Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Why Neural Emulators of the Atmosphere Lose the Butterfly

Le Duc

leduc@g.ecc.u-tokyo.ac.jp

The University of Tokyo

Izza Nur Rahman

The University of Tokyo

Yohei Sawada

The University of Tokyo

Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11764397/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Why Neural Emulators of the Atmosphere Lose the Butterfly

Le Duc[1,2], Izza Nur Rahman[1], and Yohei Sawada[1]

[1]Department of Civil Engineering, University of Tokyo, Tokyo 113-8656, Japan

[2]Meteorological Research Institute, Tsukuba 305-0052, Japan

Corresponding author: Le Duc, leduc@g.ecc.u-tokyo.ac.jp

November 17, 2026

Short title: Why neural emulators lose the butterfly

Key points

- Trained on exact model data, an emulator grows small perturbations some six hundred times too little, with nearly the same gain at every flow
- No training objective we tested restores the butterfly, and tangent matching recovers amplitude only as rotational, large-scale growth
- Averaging the truth's tangent over nearby atmospheres reproduces the deficit's kind, and the emulator's response points along that average, making unresolved detail the leading explanation

Abstract

Neural emulators forecast the weather as well as the physical models they imitate, yet they grow small disturbances far too slowly. The usual explanations blame the data: the analyses are too noisy, and too coarse in time, to contain the butterfly effect. Both can be tested only where the data are exact. We therefore train an emulator on a general circulation model made fully differentiable, so that its training data are exact, the true response to a disturbance is known rather than inferred, and every discrepancy belongs to the emulator. The defect survives. Disturbance energy grows more than two orders of magnitude less than in the model it imitates, and the failure is not weak growth alone but blindness to the weather: the true amplification varies more than tenfold between atmospheres, where the emulator's hardly varies. No training objective we tested restores it, and the one that comes closest grows the wrong structures and worsens the forecast. We trace the failure to what fitting by squared error returns: the average of the futures of the atmospheres a network cannot tell apart. Averaging the true response over such atmospheres removes its dependence on the weather, and the emulator's own response points along that average rather than along the one it was given. Neither a change of loss nor the architectures and capacity we could reach escapes this. Restoring the butterfly effect asks instead for a model that can tell such atmospheres apart, or one that samples them rather than averaging them. An accurate forecast is therefore no evidence of an accurate sensitivity.

Plain Language Summary

A famous property of the atmosphere is the butterfly effect: tiny disturbances grow until they spoil the forecast. Machine-learning weather models predict the weather well but largely fail to reproduce this growth, which matters because forecast centers rely on it to estimate uncertainty. We build a machine-learning copy of a small but complete climate model, so that we know the

exact answer, and ask whether the defect can be trained away and, if not, why. We find that no training method we tried repairs it, and that the one which comes closest does so by amplifying the wrong kind of weather structure while nearly doubling the forecast error. The evidence points to a simple reason: the model cannot tell apart atmospheres that differ only in fine detail, and training it to be as accurate as possible on average makes it predict the average of their futures. Because each of these atmospheres amplifies disturbances in its own way, the average amplifies very little. Changing the network, or the way it is built, did not escape this in any of our tests. What would escape it is a model that can tell these atmospheres apart, or one that draws from them at random instead of averaging them.

