Supporting Information for "Why Deterministic AI Weather Models Fail at Extremes and Physical Conservation: From the Perspective of Probabilistic Mean Field"

Yihui Ding[1,2], Baoheng Yao[2], Jingsong Yang[1,2,3], Wenlei Peng[1,2]

[1]State Key Laboratory of Satellite Ocean Environment Dynamics, Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou, 310012, China

[2]School of Oceanography, Shanghai Jiao Tong University, Shanghai, 200030, China

[3]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, 519082, China

Introduction

Deterministic AI weather and ocean models are far more computationally efficient and achieve lower mean-square error than state-of-the-art operational forecasts, yet they are widely criticized for smoothing small scales, underestimating extremes, and violating physical constraints. We offer a unified Markov-Process explanation: MSE-trained deterministic models learn the conditional mean $\mathbb{E}[\mathbf{x}_{t+\Delta t} \mid \mathbf{x}_t]$ rather than the full transition law $p(\mathbf{x}_{t+\Delta t} \mid \mathbf{x}_t)$, which links low RMSE to over-smoothing and underestimation of extremes. Moreover, this conditional mean represents a linear combination of potential physical states, and thus violates the basic property of nonlinear dynamics that linear combinations of solutions are in general not solutions of the governing equations, which accounts for the fundamental origin of physical-conservation violations in deterministic models. We first establish the corresponding mathematical analysis, and then support the theory with a finite-state Markov experiment and a stochastically perturbed Kelvin–Helmholtz fluid experiment using matched U-Net backbones. Both experiments show that deterministic

models can attain lower next-step RMSE, but at the cost of almost completely discarding next-step distributional information relative to models that recover the full transition law. Although our analysis focuses on deterministic models trained with an RMSE loss as a representative case, the conclusions apply to deterministic forecasting models more generally. From an information-entropy perspective, deterministic forecasts are low-entropy at each step and overconfident relative to atmospheric and oceanic variability, whereas diffusion models sample from conditional transition distributions and better match intrinsic complexity. Accordingly, distributional generative methods may be better suited to future AI weather prediction, reconstruction, and bias correction.

Mathematical Proofs in the Preliminary Chapter

Proposition 1 (MSE-optimal deterministic forecast as conditional expectation)**.** *Let* $(\mathbf{X}_t, \mathbf{X}_{t+\Delta t})$ *be generated by the Markov process and assume* $\mathbb{E}\|\mathbf{X}_{t+\Delta t}\|_2^2 < \infty$*. Consider any measurable estimator* $g : \mathcal{S} \to \mathbb{R}^d$ *and define the population MSE risk*

$$\mathcal{R}(g) = \mathbb{E}\big[\|\mathbf{X}_{t+\Delta t} - g(\mathbf{X}_t)\|_2^2\big]\,. \tag{1}$$

Then any minimizer of $\mathcal{R}(g)$ *satisfies*

$$g^\star(\mathbf{x}_t) = \mathbb{E}[\mathbf{X}_{t+\Delta t} \,|\, \mathbf{X}_t = \mathbf{x}_t] \qquad \textit{almost surely.} \tag{2}$$

Consequently, a deterministic neural model $\mathcal{M}_{\boldsymbol{\theta}}$ *trained by minimizing RMSE learns the conditional mean of the future state rather than the full conditional distribution* $p(\mathbf{x}_{t+\Delta t} \mid \mathbf{x}_t)$*.*

Proof. Fix $\mathbf{x}_t \in \mathcal{S}$ and define the posterior mean

$$\boldsymbol{\mu}(\mathbf{x}_t) = \mathbb{E}[\mathbf{X}_{t+\Delta t} \,|\, \mathbf{X}_t = \mathbf{x}_t] = \int_{\mathcal{S}} \mathbf{x}\, p(\mathbf{x} \mid \mathbf{x}_t)\, d\mathbf{x}. \tag{3}$$

Consider the conditional MSE risk

$$\mathcal{R}(g \mid \mathbf{x}_t) = \mathbb{E}\big[\|\mathbf{X}_{t+\Delta t} - g(\mathbf{x}_t)\|_2^2 \,\big|\, \mathbf{X}_t = \mathbf{x}_t\big] = \int_{\mathcal{S}} \|\mathbf{x} - g(\mathbf{x}_t)\|_2^2\, p(\mathbf{x} \mid \mathbf{x}_t)\, d\mathbf{x}. \tag{4}$$

