MUNDO: A MULTIRESOLUTION GENERATIVE FRAMEWORK FOR TEMPORAL DOWNSCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamical systems are often observed at coarse temporal resolution through sparse states or temporal aggregates, leaving the underlying fine-scale trajectory unresolved. We present MUNDO (**M**ultiresolution **UN**ified **DO**wnscaling), a generative framework for temporal downscaling. MUNDO builds on engression, which trains a conditional sampler with a strictly proper scoring rule and generates a sample in a single network evaluation, without iterative diffusion-based sampling. MUNDO extends this to a hierarchy of resolutions by factorizing the observation operator into successive coarsenings. The resulting conditionals are then sampled in reverse, with each level generating only the degrees of freedom lost between adjacent resolutions. We evaluate MUNDO on two-scale Lorenz-96, four 2D physical systems, and multivariate ERA5 reanalysis over Europe, for both temporal interpolation and disaggregation. Across datasets and tasks, MUNDO generally outperforms flow-based baselines. Single-step MUNDO achieves strong predictive skill, while recursive MUNDO generally improves ensemble calibration, with better spread-skill agreement and predictive coverage.

1 INTRODUCTION

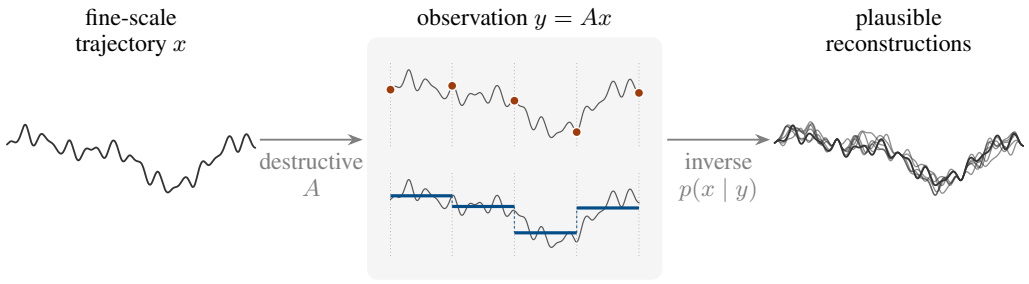

Figure 1: **Coarse observations of a dynamical system define a one-to-many inverse problem.** Subsampling (orange dots) yields interpolation, while averaging (blue bars) yields disaggregation. In both cases, the reconstruction target is the conditional law $p(x \mid y)$.

A dynamical system is rarely observed at the temporal resolution at which it evolves. Numerical models integrate at fine time steps but typically archive only a sparse subset or an aggregate of states, as storing the full trajectory is prohibitively expensive. This is especially common in weather and climate: climate simulations often retain daily or monthly means of sub-hourly dynamics (Eyring et al., 2016), while data-driven weather forecasters typically predict states at six-hour intervals (Lam et al., 2023; Kurth et al., 2023). The same loss of fine-scale temporal information occurs in other simulated and observed physical systems, including fluid and wave dynamics. Temporal downscaling reconstructs the missing information at a fraction of the cost of rerunning or training models (Brenowitz et al., 2025). It covers, often without distinction, *interpolation*, where sparse states are observed, and *disaggregation*, where aggregates are given (Figure 1). Both are instances of the same inverse problem: recovering a fine trajectory from an incomplete observation of it.

For chaotic, multiscale dynamics, this inverse problem is one-to-many: endpoints do not determine the trajectory between them, and an aggregate does not determine the fine-scale variability

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.

that produced it. Therefore, the target is a distribution over trajectories whose samples are consistent with the observation. Learned temporal downscalers are mostly deterministic and optimized for pointwise error (Harilal et al., 2023; Wang et al., 2025; Leinonen et al., 2024), so they return one reconstruction, the conditional mean, and none of the range compatible with the observation. Probabilistic methods represent that range but enforce the observation only indirectly, through conditioning or penalty terms (Ingstad et al., 2026; Bassetti et al., 2024), so a draw need not satisfy it. Posterior samplers accept arbitrary linear operators and enforce the observation through repeated data-consistency updates (Kim et al., 2025). Among these methods, none returns draws that satisfy the observation exactly under both operators without an iterative solve.

This paper presents MUNDO (**M**ultiresolution **UN**ified **DO**wnscaling), a conditional sampler for temporal downscaling under linear observation operators. MUNDO builds on engression (Shen & Meinshausen, 2025), which learns a conditional sampler with a strictly proper scoring rule and generates each draw in a single network evaluation. Such a sampler must recover all unresolved scales at once. Temporal downscaling, however, decomposes naturally across resolutions: a large temporal coarsening can be expressed as a composition of smaller coarsenings. We exploit this structure through reverse Markov learning (Shen et al., 2025), which turns the decomposition into a hierarchy of conditional samplers. Each level generates only the degrees of freedom lost between adjacent resolutions. We parameterize these as residuals in the kernel of the corresponding observation operator, so every level remains exactly consistent with its parent.

Our contributions are:

- a conditional sampler for linear observation operators, parameterized as a classical reconstruction plus a learned null-space residual. The same formulation applies to interpolation and disaggregation, enforces the observation exactly without penalty terms or iterative projection, and remains substantially cheaper to sample from than iterative diffusion- or flow-based solvers;
- a temporal multiresolution factorization based on reverse Markov learning, which factors the observation operator into successive coarsenings and samples the corresponding conditionals in reverse;
- a broad evaluation on chaotic systems, simulated physical dynamics, and multivariate ERA5 reanalysis, measuring probabilistic skill, calibration, observation consistency, and temporal structure, and comparing multiresolution sampling with direct generation of the full fine-scale trajectory.

The paper is organized as follows. Section 2 reviews related work on temporal downscaling. Section 3 states the inverse problem and the target conditional law, then describes the method. Section 4 describes the systems, baselines, and evaluation protocol, and reports the results.

2 Related work

Statistical temporal downscaling. Deterministic diurnal profiles (Förster et al., 2016; Han et al., 2023) require daily statistics such as extremes and smooth day-to-day variability (Debele et al., 2007). Denton–Cholette (Denton, 1971; Cholette, 1984) returns the smoothest record compatible with each aggregate. Quantile mapping and CDF-t (Déqué, 2007; Michelangeli et al., 2009) adjust marginals but leave chronology and aggregates unconstrained. Inflation rescales variance, while randomization adds variability with the temporal structure of a noise model (von Storch, 1999). Stochastic weather generators (Richardson, 1981; Ailliot et al., 2015; Wilks, 2010), analogs and fragments (Zorita & von Storch, 1999; Buishand & Brandsma, 2001; Pierce et al., 2014), and multiplicative cascades (Schertzer & Lovejoy, 1987; Lombardo et al., 2017) generate variability from parametric models, finite archives, or scaling laws for intermittent positive variables, respectively.

Learned temporal downscaling. Supervised temporal downscaling includes disaggregation (Kajbaf et al., 2022) and methods inspired by video frame interpolation (Jiang et al., 2018): STint (Harilal et al., 2023), TemDeep (Wang et al., 2025), and ModAFNO (Leinonen et al., 2024) reconstruct intermediate states under pointwise losses. Squared error targets the conditional mean, with the variance deficit in Equation (7). Generative super-resolution moved from adversarial networks (Leinonen et al., 2021; Stengel et al., 2020) to diffusion models, which have been used to reconstruct daily fields from monthly means (Bassetti et al., 2024) and to generate hourly ensembles from

six-hourly climate simulations (Schmidt et al., 2025). HourGlass (Ingstad et al., 2026) reconstructs hourly states between six-hourly bounds using fair CRPS and auxiliary penalties, without addressing aggregate observations.

Inverse-problem solvers. A separate line of work imposes observations on an unconditional generative prior during sampling. DDNM preserves the measurement-determined component while refining null-space content during reverse diffusion (Wang et al., 2023). FlowDPS injects likelihood-gradient updates into flow-matching dynamics (Kim et al., 2025). Both accommodate general linear operators without task-specific training, enforcing observations through iterative sampling updates rather than the generator parameterization.

3 Problem and Method

3.1 Setting and Observation Operator

Let $X_i \in \mathbb{R}^d$ denote the instantaneous state of the system and δ the target temporal resolution. A window holds N steps of length δ, and its fine record $x = (x_1, \dots, x_N) \in \mathcal{X} \subseteq \mathbb{R}^{Nd}$ is a draw from μ, the joint law of N consecutive states at resolution δ, whose density is unknown but from which we hold samples. We write x for both the random record and one of its realizations, context resolving the ambiguity.

The fine record is never observed directly, only through an observation operator that degrades it,

$$y = f(x), \qquad f : \mathbb{R}^{Nd} \to \mathbb{R}^m, \quad m < Nd, \tag{1}$$

where y is the coarse record. We take f linear and identify it with its matrix $A \in \mathbb{R}^{m \times Nd}$, $f(x) = Ax$, assumed to have full row rank m without loss of generality. Then f is surjective onto $\mathbb{R}^m$, with right inverse $A^\dagger = A^\top (AA^\top)^{-1}$, but not injective: its kernel has dimension $\dim \ker A = Nd - m > 0$. Given y, the *compatible set* is the preimage

$$\mathcal{X}_y \;=\; f^{-1}(\{y\}) \;=\; \left\{\, x \in \mathbb{R}^{Nd} \;:\; Ax = y \,\right\} \;=\; A^\dagger y \;+\; \ker A, \tag{2}$$

an affine subspace directed by $\ker A$. In particular, the true record splits as

$$x \;=\; A^\dagger y \;+\; r, \qquad r \;=\; (I - A^\dagger A)\, x \;\in\; \ker A, \tag{3}$$

into a component fixed by the observation and a residual orthogonal to it, which the observation leaves entirely undetermined: the kernel is exactly the information that f destroys. Equation (3) is the range-null space decomposition, $A^\dagger y$ carrying the range-space content and r the null-space content. It underlies null-space learning for inverse problems (Schwab et al., 2019), and is used to enforce data consistency in diffusion-based image restoration (Wang et al., 2023).

In this work, f coarsens the record in time: x is observed only at a coarser step $\Delta = k\delta$, with $k \in \mathbb{N}^*$ dividing N, so that $n = N/k$ coarse steps remain and $m = nd$, through one of two operators

$$(A_{\text{sub}}\, x)_j = x_{jk}, \qquad\qquad (A_{\text{avg}}\, x)_j = \frac{1}{k} \sum_{i \in B_j} x_i, \qquad j = 1, \dots, n, \tag{4}$$

where $B_j = \{(j-1)k + 1, \dots, jk\}$ is the j-th block of k fine steps. Reconstruction from A_{sub} is interpolation, and from A_{avg} disaggregation. Each coarse value draws on its own block, so that $AA^\top = c\, I_{nd}$, with $c = 1$ for subsampling and $c = 1/k$ for block averaging, and $A^\dagger = c^{-1} A^\top$. The kernels,

$$\begin{aligned} \ker A_{\text{sub}} &= \{\, u : u_{jk} = 0,\ j = 1, \dots, n \,\}, \\ \ker A_{\text{avg}} &= \left\{\, u : \textstyle\sum_{i \in B_j} u_i = 0,\ j = 1, \dots, n \,\right\}, \end{aligned} \tag{5}$$

both of dimension $(N - n)d$, are the sub-Δ variability the operators discard: the values between retained instants in one case, the fluctuation about the block mean in the other.

3.2 AN ILL-POSED INVERSE PROBLEM

Recovering x from y alone is ill-posed in the sense of Hadamard (1902). By surjectivity, a solution exists for every $y \in \mathbb{R}^{nd}$, but it is never unique, since $\mathcal{X}_y$ is an affine subspace of dimension $(N - n)d$ rather than a singleton (Engl et al., 1996). This non-uniqueness prevents any deterministic reconstruction from representing the full set of trajectories compatible with y. A deterministic rule $g : \mathbb{R}^{nd} \to \mathbb{R}^{Nd}$ must therefore select a single representative of $\mathcal{X}_y$. Its kernel component is a deterministic function of y, so the conditional law of its output given y is a Dirac mass and cannot represent the variability among trajectories compatible with the same observation.

3.3 THE VARIANCE DEFICIT OF THE CONDITIONAL MEAN

Minimizing expected squared error returns the conditional mean $\mathbb{E}[x \mid y]$. Since f is linear, the conditional mean remains exactly consistent with the coarse observation, as

$$f(\mathbb{E}[x \mid y]) = \mathbb{E}[f(x) \mid y] = y. \tag{6}$$

However, its variance is systematically deficient. Let $\varphi : \mathbb{R}^{Nd} \to \mathbb{R}$ be a linear functional of the fine record. The law of total variance gives

$$\mathrm{Var}[\varphi(\mathbb{E}[x \mid y])] = \mathrm{Var}[\varphi(x)] - \mathbb{E}[\mathrm{Var}[\varphi(x) \mid y]], \tag{7}$$

where the last term is the variability in $\varphi(x)$ left unresolved by the coarse observation. The conditional mean therefore underestimates the variance of $\varphi(x)$ by exactly that term. This deficit is a consequence of the quadratic criterion, not of any particular model class. It vanishes when $\varphi(x)$ is determined by y, and reaches $\mathrm{Var}[\varphi(x)]$ when $\mathbb{E}[\varphi(x) \mid y]$ is constant.

3.4 PROBABILISTIC FORMULATION

The limits of deterministic reconstruction motivate a different target. Rather than selecting a single member of $\mathcal{X}_y$, we seek the conditional law of fine records given the coarse observation. Under the law μ, we define

$$p(x \mid y) \;=\; \mu\left(x \;\middle|\; f(x) = y\right), \tag{8}$$

the conditional law of fine records given the coarse observation[1]. Reconstructing then means *sampling* $p(\cdot \mid y)$: at fixed y, generating multiple compatible records whose dispersion reflects the variability left unresolved by the observation.

3.5 MULTIRESOLUTION SAMPLING

Sampling the conditional law in (8) amounts to generating the residual in (3). For any right inverse $A^\dagger$ of A, every record compatible with y can be written as a baseline reconstruction plus an element of $\ker A$. MUNDO therefore draws

$$\hat{x} = \underbrace{A^\dagger y}_{\text{classical reconstruction}} + \underbrace{\left(I - A^\dagger A\right) G_\theta(y, \varepsilon, c)}_{\text{residual} \in \ker A}, \tag{9}$$

where G_θ is a neural network, c denotes additional conditioning, and, following engression (Shen & Meinshausen, 2025), randomness enters through ε, with one network evaluation per draw. Since $AA^\dagger = I$, the affine map $P_y(u) = A^\dagger y + (I - A^\dagger A)u$ is a projection onto $\mathcal{X}_y$. Thus, $\hat{x} = P_y(G_\theta(y, \varepsilon, c))$ satisfies $A\hat{x} = y$ by construction, without penalty terms. The residual head is initialized at zero, so an untrained level returns the baseline reconstruction $A^\dagger y$. MUNDO does not prescribe the architecture of G_θ; Section 4 specifies the architecture used for each system.

Multiresolution factorization. Following reverse Markov learning (Shen et al., 2025), we extend the single conditional sampler across a hierarchy of temporal resolutions. Both observation operators factor into a chain of full-row-rank linear coarsenings,

$$A = A_R \circ \cdots \circ A_1, \qquad x^{(r)} = A_r\, x^{(r-1)}, \qquad x^{(0)} = x, \quad x^{(R)} = y. \tag{10}$$

[1] Since the conditioning event has probability zero, $\mu(\cdot \mid f(x) = y)$ is understood as the disintegration of μ along f, defined for $f_{\#}\mu$-almost every y and satisfying $\mu = \int p(\cdot \mid y)\, \mathrm{d}(f_{\#}\mu)(y)$, with $p(\cdot \mid y)$ supported on $\mathcal{X}_y$.

Each $x^{(r)}$ is thus a record at an intermediate resolution, and each A_r defines an inverse problem of the same form as Equation (1), with its own compatible set and kernel. Because every level deterministically determines all coarser levels, conditioning on $x^{(r+1)}$ also fixes $x^{(r+2)}, \ldots, x^{(R)}$. The conditional law therefore factorizes along the hierarchy,

$$p\left(x \mid y\right) = \int \prod_{r=0}^{R-1} p\left(x^{(r)} \mid x^{(r+1)}\right) \mathrm{d}x^{(1)} \cdots \mathrm{d}x^{(R-1)}. \tag{11}$$

This factorization is exact: a single conditional spanning coarsening ratio k is replaced by a sequence of smaller conditional problems of the same form. MUNDO samples these conditionals from coarse to fine, applying Equation (9) with A_r when generating $x^{(r-1)}$ from $x^{(r)}$. Every generated level therefore reproduces its parent exactly, and consistency propagates to the original coarse observation. Approximation enters only through the learned conditional factors, not through the factorization itself. We refer to the $R = 1$ instance as single-step MUNDO and to instances with $R > 1$ as recursive MUNDO. In recursive MUNDO, each level can also condition on the last state of the preceding window at its own resolution (Appendix A.4).

Training objective. Learning a conditional sampler requires an objective whose population optimum is the target distribution. We use the energy score, the strictly proper scoring rule used in engression (Gneiting & Raftery, 2007; Shen & Meinshausen, 2025). For an ensemble of M independent draws $u^{(1)}, \ldots, u^{(M)}$ and target v, its unbiased ensemble estimator is

$$S_M\left(u^{(1:M)}, v\right) = \frac{1}{M} \sum_{m=1}^{M} \left\| u^{(m)} - v \right\| - \frac{1}{2M(M-1)} \sum_{\substack{m,m'=1 \\ m \neq m'}}^{M} \left\| u^{(m)} - u^{(m')} \right\|. \tag{12}$$

Recursive MUNDO evaluates the energy score on projected samples, whereas single-step MUNDO trains on unprojected samples and applies projection after training (Appendix A.3). Each level is trained against its true parent, so that it learns the conditional factor in (11) rather than errors propagated from coarser learned levels.

Although strictly proper, the energy score can have weak discrimination for dependence misspecification in high dimension (Scheuerer & Hamill, 2015). In some of our spatial systems, training on the energy score alone overestimates fine-scale variability (Appendix A.2). Where spatial coherence is not sufficiently captured, we augment the energy score with an auxiliary penalty,

$$\mathcal{L} = S_M + \lambda \mathcal{R}_{\text{spatial}}. \tag{13}$$

One choice is a variogram penalty (Scheuerer & Hamill, 2015). Let $\gamma_h^a(x) = |x_{a+h} - x_a|^p$ denote the order-p pairwise difference at cell a and spatial lag h, with $p = \frac{1}{2}$. An unbiased M-draw estimator is

$$V_M(u_{1:M}, v) = \sum_h w_h \sum_a \frac{1}{M(M-1)} \sum_{\substack{m,m'=1 \\ m \neq m'}}^{M} \left(\gamma_h^a(v) - \gamma_h^a(u_m)\right)\left(\gamma_h^a(v) - \gamma_h^a(u_{m'})\right). \tag{14}$$

We use lags of one, two, and four grid cells along each spatial axis. A second option is a spectral CRPS penalty, following its use in data-driven weather forecasting (Bonev et al., 2025). It evaluates probabilistic skill after transforming ensemble members and targets to the spectral domain, thereby penalizing errors in the distribution of spatial scales. We treat both the variogram and spectral terms as auxiliary spatial penalties rather than as part of the definition of MUNDO. When a spatial penalty is used, λ is set initially so that the energy and spatial terms have comparable magnitude.

End-to-end fine-tuning. Teacher forcing trains each conditional factor on the true parent, whereas at inference it receives a generated one, so errors can accumulate along the reverse chain. Therefore, we optionally fine-tune the complete hierarchy end-to-end by generating full trajectories from the coarse observation and backpropagating the final loss through all levels jointly. This exposes each level to the parent distribution encountered at inference and allows finer levels to compensate for upstream errors. During this stage, we retain the energy-score objective and replace the variogram penalty with a spectral CRPS term (Appendix A.2). Fine-tuning is used only when validation diagnostics reveal error propagation or degraded calibration in recursive MUNDO.

4 EXPERIMENTS

The experiments address three questions: (i) do generated trajectories satisfy the coarse observation; (ii) when does multiresolution sampling improve over generating the full window in a single step; and (iii) does the ensemble recover the unresolved conditional variability and fine-scale temporal structure as predictability decreases?

4.1 EXPERIMENTAL SETUP

The protocol is shared across the three tiers of Table 1. Starting from a fine-resolution record, we apply one of the observation operators in (4) and reconstruct the withheld trajectory from the resulting coarse record. Interpolation windows hold $N + 1$ states with both endpoints x_0 and x_N observed. For the two simulated tiers, the fully resolved trajectory is available and predictability can be varied systematically, allowing us to study performance as a function of both the window length N and the intrinsic decorrelation time of the dynamics. ERA5 provides the corresponding archive-based setting. Appendix B gives equations, splits, and training details.

Table 1: **Experimental systems.** We evaluate MUNDO across three tiers of increasing state dimension and realism. Each window of N steps is coarsened to a single value; the factorization lists the coarsenings applied from coarse to fine, one level per factor (2^6: six factors of two).