1 Introduction

Within a few years, machine-learning weather prediction models have changed what is expected of a numerical forecast. Trained on reanalyses, emulators such as FourCastNet [Pathak et al., 2022], Pangu-Weather [Bi et al., 2023], GraphCast [Lam et al., 2023], the operational AIFS [Lang et al., 2024] and the foundation model Aurora [Bodnar et al., 2025] now match, and in places exceed, the deterministic skill of the operational systems they imitate [Rasp et al., 2024], at a small fraction of the computational cost. It is therefore natural to ask whether they can also take over the other tasks of a numerical model: ensemble prediction, data assimilation, and the study of predictability itself, together with every use that differentiates the emulator, from the optimization of its initial conditions [Vonich and Hakim, 2024, 2026] to adjoint sensitivity and variational assimilation through the network. All of these rest on a property that deterministic skill does not measure. The atmosphere is chaotic, and a small difference between two states grows, first at small scales and then upscale, until the two states are no more alike than two random states of the climate [Lorenz, 1963, 1969, Rotunno and Snyder, 2008, Lorenz, 1982, Zhang et al., 2019]. This growth, the butterfly effect, is what an ensemble samples [Leutbecher and Palmer, 2008], what limits predictability, and what a data assimilation system must follow between observations. How tight a limit it sets is itself debated, between the real butterfly effect of Palmer et al. [2014], in which the smallest scales impose a finite horizon that no reduction of the initial error can push back, and the view of Durran and Gingrich [2014] that errors at those scales are of little practical importance beside the synoptic ones. A model that forecasts well but does not grow small perturbations correctly is a good forecaster and a poor model of forecast uncertainty.

It is now well documented that machine-learning models do not grow small perturbations correctly. Selz and Craig [2023] showed that a machine-learning model driven by tiny initial perturbations substantially underestimates their growth, and Selz and Craig [2026] extended the test to several models of different architectures and sizes, none of which reproduced the butterfly effect in a physically consistent way. In some models the perturbations showed no accelerated growth at all. In others they grew, but through numerical noise, in some cases depending on whether the computation was carried out on a central or a graphics processor. Selz and Craig [2026] attribute the failure to the training data rather than to the architecture. Analyses carry errors much larger than the perturbations of interest, so the butterfly is not contained in the data from which the models learn, and learning it would require far closer analogs than a few decades of analyses provide. Hassanzadeh et al. [2026] attribute the deficit instead to the coarse-graining of the training data, and show that reducing it (the full state at one solver step in Lorenz-96, the hourly rather than six-hourly Pangu-Weather) restores butterfly-like growth. Both explanations blame the data, one its errors and the other its resolution in time, and both can be tested only where the data can be made exact and complete. Training on a long simulation of a conventional model, a nature run that is an exactly known truth, might in principle work, although Selz and Craig [2026] judged it impractical. The same question has since been put to probabilistic machine-learning models [Chen et al., 2026]. The smoothness

of machine-learning forecasts, to which the mean-squared-error objective contributes, has also been identified as a general limitation of these models [Bonavita, 2024], and has consequences for how such forecasts must be verified [Bonavita and Geer, 2026].

These explanations raise a question that studies based on reanalyses cannot answer. Suppose that the training data were exact, taken from a nature run of the very model that the emulator imitates, so that, given the model's full state, the next state is a deterministic function of the current one. The mean of the next state given the current state is then the next state itself, and the minimizer of the mean squared error is the perfect model, butterfly included. Does the butterfly then return? If it does not, where does the loss of growth come from? A natural suspect is still an averaging, but an averaging of what, and why should it remove precisely the growth? And if averaging is the problem, can a different training objective, a different network, or a different way of building the map avoid it?

We address these questions with a twin experiment in which such exact data are available. The truth is SPEEDY [Molteni, 2003], a spectral primitive-equation general circulation model, which we have reimplemented in JAX so that it is differentiable from end to end, and the emulator is trained on a 20-year free run of that same model, a nature run in the sense of Selz and Craig [2026]. Because truth and emulator share a code base, every discrepancy is attributable to the emulator rather than to a mismatch of physics, resolution, or data. Because the truth is differentiable, its tangent linear operator and singular vectors are available exactly, so that the growth of perturbations can be compared operator by operator rather than through the shadow it casts on a forecast score. The setting is deliberately modest, with T30 resolution and eight vertical levels, and it is this modesty that allows the results to be stated sharply.

We list here the questions that the paper addresses:

- How large is the emulator's deficit in perturbation growth?
- Can any training objective, deterministic or stochastic, restore the butterfly? If one comes close, does it grow the right structures, and at what cost?
- If none does, what is it about the trained emulator that prevents it?
- And does an operational emulator show the same insensitivity to the flow?