Adding and subtracting $\boldsymbol{\mu}(\mathbf{x}_t)$ inside the squared norm yields

$$\begin{aligned}\mathcal{R}(g \mid \mathbf{x}_t) &= \int_{\mathcal{S}} \|\mathbf{x} - \boldsymbol{\mu}(\mathbf{x}_t) + \boldsymbol{\mu}(\mathbf{x}_t) - g(\mathbf{x}_t)\|_2^2 \, p(\mathbf{x} \mid \mathbf{x}_t)\, d\mathbf{x} \\ &= \int_{\mathcal{S}} \|\mathbf{x} - \boldsymbol{\mu}(\mathbf{x}_t)\|_2^2 \, p(\mathbf{x} \mid \mathbf{x}_t)\, d\mathbf{x} + \|\boldsymbol{\mu}(\mathbf{x}_t) - g(\mathbf{x}_t)\|_2^2 \\ &\quad + 2\,(\boldsymbol{\mu}(\mathbf{x}_t) - g(\mathbf{x}_t))^\top \int_{\mathcal{S}} (\mathbf{x} - \boldsymbol{\mu}(\mathbf{x}_t))\, p(\mathbf{x} \mid \mathbf{x}_t)\, d\mathbf{x}. \end{aligned} \tag{5}$$

By the definition of $\boldsymbol{\mu}(\mathbf{x}_t)$ in (3), the integral in the last term equals zero, so the cross term vanishes and

$$\mathcal{R}(g \mid \mathbf{x}_t) = \mathrm{Var}(\mathbf{X}_{t+\Delta t} \,|\, \mathbf{X}_t = \mathbf{x}_t) + \|\boldsymbol{\mu}(\mathbf{x}_t) - g(\mathbf{x}_t)\|_2^2 . \tag{6}$$

The first term is independent of g, whereas the second term is minimized if and only if

$$g(\mathbf{x}_t) = \boldsymbol{\mu}(\mathbf{x}_t) = \mathbb{E}[\mathbf{X}_{t+\Delta t} \,|\, \mathbf{X}_t = \mathbf{x}_t] . \tag{7}$$

Since (6) holds for every $\mathbf{x}_t \in \mathcal{S}$, the global minimizer of (1) is given by (2).

Finally, because $\mathcal{M}_{\boldsymbol{\theta}}$ is a deterministic function of $\mathbf{X}_t$, minimizing RMSE is a special case of minimizing (1) over functions of the form $g(\mathbf{X}_t) = \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{X}_t)$.

□

Proposition 2 (Conditional expectation is not a nonlinear physical solution)**.** *Let u_t be a random state generated by a nonlinear dynamical system*

$$\partial_t u = \mathcal{N}(u), \tag{8}$$

where $\mathcal{N}$ is nonlinear. Suppose every sample path satisfies this equation, and define the one-step conditional mean

$$\bar{u}_{t+\Delta t} = \mathbb{E}[u_{t+\Delta t} \mid u_t]. \tag{9}$$

If the conditional distribution of $u_{t+\Delta t}$ given u_t is not almost surely degenerate, then $\bar{u}_{t+\Delta t}$ is not, in general, a solution of $\partial_t u = \mathcal{N}(u)$. Instead,

$$\partial_t \bar{u}_{t+\Delta t} = \mathcal{N}(\bar{u}_{t+\Delta t}) + R_{t+\Delta t}, \qquad R_{t+\Delta t} = \mathbb{E}[\mathcal{N}(u_{t+\Delta t}) \mid u_t] - \mathcal{N}(\mathbb{E}[u_{t+\Delta t} \mid u_t]). \tag{10}$$

Moreover, $R_{t+\Delta t} \neq 0$ whenever $\mathcal{N}$ is non-affine on the support of the conditional distribution.

