Tackling Climate Change with Machine Learning • TCCML @ NeurIPS 2026

Kaikwan Lau **Gary P. T. Choi**

Department of Mathematics, CUHK
With support from the NVIDIA AI Technology Center

What Makes Diffusion-Based Weather Downscaling Work?

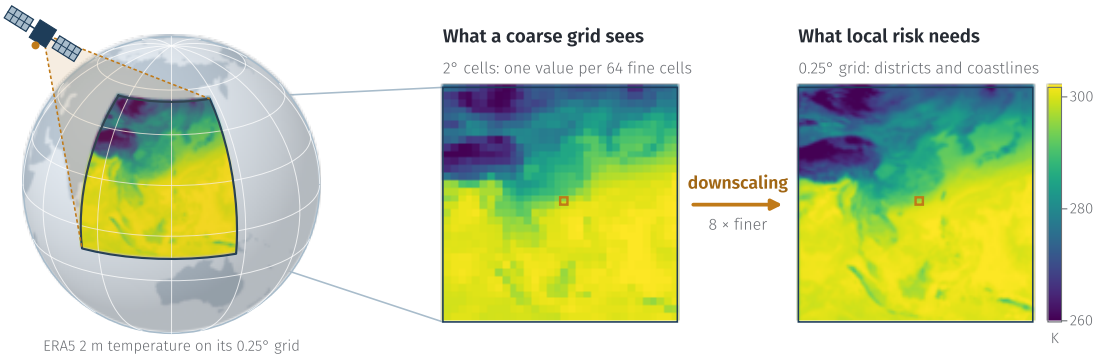
A Controlled Comparison of CorrDiff- and Patch-DM-Style Models

香港中文大學
The Chinese University of Hong Kong

NVIDIA

NEURAL INFORMATION PROCESSING SYSTEMS

Global weather data are too coarse for local risk

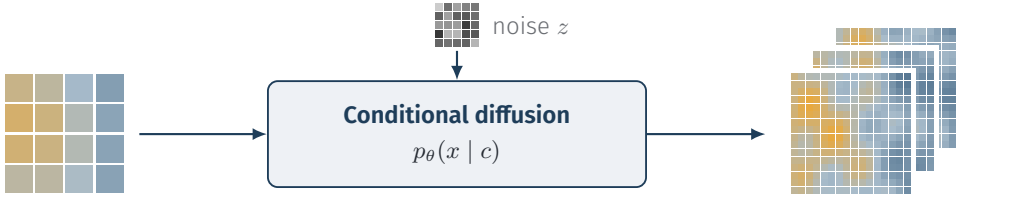


One coarse value has to stand for 64 fine grid cells.

Global models resolve tens of kilometres, while heat, wind and flood risk play out locally.

Coarse weather admits many finer fields

We want a conditional ensemble of temperature and wind fields, given a coarse atmospheric input.



Different noise draws represent unresolved spatial variability.

Research question

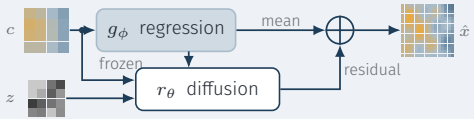
Does the best diffusion formulation depend on whether it generates a **full field** or a **residual**?

Paper: Introduction and Section 2. Diagram is conceptual.

Two generation targets, one shared framework

CorrDiff-style

A frozen regressor supplies the mean. Diffusion generates a residual.

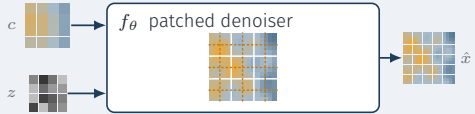


$$\hat{x} = g_\phi(c) + r_\theta(z, c, g_\phi(c))$$

- Full-domain denoising
- Two trained networks

Patch-DM-style

A patched denoiser generates the full field, sharing features across patches.



$$\hat{x} = f_\theta(z, c)$$

- Shifted feature collage; global context
- One trained network

Within each family: compare ϵ**-DDPM** with **EDM**.

Architecture and target remain coupled in the main comparison.

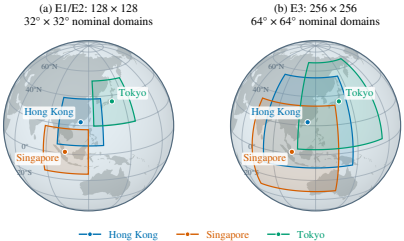
Paper: Table 1 and Section 2. These are style-matched research implementations. Schematics follow the equations; icons are conceptual.

A shared ERA5 benchmark tests three settings

Perfect-model task: coarsen ERA5, then reconstruct its original $0.25°$ grid.

	Input	Output	Years
E1 (8×)	8^2	128^2	2018–20
E2 (16×)	16^2	128^2	2018–20
E3 (8×)	8^2	256^2	2019–20

T_{2m}, u_{10m}, v_{10m}, every 6 hours
Chronological split: 80 / 10 / 10%



Overlapping domains centred on Hong Kong, Singapore and Tokyo.

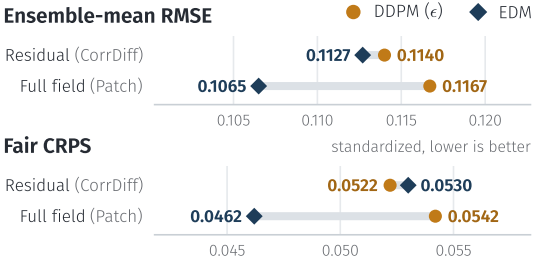
Common training and evaluation

Shared U-Net family and data pipeline. 50,000 steps per trained network; 16 ensemble members per input.

