SR (center), and single-stage SPyD (right). All three share the same coarse conditioning, so each zoom compares the same weather system. Boxes: a simulated tropical cyclone east of Luzon (minimum sea-level pressure 935 hPa, peak 10 m winds 57 m s^{-1}), an occluded extratropical cyclone in the central Atlantic, Maritime Continent deep convection, the Pacific ITCZ, and a Southern Ocean frontal band. Scale bars: 100 km.

measures variation at each spatial scale; here mesoscale precipitation power refers to variation in storm- and rainband-scale rainfall, between broad circulation patterns and individual grid cells. All three models closely match ICON's 2 m temperature distribution and spectrum. For precipitation, cBottle-SR has 30–50% too much mesoscale power and a heavier upper tail. Single-stage SPyD follows the ICON spectrum more closely, while multi-stage SPyD has too little power near the mesoscale peak. Appendix B reports a paired-error diagnostic for the same comparison.

4.3 SPyD reproduces precipitation extremes

Extreme rainfall matters for local climate impacts, so upper-tail precipitation is a key downscaling test, as in HiRO-ACE (Perkins et al., 2025). Table 1 compares precipitation percentiles pooled over 32 held-out global samples. The coarse input is 87% too low at the 99.999th percentile, so a downscaler must generate nearly all of this rare intensity. At that percentile, single-stage SPyD produces 130.03 mm/hr against ICON's 132.16 mm/hr (1.6% low); cBottle-SR is 14% high and multi-stage SPyD 16% low. Across the reported percentiles, single-stage SPyD deviates from ICON by 1.5–4%. Section 4.5 examines multi-stage SPyD further under generated coarse inputs.

4.4 Single-stage ensembles improve CRPS with near-uniform ranks

Downscaling should represent uncertainty as well as produce plausible fields. We score eight-member conditional ensembles with fair continuous ranked probability score (CRPS), a proper scoring rule for the predicted distribution; lower is better. Across 12 held-out months and 16 patches per month, single-stage SPyD has lower CRPS than cBottle-SR on all twelve channels, significantly on eight under a paired patch-level bootstrap (95% intervals) and never significantly worse (Table 2). Multi-

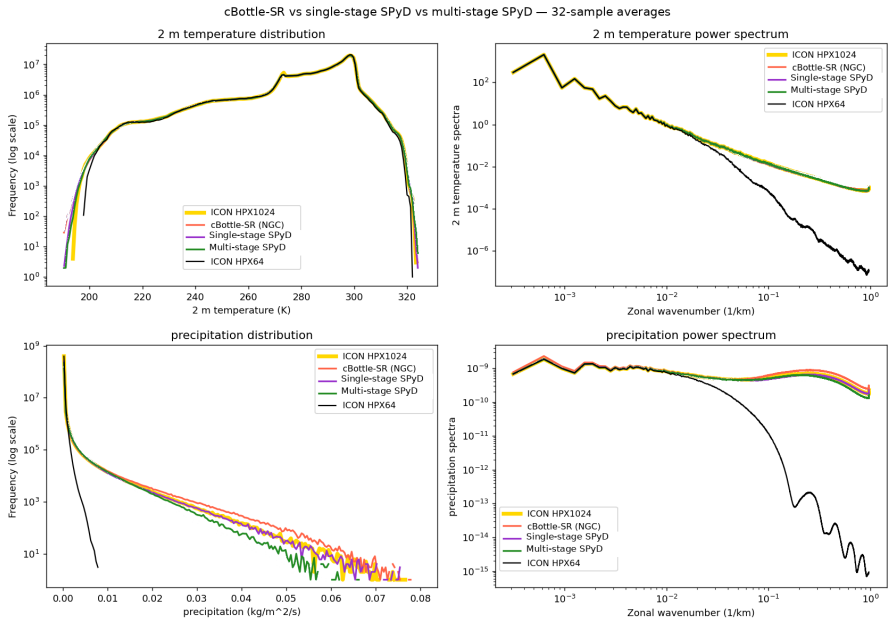

Figure 3: Pixel distributions (left) and spherical-harmonic power spectra (right), averaged over 32 held-out states, for 2 m temperature (top) and precipitation (bottom). All three models closely match ICON (gold) for temperature. For precipitation, cBottle-SR (red) has excess mesoscale power and a heavier upper tail; single-stage SPyD (purple) is closer to ICON, while multi-stage SPyD (green) has too little power near the mesoscale peak.

Table 1: Empirical precipitation percentiles (mm/hr), pooled over 32 held-out global samples (403M precipitation values). Parentheses give relative bias against ICON; smaller absolute bias is better. Bold marks the generated model closest to ICON. Naive upsampling scores the coarse input ($\mathcal{C}_4 x$) with no generation.

Percentile	ICON	Naive upsampling ($\mathcal{C}_4 x$)	cBottle-SR	Single-stage SPyD	Multi-stage SPyD
P99	2.23	2.33 (+4%)	2.15 (−4%)	**2.19** (−2%)	2.33 (+4%)
P99.9	22.49	5.79 (−74%)	25.16 (+12%)	21.62 (−4%)	**21.83** (−3%)
P99.99	72.58	10.58 (−85%)	85.20 (+17%)	**71.51** (−1.5%)	64.65 (−11%)
P99.999	132.16	16.89 (−87%)	151.30 (+14%)	**130.03** (−1.6%)	111.28 (−16%)

stage SPyD attains the best CRPS on five smooth large-scale channels, while the single-stage model remains stronger on convective channels such as precipitation. Table 2 also reports ensemble-mean RMSE, on which single-stage SPyD improves on cBottle-SR on all twelve channels; because RMSE scores only the ensemble mean and favors smooth predictions such as naive upsampling, we treat CRPS as the primary metric.

We check calibration by ranking the held-out value among the eight generated members: a nearly uniform rank histogram means the target falls throughout the ensemble at the expected frequencies. The single-stage histograms are near-uniform for temperature, winds, radiation, pressure, and SST (Figure 4). cBottle-SR's sloped histograms indicate conditional biases, while multi-stage SPyD's domed histograms indicate over-dispersion. Precipitation and sea-ice fraction contain many exact zeros, causing rank ties that make those histograms harder to interpret.

4.5 Single-stage SPyD preserves spectra in the full generative pipeline

We next replace the coarsened ICON conditioning with samples from cBottle's coarse generator, testing whether the downscalers remain stable in the full pipeline. Across 16 coarse samples spanning all seasons, single-stage SPyD keeps precipitation spectra within 3% of the reference; cBottle-SR has 36% excess fine-scale precipitation power. At its fine stage, multi-stage SPyD adds convective cells in 20–22% of quiescent pixels, compared with 0.62% at its intermediate output. Appendix F gives the sampling protocol, spectra, and the lifting-operator comparison.

Table 2: Probabilistic and deterministic ensemble scores per channel (physical units; lower is better), 12 held-out months × 16 patches × 8-member ensembles, equal conditioning information. Bold marks the best value. CRPS ranks the three generated models; naive upsampling scores the coarse input $\mathcal{C}_4 x$ as a deterministic forecast (CRPS = MAE). Ensemble-mean RMSE ranks all three reported columns, including naive upsampling.