Tier	System	State shape	Window N and factorization	Control of predictability	Backbone
Lorenz-96		36×1	64: 2^6; 128: 2^7	forcing $F \in \{10, 24, 32, 64\}$	circular CNN
AutoSim	Advection-diffusion	$64^2 \times 1$	8: 2^3; 16: 2^4; 24: $2, 2, 2, 3$; 32: 2^5	regime and N (via $N\delta/\tau$)	NATTEN (Hassani et al., 2023)
	Gray–Scott	$64^2 \times 2$			
	Navier–Stokes	$64^2 \times 3$			
	Gross–Pitaevskii	$64^2 \times 3$			
ERA5		$149 \times 281 \times 7$	24: $2, 2, 2, 3$	–	ArchesWeather (Couairon et al., 2026)

Methods. We compare single-step and recursive MUNDO with conditional flow matching (Lipman et al., 2023), FlowDPS (Kim et al., 2025), and, on the two-dimensional systems, SimVP (Gao et al., 2022) as a deterministic baseline. MUNDO, conditional flow matching, and FlowDPS are matched in parameter count on Lorenz-96 and AutoSim, while ERA5 uses architecture-specific configurations (Appendix B.4). Separate models are trained for interpolation and disaggregation. On AutoSim and ERA5, the learned baselines are projected onto the constraint set after sampling using the map in Equation (9). On Lorenz-96, we use the reference implementations. The non-learned baselines are the classical methods of Section 2: linear and natural cubic-spline interpolation for interpolation, and Denton–Cholette and CDF-t for disaggregation.

Evaluation. We evaluate separately whether a reconstruction satisfies the observation and whether it recovers the unresolved conditional distribution. A reconstruction is consistent if it preserves the retained states under interpolation and the imposed aggregates under disaggregation. In the tables, *cons.* indicates whether a method is consistent to within 10^{-4} in standardized units by construction (✓), by projection after sampling (✓$^\pi$), or not at all (×). We then assess the learned conditional distribution through the fair CRPS (Ferro, 2014), the spread-skill ratio (Rasp et al., 2024), and the mean absolute error of empirical coverage over nominal central intervals (Greenbury et al., 2026). Because these scores are marginal in time, we additionally report the temporal variogram ratio t_1 to measure short-lag temporal structure. Appendix B.5 defines all metrics. On Lorenz-96 and AutoSim, scores are averaged over three training seeds, with $\pm$ indicating the standard deviation across seeds. ERA5 uses a single seed. Appendices B.1, B.2, and B.3 detail the evaluation protocols.

4.2 RESULTS

Lorenz-96. We use the two-scale Lorenz-96 system (Lorenz, 2006) as a controlled setting in which predictability can be varied directly through the forcing F. Table 2 reports the representative case

Table 2: **Lorenz-96 at $N = 64$, $F = 32$.** †: causal chain. Underlined: second-lowest CRPS among the methods that satisfy the observation.

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 6$)†	✓	**0.2214**$\pm$0.0009	0.89	0.036	**0.99**
MUNDO ($R = 1$)	✓$^{\pi}$	<u>0.2594</u>$\pm$0.0015	**1.02**	**0.012**	0.98
Flow matching	×	0.2520$\pm$0.0026	0.84	0.045	1.12
FlowDPS	×	0.3695$\pm$0.0020	0.98	0.015	1.06
Denton–Cholette	✓	0.6777	–	–	0.35
CDF-t	×	0.7940	–	–	0.51
Interpolation					
MUNDO ($R = 6$)	✓	<u>0.1319</u>$\pm$0.0024	0.90	**0.015**	0.99
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0634**$\pm$0.0005	**0.94**	0.032	**1.00**
Flow matching	×	0.1046$\pm$0.0033	1.36	0.190	1.88
FlowDPS	×	0.4701$\pm$0.0021	1.02	0.007	1.59
Linear	✓	0.7822	–	–	0.48
Cubic spline	✓	0.8191	–	–	0.49

Table 3: **ERA5, daily-to-hourly.** Underlined: second-lowest CRPS among the methods that satisfy the observation.

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)	✓	**0.0974**	**0.87**	**0.131**	**1.07**
MUNDO ($R = 1$)	✓$^{\pi}$	<u>0.1100</u>	0.78	0.238	1.23
Flow matching	✓$^{\pi}$	0.1490	0.48	0.213	0.87
FlowDPS	✓$^{\pi}$	0.1382	0.56	0.180	0.86
SimVP	✓$^{\pi}$	0.1455	–	–	0.66
Denton–Cholette	✓	0.1864	–	–	0.53
CDF-t	×	0.4684	–	–	0.59
Interpolation					
MUNDO ($R = 4$)	✓	**0.0931**	**0.81**	0.151	1.05
MUNDO ($R = 1$)	✓$^{\pi}$	<u>0.0942</u>	0.69	0.186	1.29
Flow matching	✓$^{\pi}$	0.1399	0.53	**0.150**	0.86
FlowDPS	✓$^{\pi}$	0.3921	0.31	0.341	**1.02**
SimVP	✓$^{\pi}$	0.1240	–	–	0.81
Linear	✓	0.2204	–	–	0.55
Cubic spline	✓	0.2225	–	–	0.57

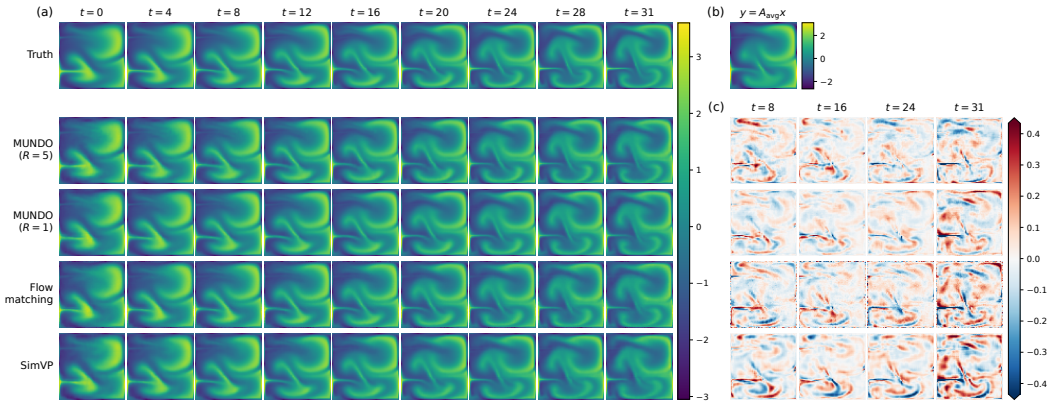

Figure 2: **Navier–Stokes disaggregation at $N = 32$.** (a) Ground truth and reconstructions at selected times. (b) Coarse observation, given by the temporal mean over the window. (c) Reconstruction errors $\hat{x} - x$ at selected times; rows follow the method order in (a).

at $N = 64$, $F = 32$, while Appendix C.1 reports the remaining forcing and window-length settings. Without a final projection, neither flow-based baseline satisfies the observation constraint on Lorenz-96, despite the consistency updates used by FlowDPS. Flow matching achieves CRPS comparable to recursive MUNDO on both tasks, but its members show excessive step-to-step noise, with almost twice the true short-lag variability under interpolation. Across the sweep, single-step MUNDO keeps the lower CRPS under interpolation in every chaotic setting, while recursion lowers it under disaggregation in every setting. As predictability decreases, the CRPS gap closes and the advantage of recursion shifts to calibration and short-lag structure, which degrade for single-step MUNDO.

AutoSim. We use four systems from AutoSim (Greenbury et al., 2026): advection-diffusion as a near-deterministic control, Gray–Scott for multimodal pattern formation, conditioned Navier–Stokes for turbulent dynamics, and Gross–Pitaevskii for dispersive wave dynamics. Table 4 reports the benchmark tier at $N = 32$, while Appendix C.2 reports the remaining window lengths and predictability regimes. In Table 4, single-step MUNDO already outperforms every baseline in CRPS on all systems and tasks, and remains the strongest variant for interpolation. Both flow-based methods overestimate short-lag variability on every system and task, while t_1 is closer to one for MUNDO. In the harder, less predictable settings, recursion generally improves ensemble calibration and, where predictability is lowest, lowers CRPS under disaggregation, while retaining realistic fine-scale temporal structure. Figures 2 and 3 illustrate representative reconstructions for disaggregation and interpolation on Navier–Stokes and Gross–Pitaevskii, respectively. Appendix C.2 also shows reconstructions on advection-diffusion and Gray–Scott (Figures 8 and 9).

Table 4: **AutoSim at $N = 32$.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 5$)	✓	0.0026±0.0000	**0.97**	0.090	**1.02**
MUNDO ($R = 1$)	✓$^\pi$	**0.0020**±0.0000	0.83	0.140	1.04
Flow matching	✓$^\pi$	0.0025±0.0000	0.77	0.185	1.20
FlowDPS	✓$^\pi$	0.0107±0.0003	0.42	**0.049**	1.22
SimVP	✓$^\pi$	0.0044	–	–	1.03
Denton–Cholette	✓	0.0464	–	–	0.90
CDF-t	×	0.0783	–	–	0.86
Interpolation					
MUNDO ($R = 5$)	✓	0.0022±0.0000	**1.04**	0.090	1.12
MUNDO ($R = 1$)	✓$^\pi$	**0.0013**±0.0000	0.94	0.160	1.04
Flow matching	✓$^\pi$	0.0021±0.0000	0.82	0.242	1.19
FlowDPS	✓$^\pi$	0.0326±0.0024	0.64	**0.060**	1.28
SimVP	✓$^\pi$	0.0044	–	–	**1.03**
Linear	✓	0.0554	–	–	0.87
Cubic spline	✓	0.0531	–	–	0.92

(b) **Gray–Scott**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 5$)	✓	0.0034±0.0000	**1.01**	**0.040**	**1.01**
MUNDO ($R = 1$)	✓$^\pi$	**0.0032**±0.0000	1.35	0.048	**1.01**
Flow matching	✓$^\pi$	0.0049±0.0001	0.92	0.147	1.09
FlowDPS	✓$^\pi$	0.0112±0.0002	1.04	0.078	1.09
SimVP	✓$^\pi$	0.0085	–	–	**1.01**
Denton–Cholette	✓	0.1144	–	–	0.93
CDF-t	×	0.1983	–	–	0.94
Interpolation					
MUNDO ($R = 5$)	✓	0.0035±0.0000	**1.03**	**0.037**	1.04
MUNDO ($R = 1$)	✓$^\pi$	**0.0024**±0.0000	0.93	0.060	1.01
Flow matching	✓$^\pi$	0.0041±0.0001	1.10	0.153	1.08
FlowDPS	✓$^\pi$	0.0399±0.0008	1.20	0.072	1.12
SimVP	✓$^\pi$	0.0076	–	–	**1.00**
Linear	✓	0.1678	–	–	0.90
Cubic spline	✓	0.1497	–	–	0.93

(c) **Navier–Stokes**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 5$)	✓	0.0159±0.0002	0.70	**0.020**	1.05
MUNDO ($R = 1$)	✓$^\pi$	**0.0129**±0.0001	0.53	0.095	**1.02**
Flow matching	✓$^\pi$	0.0241±0.0004	**1.24**	0.088	1.47
FlowDPS	✓$^\pi$	0.0374±0.0009	0.65	0.039	1.28
SimVP	✓$^\pi$	0.0286	–	–	**1.02**
Denton–Cholette	✓	0.0759	–	–	0.91
CDF-t	×	0.1467	–	–	0.96
Interpolation					
MUNDO ($R = 5$)	✓	0.0117±0.0000	0.71	**0.015**	1.08
MUNDO ($R = 1$)	✓$^\pi$	**0.0091**±0.0000	0.55	0.072	1.03
Flow matching	✓$^\pi$	0.0163±0.0003	**1.25**	0.046	1.26
FlowDPS	✓$^\pi$	0.2365±0.0050	0.49	0.191	1.80
SimVP	✓$^\pi$	0.0193	–	–	**1.02**
Linear	✓	0.0921	–	–	0.93
Cubic spline	✓	0.0806	–	–	0.95

(d) **Gross–Pitaevskii**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 5$)‡	✓	**0.1202**±0.0018	0.91	**0.047**	**1.05**
MUNDO ($R = 1$)	✓$^\pi$	0.1249±0.0014	0.85	0.049	0.91
Flow matching	✓$^\pi$	0.1658±0.0002	**0.92**	0.140	1.40
FlowDPS	✓$^\pi$	0.3174±0.0027	**1.08**	0.130	1.28
SimVP	✓$^\pi$	0.2071	–	–	0.85
Denton–Cholette	✓	0.5167	–	–	0.23
CDF-t	×	0.6526	–	–	0.38
Interpolation					
MUNDO ($R = 5$)‡	✓	0.0450±0.0005	0.93	**0.034**	1.07
MUNDO ($R = 1$)	✓$^\pi$	**0.0406**±0.0004	**0.98**	0.087	1.05
Flow matching	✓$^\pi$	0.0743±0.0012	0.91	0.173	1.30
FlowDPS	✓$^\pi$	0.2332±0.0059	0.55	0.093	1.29
SimVP	✓$^\pi$	0.0612	–	–	**1.00**
Linear	✓	0.6209	–	–	0.37
Cubic spline	✓	0.6523	–	–	0.38

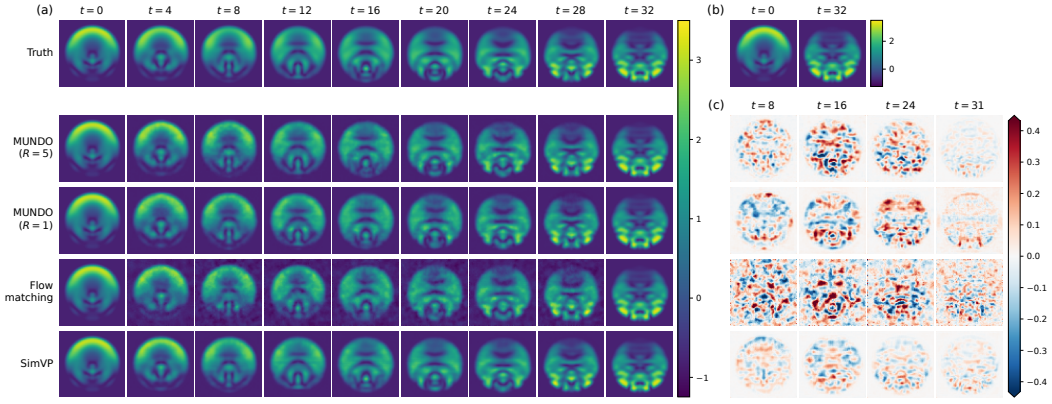

Figure 3: **Gross–Pitaevskii interpolation at $N = 32$.** (a) Ground truth and reconstructions at selected times. (b) Observed endpoint states. (c) Reconstruction errors $\hat{x} - x$ at selected times; rows follow the method order in (a).

ERA5. We evaluate daily-to-hourly downscaling on ERA5 reanalysis over Europe (Hersbach et al., 2020), using seven surface variables on a 149×281 grid. Table 3 shows that both MUNDO variants outperform every baseline in CRPS. At the baselines' parameter count, single-step MUNDO already lowers CRPS by 20–24%, while the flow-based ensembles are strongly under-dispersed. Recursion is most useful for disaggregation, where it lowers CRPS by a further 11%, improves spread-skill agreement and coverage, and brings t_1 close to one. Under interpolation, the two variants have nearly identical CRPS, but recursion again improves calibration and temporal structure. Appendix C reports per-variable scores and additional diagnostics. Figure 4 shows representative

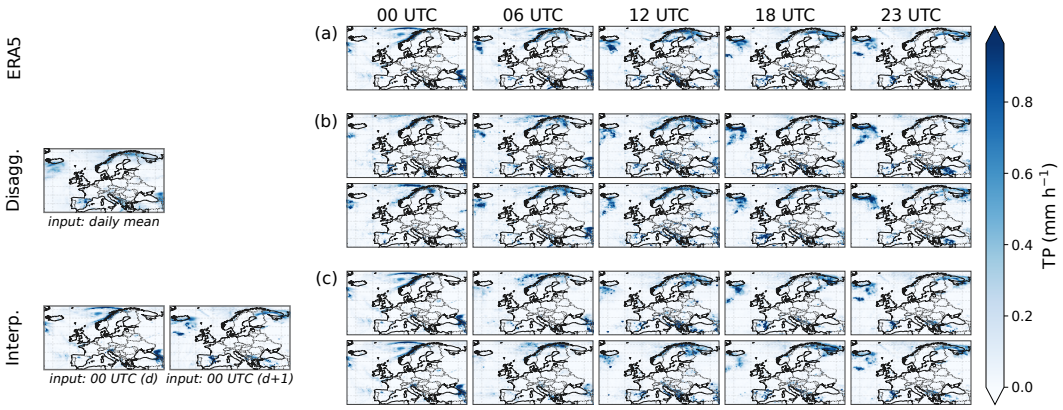

Figure 4: **ERA5 total precipitation** ($\mathrm{mm\,h^{-1}}$) **daily-to-hourly reconstructions.** (a) Ground-truth hourly precipitation at selected times. (b) Disaggregation from the observed daily mean (leftmost), with two independent recursive MUNDO members shown in the two rows. (c) Interpolation between the observed 00 UTC states of consecutive days (leftmost), with the same layout.

ERA5 total-precipitation reconstructions under both operators, with distinct members recovering plausible sub-daily structure from the same coarse observation.

5 Conclusion

Temporal interpolation and disaggregation are two instances of one inverse problem: recovering a distribution of fine trajectories compatible with a coarse observation. MUNDO addresses both with a common construction. Each sampler generates a residual in the null space of the observation operator, enforcing the observation exactly. Single-step MUNDO samples the fine trajectory directly, while recursive MUNDO uses a multiresolution factorization to decompose a large coarse-to-fine conditional into a sequence of smaller ones sampled in reverse.

Across Lorenz-96, four simulated physical systems, and ERA5 reanalysis, single-step MUNDO already achieves strong predictive skill. As predictability decreases, recursive MUNDO generally improves calibration, with better coverage, while preserving fine-scale temporal structure and, where predictability is lowest, improving CRPS under disaggregation. Under interpolation, single-step MUNDO generally keeps the lower CRPS. Extending the construction to joint spatiotemporal and more general observation operators, including data assimilation settings, is a natural next step. A second direction is to apply MUNDO to learned forecasts and climate-model archives, including settings without paired fine-resolution truth, to generate plausible fine-scale temporal realizations of climate projections.

AI use statement

In this work, we used generative AI tools to assist with software implementation, drafting, editing, and figure production. We have not used them to formulate hypotheses, develop the conceptual framework, design the experiments or evaluation, generate or preprocess data, or interpret the results. The remaining tasks requiring disclosure are not applicable here. All AI-assisted code was reviewed by the authors, and all reported numbers come from our experimental records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Pierre Ailliot, Denis Allard, Valérie Monbet, and Philippe Naveau. Stochastic weather generators: an overview of weather type models. *Journal de la société française de statistique*, 156(1):101–113, 2015.

Seth Bassetti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. Diffesm: Conditional emulation of temperature and precipitation in earth system models with 3d diffusion models. *Journal of Advances in Modeling Earth Systems*, 16(10):e2023MS004194, 2024. doi: 10.1029/2023MS004194.

Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D. Collins, Michael S. Pritchard, and Alexander Keller. FourCastNet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale. *arXiv:2507.12144v2 [cs.LG]*, 2025. doi: 10.48550/arXiv.2507.12144.

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:2505.06474v2 [physics.ao-ph]*, 2025. doi: 10.48550/arXiv.2505.06474.

T. A. Buishand and T. Brandsma. Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbour resampling. *Water Resources Research*, 37(11):2761–2776, 2001. doi: 10.1029/2001WR000291.

Pierre A. Cholette. Adjusting sub-annual series to yearly benchmarks. *Survey Methodology*, 10(1): 35–49, 1984.

Guillaume Couairon, Renu Singh, Anastase Charantonis, Christian Lessig, and Claire Monteleoni. ArchesWeatherGen: Skillful and compute-efficient probabilistic weather forecasting with machine learning. *Science Advances*, 12(17):eadx2372, 2026. ISSN 2375-2548. doi: 10.1126/sciadv.adx2372.

Bekele Debele, Raghavan Srinivasan, and Jean-Yves Parlange. Accuracy evaluation of weather data generation and disaggregation methods at finer timescales. *Advances in Water Resources*, 30(5): 1286–1300, 2007. doi: 10.1016/j.advwatres.2006.11.009.

Frank T. Denton. Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association*, 66(333):99–102, 1971. doi: 10.1080/01621459.1971.10482227.

Michel Déqué. Frequency of precipitation and temperature extremes over france in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, 57(1-2):16–26, 2007. doi: 10.1016/j.gloplacha.2006.11.030.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, 1996. doi: 10.1007/978-94-009-1740-8.

Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016. ISSN 1991-959X. doi: 10.5194/gmd-9-1937-2016.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. ISSN 00359009. doi: 10.1002/qj.2270.