The paper is organized as follows. Section 2 describes the twin: the reference atmosphere and the two emulator architectures. Section 3 sets out how the butterfly is measured, with a set of scores that separates the physical butterfly from spurious growth, and how a perturbation must be constructed to measure growth at all. Section 4 reports the training objectives and their uniform failure, including the tangent-matching objective that comes closest and what it costs. Section 5 asks why a good forecast can have the wrong sensitivity, tests the averaging explanation against the alternatives, and, in Section 5.9, measures the same quantity on an operational emulator. Section 6 summarizes the findings and discusses their implications.

2 Numerical models

The experiment is a twin: a reference atmosphere that plays the role of the truth, and a neural emulator trained to imitate it. Both are described here, and the measurements made on them are reported in the sections that follow.

2.1 The reference atmosphere

Every measurement in this paper is a comparison against a reference, and the reference is a numerical atmosphere we can differentiate. It is SPEEDY [Molteni, 2003], a hydrostatic primitive-equation general circulation model, run at T30 spectral truncation on a 96×48 Gaussian grid with eight sigma levels and a native time step of 2400 s. Its prognostic state is the two

horizontal wind components, temperature, and specific humidity on every level, together with surface pressure, which is 33 fields on the grid and about 1.5×10^5 degrees of freedom. The dynamical core is spectral, and the time integration is leapfrog with a Robert-Asselin filter, so a model state is a pair of adjacent time levels rather than a single field.

Two properties make this model, rather than a reanalysis, the right reference. First, it is a JAX re-implementation of SPEEDY's dynamical core with physics ported from ICON [Zängl et al., 2015], so that it is differentiable from end to end, in the manner that differentiable programming has made usual for Earth system models [Gelbrecht et al., 2023] and that also allows learned components to be trained inside a dynamical core [Kochkov et al., 2024]. The tangent linear operator, its adjoint, and the singular vectors of any finite-time map are then available exactly by automatic differentiation, rather than through a separately coded and separately approximated linearization. These are the instruments with which perturbation growth is measured throughout the paper, and here they carry no linearization error of their own. Second, the model is the reference by construction. No observation operator, no analysis increment, and no unknown model error stands between the dynamics we study and the dynamics we verify against, so that any deficit reported later has a single admissible source and cannot be charged to a mismatch of physics, resolution, or data.

Because the whole subject of the paper is the fidelity of physical error growth, the reference must have physics worth trusting. SPEEDY's own parameterizations are deliberately schematic, and we replace them with a differentiable stack ported from the ICON model [Zängl et al., 2015, ico, 2026]: a scale-aware Tiedtke-Bechtold mass-flux convection scheme with condensate detrainment, a total-turbulent-energy boundary layer, single-moment cloud microphysics, Monin-Obukhov surface exchange, sub-grid orographic drag, a bucket land surface, and a resolved diurnal cycle. Every scheme is differentiable, so the tangent and adjoint operators run through the full moist physics and not through a dry reduced core, which is what makes the growth we measure a growth of the atmosphere's instabilities rather than of a linearized caricature.

Faithful schemes must still produce a recognizable atmosphere, or there is no butterfly worth studying. Figure 1 shows a single instantaneous state of the reference, on a northern summer day. The relative vorticity at 500 hPa resolves the train of baroclinic waves that organizes the midlatitude storm tracks, strongest in the winter hemisphere, the rotational large scales along which synoptic forecast error first grows. The column water vapor shows the moist tropics, the dry subtropical deserts, and the plumes of moisture that the waves draw poleward. The heaviest precipitation falls along the intertropical convergence zone, in the South Asian monsoon, and in the fronts of the storm tracks, where moist convection produces the divergent small scales through which perturbation energy later cascades upscale. The two panels are not incidental illustration. They are the two regimes whose interplay is the physical butterfly, and displaying them at the outset fixes what the rest of the paper measures: a live instability of a recognizable atmosphere, and whether a surrogate trained on that atmosphere can reproduce its growth.