	CRPS ↓				RMSE (ensemble mean) ↓		
Channel	Naive upsampling	Single-stage SPyD	Multi-stage SPyD	cBottle-SR	Naive upsampling	Single-stage SPyD	cBottle-SR
cllvi	0.07329	**0.04302**	0.04937	0.04382	**0.1564**	0.1570	0.1626
clivi	0.006824	**0.004121**	0.004428	0.004272	0.01341	**0.01308**	0.01395
tas	0.5716	0.2694	**0.2671**	0.2806	0.8596	**0.5736**	0.6055
uas	0.7911	0.49119	**0.4846**	0.49120	1.086	**0.9857**	1.001
vas	0.8164	0.5072	**0.5012**	0.5094	1.122	**1.022**	1.042
rlut	8.803	**5.594**	5.686	5.698	12.41	**12.24**	12.75
rsut	21.41	**13.01**	13.90	13.11	30.40	**30.21**	31.15
pres_msl	19.37	9.387	**9.062**	10.61	28.78	**19.19**	21.96
pr	5.744e-05	**2.856e-05**	3.921e-05	2.885e-05	**2.580e-04**	2.597e-04	2.756e-04
rsds	24.90	**15.39**	16.72	15.91	**35.71**	36.02	37.55
sst	0.4204	0.1173	**0.1153**	0.1296	0.7960	**0.2728**	0.2952
sic	0.004129	**0.001132**	0.001259	0.001273	0.008642	**0.004436**	0.005336

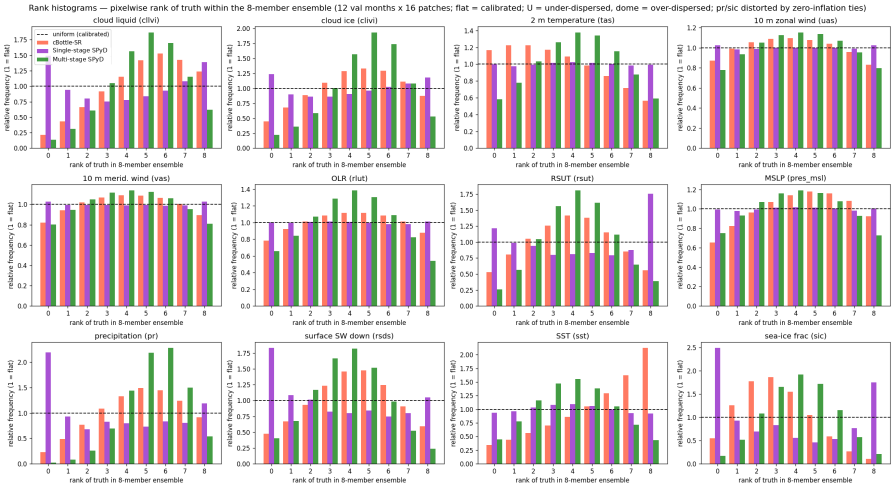

Figure 4: Pixelwise rank of the target within eight-member conditional ensembles; a flat histogram is expected under calibration. Single-stage SPyD (purple) is near-uniform on most channels; cBottle-SR (red) is sloped and multi-stage SPyD (green) is domed. Zero-inflation rank ties complicate the precipitation and sea-ice panels.

5 Discussion and Conclusion

We introduced SPyD, a Laplacian-pyramid diffusion model for kilometer-scale global climate emulation. SPyD splits spherical fields into orthogonal resolution bands built from global wavelet operators on HEALPix and noises every band on one diffusion clock, with per-channel rates calibrated from measured band statistics. The fine stage therefore trains on noisy conditioning that is an exact projection of the diffusing state. On held-out 5 km ICON data, single-stage SPyD lowers CRPS relative to cBottle-SR on all twelve channels and matches pooled extreme-precipitation percentiles to within 1.5–4% through the 99.999th. Conditioned on a pretrained coarse generator, its precipitation spectrum stays within 3% of the reference, while cBottle-SR has 36% excess fine-scale power.

Multi-stage SPyD shows that intermediate bands compose, and training its fine stage on generated conditioning is a direct way to strengthen the full pipeline. Because a finer band needs its coarser bands only once they are signal-dominated, the handoff times partition sampling into overlapping stages. A distributed sampler could run the bands concurrently on separate devices instead of in sequence. Finer grids and other storm-resolving simulations would test how far the calibrated schedules transfer.

REFERENCES

Yuval Atzmon, Maciej Bala, Yogesh Balaji, Tiffany Cai, Yin Cui, Jiaojiao Fan, Yunhao Ge, Siddharth Gururani, Jacob Huffman, Ronald Isaac, et al. Edify image: High-quality image generation with pixel space laplacian diffusion models. *arXiv preprint arXiv:2411.07126*, 2024.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Noah D Brenowitz, Tao Ge, Akshay Subramaniam, Peter Manshausen, Aayush Gupta, David M Hall, Morteza Mardani, Arash Vahdat, Karthik Kashinath, and Michael S Pritchard. Climate in a bottle: Towards a generative foundation model for the kilometer-scale global atmosphere. *arXiv preprint arXiv:2505.06474*, 2025.

Salva Rühling Cachay, Miika Aittala, Karsten Kreis, Noah Brenowitz, Arash Vahdat, Morteza Mardani, and Rose Yu. Elucidated rolling diffusion models for probabilistic forecasting of complex dynamics. *arXiv preprint arXiv:2506.20024*, 2025.

Declan J Curran, Sanaa Hobeichi, Hira Saleem, Hao Xue, and Flora D Salim. Generate the forest before the trees–a hierarchical diffusion model for climate downscaling. *arXiv preprint arXiv:2506.19391*, 2025.

Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

Krzysztof M Gorski, Eric Hivon, Anthony J Banday, Benjamin D Wandelt, Frode K Hansen, Mstvos Reinecke, and Matthia Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759–771, 2005.

Florentin Guth, Simon Coste, Valentin De Bortoli, and Stephane Mallat. Wavelet score-based generative modeling. *Advances in neural information processing systems*, 35:478–491, 2022.

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

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Cathy Hohenegger, Peter Korn, Leonidas Linardakis, René Redler, Reiner Schnur, Panagiotis Adamidis, Jiawei Bao, Swantje Bastin, Milad Behravesh, Martin Bergemann, et al. Icon-sapphire: simulating the components of the earth system and their interactions at kilometer and subkilometer scales. *Geoscientific Model Development Discussions*, 2022:1–42, 2022.

Peiyan Hu, Rui Wang, Xiang Zheng, Tao Zhang, Haodong Feng, Ruiqi Feng, Long Wei, Yue Wang, Zhi-Ming Ma, and Tailin Wu. Wavelet diffusion neural operator. In *International Conference on Learning Representations*, volume 2025, pp. 12291–12333, 2025.

Zahra Kadkhodaie, Florentin Guth, Stéphane Mallat, and Eero P Simoncelli. Learning multi-scale local conditional probability models of images. *arXiv preprint arXiv:2303.02984*, 2023.

Matthias Karlbauer, Nathaniel Cresswell-Clay, Dale R Durran, Raul A Moreno, Thorsten Kurth, Boris Bonev, Noah Brenowitz, and Martin V Butz. Advancing parsimonious deep learning weather prediction using the healpix mesh. *Journal of Advances in Modeling Earth Systems*, 16(8): e2023MS004021, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024.

Nikolay Koldunov, Tobias Kölling, Xabier Pedruzo-Bagazgoitia, Thomas Rackow, René Redler, Dmitry Sidorenko, Karl-Hermann Wieners, and Florian Andreas Ziemen. nextgems: Output of the model development cycle 3 simulations for icon and ifs. 2023.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Henry Li, Ronen Basri, and Yuval Kluger. Likelihood training of cascaded diffusion models via hierarchical volume-preserving maps. In *International Conference on Learning Representations*, volume 2024, pp. 43503–43523, 2024.

Stéphane G Mallat and Stéphane Mallat. *A wavelet tour of signal processing*, volume 2. Academic press San Diego, 1999.

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.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

W Andre Perkins, Anna Kwa, Jeremy McGibbon, Troy Arcomano, Spencer K Clark, Oliver Watt-Meyer, Christopher S Bretherton, and Lucas M Harris. Hiro-ace: Fast and skillful ai emulation and downscaling trained on a 3 km global storm-resolving model. *arXiv preprint arXiv:2512.18224*, 2025.

Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10199–10208. IEEE, 2023.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Salva Rühling Cachay, Brian Henn, Oliver Watt-Meyer, Christopher S Bretherton, and Rose Yu. Probablistic emulation of a global climate model with spherical dyffusion. *Advances in Neural Information Processing Systems*, 37:127610–127644, 2024.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Hans Segura, Xabier Pedruzo-Bagazgoitia, Philipp Weiss, Sebastian K Müller, Thomas Rackow, Junhong Lee, Edgar Dolores-Tesillos, Imme Benedict, Matthias Aengenheyster, Razvan Aguridan, et al. nextgems: entering the era of kilometer-scale earth system modeling. *Geoscientific Model Development*, 18(20):7735–7761, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Prakhar Srivastava, Ruihan Yang, Gavin Kerrigan, Gideon Dresdner, Jeremy McGibbon, Christopher Bretherton, and Stephan Mandt. Precipitation downscaling with spatiotemporal video diffusion. *Advances in Neural Information Processing Systems*, 37:56374–56400, 2024.

Bjorn Stevens, Masaki Satoh, Ludovic Auger, Joachim Biercamp, Christopher S Bretherton, Xi Chen, Peter Düben, Falko Judt, Marat Khairoutdinov, Daniel Klocke, et al. Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, 6(1):61, 2019.