Kristian Förster, Florian Hanzer, Benjamin Winter, Thomas Marke, and Ulrich Strasser. An open-source MEteoroLOgical observation time series DISaggregation tool (MELODIST v0.1.1). *Geoscientific Model Development*, 9(7):2315–2333, 2016. doi: 10.5194/gmd-9-2315-2016.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SimVP: Simpler yet better video prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3160–3170, 2022. doi: 10.1109/CVPR52688.2022.00317.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Sam F. Greenbury, Radka Jersakova, Paolo Conti, Marjan Famili, Christopher Iliffe Sprague, Edwin Brown, and Jason D. McEwen. Reliability of Probabilistic Emulation of Physical Systems. *arXiv:2606.12997v1 [cs.LG]*, 2026. doi: 10.48550/arXiv.2606.12997.

Jacques Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

Xue Han, Chunping Chang, Zhongling Guo, Shuiqing Yin, Jifeng Li, Huiru Li, Rende Wang, and Qing Li. Comparison of deterministic wind speed generation models for wind erosion modeling. *Agricultural and Forest Meteorology*, 334:109438, 2023. doi: 10.1016/j.agrformet.2023.109438.

Nidhin Harilal, Bri-Mathias Hodge, Aneesh Subramanian, and Claire Monteleoni. STint: Self-supervised temporal interpolation for geospatial data. *arXiv:2309.00059v1 [cs.CV]*, 2023. doi: 10.48550/arXiv.2309.00059.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6185–6194. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00599.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia De Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. ISSN 0035-9009, 1477-870X. doi: 10.1002/qj.3803.

Magnus Sikora Ingstad, Mariana C. A. Clare, Olav Ersland, Vera Gahlen, Håvard Homleid Haugen, Ophélia Miralles, Even M. Nordhagen, Thomas N. Nipen, Ivar A. Seierstad, John Bjørnar Bremnes, Michael Maier-Gerber, Zied Ben Bouallègue, Harrison Cook, Christian Lessig, Gert Mertes, Cathal O'Brien, Florian Pinault, Ana Prieto Nemesio, and Matthew Chantry. HourGlass: A probabilistic data-driven temporal downscaler for global and regional weather forecasting. *arXiv:2607.11457v1 [physics.ao-ph]*, 2026. doi: 10.48550/arXiv.2607.11457.

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9000–9008, 2018. doi: 10.1109/CVPR.2018.00938.

Azin Al Kajbaf, Michelle Bensi, and Kaye L. Brubaker. Temporal downscaling of precipitation from climate model projections using machine learning. *Stochastic Environmental Research and Risk Assessment*, 36(8):2173–2194, 2022. ISSN 1436-3240, 1436-3259. doi: 10.1007/s00477-022-02259-2.

Jeongsol Kim, Bryan Sangwoo Kim, and Jong Chul Ye. FlowDPS: Flow-driven posterior sampling for inverse problems. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12328–12337. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01146.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11, 2023. doi: 10.1145/3592979.3593412.

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. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adi2336.

Jussi Leinonen, Daniele Nerini, and Alexis Berne. Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, 59(9):7211–7223, 2021. doi: 10.1109/TGRS.2020.3032790.

Jussi Leinonen, Boris Bonev, Thorsten Kurth, and Yair Cohen. Modulated adaptive fourier neural operators for temporal interpolation of weather forecasts. *arXiv:2410.18904v1 [physics.ao-ph]*, 2024. doi: 10.48550/arXiv.2410.18904.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

F. Lombardo, E. Volpi, D. Koutsoyiannis, and F. Serinaldi. A theoretically consistent stochastic cascade for temporal disaggregation of intermittent rainfall. *Water Resources Research*, 53(6): 4586–4605, 2017. doi: 10.1002/2017WR020529.

Edward N. Lorenz. Predictability – a problem partly solved. In Renate Hagedorn and Tim Palmer (eds.), *Predictability of Weather and Climate*, pp. 40–58. Cambridge University Press, Cambridge, 2006. ISBN 978-0-521-84882-4. doi: 10.1017/CBO9780511617652.004.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

P.-A. Michelangeli, M. Vrac, and H. Loukos. Probabilistic downscaling approaches: Application to wind cumulative distribution functions. *Geophysical Research Letters*, 36(11), 2009. ISSN 1944-8007. doi: 10.1029/2009GL038401.

David W. Pierce, Daniel R. Cayan, and Bridget L. Thrasher. Statistical downscaling using localized constructed analogs (LOCA). *Journal of Hydrometeorology*, 15(6):2558–2585, 2014. doi: 10.1175/JHM-D-14-0082.1.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. ISSN 1942-2466, 1942-2466. doi: 10.1029/2023MS004019.

C. W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1):182–190, 1981. doi: 10.1029/WR017i001p00182.

Daniel Schertzer and Shaun Lovejoy. Physical modeling and analysis of rain and clouds by anisotropic scaling multiplicative processes. *Journal of Geophysical Research: Atmospheres*, 92(D8):9693–9714, 1987. ISSN 2156-2202. doi: 10.1029/JD092iD08p09693.

Michael Scheuerer and Thomas M. Hamill. Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. ISSN 1520-0493, 0027-0644. doi: 10.1175/MWR-D-14-00269.1.

Jonathan Schmidt, Luca Schmidt, Felix M. Strnad, Nicole Ludwig, and Philipp Hennig. A generative framework for probabilistic, spatiotemporally coherent downscaling of climate simulation. *npj Climate and Atmospheric Science*, 8(1):270, 2025. ISSN 2397-3722. doi: 10.1038/s41612-025-01157-y.

Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. Deep null space learning for inverse problems: convergence analysis and rates. *Inverse Problems*, 35(2):025008, 2019. doi: 10.1088/1361-6420/aaf14a.

Xinwei Shen and Nicolai Meinshausen. Engression: Extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 87(3): 653–677, 2025. ISSN 1369-7412, 1467-9868. doi: 10.1093/jrsssb/qkae108.

Xinwei Shen, Nicolai Meinshausen, and Tong Zhang. Reverse markov learning: Multi-step generative models for complex distributions. *arXiv:2502.13747v2 [cs.LG]*, 2025. doi: 10.48550/arXiv.2502.13747.

Karen Stengel, Andrew Glaws, Dylan Hettinger, and Ryan N. King. Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, 117(29): 16805–16815, 2020. doi: 10.1073/pnas.1918964117.

Hans von Storch. On the use of "inflation" in statistical downscaling. *Journal of Climate*, 12 (12):3505–3506, 1999. ISSN 0894-8755, 1520-0442. doi: 10.1175/1520-0442(1999)012<3505:OTUOII>2.0.CO;2.

Liwen Wang, Qian Li, Qi Lv, Xuan Peng, and Wei You. TemDeep: A self-supervised framework for temporal downscaling of atmospheric fields at arbitrary time resolutions. *Geoscientific Model Development*, 18(8):2427–2442, 2025. ISSN 1991-9603. doi: 10.5194/gmd-18-2427-2025.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations*, 2023.

Daniel S. Wilks. Use of stochastic weather generators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, 1(6):898–907, 2010. doi: 10.1002/wcc.85.

Eduardo Zorita and Hans von Storch. The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of Climate*, 12(8):2474–2489, 1999. doi: 10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2.

The appendix is organized into three sections. Appendix A reviews the engression objective and its finite-ensemble estimators, then examines spatial-coherence penalties, projection during training and sampling, and cross-window conditioning. Appendix B describes the Lorenz-96, AutoSim, and ERA5 datasets, then gives the training configurations, baseline implementations, computational costs, and metric definitions. Appendix C extends the Lorenz-96 and AutoSim results across predictability regimes and window lengths, and provides additional ERA5 diagnostics by hour and variable, including spatial spectra and diurnal cycles.

A Methodological details

A.1 Scoring rules and the engression objective

A scoring rule $S(F, v)$ evaluates a predictive distribution $F \in \mathcal{P}$ against an observation $v \in \mathbb{R}^D$. Using the loss convention of Equation (12), it is *proper* when

$$\mathbb{E}_{v\sim Q}\big[S(F, v)\big] \geq \mathbb{E}_{v\sim Q}\big[S(Q, v)\big] \tag{15}$$

for all $F, Q \in \mathcal{P}$, and *strictly proper* when equality implies $F = Q$ (Gneiting & Raftery, 2007). Thus the true distribution uniquely minimizes the population score.

The energy score with exponent $\beta \in (0, 2)$ is

$$S(P, v) = \mathbb{E}\,\|U - v\|_2^\beta - \frac{1}{2}\,\mathbb{E}\,\|U - U'\|_2^\beta\,, \tag{16}$$

where $U, U' \overset{\text{iid}}{\sim} P$. It is strictly proper on distributions with finite β-th moment. Following engression, we take $\beta = 1$ (Gneiting & Raftery, 2007; Shen & Meinshausen, 2025).

Applied conditionally, strict propriety implies, for p_Y-almost every y,

$$\mathbb{E}_{X\sim p^*(\cdot\,|\,y)}\left[S\left(p^*(\cdot \mid y), X\right)\right] \leq \mathbb{E}_{X\sim p^*(\cdot\,|\,y)}\left[S\left(p_g(\cdot \mid y), X\right)\right], \tag{17}$$

with equality only when $p_g(\cdot \mid y) = p^*(\cdot \mid y)$. Averaging over Y gives the engression objective (Shen & Meinshausen, 2025),

$$g^* \in \arg\min_{g\in\mathcal{G}} \mathbb{E}\left[\|X - g(Y, \varepsilon)\|_2 - \frac{1}{2}\,\|g(Y, \varepsilon) - g(Y, \varepsilon')\|_2\right], \tag{18}$$

where ε and ε' are independent noise draws. Under realizability, its population optimum satisfies $g^*(Y, \varepsilon) \mid Y = y \sim p^*(\cdot \mid y)$ for p_Y-almost every y. No density evaluation is required, which allows the same objective to be used at every level of (11).

Limitations. The energy score is strictly proper, but its finite-sample discrimination of spatial dependence can be weak in high dimension (Scheuerer & Hamill, 2015). Both terms of (12) depend only on Euclidean distances over the full field, so different spatial arrangements can receive similar scores even when their local or spectral structure differs. Strict propriety guarantees the correct population optimum over unrestricted distributions, but not that a restricted model will recover spatial dependence accurately under finite-sample optimization.

Finite-ensemble estimator. The ensemble estimators used during training must also avoid penalizing spread when estimated from a small number of members. With $M = 2$, the fair pairwise term in the energy score reduces to $\frac{1}{2}\|u - u'\|$, so Equation (12) is already unbiased for the corresponding population quantity. The variogram estimator in (14) uses the same cross-member construction. By contrast, the naive squared error of the ensemble-mean transformed increment is biased upward by $\frac{1}{2}\,\mathrm{Var}[\gamma_h]$, which implicitly penalizes spread. Using independent draws in the cross-member product makes the corresponding centered errors uncorrelated, so this additional variance term vanishes in expectation.

A.2 Spatial-coherence terms

We ablate the auxiliary spatial-coherence penalty term in (13). During level-wise training, we compare the energy score alone and augmented with either a variogram score or spectral CRPS. During

Table 5: **Spatial-coherence term ablation.** Gross–Pitaevskii disaggregation at $N = 24$, three seeds, with cross-window conditioning. Energy denotes the energy score, variogram the variogram score in (14), and spectral the spectral CRPS. Stage 1 is the level-wise training objective and Stage 2 the end-to-end fine-tuning objective; – denotes no fine-tuning. † marks the configuration used in the main experiments. SSR is the spread-skill ratio, s_1 is the one-cell spatial variogram ratio, Finest the finest-band spectral power ratio; all targeting one.

Arm	Stage 1	Stage 2	CRPS ↓	SSR → 1	$s_1 \to 1$	Finest → 1
MUNDO ($R = 4$)	energy	–	0.1047±0.0004	**0.85**	1.131	2.15±0.06
	energy + variogram	–	**0.1025**±0.0002	**0.85**	1.073	2.78±0.30
	energy + spectral	–	**0.1025**±0.0005	0.83	**1.055**	**1.96**±0.10
	energy + variogram	energy	0.0802±0.0008	0.92	1.089	2.75±0.19
	energy + variogram	energy + variogram	**0.0788**±0.0007	0.92	**1.057**	2.21±0.19
	energy + variogram	energy + spectral†	0.0799±0.0010	0.92	**1.057**	**1.17**±0.03
	energy	energy + spectral	0.0801±0.0000	0.92	1.083	1.21±0.05
	energy + spectral	energy + spectral	**0.0795**±0.0003	**0.93**	**1.062**	**1.12**±0.04
MUNDO ($R = 1$)	energy	–	0.0866±0.0005	0.83	1.135	2.97±0.35
	energy + variogram†	–	**0.0828**±0.0009	0.83	**1.026**	**1.08**±0.08

end-to-end fine-tuning, we compare the same three objectives. Single-step MUNDO has no fine-tuning stage, since its only parent is the observation itself. The spectral CRPS is computed independently on the real and imaginary parts of each two-dimensional Fourier coefficient (Bonev et al., 2025).

Table 5 reports the comparison on Gross–Pitaevskii disaggregation at $N = 24$. Because the ablation targets recovery of fine spatial scales, we additionally report the finest-band spectral power ratio, denoted Finest in the table, alongside the one-cell spatial variogram ratio s_1, both defined in Appendix B.5. For single-step MUNDO, the variogram term is important: without it, fine-scale spectral power is strongly overestimated and s_1 deteriorates. For recursive MUNDO, the level-wise variogram term has a smaller effect, while end-to-end fine-tuning with spectral CRPS brings the finest-band power close to the truth regardless of the level-wise objective.

A.3 Projection during training and sampling

Equation (9) defines the affine projection $P_y(u) = A^\dagger y + (I - A^\dagger A)u$ onto $\mathcal{X}_y$. Recursive MUNDO evaluates the training loss on projected samples at every level. Single-step MUNDO trains on the unprojected reconstruction $A^\dagger y + G_\theta(y, \varepsilon, c)$ and projects the generated samples with P_y after training.

Assuming finite first moments and realizability, the expected energy score is uniquely minimized when the distribution evaluated by the score equals the target conditional law (Appendix A.1). When projection is applied inside the loss, this guarantee applies to the distribution after projection. When the loss is evaluated on unprojected samples, the optimal unprojected distribution already equals the target law and is unchanged by projection. Both formulations therefore recover the target law at the population optimum. For a finite network, however, the choice of training objective can affect the generated distribution.

Table 6 compares the two training choices for single-step MUNDO on AutoSim disaggregation in the benchmark regime at $N = 32$. Applying projection inside the loss increases CRPS on all four systems, while its effects on calibration are mixed and temporal structure changes comparatively little. Single-step MUNDO retains lower CRPS than the baselines on all four systems regardless of when projection is applied.

A.4 Cross-window conditioning

The conditioning c in (9) can include two kinds of information from adjacent windows. Cross-window context is especially useful for disaggregation, where consecutive reconstructed windows do not share an observed boundary state.

Table 6: **Effect of projection during training or sampling.** AutoSim disaggregation, benchmark tier, $N = 32$.

	sampling only				inside the training loss			
System	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Advection-diffusion	0.0020	0.83	0.140	1.04	0.0023	0.95	0.135	1.01
Gray–Scott	0.0032	1.35	0.048	1.01	0.0034	1.76	0.058	1.00
Navier–Stokes	0.0129	0.53	0.095	1.02	0.0131	0.57	0.064	1.02
Gross–Pitaevskii	0.1249	0.85	0.049	0.91	0.1262	0.97	0.066	0.97

Table 7: **Cross-window conditioning ablation.** Disaggregation on the AutoSim benchmark regimes, three seeds, evaluated with the same trained models and cross-window context either masked or fed. CRPS gain is $1 - \mathrm{CRPS}_{\mathrm{fed}}/\mathrm{CRPS}_{\mathrm{masked}}$. r_{seam} is the cross-window step-size ratio defined in (33), with one indicating agreement with the truth.

		MUNDO					MUNDO ($R = 1$)				
		CRPS	SSR → 1		$r_{\mathrm{seam}} \to 1$		CRPS	SSR → 1		$r_{\mathrm{seam}} \to 1$	
System	N	gain	masked	fed	masked	fed	gain	masked	fed	masked	fed
Advection-diffusion	8	87%	0.97	0.92	2.56	1.03	83%	0.97	0.96	2.22	1.05
	16	49%	0.95	0.94	2.06	1.16	48%	0.94	0.92	1.74	1.12
	24	36%	0.91	0.93	1.88	1.24	30%	0.86	0.84	1.61	1.20
	32	23%	0.94	0.97	1.79	1.39	14%	0.85	0.83	1.49	1.28
Gray–Scott	8	31%	1.31	1.41	1.02	1.01	9%	1.20	1.22	1.01	1.01
	16	21%	1.10	1.19	1.05	1.01	9%	1.27	1.26	1.03	1.02
	24	19%	1.00	1.10	1.09	1.03	9%	1.28	1.35	1.06	1.03
	32	19%	0.86	1.01	1.17	1.07	11%	1.25	1.35	1.13	1.05
Navier–Stokes	8	59%	0.89	0.93	1.47	1.03	48%	0.82	0.84	1.33	1.03
	16	46%	0.85	0.85	2.20	1.26	42%	0.74	0.74	1.87	1.19
	24	29%	0.76	0.76	2.87	1.84	27%	0.63	0.62	2.48	1.63
	32	17%	0.70	0.70	4.03	2.97	16%	0.53	0.53	3.29	2.45
Gross–Pitaevskii	8	25%	0.79	0.74	1.04	1.01	27%	0.87	0.92	1.02	1.01
	16	15%	0.83	0.79	1.42	1.26	19%	0.89	0.91	1.28	1.18
	24	10%	0.88	0.85	1.86	1.65	12%	0.82	0.83	1.60	1.46
	32	8%	0.91	0.90	2.39	2.13	9%	0.84	0.85	1.93	1.75

Observed context. The first is observed context: the aggregates of the preceding and following windows, each paired with a presence flag and randomly dropped during training, so that the same model can operate with or without them. Table 7 evaluates the same trained models with cross-window context either provided or masked across all benchmark-regime window lengths. The benefit is largest at short windows: at $N = 8$, context reduces CRPS by 25–87% for MUNDO across the four systems and brings cross-window step sizes closer to the truth, with both effects weakening as N increases. Calibration changes comparatively little, except on Gray–Scott, where the added context noticeably shifts ensemble spread. Masking neighboring aggregates does not reverse the single-step comparison with flow matching at $N = 32$ in the benchmark regime: CRPS remains lower on all four systems (0.0023, 0.0036, 0.0153, 0.1371, versus 0.0025, 0.0049, 0.0241, 0.1658, respectively, for advection-diffusion, Gray–Scott, Navier–Stokes, and Gross–Pitaevskii).

Generated context. The second is generated context: recursive MUNDO can additionally condition each level on the last state of the preceding window at that level's resolution. During training, this state comes from the true trajectory. At inference, it comes from the preceding window generated by the same ensemble member. This train–test mismatch is addressed during the end-to-end fine-tuning of Section 3.5.

Conditioned on the observations, the generated windows form a Markov chain. Unavailable neighboring observations are masked, with presence flags indicating their availability. The model distribution q factorizes as

$$q\left(x_{1:T} \mid y_{1:T}\right) = q\left(x_1 \mid y_{1:2}\right) \prod_{t=2}^{T} q\left(x_t \mid x_{t-1}, y_{t-1:t+1}\right). \tag{19}$$

Table 8: **Generated cross-window context.** Disaggregation, three seeds. *No input*: recursive MUNDO trained without the preceding-window input. The other three columns report results for the same model trained with that input, evaluated with the input masked, filled with the true preceding state, or filled with the preceding window generated by the same ensemble member, as deployed. Lowest CRPS in each row in bold; SSR $\to 1$; t_1 the temporal variogram ratio at one step, $\to 1$.

			no input			masked			true			generated		
System	Regime	N	CRPS	SSR	t_1	CRPS	SSR	t_1	CRPS	SSR	t_1	CRPS	SSR	t_1
Advection-diffusion	benchmark	8	0.00044	0.89	1.02	0.00044	0.94	1.02	**0.00031**	0.90	1.02	0.00047	2.26	1.02
Advection-diffusion	mild	8	0.00395	1.23	1.04	0.00384	1.20	1.04	**0.00319**	1.13	1.03	0.02587	1.93	1.53
Advection-diffusion	chaotic	8	0.01514	1.18	1.07	0.01516	1.20	1.07	**0.01411**	1.24	1.09	0.02774	1.76	1.41
Advection-diffusion	benchmark	16	0.00173	0.94	1.02	0.00161	0.96	1.02	**0.00145**	0.94	1.02	0.00146	0.95	1.02
Advection-diffusion	mild	16	0.00657	1.02	1.02	0.00651	1.01	1.02	**0.00610**	1.02	1.02	0.00617	1.03	1.02
Advection-diffusion	chaotic	16	0.02107	1.09	1.03	0.02131	1.17	1.05	**0.02094**	1.17	1.05	0.02111	1.19	1.06
Gray–Scott	benchmark	8	0.00049	0.89	1.01	0.00054	0.91	1.01	**0.00046**	0.87	1.01	0.00047	0.86	1.01
Navier–Stokes	benchmark	8	0.00173	0.94	1.04	0.00184	0.94	1.04	**0.00166**	0.95	1.04	0.00167	0.95	1.04

Each window-level factor is defined by the multiresolution sampler. Writing $b_{t-1}^{(r)}$ for the last state of the preceding window at resolution r, for $t \geq 2$ we have

$$q\left(x_t \mid x_{t-1}, y_{t-1:t+1}\right) = \int \prod_{r=0}^{R-1} q_r\left(x_t^{(r)} \mid x_t^{(r+1)},\, b_{t-1}^{(r)},\, y_{t-1:t+1}\right) \mathrm{d}x_t^{(1)} \cdots \mathrm{d}x_t^{(R-1)}, \tag{20}$$

where $x_t^{(0)} = x_t$ and $x_t^{(R)} = y_t$. The first window uses the same hierarchy with generated context masked. The restriction to $b_{t-1}^{(r)}$ at each level is a modeling assumption. We reserve generated context for recursive MUNDO, where it is available at every intermediate resolution.