2.2 The emulators

We emulate the reference with a neural network that advances the physical state by six hours and is applied autoregressively to produce a forecast, in the manner of the operational data-driven models. From the normalized grid-space state, the two horizontal wind components, temperature, and specific humidity on all eight levels together with surface pressure, 33 channels on the 96×48 grid, augmented by the static and solar forcing fields, the network predicts the normalized six-hour increment, which is added back to advance the state. The fields it sees are those of the leading time level of the model's leapfrog pair. SPEEDY advances every step from both levels, the leading one and the level one native step earlier, so strictly the emulator does not see everything that determines the truth's next state. Section 5.8 tests whether this matters by training an emulator on both levels: it does not.

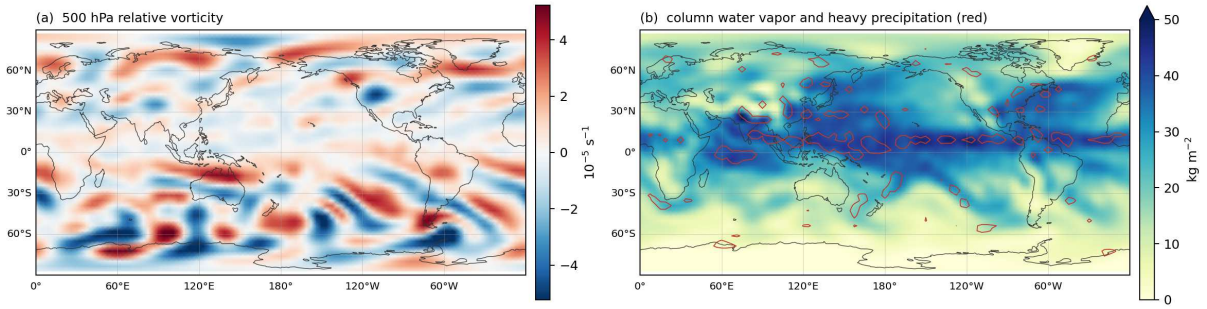

Figure 1: A single state of the calibrated SPEEDY configuration, after 200 days of free run. (a) Instantaneous relative vorticity at 500 hPa. (b) Instantaneous column water vapor, with the 10 mm day^{-1} contour of the precipitation accumulated over the preceding 24 hours (red).

We use two architectures, so that no conclusion rests on the particulars of a single network. The first is a convolutional residual network: eight pre-activation residual blocks of 128 channels with 5×5 kernels, whose padding respects the topology of the sphere, circular in longitude and edge-replicating in latitude so that the poles are not spuriously coupled, with depth rather than dilation supplying a near-global receptive field. It has 6.5 million parameters. The second is a spherical Fourier neural operator [Bonev et al., 2023], the architecture class of the operational spectral emulators, built on the reference model's own differentiable spherical-harmonic transform so that it shares the exact T30 basis. Each of its eight blocks pairs a spectral branch, which transforms the field to spectral space, mixes channels with a complex weight learned per total wavenumber and shared across zonal wavenumber in the rotationally consistent operator parameterization, and transforms back, with a pointwise grid branch, in a pre-norm residual identical in form to the convolutional block, so that the two networks differ only in their basis.

Both are trained on a 20-year free run of the reference sampled six-hourly, giving 29200 transition pairs, by minimizing the one-step mean squared error. The baseline emulator, which we call CNN, is the convolutional network trained in this way and then fine-tuned through multi-step rollouts, the standard recipe. The spherical emulator, which we call SFNO, is used as trained on the one-step error. Each emulator of Section 4 is named after the quantity its objective supervises. Holding the architecture fixed and varying only the training objective gives the family of emulators examined in Section 4, where their skill and their perturbation growth are reported.

3 How to measure a butterfly

Whether a model has a butterfly is not a yes-or-no question, and a single growth rate does not answer it. A model can grow perturbations quickly for the wrong reason. It may be running away into a numerical instability, or it may be amplifying the large-scale, rotational structures that grow fastest in the six-hour tangent. Neither is the physical butterfly, which is a small-scale, moist, divergent growth that saturates and cascades upscale to an intrinsic horizon. A useful set of scores must therefore not merely detect growth but also tell the physical butterfly from two kinds of spurious growth: a numerical instability, and the optimal, large-scale growth that singular vectors describe and with which operational ensembles are seeded.