Robbie A Watt and Laura A Mansfield. Generative diffusion-based downscaling for climate. *arXiv preprint arXiv:2404.17752*, 2024.

Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K Clark, Brian Henn, James Duncan, Noah D Brenowitz, Karthik Kashinath, Michael S Pritchard, Boris Bonev, et al. Ace: A fast, skillful learned global atmospheric model for climate prediction. *arXiv preprint arXiv:2310.02074*, 2023.

Robert L Wilby, TML Wigley, Declan Conway, PD Jones, BC Hewitson, J Main, and DS Wilks. Statistical downscaling of general circulation model output: A comparison of methods. *Water resources research*, 34(11):2995–3008, 1998.

A WAVELET TRANSFORMS ON THE HEALPIX GRID

The projectors of Section 3.2 apply a planar discrete wavelet transform to each HEALPix face. Below we review that transform, describe how the face structure, face halos, and decimation phase enter the construction, and report residuals of the implemented operators on the full sphere.

Wavelet transforms. The continuous wavelet transform correlates a signal with dilated and translated copies of a mother wavelet ψ,

$$(\mathcal{T}f)(a,b) \;=\; \frac{1}{\sqrt{a}} \int f(t)\, \psi\!\left(\frac{t-b}{a}\right) dt, \tag{10}$$

resolving f simultaneously in position and scale. Restricting to dyadic scales, an orthonormal wavelet basis $\psi_{j,n}(t) = 2^{-j/2}\psi(2^{-j}t - n)$ together with a scaling function φ generates a multiresolution analysis: nested approximation spaces $V_j = \mathrm{span}\{\varphi_{j,n}\}$ with orthogonal detail complements $V_{j-1} = V_j \oplus \mathcal{W}_j$, so any signal splits into a coarse approximation at level j plus details at levels $1, \ldots, j$ (Daubechies, 1992; Mallat & Mallat, 1999). The Daubechies family $\mathrm{db}N$ consists of the compactly supported orthonormal wavelets with filter length $2N$ and N vanishing moments; $\mathrm{db}1$ is the Haar wavelet, whose scaling function is the box function (coarsening by average pooling), and $\mathrm{db}4$, used throughout this work, has length-8 filters. On sampled data the transform is computed by the fast filter-bank recursion

$$a_{j+1}[n] = \sum_k h[k-2n]\, a_j[k], \qquad d_{j+1}[n] = \sum_k g[k-2n]\, a_j[k], \tag{11}$$

with h the scaling (lowpass) filter, g its quadrature mirror, and a_0 the input samples; orthonormality of the basis makes each recursion step an orthogonal map, inverted by its adjoint.

The matrices W and Π_j. Stacking j levels of the recursion in Eq. (11), applied separably along both axes of a two-dimensional array, defines the orthogonal matrix W of Eq. (4). Its rows are the discrete basis functions, and it maps the array to its coefficient vector $(a_j, d_j, \ldots, d_1)$. The selector Π_j is the diagonal $0/1$ matrix that retains the level-j approximation block a_j and zeroes every detail block. Hence $W^\top \Pi_j W$ reconstructs the signal from its coarse approximation alone; this is the level-j wavelet coarsening $\mathcal{C}_j$, and $I - \mathcal{C}_j$ is the band above it.

Band ranks. With $P^l_{l-j} = I - \mathcal{C}_j$ on the order-l grid, $\mathrm{rank}(P^l_{l-j}) = 12\,(4^l - 4^{l-j})$ per channel; for the $6 \to 10$ model, $12\,(4^{10} - 4^6)$. These ranks enter the handoff and calibration rules in Section 3.3. The measured band variances agree with the orthogonal decomposition in face interiors; Table 3 records the full-sphere deviations.

Nested indexing and faces. HEALPix at order l tiles the sphere with $12 \cdot 4^l$ equal-area pixels organized as twelve base faces. In nested ordering the pixel index within a face is the bit-interleaved (Morton) encoding of the face-local (x, y) coordinates. Each face is therefore a $2^l \times 2^l$ planar array, and coarsening by one order is the quadtree parent map. This alignment between grid refinement and dyadic wavelet levels lets a standard separable planar multilevel DWT realize the spherical projectors of Eq. (4): a level-j transform on the order-l grid produces approximation coefficients that live on the order-$(l-j)$ grid.

Face-halo transform. A planar transform applied to an isolated face would impose an artificial boundary at every face edge. Before the transform, we therefore pad each face with a halo of true neighbor-face content, obtained from the HEALPix adjacency. The halo width is set by the accumulated support of the multilevel filter bank, so every retained coefficient whose support lies within the face interior equals the coefficient of an ideal boundary-free transform. We apply the same padded transform to data and to noise (Section 3.2). Halos extend the coefficients at the level being transformed: a fine-level halo before analysis, and a coarse-level halo on the approximation canvas before synthesis. Face seams thus stay continuous in both directions, and the pair is inverse-consistent in face interiors. Under the orthonormal convention, the two-dimensional approximation gains a factor of 2 per level, so we normalize the decimated map by 2^{-j} to map a constant field to itself. Consequently, $\mathcal{C}_j$ preserves the global mean of every channel up to the seam residuals of Table 3.

Decimated pair and phase alignment. The construction needs two forms of the coarsening: the projector $\mathcal{C}_j$ of Eq. (4), which returns the coarse content on the fine grid, and a decimating analysis/synthesis pair $(\mathcal{A}_j, \mathcal{S}_j)$, which represents the same content on its native coarse grid for the cross-resolution handoff. The DWT is not shift-invariant: with a padding of p samples prepended to a face, the recursion in Eq. (11) becomes $a_{j+1}[n] = \sum_k h[k-2n+p]\, a_j[k]$, so the decimation lattice is $\{2n-p\}$ and the retained approximation subspace depends on $p \bmod 2$ at each level, hence on $p \bmod 2^j$ after j levels. Operators whose paddings p_c, p_d differ by an amount $\not\equiv 0 \pmod{2^j}$ therefore project onto different subspaces even though each is individually a valid wavelet coarsening. We therefore offset the face origin of the decimating pair by $(p_c - p_d) \bmod 2^j$. With this alignment, $\mathcal{S}_j\mathcal{A}_j$ and $\mathcal{C}_j$ agree to machine precision in face interiors (Table 3). Without it, the order-8 coarsening $\mathcal{C}_2$ showed $\sim$50% pointwise disagreement and a 9% near-Nyquist precipitation-power deficit.

Full-sphere operator residuals. Table 3 reports numerical residuals of the implemented operators at HPX1024, taking the worst case over both wavelet levels, Gaussian maps, and atmospheric fields. We split residuals by region: face interiors, seam strips within 128 pixels of a face edge (a conservative bound on the accumulated filter support), and corner patches. Interior residuals sit at working precision and shrink accordingly in float64; the decimated handoff is bitwise exact there. Seam and corner residuals do not depend on precision. They are structural, caused by stitching planar transforms across face frames, and bounded by 3×10^{-3} pointwise and 10^{-4} in RMS. Over the full sphere, the relative self-adjointness residual is at most 2.3×10^{-5}, the energy-conservation residual at most 9.1×10^{-4}, and the mean-preservation error at most 2.8×10^{-5}.

Table 3: Residuals of the implemented face-halo operators at HPX1024, worst case over levels $j \in \{2, 4\}$ and inputs (noise, tas, pr). Seam and corner values are identical in float32 and float64.

residual (relative)	interior max (f32 / f64)	seam max / RMS	corner max / RMS
idempotence $\|\mathcal{C}^2 x - \mathcal{C}x\|$	2×10^{-9} / 4×10^{-18}	9.6×10^{-4} / 3.6×10^{-6}	3.8×10^{-4} / 5.6×10^{-6}
repeated application $\|\mathcal{C}^5 x - \mathcal{C}x\|$	9×10^{-9} / 2×10^{-17}	2.0×10^{-3} / 1.0×10^{-5}	7.5×10^{-4} / 1.3×10^{-5}
nesting $\|\mathcal{C}_4\mathcal{C}_2 x - \mathcal{C}_4 x\|$	8×10^{-10} / 2×10^{-18}	1.8×10^{-4} / 1.9×10^{-6}	1.1×10^{-4} / 2.5×10^{-6}
handoff $\|\mathcal{S}_j\mathcal{A}_j x - \mathcal{C}_j x\|$	0 (exact)	3.0×10^{-3} / 4.1×10^{-5}	1.8×10^{-3} / 7.3×10^{-5}

B Training, Sampling, and Calibration Details

B.1 Formulation details

Rank identity. For an orthogonal projector P and $\epsilon \sim \mathcal{N}(0, I_n)$, $\mathbb{E}\|P\epsilon\|_2^2 = \mathbb{E}\,\epsilon^\top P\epsilon = \mathrm{tr}(P) = \mathrm{rank}(P)$, the normalization used for the handoff condition and per-mode amplitudes in Section 3.3.