Proof. Every sample path satisfies $\partial_t u = \mathcal{N}(u)$, so taking conditional expectation gives

$$\mathbb{E}[\partial_t u_{t+\Delta t} \mid u_t] = \mathbb{E}[\mathcal{N}(u_{t+\Delta t}) \mid u_t]. \tag{11}$$

If $\bar{u}_{t+\Delta t}$ were an exact nonlinear solution, then $\partial_t \bar{u}_{t+\Delta t} = \mathcal{N}(\bar{u}_{t+\Delta t})$, which would require $\mathcal{N}(\bar{u}_{t+\Delta t}) = \mathbb{E}[\mathcal{N}(u_{t+\Delta t}) \mid u_t]$. For nonlinear $\mathcal{N}$, this equality fails by Jensen's theorem:

$$\mathbb{E}\big[\mathcal{N}(X)\big] \geq \mathcal{N}\big(\mathbb{E}[X]\big) \tag{12}$$

unless the conditional distribution is degenerate (or $\mathcal{N}$ is affine on that support). Hence $\bar{u}_{t+\Delta t}$ carries a nonzero closure term $R_{t+\Delta t}$. □

Supplementary Table

Table S1. Markov MLP results for $n = 3$, 10, and 100 states. MLP(RMSE) and MLP(Cross-Entropy) achieve similar scalar test RMSE, but only cross-entropy training recovers the transition law (small row-wise KL).

n	Model	Params	Batch	LR	Epochs	Test RMSE	Row KL
3	MLP(RMSE)	16.9K	1024	10^{-3}	120	0.5852	8.377
3	MLP(Cross-Entropy)	17.4K	1024	10^{-3}	120	0.5850	0.0002
10	MLP(RMSE)	16.9K	1024	10^{-3}	120	0.6576	14.812
10	MLP(Cross-Entropy)	19.2K	1024	10^{-3}	120	0.6578	0.0016
100	MLP(RMSE)	264K	1024	10^{-3}	80	0.5793	14.055
100	MLP(Cross-Entropy)	366K	1024	10^{-3}	80	0.5771	0.0170

Table S2. Validation RMSE and training setup for Kelvin–Helmholtz models on a 128-by-128 grid for Det U-Net versus EDM U-Net.The EDM U-Net and Det U-Net share the same U-Net backbone ($C_b = 32$; $\sim$1.93 M parameters), and the EDM model is only slightly larger ($\sim$2.02 M) because it adds $\sim$0.05 M extra parameters for noise-level (σ) conditioning during denoising. This small overhead should not materially affect the comparison, which primarily reflects deterministic mean-field forecasting versus diffusion-based conditional sampling rather than a difference in backbone capacity.

Model	Params	Batch	LR	Epochs	Val RMSE
Det U-Net	1.97M	64	4.0e-4	300	0.0904
EDM U-Net	2.02M	64	4.0e-4	300	0.1089

Table S3. At each rollout step we compute the domain RMSE of vorticity, $\omega = \partial_x v_2 - \partial_y v_1$, between the predicted and reference fields, $\mathrm{RMSE}_\omega = \|\omega_{\mathrm{pred}} - \omega_{\mathrm{ref}}\|_2/\sqrt{N_x N_y}$, and report its time mean over the rollout window. Sample #1–#3 are three independently selected EDM U-Net diffusion members; Reduction is the percentage improvement of the best sample relative to Det U-Net, $(\overline{\mathrm{RMSE}}_{\omega,\mathrm{Det}} - \min_k \overline{\mathrm{RMSE}}_{\omega,\mathrm{Sample}\#k})/\overline{\mathrm{RMSE}}_{\omega,\mathrm{Det}} \times 100\%$.

Test member	Det U-Net	Sample #1	Sample #2	Sample #3	Reduction
3	2.302	2.327	2.306	2.330	—
14	2.835	2.768	2.823	2.802	2.4%
31	1.536	1.416	1.431	1.452	7.8%
137	1.309	1.226	1.220	1.216	7.1%

Table S4. Bulk x-momentum consistency is measured by $\varepsilon_{P_x} = |\delta P_{x,\mathrm{pred}} - \delta P_{x,\mathrm{ref}}|$, where $\delta P_x = \big[P_x(t_{\mathrm{end}}) - P_x(0)\big]/P_x(0)$ is the percent change in the domain integral $P_x(t) = \int_\Omega \rho v_x \, \mathrm{d}\mathbf{x}$ from $t = 0$ to the end of the rollout. Det U-Net and EDM U-Net are initialized from the reference state and integrated autoregressively on the periodic domain. Sample #1–#3 are three independent EDM rollouts; Reduction is the percentage improvement of the best sample relative to Det U-Net.