How a perturbation is built. How a perturbation is built decides what every score in this section measures. A perturbation of a primitive-equation model must lie on, or very near, the model's slow manifold. If it does not, the model answers with a burst of gravity waves that radiate and dissipate, and a diagnostic that reads perturbation energy cannot differentiate that

burst from growth. Noise specified on the model grid is a common choice for emulators, and a physical model must be allowed to adjust before its growth is read. What an unbalanced perturbation does to the model is visible at once. Grid-point noise of $10^{-5}\sigma$ is amplified by 115 in a single 40-minute step and by 1545 over six hours, and then goes flat at an amplitude no growth has had time to reach: the signature of a shock, not of growth. In contrast, a balanced perturbation, a scaled difference of two states on the attractor, grows by 1.6 in the first step and 7.7 over six hours, steadily and from the start.

We therefore construct every perturbation in this paper in the model's own state space, as a scaled difference of two states the dynamics actually visit,

$$\delta x \;=\; \alpha\,(x' - x), \tag{1}$$

applied to both leapfrog time levels of the base state x, with the scale α chosen so that the perturbation has a prescribed root-mean-square amplitude in units of σ. Because x and x' each satisfy the model's balance relations on their own, their difference is balanced by construction and excites no gravity-wave shock.

The scores themselves follow. Selz and Craig [2026] proposed six characteristics for telling a physical butterfly from spurious growth. We adopt four of them, and add two that a differentiable twin uniquely affords, giving the six scores set out below. Each score is computed on a small ensemble of the balanced perturbations just described, taken about states on the attractor, and every field is expressed in units of its climatological standard deviation σ so that variables of different magnitude contribute comparably. When we quote an error, a distance between states, or a width, we leave out the humidity of the three uppermost levels, where the air is nearly dry and a tiny absolute error becomes large in units of σ. Kept in, the third of these levels alone would carry more than half of the emulator's one-step error. Perturbation amplitudes and gains are computed over all fields. Leaving the three levels out changes the gains by about one percent at the median state, for the truth and for the emulators alike.

The base quantity for the amplitude and limit scores is the difference kinetic energy. For perturbation wind components $(\delta u, \delta v)$ at lead time t, averaged over the ensemble, the vertical levels, and the horizontal grid,

$$\mathrm{DKE}(t) \;=\; \tfrac{1}{2}\big\langle\, \delta u(t)^2 + \delta v(t)^2 \,\big\rangle. \tag{2}$$

Two quantities derived from this energy recur in the scores below, and it is worth stating plainly what each is. The first is the difference kinetic energy resolved in space rather than averaged over it: at every horizontal grid point, $\mathcal{F}(t) = \frac{1}{2}\langle \delta u(t)^2 + \delta v(t)^2 \rangle$ averaged over the ensemble and the vertical levels only, so that $\mathcal{F}(t)$ is a map on the sphere showing where the perturbation energy sits at lead time t. The second is the spectrum of the perturbation wind. Expanding $\delta u(t)$ and $\delta v(t)$ on every level in spherical harmonics and summing the squared coefficients of all components that share a total wavenumber ℓ, over both wind components and all levels, gives $P(\ell, t)$, the perturbation kinetic energy carried at each spatial scale, with small ℓ the planetary scales and large ℓ the smallest the grid resolves. In short, $\mathcal{F}$ records where the growth is and $P(\ell, \cdot)$ records at which scales. The migration of P in time is the cascade.

Two scores of amplitude. *Saturation fidelity* $\mathcal{S}$ asks whether the smallest perturbations reach the climatological saturation level, neither dying nor exploding. With DKE_∞ the difference energy between two independent states of the truth, the level every model is scored against,

$$\mathcal{S} \;=\; \frac{\big\langle \mathrm{DKE}(T) \big\rangle_{a \le 0.01\sigma}}{\mathrm{DKE}_\infty}, \tag{3}$$

the numerator averaged over initial amplitudes a at or below 0.01σ at the end of a long integration T. A physical butterfly gives $\mathcal{S} \approx 1$, an emulator that amplifies too weakly gives $\mathcal{S} \ll 1$,

and a numerical instability gives $\mathcal{S} \gg 1$. The score is bounded on both sides, which is what lets it flag the numerical instability that a bare growth rate rewards.