Preconditioning with the rank correction. Because the band noise occupies $\mathrm{rank}(P)$ of n dimensions, its per-pixel standard deviation is $\sigma_{\mathrm{eff},c}(t) = A_c\, t\, \sqrt{\mathrm{rank}(P)/n}$, and the EDM coefficients use it in place of a scalar noise level, per channel:

$$c_{\mathrm{in}} = \frac{1}{\sqrt{\sigma^2_{\mathrm{data},c} + \sigma^2_{\mathrm{eff},c}}},\quad c_{\mathrm{skip}} = \frac{\sigma^2_{\mathrm{data},c}}{\sigma^2_{\mathrm{data},c} + \sigma^2_{\mathrm{eff},c}},\quad c_{\mathrm{out}} = \frac{\sigma_{\mathrm{eff},c}\,\sigma_{\mathrm{data},c}}{\sqrt{\sigma^2_{\mathrm{data},c} + \sigma^2_{\mathrm{eff},c}}},\quad c_{\mathrm{noise}} = \tfrac{1}{4}\log t. \tag{12}$$

For the generated band at stage k, the channelwise EDM loss weight is

$$\lambda_{k,c}(t) = c_{\mathrm{out},c}(t)^{-2} = \frac{\sigma^2_{\mathrm{data},c} + \sigma^2_{\mathrm{eff},c}(t)}{\sigma^2_{\mathrm{data},c}\,\sigma^2_{\mathrm{eff},c}(t)}, \tag{13}$$

where the stage index is suppressed on $c_{\mathrm{out},c}$, $\sigma_{\mathrm{data},c}$, and $\sigma_{\mathrm{eff},c}$. The conditioning input is scaled analogously with its own statistics, $c_{\mathrm{cond}} = \left(\sigma^2_{\mathrm{coarse,c}} + (A_{\mathrm{lr}}t)^2\,\mathrm{rank}(P_{\mathrm{coarse}})/n\right)^{-1/2}$. A pretrained EDM coarse model composes with the pyramid through the time mapping $\sigma_{\mathrm{EDM}} = A_{\mathrm{lr}}\, t$, which places its scalar noise level on the pyramid's clock.

Linear-schedule reduction. With $\sigma_k(t) = A_k t$, every band in Eq. (3) has $d \log \sigma_k / dt = 1/t$. Summing over bands and using $\sum_k P_k = I$ gives $\dot{x} = (x - \mathcal{D}(x,t))/t$ for the joint full-state denoiser. This identity describes the joint formulation; the separately trained conditional stages are sampled in sequence (Algorithm 1).

Joint-noising interpretation. Each conditional stage uses a two-subspace instance of the forward noising in Section 3.1, on the retained state $(P_{\text{band}} + P_{\text{coarse}})x(0)$. Under the ideal projector identities, $P_{\text{band}} P_{\text{coarse}}^\top = 0$, so the projected Gaussian noises are independent; the atmospheric bands need not be. Projecting one global Gaussian map or independent maps into the two subspaces gives the same joint noise distribution. Conditioning on clean coarse fields at $t > 0$, as in a standard cascade, corresponds to a different noising construction. Each trained stage therefore realizes the joint forward process exactly on its two subspaces, and Section 4.5 evaluates the composed stages under generated coarse conditioning. Residuals of the implemented spherical operators are reported in Appendix A.

B.2 Calibration of noise rates and handoff times

Sequential schedule construction. Consider channel c of band k under an ideal projector, and let $s_{k,c} = [x_k(0)]_c$ and $y_{k,c}(t) = s_{k,c} + A_{k,c} t [P_k \epsilon]_c$. Independence and zero mean of ϵ give

$$\frac{\mathbb{E}\|y_{k,c}(t)\|_2^2}{\mathbb{E}\|s_{k,c}\|_2^2} - 1 = \frac{\mathbb{E}\|y_{k,c}(t) - s_{k,c}\|_2^2}{\mathbb{E}\|s_{k,c}\|_2^2} = r_{k,c}^2(t). \tag{14}$$

The rate in Section 3.3 gives $r_{k,c}(t) = m b_{k,c} t / t_k^\star$; signal and noise have equal energy at $t_k^\star/(m b_{k,c})$, not at $t_k^\star$. In decibels, $\mathrm{SNR}_{k,c}(t) = -20 \log_{10} r_{k,c}(t)$. With $m = 10$ and $1 \le b_{k,c} \le 2$, the evaluated $t_{\min} = 10^{-5} t^\star$ corresponds to $+74$ to $+80$ dB, while $t^\star$ corresponds to -26 to -20 dB. For the $6 \to 8$ and direct $6 \to 10$ samplers, $t_{\max} = 10 t^\star$ gives -46 to -40 dB; for $8 \to 10$, $t_{\max} = t^\star$ gives -26 to -20 dB. These are prescribed forward-noise margins, not reconstruction accuracies.

The coarse EDM endpoint fixes the clock scale: $t_{6,\text{end}} = \sigma_{6,\min}/A_{\text{lr}} = 0.02/200 = 10^{-4}$. We use an overlap factor $\kappa_8 = 10$ to place the first detail-band reference at $t^\star_{6\to 8} = \kappa_8 t_{6,\text{end}} = 10^{-3}$, then measure that band's statistics and set $A_{6\to 8,c}$. The direct $6 \to 10$ stage uses the same reference but its own combined-band statistics. In the implemented HPX1024 projected-noise convention, the coarse condition at this handoff has per-pixel noise amplitude $A_{\text{lr}} t^\star \sqrt{d_0/N_{\text{pix}}} = 200(10^{-3})/16 = 0.0125$. Its measured per-pixel signal amplitudes are 0.84–1.31 in normalized units, so its noise-to-signal amplitude ratio stays below 0.015 in every channel. To place a subsequent handoff, we require every preceding band to retain at most a fraction η_{handoff} of noise energy:

$$t^\star_{k+1} \le \sqrt{\eta_{\text{handoff}}} \min_{j \le k,\, c} \frac{\ell_{j,c}}{A_{j,c}}. \tag{15}$$

For the ideal capped-rate $6 \to 8 \to 10$ schedule, checking the intermediate band with $\eta_{\text{handoff}} = 0.04$ and $\max_c b_{8,c} = 2$ gives $t^\star_{8\to 10} = 10^{-2} t^\star_{6\to 8} = 10^{-5}$. The final-band rates then follow from its own measured statistics. The main-text rule Section 3.3 corresponds to $\eta_{\text{handoff}} = 1$.

The endpoints use separate criteria: $r_{k,c}(t_{\min,k}) \le \sqrt{\eta_{\text{clean}}}$ for every channel and $t_{\max,k} = q_k t_k^\star$ with $q_k \ge 1$. Thus $t_{\min,k} \le \sqrt{\eta_{\text{clean}}} \min_c(\ell_{k,c}/A_{k,c})$. The evaluated samplers take $t_{\min,k} = 10^{-5} t_k^\star$, giving $r_{k,c}^2(t_{\min,k}) \le 4 \times 10^{-8}$ under the capped-rate rule, and use $q_k = 10$ for $6 \to 8$ and direct $6 \to 10$ but $q_k = 1$ for $8 \to 10$ (Table 5). These conditions constrain the times without uniquely determining them.

Calibrated constants. Table 4 lists the constants of the $6 \to 10$ band on HPX1024, measured under the operators of Section 3.2. The rates follow the rule in Section 3.3 with $m = 10$ and $t^\star = 10^{-3}$. The margin m and the tail cap $b_{k,c} \le 2$ are shared across all channels and bands; we did not ablate them.