Table 8 compares conditioning on true and generated preceding-window context. The two give CRPS within about 2% of each other in most settings, except on advection-diffusion at $N = 8$, where feeding back the generated state degrades CRPS and calibration.

Recursive variants. Therefore, recursive MUNDO has four variants under disaggregation: with or without generated context, each with or without end-to-end fine-tuning. Under interpolation, enforcing the shared observed endpoint makes neighboring windows agree at their junction, so we do not use generated cross-window context and only vary fine-tuning. The tables report the plain or the fine-tuned variant, selected for each AutoSim system, predictability regime, and window length on validation diagnostics only, and scored on the same test windows as every other row. The causal chain conditions on its own output for the preceding window, so it is scored on chains, every window but the first of each test trajectory; Table 9 compares the four variants and single-step MUNDO on those windows for $N \in \{24, 32\}$ in the benchmark and chaotic regimes. On those windows the fine-tuned chain is the best of the four variants in every setting, lowering the CRPS of plain recursive MUNDO by 13–35% on Navier–Stokes and Gross–Pitaevskii; against single-step MUNDO it is ahead on Gross–Pitaevskii and chaotic advection-diffusion, level on Navier–Stokes, and behind on Gray–Scott and benchmark advection-diffusion. On Lorenz-96, we instead fix the variant in advance: the causal chain for disaggregation and the plain sampler for interpolation, both without fine-tuning.

Table 9: **Recursive MUNDO variants on AutoSim.** CRPS and coverage MAE on the test trajectories, three seeds. Disaggregation: the four variants and single-step MUNDO on the same chain windows (every window but the first of each test trajectory) and members; plain and fine-tuned, without and with the generated context of the causal chain. Interpolation: plain against fine-tuned and single-step MUNDO, on the standard evaluation windows. *: the variant reported in the main tables, plain or fine-tuned, selected on the validation trajectories.

			plain		fine-tuned		causal chain		causal chain, fine-tuned		single-step	
System	Regime	N	CRPS ↓	cov. MAE ↓	CRPS ↓	cov. MAE ↓	CRPS ↓	cov. MAE ↓	CRPS ↓	cov. MAE ↓	CRPS ↓	cov. MAE ↓
Disaggregation												
Advection-diffusion	benchmark	24	0.0022*	0.114	0.0020	0.118	0.0020	0.117	0.0018	0.125	0.0017	0.158
Advection-diffusion	benchmark	32	0.0027*	0.093	0.0026	0.100	0.0026	0.100	0.0025	0.108	0.0021	0.142
Advection-diffusion	chaotic	24	0.0247	0.014	0.0230*	0.010	0.0242	0.016	0.0225	0.014	0.0344	0.043
Advection-diffusion	chaotic	32	0.0281	0.026	0.0265*	0.010	0.0283	0.032	0.0256	0.013	0.0412	0.042
Gray–Scott	benchmark	24	0.0014*	0.048	0.0014	0.051	0.0014	0.048	0.0013	0.051	0.0013	0.052
Gray–Scott	benchmark	32	0.0019*	0.044	0.0019	0.049	0.0018	0.044	0.0018	0.049	0.0016	0.053
Gray–Scott	chaotic	24	0.0100	0.061	0.0091*	0.068	0.0088	0.059	0.0082	0.069	0.0070	0.043
Gray–Scott	chaotic	32	0.0183	0.054	0.0163*	0.058	0.0160	0.055	0.0146	0.061	0.0115	0.037
Navier–Stokes	benchmark	24	0.0084	0.023	0.0081*	0.024	0.0071	0.038	0.0069	0.038	0.0071	0.048
Navier–Stokes	benchmark	32	0.0153*	0.019	0.0147	0.019	0.0127	0.019	0.0123	0.021	0.0125	0.095
Navier–Stokes	chaotic	24	0.0574	0.028	0.0541*	0.037	0.0503	0.027	0.0484	0.035	0.0519	0.017
Navier–Stokes	chaotic	32	0.0869	0.027	0.0823*	0.034	0.0777	0.020	0.0754	0.024	0.0746	0.021
Gross–Pitaevskii	benchmark	24	0.1146	0.068	0.0899*	0.055	0.1002	0.074	0.0748	0.078	0.0938	0.048
Gross–Pitaevskii	benchmark	32	0.1523	0.072	0.1291*	0.042	0.1384	0.077	0.1130	0.062	0.1345	0.044
Interpolation												
Advection-diffusion	benchmark	24	0.0015*	0.099	0.0014	0.118	–	–	–	–	0.0010	0.172
Advection-diffusion	benchmark	32	0.0022*	0.090	0.0020	0.106	–	–	–	–	0.0013	0.160
Advection-diffusion	chaotic	24	0.0222	0.036	0.0207*	0.036	–	–	–	–	0.0185	0.048
Advection-diffusion	chaotic	32	0.0285	0.041	0.0266*	0.032	–	–	–	–	0.0213	0.047
Gray–Scott	benchmark	24	0.0030*	0.044	0.0029	0.039	–	–	–	–	0.0022	0.059
Gray–Scott	benchmark	32	0.0035*	0.037	0.0034	0.039	–	–	–	–	0.0024	0.060
Gray–Scott	chaotic	24	0.0115	0.037	0.0100*	0.041	–	–	–	–	0.0074	0.041
Gray–Scott	chaotic	32	0.0203	0.031	0.0169*	0.040	–	–	–	–	0.0112	0.032
Navier–Stokes	benchmark	24	0.0074*	0.017	0.0071	0.020	–	–	–	–	0.0063	0.048
Navier–Stokes	benchmark	32	0.0117*	0.015	0.0112	0.022	–	–	–	–	0.0091	0.072
Navier–Stokes	chaotic	24	0.0506	0.028	0.0427*	0.033	–	–	–	–	0.0370	0.022
Navier–Stokes	chaotic	32	0.0842	0.024	0.0663*	0.036	–	–	–	–	0.0650	0.028
Gross–Pitaevskii	benchmark	24	0.0320	0.072	0.0253*	0.055	–	–	–	–	0.0226	0.082
Gross–Pitaevskii	benchmark	32	0.0660	0.056	0.0450*	0.034	–	–	–	–	0.0406	0.087

B Experimental details

B.1 Lorenz-96

The two-scale Lorenz-96 model (Lorenz, 2006) couples K slow variables X_k to J fast variables $Y_{j,k}$ per slow variable,

$$\frac{d}{dt}X_k = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k}, \tag{21}$$

$$\frac{d}{dt}Y_{j,k} = -cbY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + \frac{hc}{b}X_k, \tag{22}$$

with $h = 1$, $b = 10$, and $c = 10$. We use $K = 36$ and $J = 32$. We vary the forcing over $F \in \{10, 24, 32, 64\}$ to span regimes of increasing chaoticity, with $F = 10$ non-chaotic and $F = 64$ strongly chaotic. The system is integrated with fourth-order Runge–Kutta at a step of 10^{-3} time units for 1000 time units after a 100-unit spin-up. Only the slow variables are observed and reconstructed. The fast variables remain latent and act as unresolved forcing on the slow dynamics.

The target resolution is $\delta = 0.01$ time units, with windows of $N \in \{64, 128\}$ steps. The trajectory is split sequentially into 70% training (time units 100–800), 15% validation (800–950), and 15% test (950–1100), with no window crossing a split boundary. Training samples 50 000 windows from the training interval. Evaluation uses 50 test windows, at least 0.6 time units apart, with 256 ensemble members per window. The causal chain uses generated predecessors and is evaluated on 25 chains of contiguous windows with 128 members and the same minimum separation between chains. Each chain contains eight windows at $N = 64$ or four at $N = 128$; excluding the first, which has no generated predecessor, leaves 175 and 75 scored windows, respectively.

B.2 AUTOSIM

We generate four 64×64 physical systems with AutoSim (Greenbury et al., 2026), spanning transport, reaction-diffusion, turbulent flow, and nonlinear dispersive dynamics. Their governing equations are

$$\textbf{Advection-diffusion} \qquad \partial_t \omega = \nu \nabla^2 \omega - \mu\, \mathbf{u} \cdot \nabla \omega, \tag{23}$$

$$\textbf{Gray–Scott} \qquad \begin{aligned} \partial_t u &= \delta_u \nabla^2 u - uv^2 + F(1-u), \\ \partial_t v &= \delta_v \nabla^2 v + uv^2 - (F+k)\, v, \end{aligned} \tag{24}$$

$$\textbf{Conditioned Navier–Stokes} \qquad \begin{aligned} \partial_t s &= -\mathbf{u} \cdot \nabla s, \\ \partial_t \mathbf{u} &= -(\mathbf{u} \cdot \nabla)\mathbf{u} - \nabla p + \nu \nabla^2 \mathbf{u} + \beta s\, \mathbf{e}_y, \\ \nabla \cdot \mathbf{u} &= 0, \end{aligned} \tag{25}$$

$$\textbf{Gross–Pitaevskii} \qquad i\, \partial_t \psi = \left[-\tfrac{1}{2} \nabla^2 + V(\mathbf{x}, t) + g|\psi|^2 \right] \psi. \tag{26}$$

The reconstructed fields are vorticity for advection-diffusion, the two reactants for Gray–Scott, passive scalar and velocity components for conditioned Navier–Stokes, and density together with the real and imaginary parts of ψ for Gross–Pitaevskii. In advection-diffusion, $\mathbf{u}$ is the velocity induced by ω through its streamfunction. In Gross–Pitaevskii, V is a box trap with an oscillating obstacle.

We generate 200 training trajectories and 20 validation and test trajectories per system, with splits made by trajectory. We evaluate windows of $N \in \{8, 16, 24, 32\}$ steps, with $M = 64$ ensemble members, on 200 test windows per system (180 for interpolation) spread over the test trajectories. In addition to the benchmark configuration, we use mild and chaotic regimes with progressively shorter decorrelation times. For Gray–Scott, these regimes are obtained by increasing the snapshot stride rather than changing the underlying dynamics. Table 10 gives the simulation parameters.

Table 10: **Generation parameters.** Systems in the order of (23)–(26). Span denotes trajectory length and δ the saved snapshot interval, i.e., the target resolution, both in simulation units. Parameters not listed take the values of AutoSim's shipped configurations.

System	Parameter	Benchmark	Mild	Chaotic
Advection-diffusion	advection μ	$[0.5, 2]$	$[1, 4]$	$[2, 8]$
	diffusivity ν	$[10^{-4}, 10^{-2}]$	$[5, 500] \cdot 10^{-5}$	$[2.5, 250] \cdot 10^{-5}$
	snapshot step δ		0.25	
	span		80	
Gray–Scott	snapshot step δ	4	8	16
	span	1280	2560	5120
Navier–Stokes	buoyancy β	$[0.2, 0.8]$	$[1.6, 6.4]$	$[3.2, 12.8]$
	snapshot step δ		0.261	
	span		85.9	
Gross–Pitaevskii	obstacle frequency ω	$[0.5, 2]$	$[1.5, 3]$	–
	interaction g	$[80, 200]$	$[160, 400]$	–
	snapshot step δ		0.05	
	span		17.55	

B.3 ERA5

We use hourly ERA5 reanalysis (Hersbach et al., 2020) at $0.25°$ on a grid covering Europe ($25°$W–$45°$E, $35°$N–$72°$N, 149×281 points), on seven surface fields: 10m u and v wind, 2m temperature and dewpoint, mean sea level pressure, total column water and total precipitation. Splits are by year: 1979–2015 for training, 2016–2018 for validation, 2019–2022 for test. The conditioning c in (9) includes adjacent-window observations and calendar information (hour of day and day of year), together with static orography, land-sea mask, latitude, and longitude fields as in ArchesWeather (Couairon et al., 2026). Evaluation uses 96 non-overlapping test windows evenly spaced over 2019–2022, with $M = 24$ ensemble members.

Table 11: **Training configuration and computational cost.** ES denotes the energy score, VS the variogram penalty, SC the spectral CRPS, and NFE the number of network evaluations per generated member. Budget is in epochs (ep.) on Lorenz-96 and in optimizer steps elsewhere. The MUNDO training budget is reported per sampler level; end-to-end fine-tuning updates the complete hierarchy jointly. [a]Sampling measured on eight CPU cores. [b]FlowDPS additionally applies a data-consistency projection at every solver step.

		Model			Training				Cost		
Tier	Arm	Backbone	Width	Params	Objective	Budget	Batch	lr	Wall (h)	NFE	ms/member
Lorenz-96 $F = 32$, $N = 64$	MUNDO ($R = 6$)	circ. CNN	256	6.18 M	ES	60 ep.	256	10^{-3}	0.10	6	1.6[a]
	MUNDO ($R = 1$)	circ. CNN	620	6.18 M	ES	60 ep.	256	10^{-3}	0.06	1	1.4[a]
	Flow matching	circ. CNN	618	6.18 M	flow matching	60 ep.	256	10^{-3}	0.06	56	2.6
	FlowDPS	circ. CNN	618	6.18 M	flow matching	60 ep.	256	10^{-3}	0.05	56[b]	2.7
Gross–Pitaevskii $N = 32$	MUNDO ($R = 5$)	NAT U-Net	37	12.7 M	ES+VS $\to$ ES+SC	40k $\to$ 8k	16	$3{\cdot}10^{-4} \to 10^{-4}$	3.1 + 0.6	5	1.3
	MUNDO ($R = 1$)	NAT U-Net	85	12.6 M	ES+VS	40k	16	$3{\cdot}10^{-4}$	1.4	1	0.7
	Flow matching	NAT U-Net	85	12.4 M	flow matching	40k	16	$3{\cdot}10^{-4}$	0.72	56	39
	FlowDPS	NAT U-Net	85	12.4 M	flow matching	40k	16	$3{\cdot}10^{-4}$	0.70	56[b]	40
	SimVP	SimVP-gSTA	–	9.7 M	MSE	40k	16	$3{\cdot}10^{-4}$	0.36	1	2.4
ERA5 $N = 24$	MUNDO ($R = 4$)	ArchesWeather	96	4×57.4 M	ES+VS	40k	4	$3{\cdot}10^{-4}$	25.5	4	507
	MUNDO ($R = 1$)	ArchesWeather	96	57.4 M	ES+VS	40k	4	$3{\cdot}10^{-4}$	6.9	1	127
	Flow matching	ArchesWeather	96	57.4 M	flow matching	40k	4	$3{\cdot}10^{-4}$	3.4	25	3150
	FlowDPS	ArchesWeather	96	57.4 M	flow matching	40k	4	$3{\cdot}10^{-4}$	13.2	25[b]	3160
	SimVP	SimVP-gSTA	–	7.3 M	MSE	40k	4	$3{\cdot}10^{-4}$	2.6	1	6

B.4 Training configuration and baselines

For each tier, single-step and recursive MUNDO, conditional flow matching, and FlowDPS use the same backbone and are matched in parameter count where feasible; SimVP retains its original architecture. Separate models are trained for interpolation and disaggregation. All learned arms use AdamW (Loshchilov & Hutter, 2019). AutoSim and ERA5 use cosine decay with 500 warm-up steps. The MUNDO training loss is estimated from $M = 2$ draws. Table 11 summarizes the training configuration and sampling cost. For AutoSim, it uses Gross–Pitaevskii at $N = 32$ as a representative setting.

Capacity on ERA5. On ERA5, recursive MUNDO uses four 57.4 M-parameter ArchesWeather networks, whereas single-step MUNDO and the flow-based baselines each use one. Training a capacity-matched baseline was beyond our compute budget, so the recursive comparison does not separate the effect of the hierarchy from additional capacity. Single-step MUNDO provides the comparison at matched parameter count. Training the four recursive levels takes 25.5 h in total, compared with 6.9 h for single-step MUNDO.

On AutoSim and ERA5, every learned baseline is projected onto the constraint set after sampling using the map in Equation (9). The same projection is used by FlowDPS as its consistency update at each sampling step. Its prior is trained with the same flow-matching loss and does not see the window observation, which enters only through this projection at each step. Conditional flow matching is the conditional counterpart and has no consistency step. Network evaluations per member are reported in Table 11. Lorenz-96 instead uses the reference implementation of Kim et al. (2025): conditional flow matching has no consistency mechanism, while FlowDPS applies three gradient steps to the range-space residual.

B.5 Metrics

Let $x = (x_1, \dots, x_N)$ be the withheld fine record of a test window, $y = Ax$ its observation, and $u^{(1)}, \dots, u^{(M)}$ the ensemble members drawn by a method from that observation. Each quantity below is computed per window and averaged over windows, variables and pixels, unless it is defined by pooling windows.

Accuracy. The CRPS is estimated in its fair form (Ferro, 2014),

$$\mathrm{CRPS}\left(u^{1:M}, v\right) = \frac{1}{M}\sum_{m=1}^{M}|u^{(m)} - v| - \frac{1}{2M\,(M-1)}\sum_{m\neq m'}|u^{(m)} - u^{(m')}|, \tag{27}$$

applied at each position i to each variable of x_i.

Calibration. For a calibrated ensemble, the rank of x_i among $u_i^{(1)}, \dots, u_i^{(M)}$ is uniform on $\{1, \dots, M+1\}$, per position and per variable. We summarize its histogram by the realized coverage of central member intervals and its mean absolute error over a grid of nominal levels (Greenbury et al., 2026),

$$\widehat{c}(\alpha) = \left\langle \mathbf{1}\{\widehat{q}_{(1-\alpha)/2}(u_i^{1:M}) \le x_i \le \widehat{q}_{(1+\alpha)/2}(u_i^{1:M})\} \right\rangle, \qquad \text{cov. MAE} = \frac{1}{|A|} \sum_{\alpha \in A} \left|\widehat{c}(\alpha) - \alpha\right|, \tag{28}$$

with $A = \{0.1, 0.2, \dots, 0.9, 0.95, 0.99\}$ and $\widehat{q}$ the member quantile at Weibull plotting positions.

We also report the spread-skill ratio with a finite-ensemble correction (Rasp et al., 2024),

$$\text{SSR} = \sqrt{\frac{M+1}{M}} \, \frac{\sqrt{\left\langle \text{var}(u_i^{1:M}) \right\rangle}}{\sqrt{\left\langle (\overline{u}_i - x_i)^2 \right\rangle}}, \tag{29}$$

where var is the unbiased member variance and $\langle \cdot \rangle$ averages over positions and windows. A value close to 1 indicates agreement between ensemble spread and ensemble-mean error. Values below 1 indicate insufficient spread relative to that error, and values above 1 indicate excess spread.

Consistency. A reconstruction is consistent when it reproduces the coarse observation exactly: retained states under interpolation and imposed aggregates under disaggregation. The consistency error is

$$\epsilon_{\text{cons}} = \begin{cases} \max\limits_{j,m} \left\| (A_{\text{avg}} u^{(m)})_j - y_j \right\|_\infty, & \text{disaggregation}, \\ \max\limits_{j,m} \left\| u_{i_j}^{(m)} - y_j \right\|_\infty, & \text{interpolation}, \end{cases} \tag{30}$$

where i_j is the position of the j-th stored state. A method passes if the corresponding error is below 10^{-4} in standardized units. For the methods marked ✓$^\pi$, ϵ_{cons} is measured after the projection of Appendix B.4.

Temporal coherence. We measure temporal coherence with the ratio of order-p temporal differences

$$t_h = \frac{\langle |u_{t+h} - u_t|^p \rangle}{\langle |x_{t+h} - x_t|^p \rangle}, \qquad p = \tfrac{1}{2}, \tag{31}$$

where the averages are taken over ensemble members, valid time pairs, windows, variables, and spatial positions. We report t_1, with one indicating agreement with the truth. Values above one indicate larger short-lag increments than in the reference, and values below one indicate smaller increments.

Spatial coherence. Using the order-p pairwise differences γ_h^a of (14), we measure spatial coherence with the ratio

$$s_h = \frac{\langle \gamma_h^a(u) \rangle}{\langle \gamma_h^a(x) \rangle}, \qquad p = \tfrac{1}{2}, \tag{32}$$

where the averages are taken over ensemble members, frames, windows, variables, and spatial positions. We report s_1, with one indicating agreement with the truth. Values above one indicate fields that are too rough at the grid scale, and values below one fields that are too smooth.