Tangent amplification σ_1 is the leading gain of the six-hour map over a set of balanced probe directions. Writing M for that map and $J = \partial M/\partial x$ for its linearization in σ-normalized coordinates, restricted to the span of the probes,

$$\sigma_1 \;=\; \max_{\|\delta x\|=1} \|J\,\delta x\|. \tag{4}$$

The probes are the balanced perturbations of Equation 1, 12 at each state for the flow comparison of Section 5.2 and eight in the averaging experiment of Section 5.4, and both models are read on the same grid, the truth through its projection onto the physical fields. Because the twin is differentiable, we evaluate $J\,\delta x$ exactly by automatic differentiation, a Jacobian-vector product with no finite-difference step. The norm is Euclidean in σ-normalized coordinates, which weights each variable by its climatological variance so that variables of different units are comparable. Over a finite set of probes, σ_1 is a lower bound on the norm of the six-hour tangent. They describe how an infinitesimal perturbation would grow, not how a perturbation of finite size responds, which Section 5 measures separately. This score is a reason the twin must be differentiable: only when the truth is a known operator can its tangent be compared directly, rather than through the shadow it casts in a forecast score.

Three scores of structure, which separate the butterfly from spurious growth. *Optimal-growth correlation* ρ_{op} asks whether the growth occupies the same places as large-amplitude, synoptic growth, the optimal growth that singular vectors describe. Let $\mathcal{F}_{\mathrm{small}}$ be the early-lead energy map $\mathcal{F}$ from perturbations of amplitude 0.001σ and $\mathcal{F}_{\mathrm{op}}$ the same map from perturbations of amplitude 1σ. Then

$$\rho_{\mathrm{op}} \;=\; \mathrm{corr}\big(\mathcal{F}_{\mathrm{small}},\, \mathcal{F}_{\mathrm{op}}\big) \tag{5}$$

is their centered spatial correlation. The physical butterfly grows where the large-amplitude flow does not, so that $\rho_{\mathrm{op}} \approx 0$, whereas optimal growth grows the same synoptic structures at every amplitude, so that ρ_{op} approaches one.

Spectral cascade $\Delta\bar{\ell}$ records the direction in which energy moves across scales. With the spectral centroid and its shift between an early and a late lead,

$$\bar{\ell}(t) \;=\; \frac{\sum_\ell \ell\, P(\ell,t)}{\sum_\ell P(\ell,t)}, \qquad \Delta\bar{\ell} \;=\; \bar{\ell}(t_{\mathrm{early}}) - \bar{\ell}(t_{\mathrm{late}}), \tag{6}$$

a positive $\Delta\bar{\ell}$ means the centroid falls and energy migrates from small to large scales, the upscale cascade of the physical butterfly. Optimal growth cascades the other way, $\Delta\bar{\ell} < 0$. This replaces Selz and Craig's static peak wavelength at short lead with the direction of the cascade, which is the discriminating feature and does not depend on where the resolved small-scale limit happens to lie.

Divergent fraction f_{div} asks whether the growing wind is divergent, as moist convective outflow is, or rotational, as balanced synoptic growth is. Helmholtz-decompose the perturbation wind into rotational and divergent parts, $(\delta u, \delta v) = (\delta u_\psi, \delta v_\psi) + (\delta u_\chi, \delta v_\chi)$, and take the divergent share of the kinetic energy,

$$f_{\mathrm{div}} \;=\; \frac{\langle \delta u_\chi^2 + \delta v_\chi^2 \rangle}{\langle \delta u_\chi^2 + \delta v_\chi^2 \rangle + \langle \delta u_\psi^2 + \delta v_\psi^2 \rangle}. \tag{7}$$

The value one half marks equal divergent and rotational shares and serves as a reference, not a threshold. The truth's early growth is divergence rich, whereas its later growth, after the cascade has reached the balanced large scales, falls below one half. The score is therefore compared between models in the same window, never against one half alone.