B.3 Training

Each stage trains on 128×128 patches extracted from globally decomposed fields, with a learned position embedding. SongUNet backbones have $\sim$330M parameters. We use Adam at learning rate 10^{-4} with cosine decay over 126k steps and an exponential moving average of weights with

Table 4: Per-channel band standard deviation, per-mode amplitude scaled by the tail factor, and tail factor (normalized units). The effective rate uses $\min(\beta_c, 2)$. Channels: `cllvi`/`clivi` cloud liquid/ice water path (kg m^{-2}), `tas` 2 m temperature (K), `uas`/`vas` 10 m zonal/meridional wind (m s^{-1}), `rlut` outgoing longwave radiation and `rsut` reflected shortwave at TOA (W m^{-2}), `pres_msl` mean sea-level pressure (Pa), `pr` precipitation (kg m^{-2} s^{-1}), `rsds` downwelling surface shortwave (W m^{-2}), `sst` sea-surface temperature (K), `sic` sea-ice fraction (dimensionless).

	cllvi	clivi	tas	uas	vas	rlut	rsut	pres_msl	pr	rsds	sst	sic
$\sigma_{\mathrm{data},c}$	1.43	0.78	0.078	0.224	0.257	0.365	0.429	0.037	4.35	0.218	0.208	0.244
$\ell_c\,\beta_c$	5.28	3.02	0.209	0.413	0.481	0.753	1.08	0.115	16.1	0.559	0.816	1.06
β_c	3.69	3.86	2.68	1.84	1.87	2.06	2.51	3.07	3.70	2.56	3.92	4.33

a half-life of 500 kimg. Training noise is $t \sim \mathrm{LogUniform}(10^{-8}, 10^{-2})$ with $t^\star = 10^{-3}$ for the $6\to10$ and $6\to8$ stages; the $8\to10$ stage uses $t \in [10^{-10}, 10^{-5}]$ with $t^\star = 10^{-5}$. The reference $t^\star$ sets the band noise rate, while $t_{\min}$ and $t_{\max}$ are chosen separately by the endpoint criteria above. The conditioning noise rate is $A_{\mathrm{lr}} = 200$. We also trained a variant of the $6\to8$ stage that represents the intermediate band on its native HPX256 grid; all reported cascade results use the HPX1024 representation.

Each stage trains on 16 H100 GPUs (2 nodes) with a global batch of 1024 patches (64 per GPU), in float32 with TF32 matrix products. Throughput is ≈10k optimizer steps per 4 hours of wall-clock time, so the full $6\to10$ training run (126k steps) costs ≈810 GPU-hours. The cascade's first stage ($6\to8$) uses the same configuration; the $8\to10$ stage is fine-tuned for 17.5k steps from an earlier $8\to10$ model trained for 65k steps.

B.4 Sampling

Algorithm 1 gives the multi-resolution sampler. Here $P_k = P^{a_k}_{a_{k-1}}$ and $P_{\mathrm{coarse},k}$ projects onto the retained coarser content on the stage's working grid. Setting $K = 1$ with $(a_0, a_1) = (6, 10)$ gives the single-stage sampler; $(a_0, a_1, a_2) = (6, 8, 10)$ gives the multi-stage sampler. The working grid may be finer than the band's nominal order, as in the main $6\to8$ stage.

All models use the Heun sampler with 18 steps and 128^2 patches with a 32-pixel overlap, in batches of 128 patches. Averaging overlapping denoiser predictions maintains one global trajectory (Section 3.4). Each model starts sampling at its trained $t_{\max}$: $10t^\star$ for the $6 \to 10$ and $6 \to 8$ stages and $t^\star$ for $8 \to 10$. When stochastic churn (Karras et al., 2022) is enabled, it injects globally projected band noise. Table 5 lists the per-model settings.

Table 5: Inference protocols. Every model starts at its trained window top. The multi-stage SPyD rows describe the HPX1024 implementation.

model	t window	start	S_{churn}	notes
Single-stage SPyD $6\to10$	$[10^{-8}, 10^{-2}]$	10^{-2}	0	$t^\star = 10^{-3}$
Multi-stage SPyD $6\to8$	$[10^{-8}, 10^{-2}]$	10^{-2}	0	$t^\star = 10^{-3}$; output on HPX1024
Multi-stage SPyD $8\to10$	$[10^{-10}, 10^{-5}]$	10^{-5}	20	$t^\star = 10^{-5}$
cBottle-SR	$\sigma \in [0.02, 800]$	800	0	official released protocol

B.5 Evaluation

Evaluation protocol. Spectra, pixel distributions, and extreme percentiles pool one sample per model at each of 32 held-out timestamps, selected quasi-uniformly across the held-out period (403M pixels per channel). Each model receives the coarsened truth from the same timestep; cBottle-SR outputs follow its official protocol at those timestamps.

Paired-error diagnostic. For an ideal sampler, a generated sample and the target are conditionally i.i.d. given the same coarse input. Their expected band MSE then equals twice the mean conditional band variance, which is at most twice the marginal band variance, so the ideal score normalized

Algorithm 1 Multi-resolution SPyD sampling with global multi-diffusion

Require: Coarse field c_0 at order a_0; orders $a_0 < \cdots < a_K$; stage denoisers $\mathcal{D}^{a_k}_{a_{k-1}}$ and decreasing time grids $t_{k,0} > \cdots > t_{k,S_k} = 0$

for $k = 1, \ldots, K$ **do**
  Lift c_{k-1} to the working grid of stage k; call the result c
  Draw global Gaussian maps and project them: $\epsilon_b = P_k\epsilon$, $\epsilon_c = P_{\mathrm{coarse},k}\epsilon'$
  $y \leftarrow A_k t_{k,0}\epsilon_b$ ▷ one global trajectory for the new band
  for $i = 0, \ldots, S_k - 1$ **do**
    Optionally apply globally projected band churn to $(y, t_{k,i})$, giving $(y, \widehat{t})$; otherwise $\widehat{t} = t_{k,i}$
    $v_{\widehat{t}} \leftarrow c + A_{\mathrm{lr},k}\widehat{t}\,\epsilon_c$
    $D \leftarrow \overline{\mathcal{D}}^{a_k}_{a_{k-1}}(y, v_{\widehat{t}}, \widehat{t})$ ▷ patch denoising and overlap averaging, Eq. (9)
    $d \leftarrow (y - D)/\widehat{t}$; $\widetilde{y} \leftarrow y + (t_{k,i+1} - \widehat{t})d$ ▷ Euler predictor
    if $t_{k,i+1} > 0$ **then**
      $v_{\mathrm{next}} \leftarrow c + A_{\mathrm{lr},k}t_{k,i+1}\epsilon_c$
      $D' \leftarrow \overline{\mathcal{D}}^{a_k}_{a_{k-1}}(\widetilde{y}, v_{\mathrm{next}}, t_{k,i+1})$; $d' \leftarrow (\widetilde{y} - D')/t_{k,i+1}$
      $y \leftarrow y + \frac{1}{2}(t_{k,i+1} - \widehat{t})(d + d')$ ▷ Heun corrector
    else
      $y \leftarrow \widetilde{y}$
    end if
  end for
  $c_k \leftarrow c + y$ ▷ retain the previous bands and add the generated band
end for
return c_K

by that variance is at most 2. cBottle-SR scores 2.29 on precipitation, above this bound; the SPyD models score 1.6–1.8.

Ensemble-score interpretation. The single-stage rank histograms deviate by 2–3% per bin for temperature, winds, radiation, pressure, and SST. cBottle-SR's sloped histograms indicate a warm bias in 2 m temperature and a cold bias in fine-scale SST. The cascade's domed histograms are consistent with over-dispersion, possibly from stacking the variance of two stages. Naive upsampling's lowest ensemble-mean RMSE on precipitation and downwelling shortwave (Table 2) shows how much deterministic accuracy comes from the shared coarse input; CRPS separates the generators better on these channels.

C Global Fields for the Remaining Channels

Figures 5 and 6 show the remaining ten channels in the format of Figure 2, for a different held-out state (2024-03-06T12:30). As in Figure 2, all three models share the same coarse conditioning, so each panel compares the same weather.

D Per-Noise-Level Generalization Diagnostic

Brenowitz et al. (2025) report high-noise overfitting in the cBottle coarse global generator and argue that super-resolution is less susceptible because it learns more local structure. We test whether single-stage SPyD overfits its training period with a per-noise-level diagnostic rather than an aggregate loss. For each channel, we compute the ratio of training to validation denoising loss at fixed noise levels, using 6 global states per split and 256 patches per state, for checkpoints spanning the training run (Figure 7).