Test member	Det U-Net	Sample #1	Sample #2	Sample #3	Reduction
3	3.24	2.51	2.38	4.46	26.5%
14	4.73	0.71	0.40	0.30	93.7%
31	2.30	0.31	0.63	0.46	86.7%
137	2.81	0.03	1.23	0.57	99.1%

Supplementary Figure

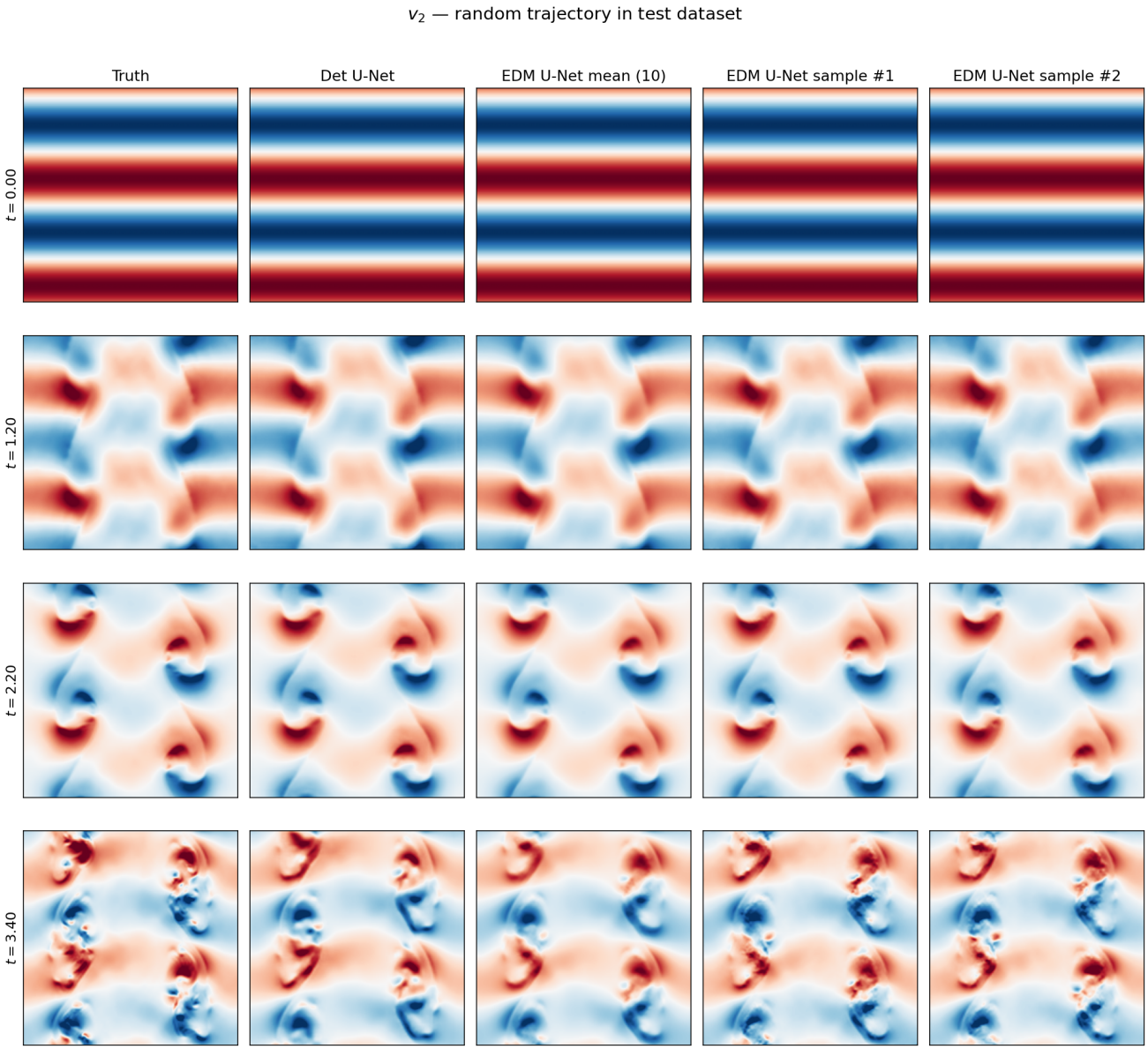

Figure S2. Autoregressive rollout of v_2 for test member 031 (Truth, Det U-Net, EDM U-Net mean over 10 samples, and two EDM samples).