Paper: Figure 1, Section 2 and Appendix A. E3 also changes years, domain, batch size and patch configuration.

CRPS ranking depends on the configuration

At E1, EDM improves RMSE in both families. The fair-CRPS ordering reverses.



The target matters.
Residual target: **DDPM** has lower CRPS.
Full-field target: **EDM** has lower CRPS.

E1 examples: single members, one test date

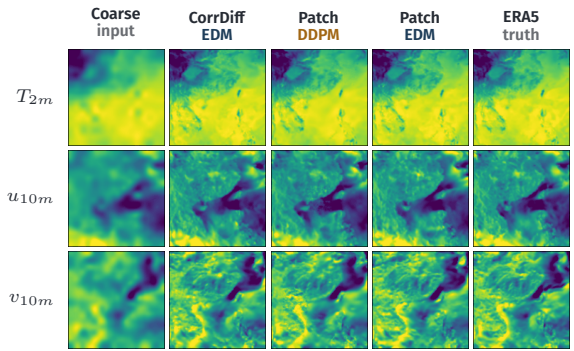
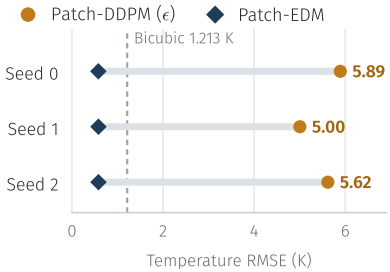
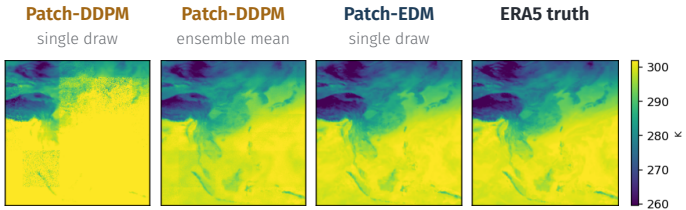


Chart: Table 2, E1 (8× to 128²), means of three trained and evaluated runs. Maps: camera-ready Appendix K, E1 seed 0, one member per model; row colour scales not shown.

Noise prediction fails at the larger setting

At E3, patched full-field ϵ-DDPM fails in all three runs despite low late training losses.





Patch-aligned offsets in the single sample.

Mean temperature RMSE

5.505 ± 0.454 K for Patch-DDPM versus **0.581 ± 0.001 K** for Patch-EDM.

Paper: Table 13, Figure 2 and its ensemble-mean panels (cropped). Right: E3 Hong Kong, seed 0, first test frame, one colour scale.

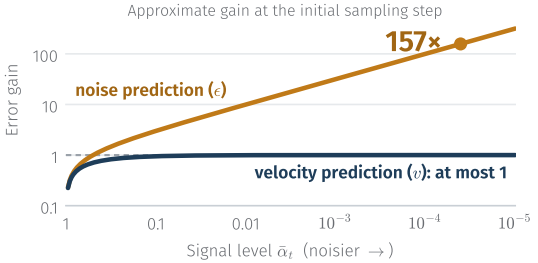
Noise errors can amplify during reconstruction

At high noise, the clean signal coefficient becomes very small.

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$$

$$\hat{x}_0 - x_0 = -\underbrace{\sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}}_{\text{error amplification}} \delta$$

Noise-prediction error: $\hat{\epsilon} = \epsilon + \delta$.



A testable explanation

Velocity prediction has a reconstruction-error gain bounded by one.
This motivates an intervention; it does not identify the entire failure mechanism.

Paper: Appendix G. Reconstruction identity and tested schedule; the gain is a sensitivity calculation. Chart: the identity over $\bar{\alpha}_t$, no schedule assumed.

Velocity prediction repairs temperature accuracy

Retrain the same patched architecture with a velocity-prediction objective.

Full-field model	NFE	T RMSE (K)	CRPS$_z$	SSR
Patch-DDPM (ϵ)	50	**5.892**	.1623	1.49
Patch-v	50	**.597**	.0467	.75
Patch-EDM	99	.580	.0454	.94

Fixed: architecture, noise schedule, DDIM grid, training steps and sampling budget.

What the intervention changes

Both reconstruction conditioning and loss weighting. Temperature improves, while the reported spread–skill ratio remains below one.

Paper: Table 3 and Appendix G. E3 seed 0 only. Bars are to scale. Diffusion velocity is distinct from northward wind.

More DDIM steps do not remove the failure

More sampling effort

Patch-DDPM at E3, seed 0
50 to 100 DDIM steps

5.89 K → 5.79 K

Temperature failure remains.

Less sampling effort

Patch-EDM at E1, seed 0
50 to 25 Heun steps

99 → 49 NFE

RMSE_z remains 0.1065.



DDIM Patch-DDPM — 50 steps → **50 NFE**

Heun Patch-EDM — 50 steps → **99 NFE**

■ one network evaluation

Heun's last step uses one

Equal solver steps do not mean equal compute.
Default NFE differences are a budget choice, not an inherent efficiency ranking.

Paper: Section 3 and Appendix F. NFE = network function evaluations. Controls apply to the recorded checkpoints. Squares show the per-step pattern.

Reliable generation needs more than low loss

Prediction parameterization matters

Low late training loss can coexist with inaccurate generation.

The controlled repair is the central result

At E3 seed 0, temperature RMSE falls from **5.89 K** to **0.597 K**.

Thank you.

We acknowledge NVAITC support and the CUHK Vice-Chancellor Early Career Professorship Research Startup Fund.

香港中文大學
The Chinese University of Hong Kong

NVIDIA

NEURAL INFORMATION PROCESSING SYSTEMS



Paper: Conclusions and acknowledgments.