The channel-mean ratio stays between about 0.9 and 1.0 across checkpoints and noise levels (Figure 7, left). The per-checkpoint minimum over channels stays between about 0.8 and 1.0 at noise levels near and above the handoff time $t^\star$. At the smallest noise levels, four to five decades below $t^\star$, where

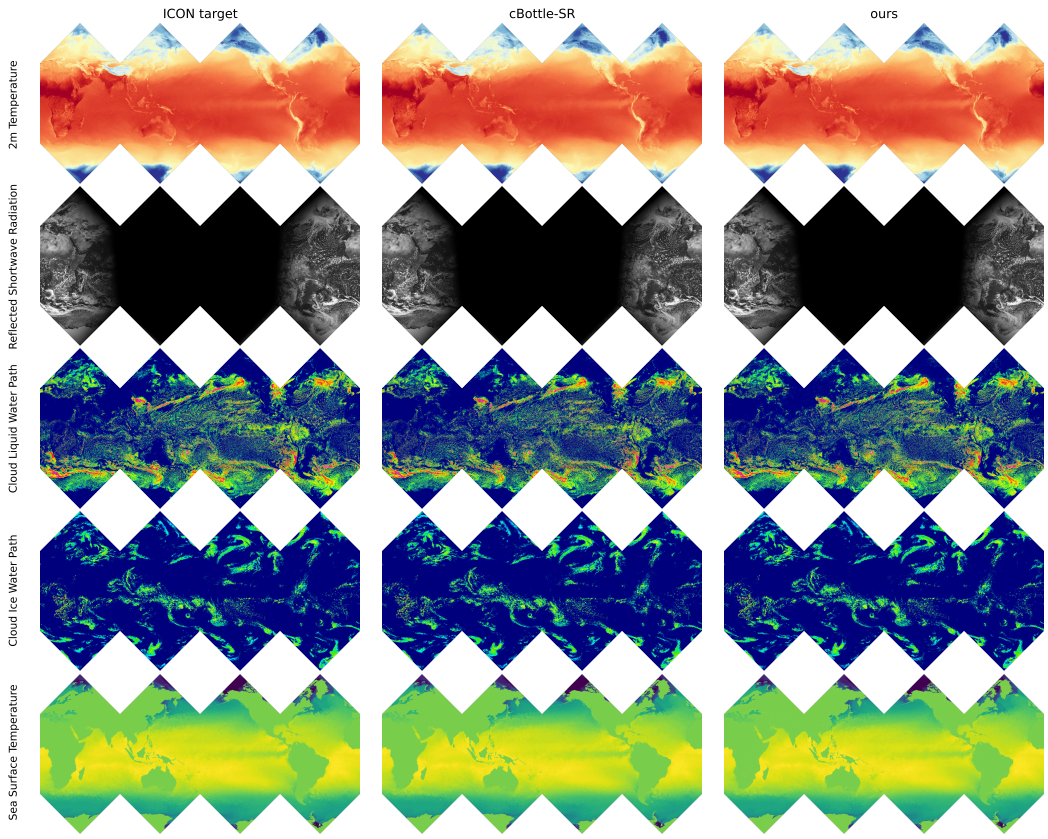

Figure 5: Global fields for one shared held-out conditioning (2024-03-06T12:30:00), rendered as a flat map of the full HEALPix sphere: 2 m temperature, reflected shortwave, cloud liquid and ice water paths (decibel scale), and SST; ICON (left), cBottle-SR (center), single-stage SPyD (right).

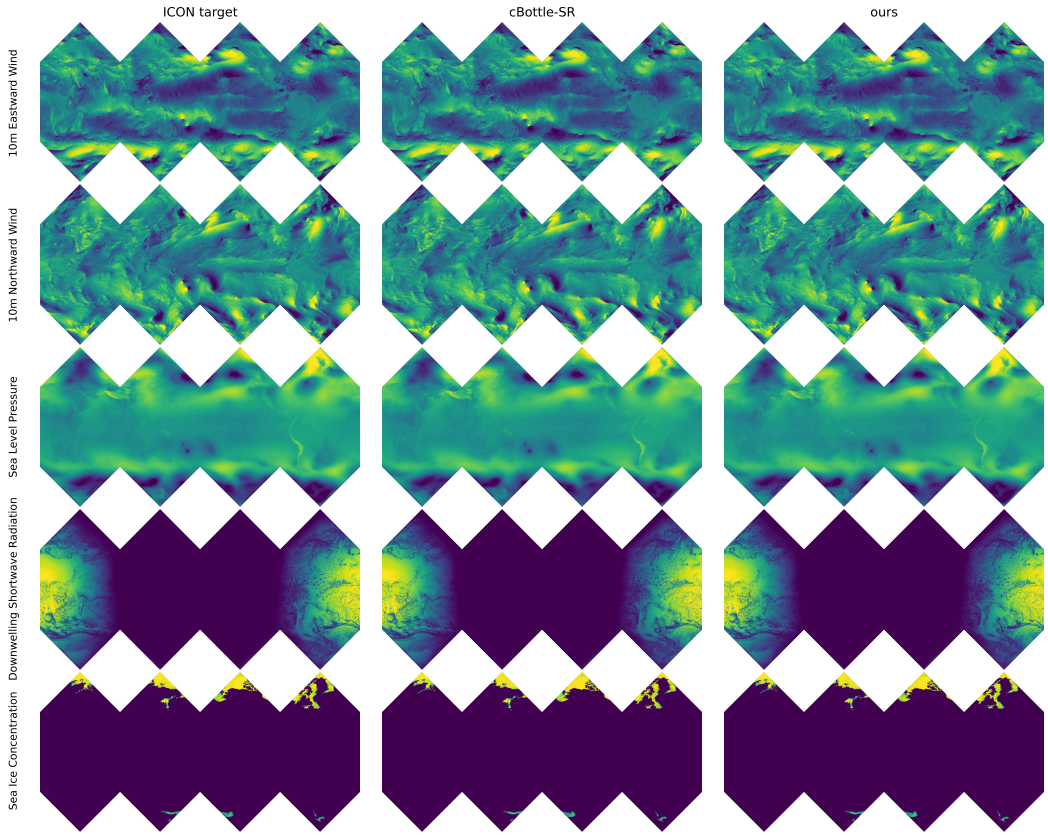

Figure 6: As Figure 5, for 10 m winds, sea-level pressure, downwelling shortwave, and sea-ice fraction.

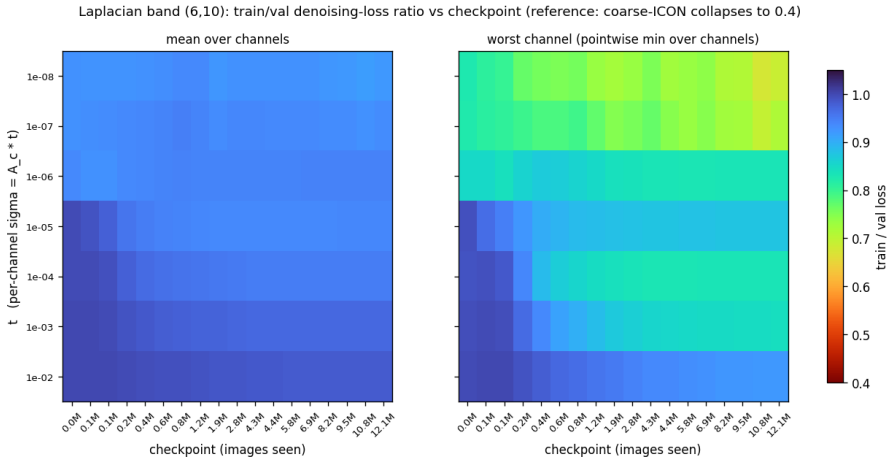

Figure 7: Train/validation denoising-loss ratios for the $6 \to 10$ model across noise levels and training checkpoints (6 global states per split, 256 patches per state). Left: channel mean; right: minimum across channels. The horizontal axis reports training images seen.

the band is signal-dominated, it decreases with training and reaches about 0.7 at the final checkpoint; the channel attaining this minimum varies across checkpoints (Figure 7, right).

The ratio of training to validation band variance reproduces the offsets from one to within a few percent per channel, including channels whose offset changes sign because their training period is the more variable one. After dividing out this variance ratio, the loss ratios are ≈ 1.0 at all noise levels above the handoff time. The same diagnostic applied to cBottle-SR gives offsets that vary little with noise level.