Fine-scale spectral power. For the two-dimensional systems, we additionally report the ratio of generated to true power in the finest resolved spatial band. On the periodic AutoSim domains, spectra are computed along one periodic axis without windowing, and the finest band is the upper half of the resolved wavenumber range, $[K/2, K]$, with K the Nyquist wavenumber. The reported ratio sums power over this band and averages over ensemble members, frames, windows, and the orthogonal spatial axis. A value of one indicates that the reconstruction carries the correct amount of power at the finest resolved scales. This metric is insensitive to where that power is located and is therefore used only as a complementary diagnostic in the spatial-coherence ablation of Appendix A.2.

Cross-window continuity. Within-window metrics do not detect artifacts at the junction between consecutive reconstructions, so we evaluate stitched chains of windows. Let $\mathcal{B}$ denote the time steps crossing a window boundary, and let $\tilde{u}$ and $\tilde{x}$ be the stitched reconstruction and truth. We define the seam ratio

$$r_{\text{seam}} = \frac{\left\langle |\tilde{u}_{t+1} - \tilde{u}_t| \right\rangle_{t \in \mathcal{B}}}{\left\langle |\tilde{x}_{t+1} - \tilde{x}_t| \right\rangle_{t \in \mathcal{B}}}, \tag{33}$$

where the average is taken over boundary steps, chains, variables, and spatial positions. A value of one indicates that cross-window steps have the correct magnitude, but not necessarily the correct direction or value, so the metric can be improved by compensating errors.

C Additional results

C.1 Lorenz-96

Tables 12 and 13 extend the main Lorenz-96 results to the remaining forcings and to windows of $N = 128$ steps. Each panel reports two measures of how much the dynamics decorrelate within a window of duration $T = N\delta$: the ratio $N\delta/\tau$, where τ is the e-folding decorrelation time of the slow variables, and λT, where λ is the leading Lyapunov exponent, so that λT counts the e-foldings of a small perturbation over the window. Both grow with F and N, from the non-chaotic $F = 10$ ($\lambda = 0$) to about thirteen e-foldings at $F = 64$, $N = 128$.

Under disaggregation, recursive MUNDO attains the lowest CRPS of all methods in five of the eight settings, with the largest margin at $F = 24$, $N = 64$. In the other three, $F = 64$ at both window lengths and $F = 24$ at $N = 128$, the learned methods converge in CRPS and flow matching is slightly lower, although it does not satisfy the observation. Recursive MUNDO also has a lower coverage error than single-step MUNDO in six of the eight settings, and keeps the temporal variogram ratio close to one where single-step MUNDO falls below it. Under interpolation, recursion improves CRPS only in the non-chaotic regime. For $F \geq 24$, single-step MUNDO has the lower CRPS at both window lengths, while recursive MUNDO retains the lower coverage error in every setting. At low forcing, flow matching is strongly overdispersed under interpolation and overestimates short-lag variability. The classical baselines satisfy the constraint but remain far behind in CRPS, and their t_1 below one reflects the missing short-lag variability of smooth reconstructions.

Figures 5 and 6 show one member of each method at $F = 32$, $N = 64$ over four consecutive windows, and Figure 7 the ensembles at one slow variable. The recursive and single-step MUNDO members are smooth in time, like the reference trajectory, whereas both flow-based methods are noisy from one step to the next, most visibly for flow matching under interpolation, which is the excess short-lag variability that t_1 measures. FlowDPS also misses the observation.

C.2 AutoSim

Tables 14–24 report the remaining AutoSim configurations: the benchmark tier at $N \in \{8, 16, 24\}$, and the mild and chaotic tiers at $N \in \{8, 16, 24, 32\}$. The mild and chaotic tiers shorten the decorrelation time of the dynamics, through stronger advection, buoyancy, or obstacle forcing, or through a coarser snapshot stride for Gray–Scott (Table 10), so that a window of fixed length spans more decorrelation times. Panel captions report the corresponding $N\delta/\tau$, defined as for Lorenz-96. In every table, the recursive MUNDO variant, plain or fine-tuned, is selected on the validation trajectories (Appendix A.4). Gross–Pitaevskii has no chaotic tier. Figures 8 and 9 show reconstructions on the two systems not illustrated in the main text, with FlowDPS and SimVP added to the rows.

Under interpolation, single-step MUNDO attains the lowest CRPS in most settings, whereas recursive MUNDO usually attains the lower coverage error. Under disaggregation, single-step MUNDO keeps the lower CRPS in most settings, and recursion lowers it where predictability is lowest: on chaotic advection-diffusion at $N \geq 16$ by 19–39%, and on Gross–Pitaevskii at $N \geq 24$ in the benchmark tier and $N \geq 16$ in the mild tier by 4–14%. Recursion is the better calibrated sampler: its spread-skill ratio is closer to one in 37 of the 44 settings and its coverage error lower in 23, and it corrects the dispersion of single-step MUNDO where it fails, namely overdispersion on Gray–Scott in the benchmark and mild tiers and underdispersion on chaotic advection-diffusion at $N \geq 16$. On advection-diffusion, the near-deterministic control, flow matching has the lowest disaggregation

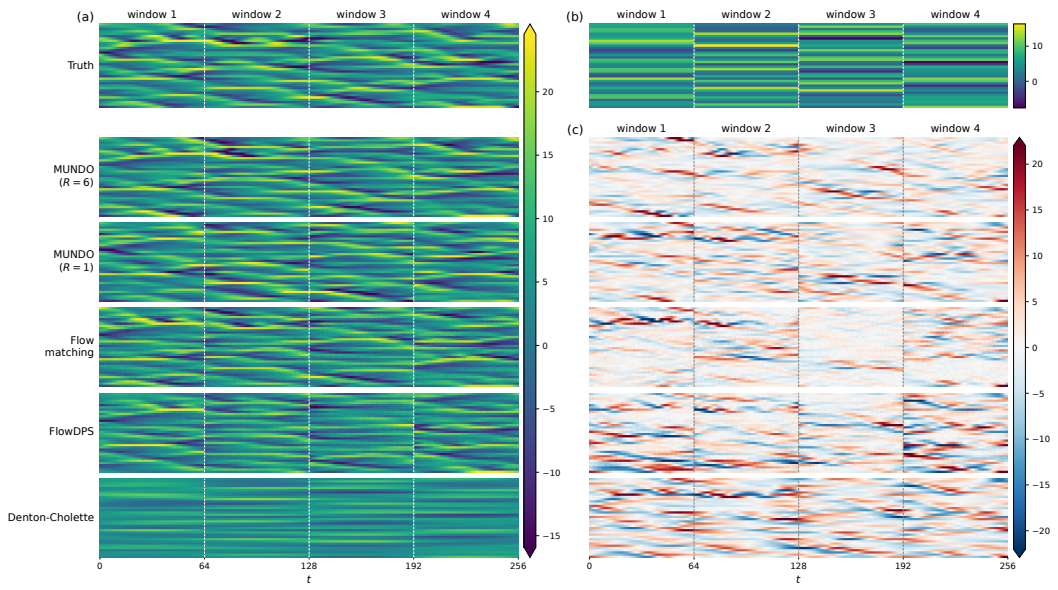

Figure 5: **Lorenz-96 disaggregation at** $F = 32$**,** $N = 64$, four consecutive windows, one member per method. (a) Truth and reconstructions, slow variables against time. (b) Observed window means. (c) Reconstruction errors $\hat{x} - x$; rows follow the method order in (a). Dashed lines mark the window boundaries.

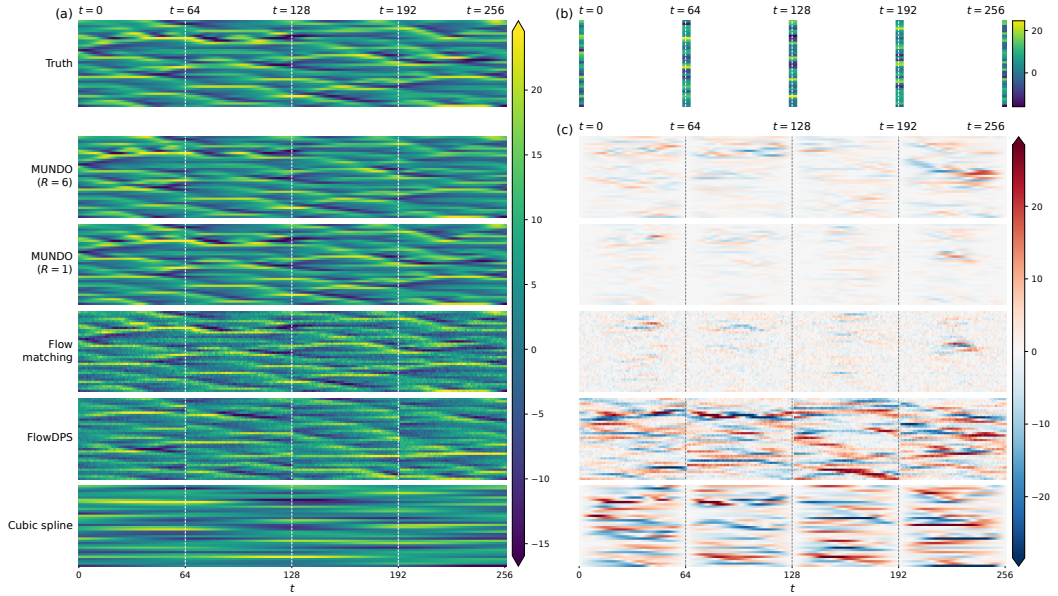

Figure 6: **Lorenz-96 interpolation at** $F = 32$**,** $N = 64$, four consecutive windows, one member per method. (a) Truth and reconstructions, slow variables against time. (b) Observed endpoint states, widened for visibility. (c) Reconstruction errors $\hat{x} - x$; rows follow the method order in (a). Dashed lines mark the observed endpoints.

Table 12: **Lorenz-96 at $N = 64$, $F \in \{10, 24, 64\}$.** †: causal chain. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) $F = 10$ ($N\delta/\tau = 2.1$, $\lambda T = 0.0$; non-chaotic).

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 6$)†	✓	**0.0080**$\pm$0.0001	**1.02**	**0.015**	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	0.0089$\pm$0.0005	**1.02**	0.019	**1.00**
Flow matching	×	0.0113$\pm$0.0000	1.79	0.188	1.14
FlowDPS	×	0.0159$\pm$0.0034	1.35	0.089	1.20
Denton–Cholette	✓	0.3706	–	–	0.76
CDF-t	×	0.2722	–	–	0.93
Interpolation					
MUNDO ($R = 6$)	✓	**0.0065**$\pm$0.0002	**1.05**	**0.045**	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	0.0071$\pm$0.0002	1.34	0.082	**1.00**
Flow matching	×	0.0412$\pm$0.0012	6.55	0.365	2.27
FlowDPS	×	0.2078$\pm$0.0098	1.78	0.203	1.95
Linear	✓	0.4083	–	–	0.85
Cubic spline	✓	0.3682	–	–	0.86

(b) $F = 24$ ($N\delta/\tau = 3.9$, $\lambda T = 1.8$).

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 6$)†	✓	**0.0825**$\pm$0.0019	0.76	0.071	0.99
MUNDO ($R = 1$)	✓$^{\pi}$	0.1174$\pm$0.0011	**0.78**	**0.034**	**1.00**
Flow matching	×	0.1234$\pm$0.0006	0.87	0.014	1.17
FlowDPS	×	0.2318$\pm$0.0010	1.19	0.046	1.15
Denton–Cholette	✓	0.6377	–	–	0.43
CDF-t	×	0.7423	–	–	0.58
Interpolation					
MUNDO ($R = 6$)	✓	0.0428$\pm$0.0006	0.85	**0.020**	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0262**$\pm$0.0004	**1.12**	0.047	**1.00**
Flow matching	×	0.0692$\pm$0.0012	2.25	0.257	2.05
FlowDPS	×	0.4160$\pm$0.0025	1.15	0.044	1.75
Linear	✓	0.7279	–	–	0.56
Cubic spline	✓	0.7710	–	–	0.57

(c) $F = 64$ ($N\delta/\tau = 8.3$, $\lambda T = 6.5$).

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 6$)†	✓	**0.4821**$\pm$0.0018	**1.01**	**0.023**	**1.02**
MUNDO ($R = 1$)	✓$^{\pi}$	0.4829$\pm$0.0002	1.02	0.049	0.80
Flow matching	×	0.4728$\pm$0.0003	0.89	0.045	1.00
FlowDPS	×	0.4909$\pm$0.0011	0.86	0.058	0.96
Denton–Cholette	✓	0.7141	–	–	0.25
CDF-t	×	0.8570	–	–	0.40
Interpolation					
MUNDO ($R = 6$)	✓	0.4038$\pm$0.0063	0.91	**0.023**	0.93
MUNDO ($R = 1$)	✓$^{\pi}$	**0.2731**$\pm$0.0023	**0.93**	0.062	**0.97**
Flow matching	×	0.3622$\pm$0.0056	0.97	0.037	1.54
FlowDPS	×	0.5191$\pm$0.0007	0.93	0.031	1.34
Linear	✓	0.8661	–	–	0.38
Cubic spline	✓	0.9093	–	–	0.39

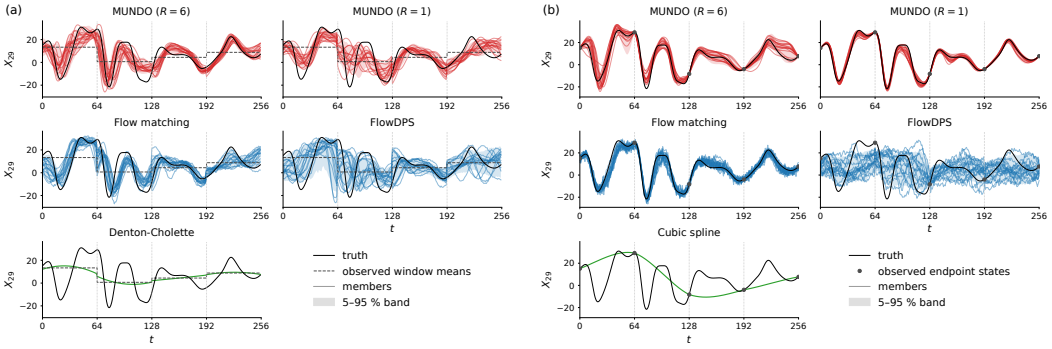

Figure 7: **Lorenz-96 at $F = 32$, $N = 64$, one slow variable** over the four windows of Figures 5 and 6. (a) Disaggregation, (b) interpolation. Each panel shows the truth, the observation, 16 of the members as thin lines and the 5–95 % range over all members as a band. The classical baselines show their single reconstruction. The variable is the one with the largest variance over the run.

CRPS in the chaotic tier and at $N \geq 24$ in the mild tier, although its ensembles are underdispersed there. Recursive MUNDO also slightly overestimates short-lag variability on Navier–Stokes interpolation, where single-step MUNDO keeps t_1 closer to one. Among the baselines, FlowDPS is the weakest learned method throughout. SimVP satisfies the observation and matches the short-lag structure on Gray–Scott and Navier–Stokes, but, being deterministic, its CRPS is about 1.2 to 10 times that of MUNDO. The classical baselines are the most oversmoothed, particularly on Gross–Pitaevskii.

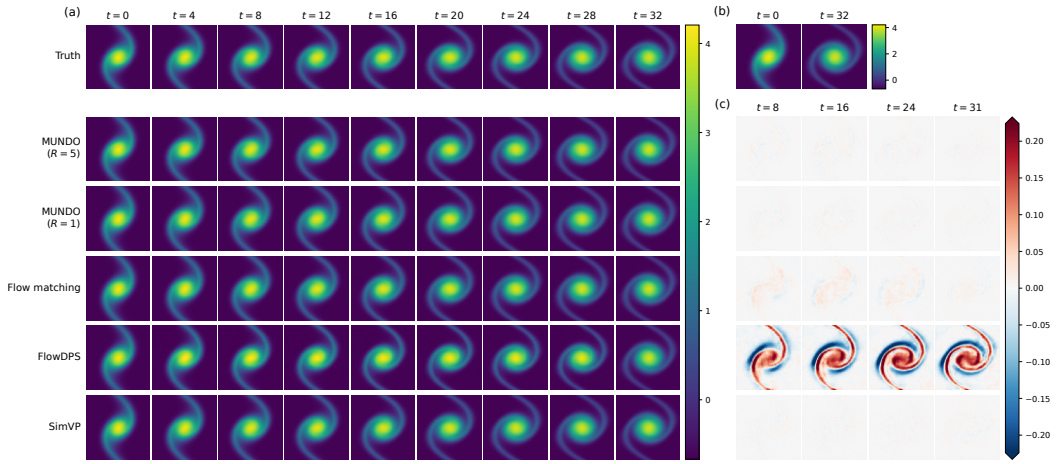

Figure 8: **Advection-diffusion interpolation at** $N = 32$, benchmark tier. (a) Ground truth and reconstructions at selected times. (b) Observed endpoint states. (c) Reconstruction errors $\hat{x} - x$ at selected times; rows follow the method order in (a).

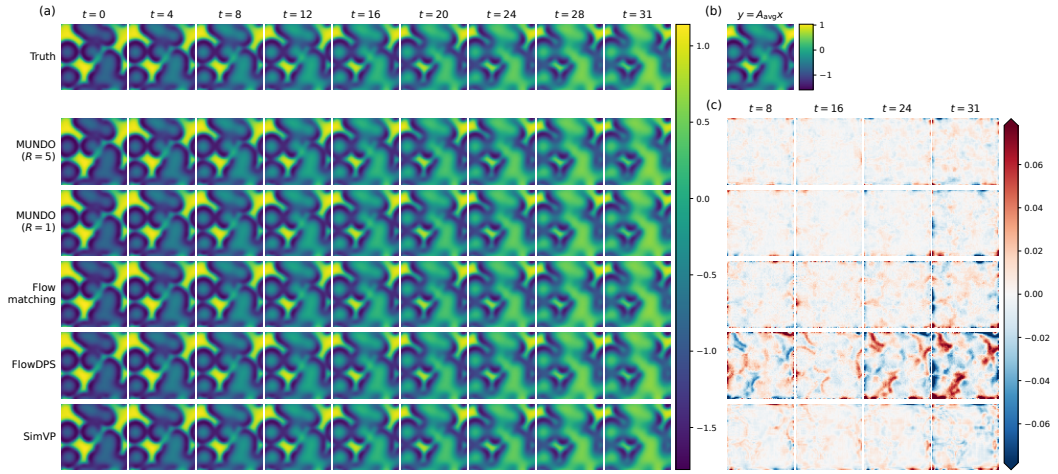

Figure 9: **Gray–Scott disaggregation at** $N = 32$, benchmark tier. (a) Ground truth and reconstructions at selected times. (b) Coarse observation, given by the temporal mean over the window. (c) Reconstruction errors $\hat{x} - x$ at selected times; rows follow the method order in (a).

Table 13: **Lorenz-96 at $N = 128$, $F \in \{10, 24, 32, 64\}$.** †: causal chain. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) $F = 10$ ($N\delta/\tau = 4.2$, $\lambda T = 0.0$; non-chaotic).

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	t_1 → 1
Disaggregation					
MUNDO ($R = 7$)†	✓	**0.0153**$\pm$0.0014	**1.00**	**0.014**	**1.00**
MUNDO ($R = 1$)	✓$^\pi$	0.0171$\pm$0.0018	1.08	0.024	**1.00**
Flow matching	×	0.0202$\pm$0.0007	2.35	0.236	1.46
FlowDPS	×	0.0981$\pm$0.0028	2.31	0.248	1.38
Denton–Cholette	✓	0.8633	–	–	0.32
CDF-t	×	1.1244	–	–	0.57
Interpolation					
MUNDO ($R = 7$)	✓	**0.0141**$\pm$0.0012	**1.15**	**0.039**	**1.00**
MUNDO ($R = 1$)	✓$^\pi$	0.0186$\pm$0.0039	1.33	0.047	**1.00**
Flow matching	×	0.0717$\pm$0.0020	7.33	0.372	3.00
FlowDPS	×	0.3866$\pm$0.0038	1.29	0.097	2.54
Linear	✓	0.9768	–	–	0.55
Cubic spline	✓	1.0875	–	–	0.56

(b) $F = 24$ ($N\delta/\tau = 7.9$, $\lambda T = 3.7$).

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	t_1 → 1
Disaggregation					
MUNDO ($R = 7$)†	✓	**0.3275**$\pm$0.0022	0.98	**0.012**	**0.99**
MUNDO ($R = 1$)	✓$^\pi$	0.3412$\pm$0.0041	**1.01**	0.013	0.93
Flow matching	×	0.3127$\pm$0.0022	0.93	0.011	1.34
FlowDPS	×	0.4510$\pm$0.0010	1.03	0.007	1.27
Denton–Cholette	✓	0.7441	–	–	0.25
CDF-t	×	0.8923	–	–	0.39
Interpolation					
MUNDO ($R = 7$)	✓	0.2441$\pm$0.0031	**0.95**	**0.016**	0.97
MUNDO ($R = 1$)	✓$^\pi$	**0.1578**$\pm$0.0016	0.86	0.042	**1.00**
Flow matching	×	0.2103$\pm$0.0015	1.86	0.255	3.33
FlowDPS	×	0.5132$\pm$0.0006	1.06	0.016	2.29
Linear	✓	0.8944	–	–	0.39
Cubic spline	✓	0.9411	–	–	0.40

(c) $F = 32$ ($N\delta/\tau = 9.8$, $\lambda T = 6.2$).