One score of the limit. The last score asks the question the butterfly effect is named for: does a smaller initial error always yield more forecast time, or does the yielding stop? Let $\tau(a)$ be the lead at which a perturbation of initial amplitude a first reaches a quarter of its saturation level, measured across three decades of a. If the growth rate does not depend on amplitude, every tenfold reduction of a yields the same extra stretch of τ, and τ keeps climbing without bound. If instead the smallest errors grow fastest, each further reduction yields less than the one before and τ flattens: that flattening is the horizon. The *predictability-limit index* Π simply compares what the small-amplitude end of the range bought, $\Delta\tau_{\text{small}}$, with what the large-amplitude end bought, $\Delta\tau_{\text{large}}$,

$$\Pi \;=\; 1 - \frac{\Delta\tau_{\text{small}}}{\Delta\tau_{\text{large}}}, \tag{8}$$

so $\Pi \approx 1$ means the last decade of refinement bought nothing and a horizon has been reached within the range tested, while $\Pi \approx 0$ means it bought as much as the first and none has. The index speaks only about the decades tested. A smooth model of finite resolution has no horizon at arbitrarily small amplitude, and no finite experiment can rule one out.

For the truth, saturation fidelity is near one, the tangent amplifies a broad subspace, the early growth is divergent and uncorrelated with the optimal-growth pattern, the centroid falls, and the limit index is near one: the profile of a genuine butterfly. The sections that follow ask how much of that profile an emulator trained on the very same atmosphere can be made to reproduce.

4 Can a loss restore a butterfly?

Whether an emulator grows perturbations the way the atmosphere does may depend on how it is trained. Mean squared error is the standard objective, and if the missing growth is a property of that choice then a different loss should restore it. We test this directly, holding the architecture fixed and varying only the loss. Write M_θ for the emulator's normalized six-hour map and $y = M_\star(x)$ for the truth's next state. Most of the objectives we train have the form

$$\mathcal{L} \;=\; \underbrace{\mathbb{E}\,\big\|M_\theta(x) - y\big\|^2}_{\mathcal{L}_0} + \lambda\,\mathcal{P}, \tag{9}$$

a mean-squared-error anchor $\mathcal{L}_0$ that preserves forecast skill and a penalty $\mathcal{P}$ that is the object of the experiment.

The anchor is also the quantity by which we judge an emulator as a forecast model, and we name it here because it is quoted throughout the paper. The one-step error is

$$e_1 \;=\; \Big(\mathbb{E}\,\big\|M_\theta(x) - M_\star(x)\big\|^2\Big)^{1/2}, \tag{10}$$

the root-mean-square difference between the emulator's state after one of its own steps, six hours, and the truth's, over held-out states, in the normalized coordinates of Section 3, so that it is a fraction of the climatological spread of the field. It is $\mathcal{L}_0$ itself, up to the square root: the emulator is trained to make it small, and the question of this paper is what else is lost while it is being made small. The exception is the stochastic objective, which replaces the anchor with a probabilistic score. We score each trained emulator on the six measures of Section 3. Figure 2 collects every score of every objective, together with the forecast skill each one costs, and the subsections that follow take them one at a time.



Figure 2: Every objective against every score. Rows: the truth and the eleven trained objectives, each named after the quantity its loss supervises. Columns: the growth factor of balanced perturbation energy over 12 days at an initial amplitude of 0.001σ; the six scores of Section 3, saturation fidelity $\mathcal{S}$, predictability-limit index Π, tangent amplification σ_1, optimal-growth correlation ρ_{op}, divergent fraction f_{div} and spectral cascade $\Delta\bar{\ell}$; the one-step zonal-wind error, which is the forecast skill; and how many of the six the objective meets. A cell is dark where it reaches the butterfly's own value, pale where it is part of the way and white where it is not, on the rule printed under each heading, of which the two in the structure columns are the thresholds the instrument itself prints. The word none marks a score left undefined by a perturbation that never reaches a quarter of saturation within the 45 days measured.