E Additional Storm Cases

We selected the four deepest storms in the simulation by minimum track sea-level pressure. For each, we condition all models on the same coarse input and show precipitation, outgoing longwave radiation, mean sea-level pressure, and 10 m wind in a window centered on the storm. Figure 8 shows the strongest tropical cyclone, and Figures 9 to 11 show the next three: a western-Pacific typhoon, a southern-hemisphere cyclone, and an extratropical comma system. In all four cases, the SPyD models reproduce the eyewall and rainband organization where cBottle-SR saturates the convective core, and all models capture the pressure and wind fields accurately. All models also share one deficiency in every case: their fine-scale structure is isotropic where ICON has shear-aligned filaments.

F Full Generative Pipeline Evaluation

We draw 16 samples from the cBottle coarse model, spanning all seasons, and use them in place of coarsened ICON fields as conditioning. The samples are lifted to the SPyD conditioning space with the phase-aligned synthesis of Appendix A, while cBottle-SR receives the same fields through its native pathway. Because generated coarse fields have no paired ICON target, we compare 16-sample averages against unpaired held-out ICON statistics (Figure 12). To measure rain that a model generates where the reference has almost none, we also define a *spurious-cell rate*: the fraction of pixels exceeding $0.1\,\mathrm{mm\,hr^{-1}}$ inside quiescent regions, taken as the quarter of 128^2 patches with the lowest reference-field standard deviation. ICON itself scores 0.09% on this metric.

Under this protocol, the single-stage SPyD precipitation spectrum stays within 3% of the reference, and its spurious-cell rate rises from 0.4% to 0.5–1.2%. cBottle-SR's mesoscale excess persists and grows to $+36\%$ at fine scales. Multi-stage SPyD produces 20–22% spurious convective cells under the same conditioning, consistently across the two lifting operators described below, while its

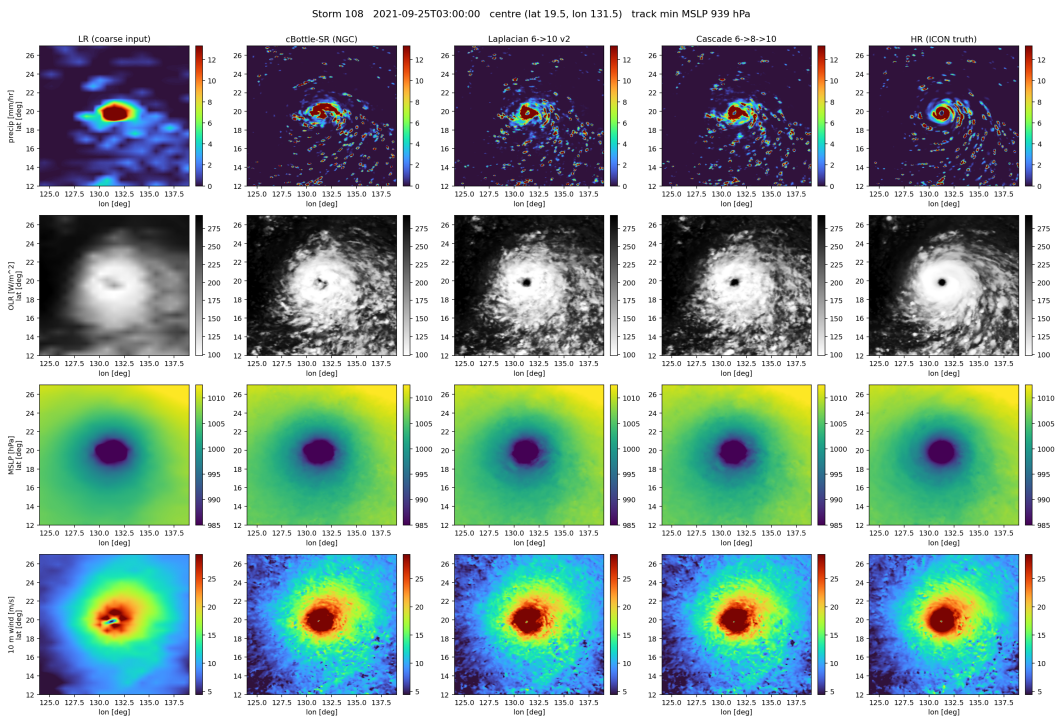

Figure 8: Storm-centered comparison for the strongest tropical cyclone in the simulation (rows: precipitation, OLR, MSLP, 10 m wind; columns: coarse input, cBottle-SR, single-stage SPyD, multi-stage SPyD, ICON target). Both SPyD models resolve the eye and eyewall; cBottle-SR saturates the core.

intermediate output scores 0.62%; the fine stage therefore raises the rate by about $35\times$. Training the fine stage on generated conditioning is a natural way to reduce this sensitivity.

We test two lifting operators because the coarse model and the SPyD stages do not share a multiresolution basis: the coarse model is trained under block averaging, a Haar (db1) coarsening, while the SPyD stages use the db4 coarsening of Section 3.2. The phase-aligned wavelet lift transmits texture artifacts from the coarse model. The bilinear lift does not suppress the cascade's amplification, and it removes power near the conditioning band edge that the SPyD stages cannot regenerate, which appears as a dip in the spectrum at that edge. cBottle-SR is less sensitive to the choice of lift, consistent with its weaker constraint to preserve coarse input structure. The SPyD models instead retain the coarse input and inherit what the coarse model supplies: under generated conditioning, their fine-scale temperature power falls to 0.55 of the reference because the coarse model undersupplies it. Their pipeline fidelity therefore depends on the coarse generator; training the coarse generator in the same wavelet basis could reduce this mismatch, and we leave it to future work.

Reproducibility Statement

Sections 3.1 to 3.4 specify the model, and Appendix A gives the construction of the spherical wavelet operators together with their measured numerical residuals. Appendix B gives the calibration rule, the calibrated per-channel constants (Table 4), the training configuration and compute, the sampling algorithm (Algorithm 1) with per-model inference settings (Table 5), and the evaluation protocols. All experiments use the publicly available nextGEMS ICON Cycle 3 simulation (Koldunov et al., 2023) and the released cBottle-SR and cBottle coarse-generator checkpoints (Brenowitz et al., 2025). We will release our code upon publication.

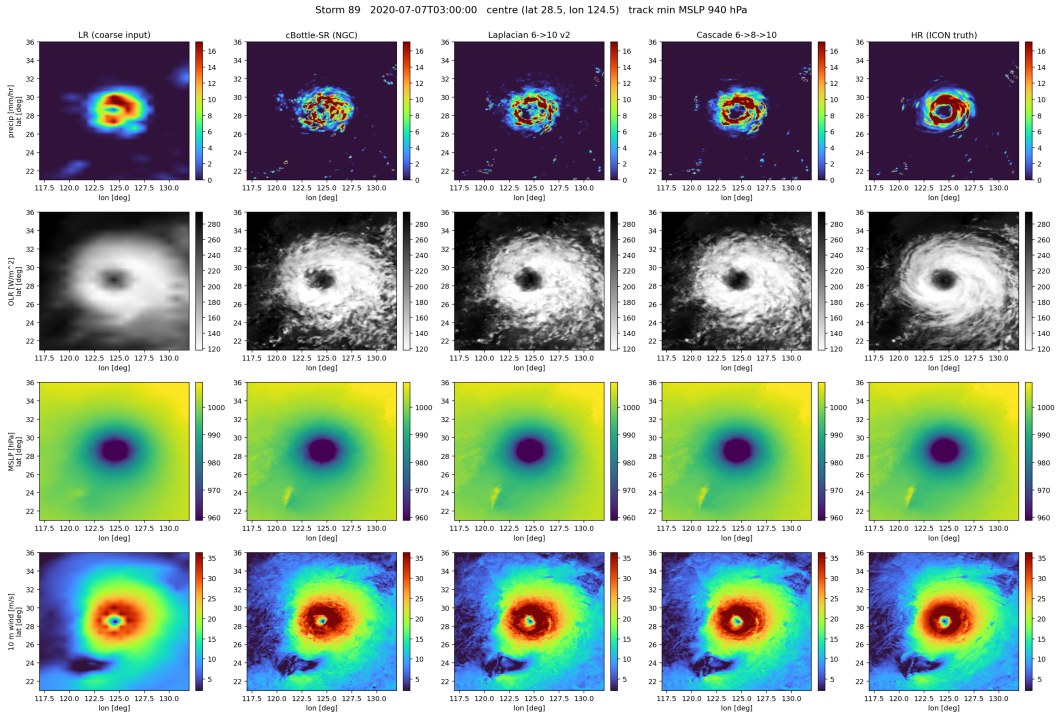

Figure 9: As Figure 8, for a western-Pacific typhoon.