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	t_1 → 1
Disaggregation					
MUNDO ($R = 7$)†	✓	**0.4810**$\pm$0.0023	**1.00**	**0.021**	**0.99**
MUNDO ($R = 1$)	✓$^\pi$	0.4986$\pm$0.0010	0.98	0.046	0.82
Flow matching	×	0.4915$\pm$0.0004	0.89	0.043	1.23
FlowDPS	×	0.5127$\pm$0.0010	0.93	0.029	1.18
Denton–Cholette	✓	0.7565	–	–	0.21
CDF-t	×	0.9134	–	–	0.37
Interpolation					
MUNDO ($R = 7$)	✓	0.3927$\pm$0.0042	**0.96**	**0.004**	0.95
MUNDO ($R = 1$)	✓$^\pi$	**0.2970**$\pm$0.0008	0.91	0.059	**0.96**
Flow matching	×	0.3162$\pm$0.0017	1.30	0.167	2.91
FlowDPS	×	0.5366$\pm$0.0020	0.99	0.011	2.02
Linear	✓	0.9172	–	–	0.34
Cubic spline	✓	0.9626	–	–	0.35

(d) $F = 64$ ($N\delta/\tau = 16.6$, $\lambda T = 13.1$).

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	t_1 → 1
Disaggregation					
MUNDO ($R = 7$)†	✓	**0.5377**$\pm$0.0009	**1.01**	**0.032**	1.08
MUNDO ($R = 1$)	✓$^\pi$	0.5495$\pm$0.0060	1.04	0.099	**0.93**
Flow matching	×	0.5339$\pm$0.0001	0.88	0.050	1.07
FlowDPS	×	0.5376$\pm$0.0005	0.83	0.071	1.01
Denton–Cholette	✓	0.7628	–	–	0.15
CDF-t	×	0.9584	–	–	0.29
Interpolation					
MUNDO ($R = 7$)	✓	0.5024$\pm$0.0013	0.92	**0.029**	**1.01**
MUNDO ($R = 1$)	✓$^\pi$	**0.4624**$\pm$0.0046	**0.98**	0.052	0.76
Flow matching	×	0.5058$\pm$0.0048	0.96	0.010	1.61
FlowDPS	×	0.5520$\pm$0.0006	0.91	0.041	1.41
Linear	✓	0.9475	–	–	0.27
Cubic spline	✓	0.9934	–	–	0.28

C.3 ERA5

C.3.1 Hourly scores per variable

Figures 10 and 11 resolve the fair CRPS by hour of day and by variable, in physical units. Under disaggregation, the ordering of Table 3 holds at almost every hour and variable, and the error varies little across the day except for precipitation, which peaks with afternoon convection. Under interpolation, error vanishes at the observed states and peaks near the middle of the window, where the classical interpolants are worst; flow matching falls furthest behind MUNDO there, especially for wind and pressure.

C.3.2 Power spectra

The spatial power spectrum measures how variance is distributed across spatial scales, so over-smoothing or excess small-scale noise appears directly as missing or excess fine-scale power. Figures 12 and 13 show the resulting spectra. Under disaggregation, both MUNDO variants remain close to ERA5 across variables, while flow matching and SimVP lose fine-scale power, mildly for winds and dewpoint and most strongly for precipitation, where they remain below the reference across scales. Under interpolation, SimVP undershoots across variables, and for precipitation all methods lose power below a few hundred kilometers, with recursive and single-step MUNDO closest to ERA5.

C.3.3 Diurnal cycle

The diurnal cycle is the dominant sub-daily signal of surface temperature and is absent from a daily mean, so recovering its amplitude and phase tests whether the generated hours carry physically timed structure rather than noise. Figures 14 and 15 show the mean diurnal cycle of 2m temperature, as anomalies from each day's mean, at a continental pixel, a maritime pixel, and averaged over the domain. Both MUNDO variants reproduce the amplitude and phase of the mean cycle on both tasks, including the nearly flat maritime cycle.

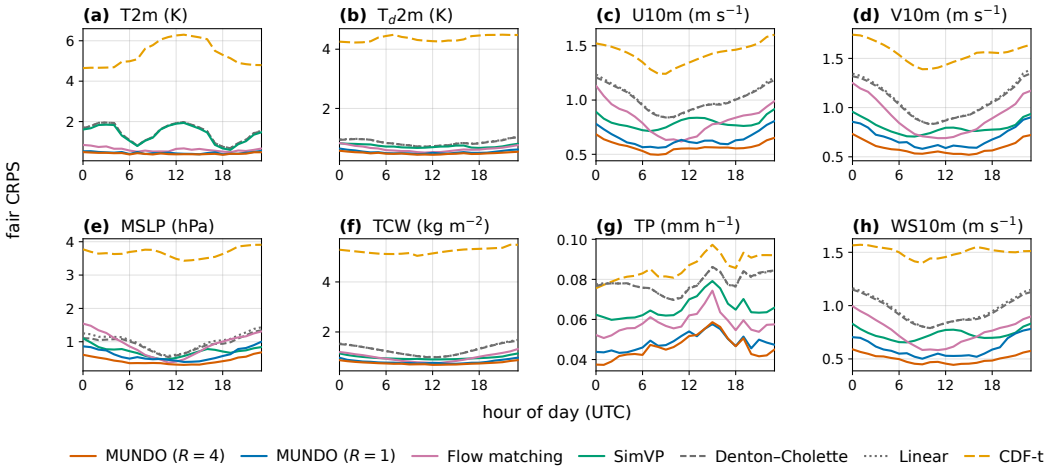

Figure 10: **Fair CRPS by hour of day under disaggregation.** Per variable in physical units, averaged over the test windows. Wind speed (h) is computed from the two wind components.

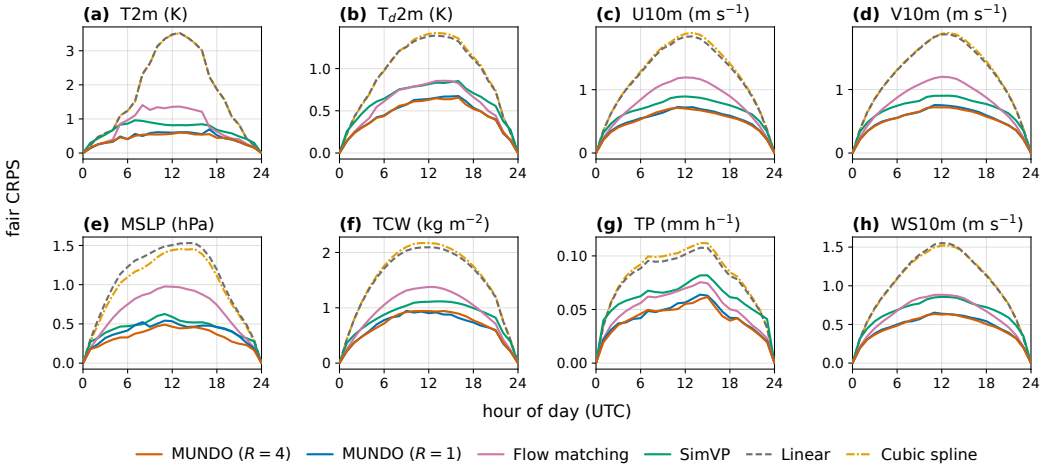

Figure 11: **Fair CRPS by hour of day under interpolation.** Same construction as Figure 10.

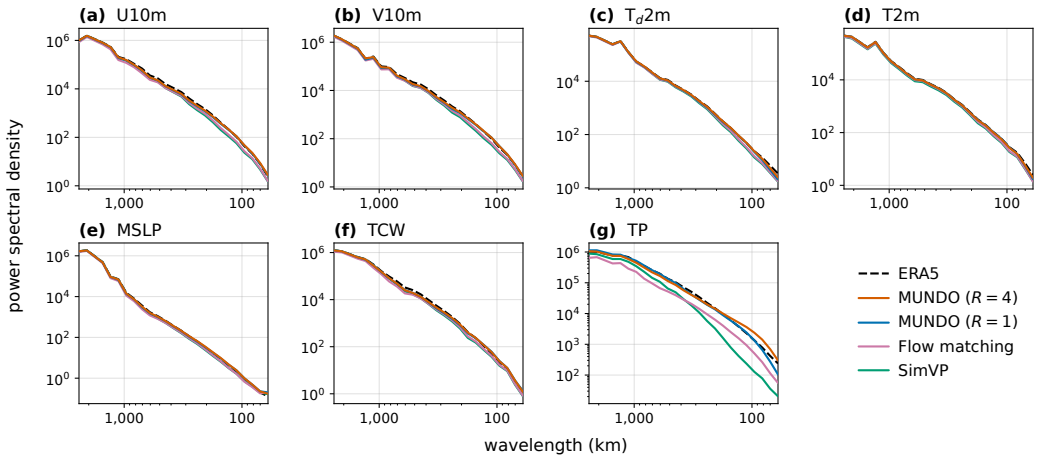

Figure 12: **Spatial power spectra under disaggregation.** Power spectral density per variable, averaged over test windows and ensemble members, against ERA5 (dashed).

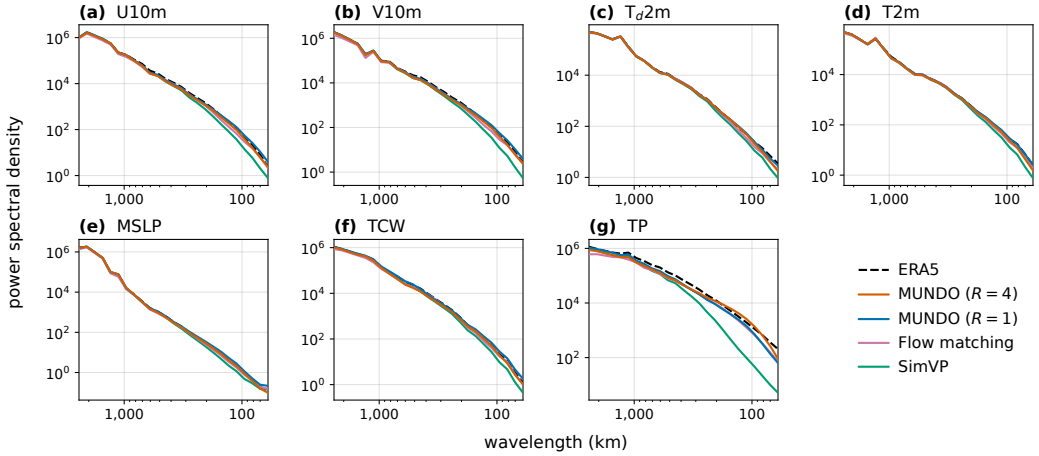

Figure 13: **Spatial power spectra under interpolation.** Same construction as Figure 12.

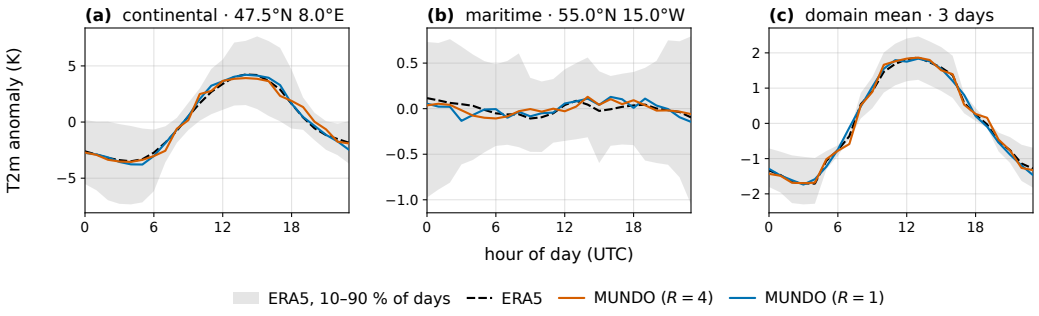

Figure 14: **Diurnal cycle of 2m temperature under disaggregation.** Shown as anomalies from each day's mean, at a continental pixel, a maritime pixel, and averaged over the domain. Lines are the average over the test days, one member per day; shading is the 10–90% range of ERA5's day-to-day cycles.

Table 14: **AutoSim at $N = 32$, most chaotic tier of each system.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion** ($N\delta/\tau = 4.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	t_1 → 1
Disaggregation					
MUNDO ($R = 5$)‡	✓	0.0263$\pm$0.0008	**1.02**	**0.011**	0.94
MUNDO ($R = 1$)	✓$^{\pi}$	0.0432$\pm$0.0063	0.69	0.038	0.71
Flow matching	✓$^{\pi}$	**0.0221**$\pm$0.0000	0.74	0.046	**0.95**
FlowDPS	✓$^{\pi}$	0.0603$\pm$0.0030	0.53	0.158	0.68
SimVP	✓$^{\pi}$	0.0331	–	–	0.78
Denton–Cholette	✓	0.1041	–	–	0.26
CDF-t	×	0.1181	–	–	0.28
Interpolation					
MUNDO ($R = 5$)‡	✓	0.0266$\pm$0.0003	0.97	0.032	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0213**	**0.99**	0.047	0.87
Flow matching	✓$^{\pi}$	**0.0213**$\pm$0.0001	0.77	**0.025**	0.96
FlowDPS	✓$^{\pi}$	0.0765$\pm$0.0027	0.67	0.092	0.76
SimVP	✓$^{\pi}$	0.0353	–	–	0.81
Linear	✓	0.1348	–	–	0.36
Cubic spline	✓	0.1412	–	–	0.37

(b) **Gray–Scott** ($N\delta/\tau = 3.2$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	t_1 → 1
Disaggregation					
MUNDO ($R = 5$)‡	✓	0.0199$\pm$0.0006	**1.03**	0.047	1.01
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0146**$\pm$0.0001	0.82	0.025	1.01
Flow matching	✓$^{\pi}$	0.0176$\pm$0.0001	0.88	0.081	1.02
FlowDPS	✓$^{\pi}$	0.1865$\pm$0.0031	1.05	**0.017**	0.97
SimVP	✓$^{\pi}$	0.0309	–	–	**1.00**
Denton–Cholette	✓	0.6017	–	–	0.58
CDF-t	×	0.6330	–	–	0.62
Interpolation					
MUNDO ($R = 5$)‡	✓	0.0169$\pm$0.0001	**0.95**	0.040	1.02
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0112**$\pm$0.0002	0.77	**0.032**	1.01
Flow matching	✓$^{\pi}$	0.0142$\pm$0.0002	0.90	0.093	1.02
FlowDPS	✓$^{\pi}$	0.2287$\pm$0.0095	0.90	0.052	0.97
SimVP	✓$^{\pi}$	0.0214	–	–	**1.00**
Linear	✓	0.6696	–	–	0.66
Cubic spline	✓	0.6924	–	–	0.69

(c) **Navier–Stokes** ($N\delta/\tau = 1.5$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	t_1 → 1
Disaggregation					
MUNDO ($R = 5$)‡	✓	0.1076$\pm$0.0008	**0.82**	**0.027**	1.24
MUNDO ($R = 1$)	✓$^{\pi}$	0.0977$\pm$0.0038	0.76	**0.027**	0.96
Flow matching	✓$^{\pi}$	**0.0951**$\pm$0.0002	0.73	0.041	1.29
FlowDPS	✓$^{\pi}$	0.1418$\pm$0.0006	0.51	0.067	**1.02**
SimVP	✓$^{\pi}$	0.1352	–	–	0.94
Denton–Cholette	✓	0.1991	–	–	0.63
CDF-t	×	0.3357	–	–	0.73
Interpolation					
MUNDO ($R = 5$)‡	✓	0.0663$\pm$0.0016	**0.79**	0.036	1.26
MUNDO ($R = 1$)	✓$^{\pi}$	0.0650$\pm$0.0036	0.77	**0.028**	1.02
Flow matching	✓$^{\pi}$	**0.0600**$\pm$0.0002	0.71	0.051	1.19
FlowDPS	✓$^{\pi}$	0.2544$\pm$0.0021	0.68	0.071	1.47
SimVP	✓$^{\pi}$	0.0778	–	–	**1.00**
Linear	✓	0.2409	–	–	0.70
Cubic spline	✓	0.2387	–	–	0.72

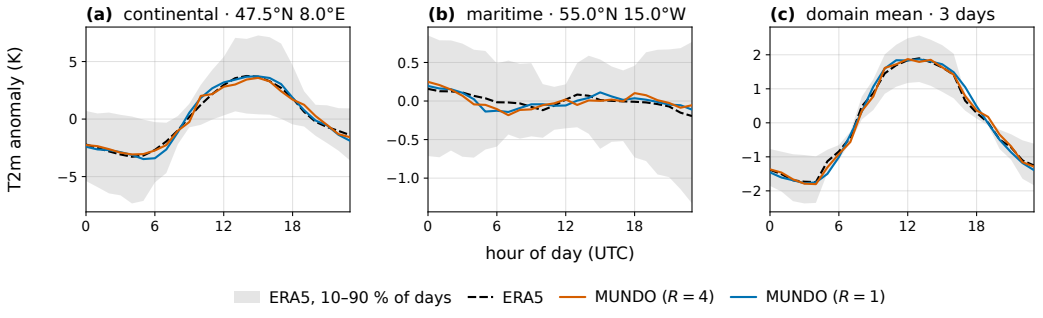

Figure 15: **Diurnal cycle of 2m temperature under interpolation.** Same construction as Figure 14.

Table 15: **AutoSim at $N = 24$.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)	✓	0.0020$\pm$0.0000	**0.93**	0.114	**1.02**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0016**$\pm$0.0000	0.84	0.158	1.05
Flow matching	✓$^{\pi}$	0.0022$\pm$0.0000	0.78	0.179	1.17
FlowDPS	✓$^{\pi}$	0.0096$\pm$0.0004	0.48	**0.050**	1.18
SimVP	✓$^{\pi}$	0.0037	–	–	1.04
Denton–Cholette	✓	0.0310	–	–	0.93
CDF-t	×	0.0660	–	–	0.90
Interpolation					
MUNDO ($R = 4$)	✓	0.0015$\pm$0.0000	0.99	0.099	1.09
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0010**$\pm$0.0000	**1.00**	0.172	**1.04**
Flow matching	✓$^{\pi}$	0.0017$\pm$0.0000	0.87	0.245	1.17
FlowDPS	✓$^{\pi}$	0.0205$\pm$0.0013	0.65	**0.027**	1.22
SimVP	✓$^{\pi}$	0.0036	–	–	**1.04**
Linear	✓	0.0387	–	–	0.91
Cubic spline	✓	0.0350	–	–	0.95

(b) **Gray–Scott**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)	✓	**0.0028**$\pm$0.0000	1.10	**0.043**	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0028**$\pm$0.0000	1.35	0.048	1.01
Flow matching	✓$^{\pi}$	0.0039$\pm$0.0000	0.91	0.129	1.06
FlowDPS	✓$^{\pi}$	0.0094$\pm$0.0003	**0.94**	0.071	1.07
SimVP	✓$^{\pi}$	0.0081	–	–	1.01
Denton–Cholette	✓	0.0770	–	–	0.91
CDF-t	×	0.1805	–	–	0.95
Interpolation					
MUNDO ($R = 4$)	✓	0.0030$\pm$0.0001	1.04	**0.044**	1.02
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0022**$\pm$0.0000	**1.01**	0.059	1.01
Flow matching	✓$^{\pi}$	0.0033$\pm$0.0000	**1.01**	0.141	1.06
FlowDPS	✓$^{\pi}$	0.0337$\pm$0.0005	1.15	0.067	1.09
SimVP	✓$^{\pi}$	0.0071	–	–	**1.00**
Linear	✓	0.1168	–	–	0.92
Cubic spline	✓	0.0994	–	–	0.94

(c) **Navier–Stokes**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	0.0088$\pm$0.0000	0.75	**0.020**	1.05
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0077**$\pm$0.0000	0.62	0.052	1.02
Flow matching	✓$^{\pi}$	0.0131$\pm$0.0002	**0.79**	0.029	1.13
FlowDPS	✓$^{\pi}$	0.0226$\pm$0.0001	0.60	0.067	1.11
SimVP	✓$^{\pi}$	0.0200	–	–	1.02
Denton–Cholette	✓	0.0434	–	–	0.95
CDF-t	×	0.1230	–	–	**0.99**
Interpolation					
MUNDO ($R = 4$)	✓	0.0074$\pm$0.0000	0.76	**0.017**	1.06
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0063**$\pm$0.0000	0.62	0.048	1.03
Flow matching	✓$^{\pi}$	0.0092$\pm$0.0001	**0.78**	0.027	1.12
FlowDPS	✓$^{\pi}$	0.1420$\pm$0.0070	0.42	0.219	1.25
SimVP	✓$^{\pi}$	0.0139	–	–	**1.02**
Linear	✓	0.0591	–	–	0.96
Cubic spline	✓	0.0453	–	–	**0.98**