64

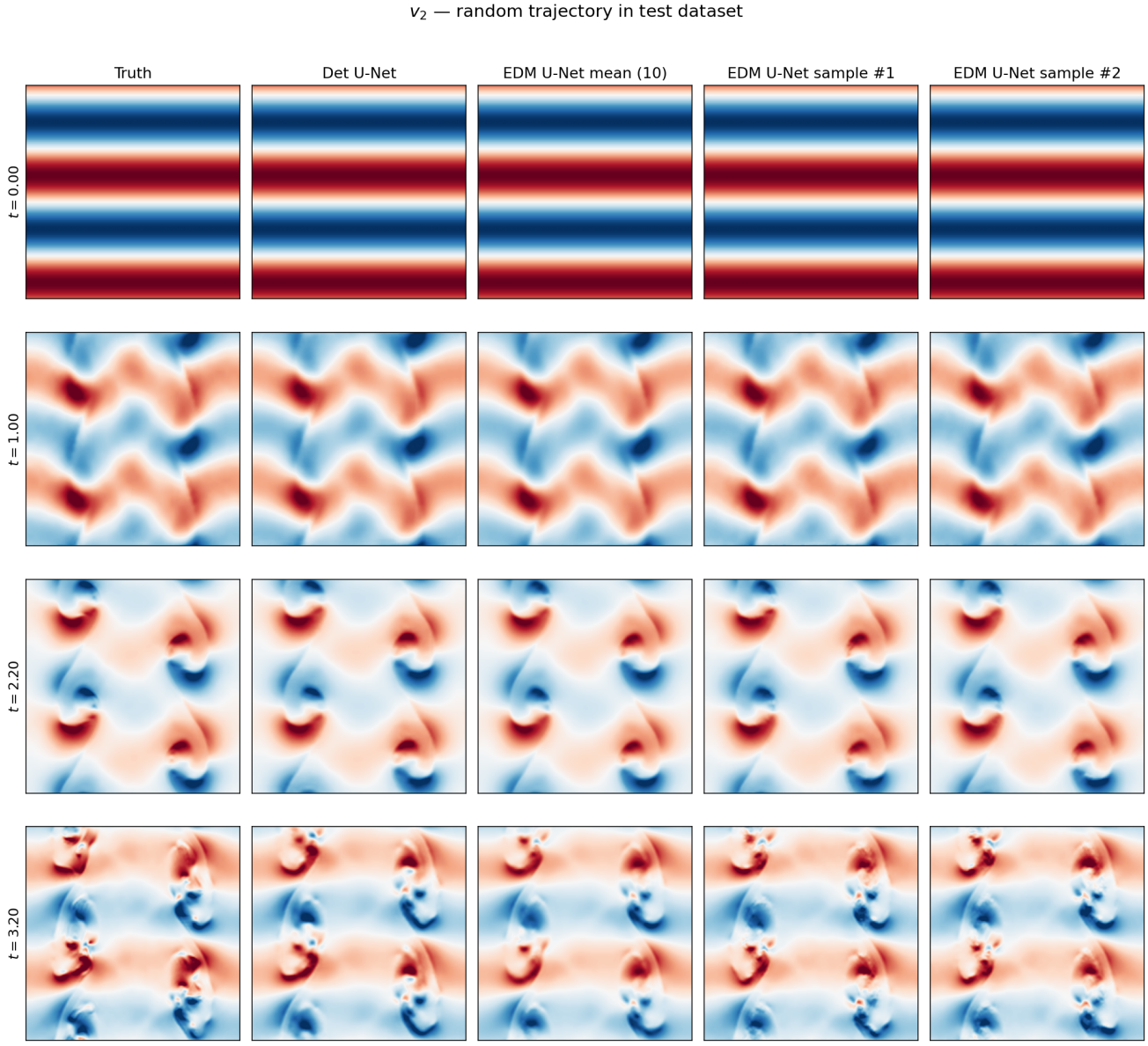

Figure S3. Same as Figure S2, but for test member 137.