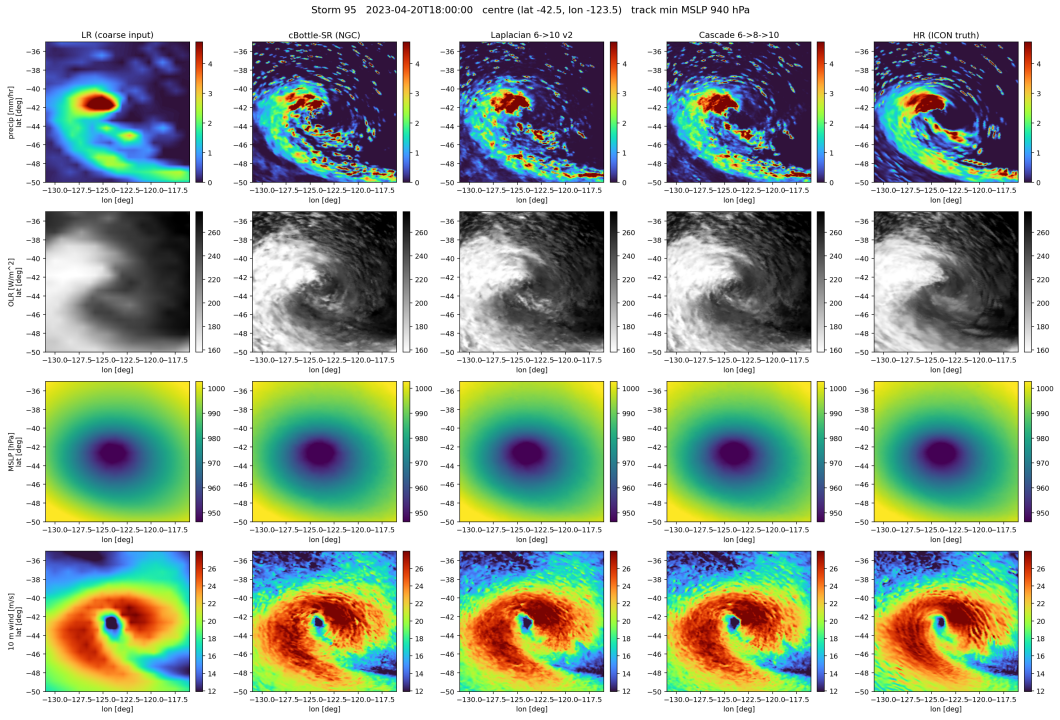

Figure 10: As Figure 8, for a southern-hemisphere cyclone.

AI USE DISCLOSURE

Generative AI tools assisted with editing and polishing the manuscript's writing, including its organization and readability. The authors did not use Generative AI tools for main ideation.

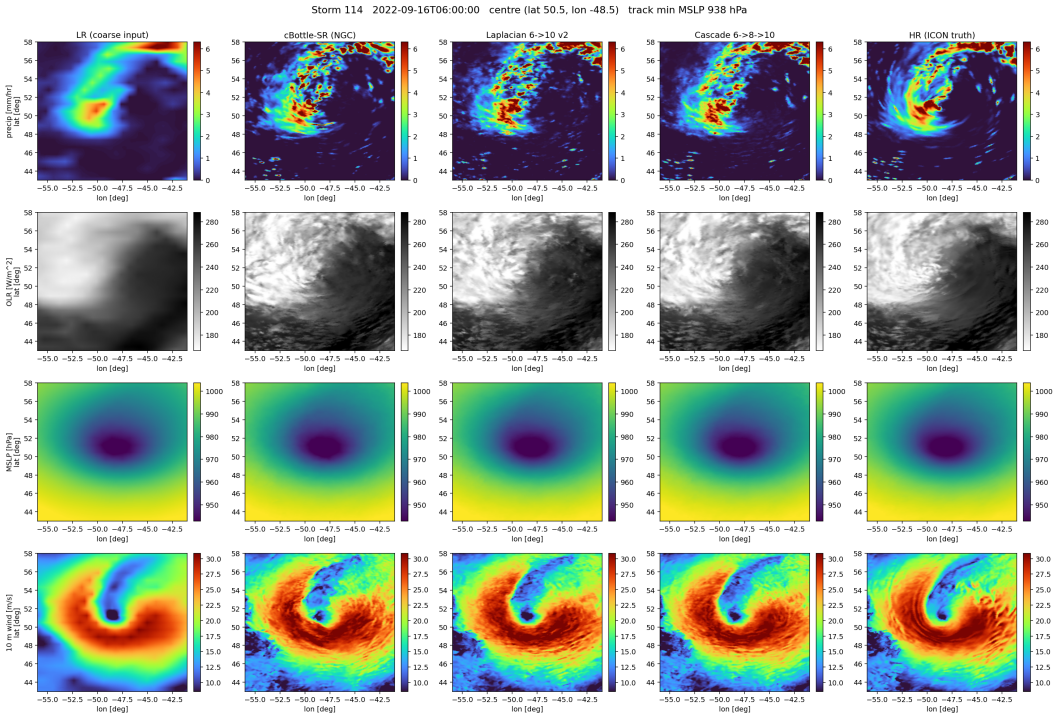

Figure 11: As Figure 8, for an extratropical comma system.

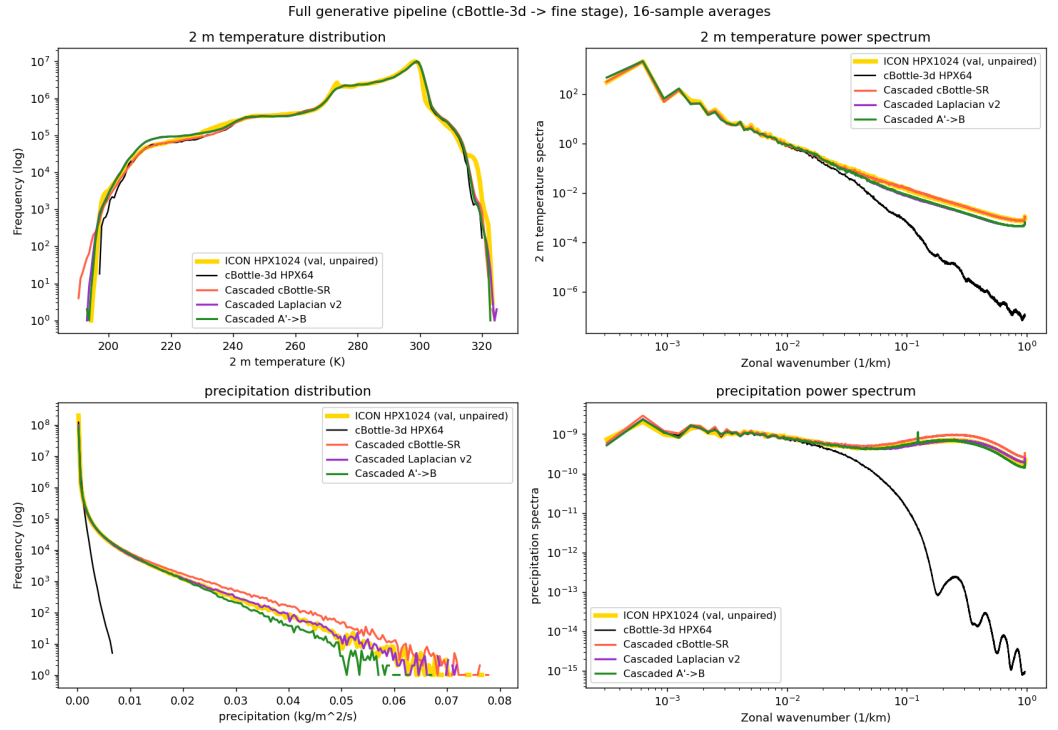

Figure 12: Full generative pipeline (cBottle coarse model → fine stage), 16-sample averages against unpaired held-out ICON statistics. The single-stage SPyD precipitation spectrum stays on the reference through the complete pipeline; cBottle-SR retains excess fine-scale power.