(d) **Gross–Pitaevskii**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	**0.0799**$\pm$0.0010	**0.92**	0.065	1.06
MUNDO ($R = 1$)	✓$^{\pi}$	0.0828$\pm$0.0008	0.83	**0.057**	**0.98**
Flow matching	✓$^{\pi}$	0.1085$\pm$0.0010	0.89	0.092	1.06
FlowDPS	✓$^{\pi}$	0.2448$\pm$0.0010	1.12	0.117	1.03
SimVP	✓$^{\pi}$	0.1536	–	–	0.91
Denton–Cholette	✓	0.4941	–	–	0.29
CDF-t	×	0.6127	–	–	0.43
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0253$\pm$0.0003	0.92	**0.055**	1.06
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0226**$\pm$0.0004	**0.95**	0.082	1.06
Flow matching	✓$^{\pi}$	0.0305$\pm$0.0002	0.82	0.113	1.07
FlowDPS	✓$^{\pi}$	0.1385$\pm$0.0016	0.49	0.094	0.91
SimVP	✓$^{\pi}$	0.0388	–	–	**1.02**
Linear	✓	0.5915	–	–	0.41
Cubic spline	✓	0.6193	–	–	0.43

Table 16: **AutoSim at $N = 16$.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)	✓	<u>0.0016</u>$\pm$0.0000	**0.94**	0.132	**1.03**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0012**$\pm$0.0000	0.92	0.175	1.05
Flow matching	✓$^{\pi}$	0.0019$\pm$0.0000	0.81	0.148	1.15
FlowDPS	✓$^{\pi}$	0.0079$\pm$0.0004	0.49	**0.053**	1.15
SimVP	✓$^{\pi}$	0.0040	–	–	1.08
Denton–Cholette	✓	0.0149	–	–	0.95
CDF-t	×	0.0551	–	–	0.95
Interpolation					
MUNDO ($R = 4$)	✓	<u>0.0011</u>$\pm$0.0000	**0.94**	0.117	1.08
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0006**$\pm$0.0000	**1.06**	0.199	1.04
Flow matching	✓$^{\pi}$	0.0012$\pm$0.0000	0.91	0.247	1.15
FlowDPS	✓$^{\pi}$	0.0134$\pm$0.0008	0.64	**0.040**	1.18
SimVP	✓$^{\pi}$	0.0028	–	–	1.03
Linear	✓	0.0207	–	–	0.95
Cubic spline	✓	0.0162	–	–	**0.98**

(b) **Gray–Scott**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)	✓	**0.0023**$\pm$0.0000	1.19	**0.053**	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	<u>0.0025</u>$\pm$0.0000	1.26	0.057	1.01
Flow matching	✓$^{\pi}$	0.0030$\pm$0.0000	**0.99**	0.133	1.05
FlowDPS	✓$^{\pi}$	0.0074$\pm$0.0001	0.85	0.088	1.06
SimVP	✓$^{\pi}$	0.0081	–	–	1.01
Denton–Cholette	✓	0.0419	–	–	0.91
CDF-t	×	0.1559	–	–	0.97
Interpolation					
MUNDO ($R = 4$)	✓	<u>0.0026</u>$\pm$0.0000	1.10	**0.048**	1.02
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0020**$\pm$0.0000	1.03	0.064	1.01
Flow matching	✓$^{\pi}$	0.0027$\pm$0.0000	**1.01**	0.139	1.04
FlowDPS	✓$^{\pi}$	0.0245$\pm$0.0003	1.18	0.090	1.08
SimVP	✓$^{\pi}$	0.0068	–	–	**1.00**
Linear	✓	0.0687	–	–	0.95
Cubic spline	✓	0.0555	–	–	0.96

(c) **Navier–Stokes**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	<u>0.0043</u>$\pm$0.0000	**0.85**	0.039	1.05
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0037**$\pm$0.0000	0.74	**0.022**	1.03
Flow matching	✓$^{\pi}$	0.0075$\pm$0.0001	0.65	0.037	1.06
FlowDPS	✓$^{\pi}$	0.0142$\pm$0.0001	0.59	0.064	1.07
SimVP	✓$^{\pi}$	0.0127	–	–	1.03
Denton–Cholette	✓	0.0161	–	–	**0.98**
CDF-t	×	0.1086	–	–	**1.02**
Interpolation					
MUNDO ($R = 4$)‡	✓	<u>0.0043</u>$\pm$0.0000	**0.79**	**0.022**	1.07
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0038**$\pm$0.0000	0.75	0.027	1.03
Flow matching	✓$^{\pi}$	0.0051$\pm$0.0000	0.70	0.028	1.07
FlowDPS	✓$^{\pi}$	0.0627$\pm$0.0017	0.62	0.118	1.13
SimVP	✓$^{\pi}$	0.0092	–	–	1.02
Linear	✓	0.0292	–	–	0.98
Cubic spline	✓	0.0170	–	–	**1.00**

(d) **Gross–Pitaevskii**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	<u>0.0509</u>$\pm$0.0004	**0.96**	**0.073**	1.06
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0490**$\pm$0.0005	0.91	0.079	1.03
Flow matching	✓$^{\pi}$	0.0652$\pm$0.0005	0.89	0.081	1.02
FlowDPS	✓$^{\pi}$	0.1430$\pm$0.0008	1.17	0.106	**1.00**
SimVP	✓$^{\pi}$	0.1023	–	–	0.95
Denton–Cholette	✓	0.4324	–	–	0.41
CDF-t	×	0.5194	–	–	0.52
Interpolation					
MUNDO ($R = 4$)‡	✓	<u>0.0149</u>$\pm$0.0001	**0.94**	**0.060**	1.05
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0117**$\pm$0.0001	0.90	0.087	1.05
Flow matching	✓$^{\pi}$	0.0151$\pm$0.0002	0.81	0.094	1.06
FlowDPS	✓$^{\pi}$	0.0781$\pm$0.0002	0.57	0.061	0.94
SimVP	✓$^{\pi}$	0.0242	–	–	**1.03**
Linear	✓	0.5174	–	–	0.51
Cubic spline	✓	0.5396	–	–	0.53

Table 17: **AutoSim at $N = 8$.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)‡	✓	**0.0004**$\pm$0.0000	0.89	0.178	1.02
MUNDO ($R = 1$)	✓$^{\pi}$	0.0005$\pm$0.0000	0.96	0.219	1.05
Flow matching	✓$^{\pi}$	0.0021$\pm$0.0000	**1.02**	0.096	1.11
FlowDPS	✓$^{\pi}$	0.0049$\pm$0.0001	0.48	**0.059**	1.14
SimVP	✓$^{\pi}$	0.0027	–	–	1.05
Denton–Cholette	✓	0.0028	–	–	0.98
CDF-t	×	0.0493	–	–	**0.99**
Interpolation					
MUNDO ($R = 3$)	✓	0.0005$\pm$0.0000	**1.02**	0.213	1.06
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0003**$\pm$0.0000	1.18	0.250	1.05
Flow matching	✓$^{\pi}$	0.0008$\pm$0.0000	0.78	0.224	1.11
FlowDPS	✓$^{\pi}$	0.0078$\pm$0.0001	0.72	**0.117**	1.15
SimVP	✓$^{\pi}$	0.0017	–	–	1.03
Linear	✓	0.0065	–	–	0.98
Cubic spline	✓	0.0032	–	–	**1.00**

(b) **Gray–Scott**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)‡	✓	**0.0011**$\pm$0.0000	1.84	0.092	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	0.0012$\pm$0.0000	1.22	**0.091**	1.01
Flow matching	✓$^{\pi}$	0.0015$\pm$0.0000	**0.96**	0.150	1.03
FlowDPS	✓$^{\pi}$	0.0032$\pm$0.0000	0.82	0.117	1.05
SimVP	✓$^{\pi}$	0.0049	–	–	1.01
Denton–Cholette	✓	0.0167	–	–	0.96
CDF-t	×	0.1415	–	–	1.01
Interpolation					
MUNDO ($R = 3$)	✓	0.0013$\pm$0.0000	1.12	**0.056**	1.01
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0011**$\pm$0.0000	1.12	0.122	1.01
Flow matching	✓$^{\pi}$	0.0014$\pm$0.0000	**0.99**	0.176	1.03
FlowDPS	✓$^{\pi}$	0.0151$\pm$0.0001	1.15	0.104	1.05
SimVP	✓$^{\pi}$	0.0047	–	–	**1.00**
Linear	✓	0.0275	–	–	0.99
Cubic spline	✓	0.0213	–	–	0.99

(c) **Navier–Stokes**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)	✓	**0.0018**$\pm$0.0000	**0.93**	0.067	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0018**$\pm$0.0000	0.84	0.049	1.03
Flow matching	✓$^{\pi}$	0.0037$\pm$0.0000	0.60	0.052	1.04
FlowDPS	✓$^{\pi}$	0.0070$\pm$0.0000	0.61	**0.041**	1.05
SimVP	✓$^{\pi}$	0.0066	–	–	1.02
Denton–Cholette	✓	0.0033	–	–	**0.99**
CDF-t	×	0.1039	–	–	1.04
Interpolation					
MUNDO ($R = 3$)	✓	0.0025$\pm$0.0000	0.90	0.046	1.05
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0023**$\pm$0.0000	0.88	0.069	1.03
Flow matching	✓$^{\pi}$	0.0026$\pm$0.0000	0.67	**0.033**	1.05
FlowDPS	✓$^{\pi}$	0.0259$\pm$0.0001	**1.01**	0.041	1.05
SimVP	✓$^{\pi}$	0.0053	–	–	1.02
Linear	✓	0.0088	–	–	**1.00**
Cubic spline	✓	0.0036	–	–	1.01

(d) **Gross–Pitaevskii**

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)‡	✓	0.0169$\pm$0.0003	**0.96**	0.094	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0152**$\pm$0.0000	0.92	0.087	1.04
Flow matching	✓$^{\pi}$	0.0194$\pm$0.0001	0.87	0.091	1.03
FlowDPS	✓$^{\pi}$	0.0641$\pm$0.0008	0.82	**0.056**	0.96
SimVP	✓$^{\pi}$	0.0408	–	–	**1.02**
Denton–Cholette	✓	0.2503	–	–	0.65
CDF-t	×	0.2742	–	–	0.70
Interpolation					
MUNDO ($R = 3$)‡	✓	0.0054$\pm$0.0001	**0.93**	0.079	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0041**$\pm$0.0000	0.88	0.084	1.04
Flow matching	✓$^{\pi}$	0.0057$\pm$0.0000	0.74	0.080	1.05
FlowDPS	✓$^{\pi}$	0.0359$\pm$0.0001	0.81	**0.056**	**1.01**
SimVP	✓$^{\pi}$	0.0135	–	–	1.03
Linear	✓	0.3259	–	–	0.70
Cubic spline	✓	0.3156	–	–	0.74

Table 18: **AutoSim at $N = 32$, mild tier of each system.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion** ($N\delta/\tau = 2.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 5$)	✓	0.0091±0.0001	**1.02**	**0.022**	**1.02**
MUNDO ($R = 1$)	✓$^\pi$	0.0082±0.0001	0.54	0.049	**0.98**
Flow matching	✓$^\pi$	**0.0073**±0.0001	0.77	0.073	1.08
FlowDPS	✓$^\pi$	0.0268±0.0006	0.42	0.094	0.97
SimVP	✓$^\pi$	0.0127	–	–	0.96
Denton–Cholette	✓	0.0804	–	–	0.60
CDF-t	×	0.0970	–	–	0.58
Interpolation					
MUNDO ($R = 5$)‡	✓	0.0079±0.0001	0.92	0.061	1.07
MUNDO ($R = 1$)	✓$^\pi$	**0.0057**±0.0000	**0.95**	0.081	**1.00**
Flow matching	✓$^\pi$	0.0067±0.0000	0.78	0.109	1.08
FlowDPS	✓$^\pi$	0.0440±0.0045	0.68	**0.037**	1.01
SimVP	✓$^\pi$	0.0135	–	–	0.98
Linear	✓	0.0996	–	–	0.62
Cubic spline	✓	0.1021	–	–	0.65

(b) **Gray–Scott** ($N\delta/\tau = 1.6$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 5$)	✓	0.0088±0.0002	0.79	0.043	1.01
MUNDO ($R = 1$)	✓$^\pi$	**0.0066**±0.0001	1.17	0.042	1.01
Flow matching	✓$^\pi$	0.0087±0.0002	**0.89**	0.120	1.04
FlowDPS	✓$^\pi$	0.0440±0.0004	0.65	**0.028**	1.02
SimVP	✓$^\pi$	0.0147	–	–	**1.00**
Denton–Cholette	✓	0.2825	–	–	0.91
CDF-t	×	0.3253	–	–	0.89
Interpolation					
MUNDO ($R = 5$)‡	✓	0.0061±0.0000	1.06	**0.048**	1.03
MUNDO ($R = 1$)	✓$^\pi$	**0.0044**±0.0000	0.86	0.053	1.01
Flow matching	✓$^\pi$	0.0063±0.0001	**0.97**	0.130	1.04
FlowDPS	✓$^\pi$	0.0692±0.0058	0.86	0.084	1.03
SimVP	✓$^\pi$	0.0111	–	–	**1.00**
Linear	✓	0.3388	–	–	0.84
Cubic spline	✓	0.3232	–	–	0.90

(c) **Navier–Stokes** ($N\delta/\tau = 1.3$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 5$)‡	✓	0.0757±0.0022	0.75	**0.021**	1.18
MUNDO ($R = 1$)	✓$^\pi$	0.0718±0.0030	0.72	0.032	**0.99**
Flow matching	✓$^\pi$	**0.0704**±0.0003	**0.79**	0.052	1.29
FlowDPS	✓$^\pi$	0.1130±0.0016	0.52	0.060	1.06
SimVP	✓$^\pi$	0.1014	–	–	0.97
Denton–Cholette	✓	0.1856	–	–	0.71
CDF-t	×	0.3014	–	–	0.79
Interpolation					
MUNDO ($R = 5$)‡	✓	0.0460±0.0010	0.68	0.023	1.16
MUNDO ($R = 1$)	✓$^\pi$	**0.0359**±0.0008	0.63	**0.019**	1.02
Flow matching	✓$^\pi$	0.0447±0.0001	**0.73**	0.053	1.19
FlowDPS	✓$^\pi$	0.2724±0.0166	0.65	0.098	1.56
SimVP	✓$^\pi$	0.0583	–	–	**1.01**
Linear	✓	0.2257	–	–	0.75
Cubic spline	✓	0.2222	–	–	0.77

(d) **Gross–Pitaevskii** ($N\delta/\tau = 4.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 5$)‡	✓	**0.1324**±0.0045	0.95	0.058	**1.02**
MUNDO ($R = 1$)	✓$^\pi$	0.1532±0.0098	0.85	**0.052**	0.86
Flow matching	✓$^\pi$	0.1726±0.0012	0.91	0.125	1.33
FlowDPS	✓$^\pi$	0.3262±0.0013	**1.04**	0.109	1.15
SimVP	✓$^\pi$	0.2059	–	–	0.84
Denton–Cholette	✓	0.5429	–	–	0.17
CDF-t	×	0.6883	–	–	0.32
Interpolation					
MUNDO ($R = 5$)‡	✓	**0.0583**±0.0020	0.93	**0.033**	1.07
MUNDO ($R = 1$)	✓$^\pi$	0.0823±0.0110	1.06	0.084	**0.97**
Flow matching	✓$^\pi$	0.0950±0.0010	**0.95**	0.150	1.27
FlowDPS	✓$^\pi$	0.2807±0.0068	0.52	0.088	1.21
SimVP	✓$^\pi$	0.0791	–	–	0.95
Linear	✓	0.6567	–	–	0.31
Cubic spline	✓	0.6906	–	–	0.32

Table 19: **AutoSim at $N = 24$, mild tier of each system.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion** ($N\delta/\tau = 1.5$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)	✓	0.0073±0.0001	**1.01**	**0.039**	1.02
MUNDO ($R = 1$)	✓$^\pi$	0.0063±0.0001	0.67	0.067	**1.00**
Flow matching	✓$^\pi$	**0.0062**±0.0000	0.76	0.077	1.06
FlowDPS	✓$^\pi$	0.0242±0.0012	0.39	0.094	0.94
SimVP	✓$^\pi$	0.0114	–	–	0.98
Denton–Cholette	✓	0.0699	–	–	0.66
CDF-t	×	0.0868	–	–	0.64
Interpolation					
MUNDO ($R = 4$)	✓	0.0058±0.0001	**1.03**	0.065	1.05
MUNDO ($R = 1$)	✓$^\pi$	**0.0045**±0.0001	0.96	0.088	**1.01**
Flow matching	✓$^\pi$	0.0054±0.0000	0.77	0.110	1.07
FlowDPS	✓$^\pi$	0.0359±0.0029	0.62	**0.032**	0.98
SimVP	✓$^\pi$	0.0119	–	–	**0.99**
Linear	✓	0.0819	–	–	0.68
Cubic spline	✓	0.0846	–	–	0.72

(b) **Gray–Scott** ($N\delta/\tau = 1.2$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)	✓	0.0059±0.0001	1.12	**0.045**	**1.00**
MUNDO ($R = 1$)	✓$^\pi$	**0.0056**±0.0001	1.56	0.046	1.01
Flow matching	✓$^\pi$	0.0067±0.0002	0.89	0.108	1.03
FlowDPS	✓$^\pi$	0.0211±0.0007	**1.04**	0.052	1.02
SimVP	✓$^\pi$	0.0124	–	–	**1.00**
Denton–Cholette	✓	0.1887	–	–	0.95
CDF-t	×	0.2624	–	–	0.94
Interpolation					
MUNDO ($R = 4$)	✓	0.0047±0.0001	**1.01**	**0.036**	1.01
MUNDO ($R = 1$)	✓$^\pi$	**0.0035**±0.0000	0.93	0.057	1.01
Flow matching	✓$^\pi$	0.0048±0.0000	0.96	0.119	1.03
FlowDPS	✓$^\pi$	0.0355±0.0007	1.11	0.100	1.04
SimVP	✓$^\pi$	0.0098	–	–	**1.00**
Linear	✓	0.2436	–	–	0.89
Cubic spline	✓	0.2213	–	–	0.95

(c) **Navier–Stokes** ($N\delta/\tau = 1.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	0.0508±0.0004	**0.74**	**0.023**	1.09
MUNDO ($R = 1$)	✓$^\pi$	**0.0449**±0.0004	0.70	0.032	**1.01**
Flow matching	✓$^\pi$	0.0453±0.0004	0.70	**0.023**	1.08
FlowDPS	✓$^\pi$	0.0820±0.0009	0.58	0.090	0.97
SimVP	✓$^\pi$	0.0742	–	–	**0.99**
Denton–Cholette	✓	0.1434	–	–	0.78
CDF-t	×	0.2630	–	–	0.86
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0285±0.0004	**0.69**	0.024	1.08
MUNDO ($R = 1$)	✓$^\pi$	**0.0238**±0.0001	0.63	**0.019**	**1.02**
Flow matching	✓$^\pi$	0.0273±0.0001	0.65	0.025	1.08
FlowDPS	✓$^\pi$	0.2600±0.0071	0.55	0.151	1.37
SimVP	✓$^\pi$	0.0431	–	–	**1.02**
Linear	✓	0.1711	–	–	0.81
Cubic spline	✓	0.1647	–	–	0.83

(d) **Gross–Pitaevskii** ($N\delta/\tau = 3.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	**0.0939**±0.0014	0.92	0.063	**1.03**
MUNDO ($R = 1$)	✓$^\pi$	0.1000±0.0021	0.82	**0.050**	0.92
Flow matching	✓$^\pi$	0.1202±0.0003	0.86	0.082	**1.03**
FlowDPS	✓$^\pi$	0.3034±0.0007	**1.07**	0.096	0.93
SimVP	✓$^\pi$	0.1631	–	–	0.87
Denton–Cholette	✓	0.5349	–	–	0.21
CDF-t	×	0.6792	–	–	0.37
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0343±0.0007	**0.97**	**0.042**	1.05
MUNDO ($R = 1$)	✓$^\pi$	**0.0332**±0.0004	1.10	0.084	1.04
Flow matching	✓$^\pi$	0.0476±0.0004	0.82	0.091	1.05
FlowDPS	✓$^\pi$	0.1743±0.0062	0.47	0.095	0.91
SimVP	✓$^\pi$	0.0529	–	–	**0.99**
Linear	✓	0.6584	–	–	0.35
Cubic spline	✓	0.6888	–	–	0.36

Table 20: **AutoSim at $N = 16$, mild tier of each system.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion** ($N\delta/\tau = 1.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	0.0056$_{\pm 0.0001}$	**0.92**	**0.056**	1.02
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0044**$_{\pm 0.0001}$	0.90	0.097	**1.01**
Flow matching	✓$^{\pi}$	0.0046$_{\pm 0.0000}$	0.77	0.065	1.06
FlowDPS	✓$^{\pi}$	0.0191$_{\pm 0.0018}$	0.37	0.090	0.95
SimVP	✓$^{\pi}$	0.0110	–	–	1.02
Denton–Cholette	✓	0.0464	–	–	0.77
CDF-t	×	0.0659	–	–	0.75
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0038$_{\pm 0.0000}$	0.95	0.082	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0028**$_{\pm 0.0000}$	**0.97**	0.108	1.02
Flow matching	✓$^{\pi}$	0.0036$_{\pm 0.0000}$	0.79	0.124	1.06
FlowDPS	✓$^{\pi}$	0.0248$_{\pm 0.0012}$	0.55	**0.021**	0.97
SimVP	✓$^{\pi}$	0.0095	–	–	**1.01**
Linear	✓	0.0561	–	–	0.77
Cubic spline	✓	0.0556	–	–	0.80

(b) **Gray–Scott** ($N\delta/\tau = 0.8$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)	✓	**0.0039**$_{\pm 0.0000}$	1.20	**0.042**	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	0.0040$_{\pm 0.0001}$	1.36	0.047	**1.00**
Flow matching	✓$^{\pi}$	0.0048$_{\pm 0.0001}$	0.92	0.092	1.02
FlowDPS	✓$^{\pi}$	0.0137$_{\pm 0.0002}$	**0.99**	0.054	1.02
SimVP	✓$^{\pi}$	0.0112	–	–	**1.00**
Denton–Cholette	✓	0.0970	–	–	0.97
CDF-t	×	0.2008	–	–	0.97
Interpolation					
MUNDO ($R = 4$)	✓	0.0036$_{\pm 0.0000}$	1.11	**0.033**	1.01
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0029**$_{\pm 0.0000}$	0.98	0.055	**1.00**
Flow matching	✓$^{\pi}$	0.0036$_{\pm 0.0000}$	**0.99**	0.106	1.02
FlowDPS	✓$^{\pi}$	0.0240$_{\pm 0.0007}$	1.18	0.106	1.03
SimVP	✓$^{\pi}$	0.0100	–	–	**1.00**
Linear	✓	0.1430	–	–	0.94
Cubic spline	✓	0.1194	–	–	0.98

(c) **Navier–Stokes** ($N\delta/\tau = 0.6$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)	✓	0.0246$_{\pm 0.0004}$	**0.83**	0.040	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0197**$_{\pm 0.0002}$	0.71	**0.017**	1.02
Flow matching	✓$^{\pi}$	0.0218$_{\pm 0.0002}$	0.71	0.018	1.04
FlowDPS	✓$^{\pi}$	0.0425$_{\pm 0.0001}$	0.60	0.071	**0.99**
SimVP	✓$^{\pi}$	0.0399	–	–	1.03
Denton–Cholette	✓	0.0863	–	–	0.87
CDF-t	×	0.2250	–	–	0.94
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0156$_{\pm 0.0002}$	**0.74**	0.035	1.06
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0130**$_{\pm 0.0002}$	0.67	0.023	**1.02**
Flow matching	✓$^{\pi}$	0.0151$_{\pm 0.0001}$	0.64	**0.021**	1.04
FlowDPS	✓$^{\pi}$	0.1731$_{\pm 0.0051}$	0.44	0.173	1.24
SimVP	✓$^{\pi}$	0.0271	–	–	**1.02**
Linear	✓	0.1048	–	–	0.88
Cubic spline	✓	0.0960	–	–	0.91

(d) **Gross–Pitaevskii** ($N\delta/\tau = 2.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	**0.0636**$_{\pm 0.0004}$	**0.95**	0.066	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	0.0667$_{\pm 0.0004}$	0.78	**0.051**	0.97
Flow matching	✓$^{\pi}$	0.0841$_{\pm 0.0010}$	0.85	0.064	**1.00**
FlowDPS	✓$^{\pi}$	0.2612$_{\pm 0.0031}$	1.09	0.085	0.93
SimVP	✓$^{\pi}$	0.1246	–	–	0.90
Denton–Cholette	✓	0.5177	–	–	0.30
CDF-t	×	0.6558	–	–	0.46
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0198$_{\pm 0.0000}$	0.96	**0.050**	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0174**$_{\pm 0.0010}$	**0.97**	0.074	1.05
Flow matching	✓$^{\pi}$	0.0219$_{\pm 0.0001}$	0.83	0.076	1.04
FlowDPS	✓$^{\pi}$	0.0938$_{\pm 0.0009}$	0.55	0.079	0.91
SimVP	✓$^{\pi}$	0.0330	–	–	**1.01**
Linear	✓	0.6296	–	–	0.42
Cubic spline	✓	0.6634	–	–	0.44

Table 21: **AutoSim at $N = 8$, mild tier of each system.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion** ($N\delta/\tau = 0.5$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)‡	✓	0.0031$\pm$0.0001	0.89	0.112	1.02
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0021**$\pm$0.0001	**0.99**	0.145	1.03
Flow matching	✓$^{\pi}$	0.0027$\pm$0.0000	0.76	**0.054**	1.05
FlowDPS	✓$^{\pi}$	0.0120$\pm$0.0015	0.45	0.065	0.96
SimVP	✓$^{\pi}$	0.0063	–	–	**1.01**
Denton–Cholette	✓	0.0196	–	–	0.88
CDF-t	×	0.0432	–	–	0.87
Interpolation					
MUNDO ($R = 3$)‡	✓	0.0016$\pm$0.0000	0.97	0.142	1.03
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0011**$\pm$0.0000	**0.98**	0.164	**1.02**
Flow matching	✓$^{\pi}$	0.0017$\pm$0.0000	0.80	0.134	1.05
FlowDPS	✓$^{\pi}$	0.0182$\pm$0.0025	0.57	**0.027**	0.97
SimVP	✓$^{\pi}$	0.0064	–	–	**1.02**
Linear	✓	0.0268	–	–	0.88
Cubic spline	✓	0.0239	–	–	0.92

(b) **Gray–Scott** ($N\delta/\tau = 0.4$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)‡	✓	**0.0018**$\pm$0.0000	1.89	0.063	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	0.0019$\pm$0.0001	1.43	**0.059**	**1.00**
Flow matching	✓$^{\pi}$	0.0020$\pm$0.0000	**1.00**	0.092	1.01
FlowDPS	✓$^{\pi}$	0.0053$\pm$0.0002	**1.00**	0.063	1.02
SimVP	✓$^{\pi}$	0.0068	–	–	**1.00**
Denton–Cholette	✓	0.0309	–	–	0.95
CDF-t	×	0.1689	–	–	0.98
Interpolation					
MUNDO ($R = 3$)	✓	0.0018$\pm$0.0000	1.23	**0.036**	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0016**$\pm$0.0001	1.16	0.073	**1.00**
Flow matching	✓$^{\pi}$	0.0019$\pm$0.0000	**1.02**	0.089	1.01
FlowDPS	✓$^{\pi}$	0.0147$\pm$0.0000	1.29	0.114	1.02
SimVP	✓$^{\pi}$	0.0067	–	–	**1.00**
Linear	✓	0.0568	–	–	0.97
Cubic spline	✓	0.0409	–	–	0.98

(c) **Navier–Stokes** ($N\delta/\tau = 0.3$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)	✓	0.0064$\pm$0.0002	**0.89**	0.071	1.03
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0053**$\pm$0.0000	0.77	0.041	**1.02**
Flow matching	✓$^{\pi}$	0.0076$\pm$0.0001	0.70	**0.028**	1.03
FlowDPS	✓$^{\pi}$	0.0160$\pm$0.0001	0.62	0.044	**1.02**
SimVP	✓$^{\pi}$	0.0159	–	–	**1.02**
Denton–Cholette	✓	0.0261	–	–	0.96
CDF-t	×	0.1910	–	–	1.03
Interpolation					
MUNDO ($R = 3$)‡	✓	0.0053$\pm$0.0000	0.78	0.041	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0047**$\pm$0.0000	0.74	0.042	1.03
Flow matching	✓$^{\pi}$	0.0057$\pm$0.0000	0.66	0.024	1.03
FlowDPS	✓$^{\pi}$	0.0345$\pm$0.0001	**0.81**	**0.022**	1.03
SimVP	✓$^{\pi}$	0.0123	–	–	1.02
Linear	✓	0.0393	–	–	0.96
Cubic spline	✓	0.0292	–	–	**0.99**

(d) **Gross–Pitaevskii** ($N\delta/\tau = 1.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)‡	✓	0.0216$\pm$0.0000	**0.98**	0.083	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0193**$\pm$0.0002	0.97	0.077	1.03
Flow matching	✓$^{\pi}$	0.0277$\pm$0.0001	0.86	0.071	**1.02**
FlowDPS	✓$^{\pi}$	0.0747$\pm$0.0003	0.81	**0.050**	0.93
SimVP	✓$^{\pi}$	0.0546	–	–	**0.98**
Denton–Cholette	✓	0.3222	–	–	0.60
CDF-t	×	0.3610	–	–	0.68
Interpolation					
MUNDO ($R = 3$)‡	✓	0.0070$\pm$0.0002	**0.95**	0.070	1.03
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0052**$\pm$0.0001	0.93	0.072	1.03
Flow matching	✓$^{\pi}$	0.0068$\pm$0.0001	0.75	0.060	1.04
FlowDPS	✓$^{\pi}$	0.0467$\pm$0.0006	0.67	**0.049**	0.96
SimVP	✓$^{\pi}$	0.0180	–	–	**1.02**
Linear	✓	0.4062	–	–	0.66
Cubic spline	✓	0.4028	–	–	0.70

Table 22: **AutoSim at $N = 24$, most chaotic tier of each system.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion** ($N\delta/\tau = 3.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	<u>0.0228</u>$\pm$0.0003	**0.99**	**0.011**	**0.96**
MUNDO ($R = 1$)	✓$^{\pi}$	0.0349$\pm$0.0015	0.58	0.040	0.77
Flow matching	✓$^{\pi}$	**0.0204**$\pm$0.0002	0.78	0.043	**0.96**
FlowDPS	✓$^{\pi}$	0.0573$\pm$0.0023	0.41	0.200	0.65
SimVP	✓$^{\pi}$	0.0314	–	–	0.82
Denton–Cholette	✓	0.0964	–	–	0.33
CDF-t	×	0.1088	–	–	0.35
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0207$\pm$0.0001	0.95	0.036	**0.99**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0185**$\pm$0.0001	**0.99**	0.048	0.90
Flow matching	✓$^{\pi}$	<u>0.0190</u>$\pm$0.0000	0.81	**0.025**	0.97
FlowDPS	✓$^{\pi}$	0.0763$\pm$0.0034	0.59	0.118	0.71
SimVP	✓$^{\pi}$	0.0335	–	–	0.85
Linear	✓	0.1230	–	–	0.41
Cubic spline	✓	0.1275	–	–	0.42

(b) **Gray–Scott** ($N\delta/\tau = 2.4$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	0.0126$\pm$0.0001	0.79	0.054	1.01
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0100**$\pm$0.0000	0.76	0.031	**1.00**
Flow matching	✓$^{\pi}$	<u>0.0121</u>$\pm$0.0001	0.79	0.076	1.01
FlowDPS	✓$^{\pi}$	0.0883$\pm$0.0066	**0.89**	**0.017**	0.95
SimVP	✓$^{\pi}$	0.0221	–	–	**1.00**
Denton–Cholette	✓	0.4740	–	–	0.74
CDF-t	×	0.4830	–	–	0.74
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0100$\pm$0.0001	**0.95**	0.041	1.01
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0074**$\pm$0.0001	0.72	0.041	**1.00**
Flow matching	✓$^{\pi}$	<u>0.0092</u>$\pm$0.0001	0.87	0.086	1.01
FlowDPS	✓$^{\pi}$	0.1219$\pm$0.0094	0.82	**0.018**	0.99
SimVP	✓$^{\pi}$	0.0162	–	–	**1.00**
Linear	✓	0.5210	–	–	0.78
Cubic spline	✓	0.5255	–	–	0.82

(c) **Navier–Stokes** ($N\delta/\tau = 1.1$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	0.0740$\pm$0.0008	0.77	0.027	1.13
MUNDO ($R = 1$)	✓$^{\pi}$	<u>0.0715</u>$\pm$0.0028	**0.79**	**0.016**	**0.99**
Flow matching	✓$^{\pi}$	**0.0634**$\pm$0.0001	0.70	0.021	1.05
FlowDPS	✓$^{\pi}$	0.1084$\pm$0.0005	0.57	0.093	0.92
SimVP	✓$^{\pi}$	0.1002	–	–	0.96
Denton–Cholette	✓	0.1567	–	–	0.72
CDF-t	×	0.2982	–	–	0.81
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0427$\pm$0.0003	**0.73**	0.033	1.12
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0370**$\pm$0.0006	0.70	**0.022**	1.03
Flow matching	✓$^{\pi}$	<u>0.0382</u>$\pm$0.0002	0.64	0.025	1.07
FlowDPS	✓$^{\pi}$	0.2572$\pm$0.0160	0.60	0.122	1.33
SimVP	✓$^{\pi}$	0.0581	–	–	**1.01**
Linear	✓	0.1909	–	–	0.76
Cubic spline	✓	0.1878	–	–	0.78

Table 23: **AutoSim at $N = 16$, most chaotic tier of each system.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion** ($N\delta/\tau = 2.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	0.0186$\pm$0.0001	**0.95**	**0.012**	0.95
MUNDO ($R = 1$)	✓$^{\pi}$	0.0229$\pm$0.0035	0.66	0.020	0.86
Flow matching	✓$^{\pi}$	**0.0159**$\pm$0.0000	0.79	0.047	**0.97**
FlowDPS	✓$^{\pi}$	0.0448$\pm$0.0025	0.33	0.200	0.68
SimVP	✓$^{\pi}$	0.0292	–	–	0.85
Denton–Cholette	✓	0.0818	–	–	0.43
CDF-t	×	0.0933	–	–	0.44
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0153$\pm$0.0002	0.94	0.046	**1.01**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0130**$\pm$0.0003	**0.97**	0.054	0.95
Flow matching	✓$^{\pi}$	0.0135$\pm$0.0000	0.82	**0.025**	**0.99**
FlowDPS	✓$^{\pi}$	0.0532$\pm$0.0004	0.49	0.133	0.72
SimVP	✓$^{\pi}$	0.0282	–	–	0.89
Linear	✓	0.1044	–	–	0.48
Cubic spline	✓	0.1081	–	–	0.50

(b) **Gray–Scott** ($N\delta/\tau = 1.6$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	0.0087$\pm$0.0000	**0.96**	0.053	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0070**$\pm$0.0002	0.83	0.038	**1.00**
Flow matching	✓$^{\pi}$	0.0080$\pm$0.0001	0.78	0.059	1.01
FlowDPS	✓$^{\pi}$	0.0391$\pm$0.0029	0.71	**0.019**	0.98
SimVP	✓$^{\pi}$	0.0168	–	–	**1.00**
Denton–Cholette	✓	0.2732	–	–	0.92
CDF-t	×	0.2774	–	–	0.88
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0062$\pm$0.0001	**0.97**	**0.037**	1.01
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0050**$\pm$0.0001	0.74	0.045	**1.00**
Flow matching	✓$^{\pi}$	0.0055$\pm$0.0000	0.83	0.068	1.01
FlowDPS	✓$^{\pi}$	0.0639$\pm$0.0035	0.54	0.076	0.98
SimVP	✓$^{\pi}$	0.0123	–	–	**1.00**
Linear	✓	0.3304	–	–	0.88
Cubic spline	✓	0.3075	–	–	0.94

(c) **Navier–Stokes** ($N\delta/\tau = 0.8$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 4$)‡	✓	0.0402$\pm$0.0007	**0.78**	0.048	1.09
MUNDO ($R = 1$)	✓$^{\pi}$	0.0338$\pm$0.0001	0.68	0.018	**1.02**
Flow matching	✓$^{\pi}$	**0.0336**$\pm$0.0003	0.69	**0.017**	1.03
FlowDPS	✓$^{\pi}$	0.0642$\pm$0.0006	0.59	0.078	0.95
SimVP	✓$^{\pi}$	0.0594	–	–	**1.02**
Denton–Cholette	✓	0.1026	–	–	0.82
CDF-t	×	0.2629	–	–	0.91
Interpolation					
MUNDO ($R = 4$)‡	✓	0.0239$\pm$0.0002	**0.72**	0.041	1.08
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0198**$\pm$0.0001	0.66	0.029	1.03
Flow matching	✓$^{\pi}$	0.0215$\pm$0.0002	0.62	**0.021**	1.04
FlowDPS	✓$^{\pi}$	0.1936$\pm$0.0142	0.48	0.161	1.24
SimVP	✓$^{\pi}$	0.0371	–	–	**1.02**
Linear	✓	0.1245	–	–	0.84
Cubic spline	✓	0.1182	–	–	0.87

Table 24: **AutoSim at $N = 8$, most chaotic tier of each system.** ‡: end-to-end fine-tuned. Underlined: second-lowest CRPS among the methods that satisfy the observation.

(a) **Advection-diffusion** ($N\delta/\tau = 1.0$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)‡	✓	0.0121$\pm$0.0002	**0.94**	**0.041**	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	<u>0.0098</u>$\pm$0.0002	0.87	0.053	0.98
Flow matching	✓$^{\pi}$	**0.0097**$\pm$0.0000	0.79	0.061	0.98
FlowDPS	✓$^{\pi}$	0.0326$\pm$0.0012	0.34	0.179	0.71
SimVP	✓$^{\pi}$	0.0213	–	–	0.88
Denton–Cholette	✓	0.0585	–	–	0.58
CDF-t	×	0.0689	–	–	0.58
Interpolation					
MUNDO ($R = 3$)‡	✓	0.0079$\pm$0.0002	0.94	0.072	1.01
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0059**$\pm$0.0002	**0.98**	0.090	**1.00**
Flow matching	✓$^{\pi}$	<u>0.0062</u>$\pm$0.0001	0.81	**0.032**	**1.00**
FlowDPS	✓$^{\pi}$	0.0401$\pm$0.0007	0.45	0.119	0.72
SimVP	✓$^{\pi}$	0.0200	–	–	0.95
Linear	✓	0.0753	–	–	0.62
Cubic spline	✓	0.0772	–	–	0.66

(b) **Gray–Scott** ($N\delta/\tau = 0.8$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)‡	✓	<u>0.0028</u>$\pm$0.0000	1.48	0.046	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0027**$\pm$0.0000	1.35	0.043	**1.00**
Flow matching	✓$^{\pi}$	0.0029$\pm$0.0000	0.90	0.042	**1.00**
FlowDPS	✓$^{\pi}$	0.0095$\pm$0.0002	**1.07**	**0.036**	**1.00**
SimVP	✓$^{\pi}$	0.0081	–	–	**1.00**
Denton–Cholette	✓	0.0825	–	–	0.99
CDF-t	×	0.1251	–	–	0.97
Interpolation					
MUNDO ($R = 3$)	✓	<u>0.0024</u>$\pm$0.0000	1.06	**0.032**	**1.00**
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0022**$\pm$0.0000	**1.05**	0.058	**1.00**
Flow matching	✓$^{\pi}$	<u>0.0024</u>$\pm$0.0000	0.92	0.048	**1.00**
FlowDPS	✓$^{\pi}$	0.0158$\pm$0.0001	1.31	0.117	1.01
SimVP	✓$^{\pi}$	0.0074	–	–	**1.00**
Linear	✓	0.1295	–	–	0.96
Cubic spline	✓	0.0995	–	–	1.01

(c) **Navier–Stokes** ($N\delta/\tau = 0.4$)

	cons.	CRPS ↓	SSR → 1	cov. MAE ↓	$t_1 \to 1$
Disaggregation					
MUNDO ($R = 3$)‡	✓	<u>0.0107</u>$\pm$0.0002	**0.80**	0.079	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0088**$\pm$0.0001	0.72	0.042	1.03
Flow matching	✓$^{\pi}$	0.0108$\pm$0.0001	0.67	**0.023**	1.02
FlowDPS	✓$^{\pi}$	0.0224$\pm$0.0002	0.58	0.048	**1.01**
SimVP	✓$^{\pi}$	0.0224	–	–	1.02
Denton–Cholette	✓	0.0379	–	–	0.93
CDF-t	×	0.2276	–	–	1.02
Interpolation					
MUNDO ($R = 3$)‡	✓	<u>0.0076</u>$\pm$0.0001	**0.76**	0.051	1.04
MUNDO ($R = 1$)	✓$^{\pi}$	**0.0067**$\pm$0.0001	0.71	0.048	1.03
Flow matching	✓$^{\pi}$	<u>0.0076</u>$\pm$0.0000	0.64	0.021	1.03
FlowDPS	✓$^{\pi}$	0.0404$\pm$0.0002	0.72	**0.014**	1.03
SimVP	✓$^{\pi}$	0.0163	–	–	**1.02**
Linear	✓	0.0520	–	–	0.94
Cubic spline	✓	0.0431	–	–	0.97