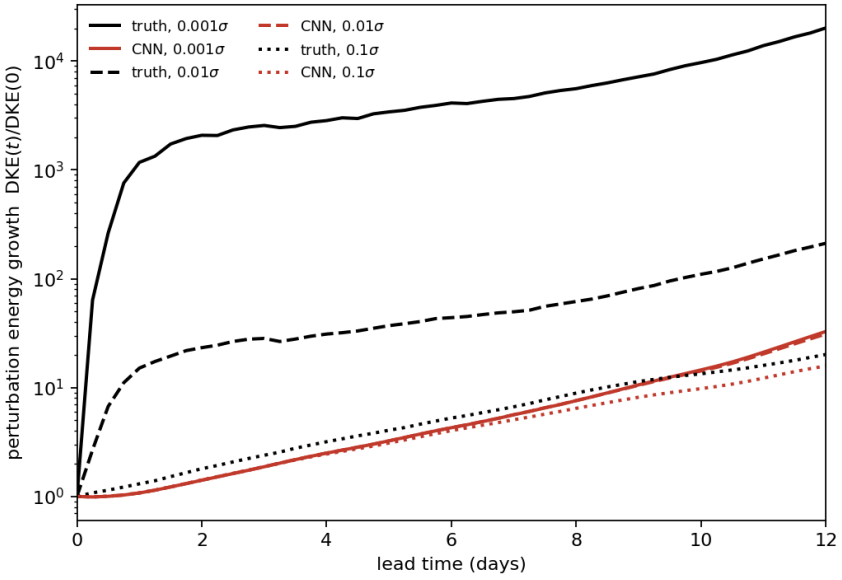

Figure 3: Growth factor of balanced perturbation energy against lead time, for the truth (black) and the baseline mean-squared-error emulator CNN (red), at initial amplitudes of 0.001σ (solid), 0.01σ (dashed), and 0.1σ (dotted).

4.1 The baseline: mean squared error

The baseline emulator CNN is trained by mean squared error alone, Equation 9 with $\lambda = 0$, followed by rollout fine-tuning, and it is our reference failure (Figure 3). Its 12-day growth of balanced perturbation energy at an initial amplitude of 0.001σ is a factor of 32.8 against the truth's 20147 (Figure 2), so its saturation fidelity $\mathcal{S}$ sits at 0.53 after 45 days against the truth's 0.98: the smallest perturbations reach about half the saturation level, and take the whole 45 days to do it. The more telling feature is the amplitude dependence. The truth grows fastest at the smallest amplitude, a factor of 20147 at 0.001σ falling to 20.1 at 0.1σ, the signature of a system whose smallest perturbations grow fastest and then saturate. CNN's three amplitudes instead collapse onto a single curve near a factor of thirty, so it has no small-perturbation regime, and predictability keeps growing as the amplitude shrinks rather than approaching a limit: its index is $\Pi = 0.01$ against the truth's 1.19 over the same three amplitudes. The tangent amplification, computed exactly, is $\sigma_1 = 1.04$ for CNN, and nearly the same at every flow, where the truth's varies with the flow from 1.3 to 20 (Section 5.2). Whether any other objective can escape this is the question of the rest of the section.

4.2 Spectral loss

The first alternative pins what mean squared error leaves loose. The spectral objective spec adds the penalty

$$\mathcal{P}_{\text{spec}} = \sum_{\ell} \big(\log P_\theta(\ell) - \log P_\star(\ell)\big)^2, \tag{11}$$

matching the emulator's kinetic-energy power spectrum $P_\theta(\ell)$ to the truth's at every total wavenumber ℓ, on the hypothesis that a model with the right variance at every scale would grow perturbations correctly. It does the opposite. spec grows least of all the single objectives, a 12-day growth factor of 3.1, and it has the lowest saturation fidelity $\mathcal{S}$ of the whole family (Figure 2). Matching the spectrum fixes the climatological variance while leaving the perturbation dynamics untouched, a sharp illustration that variance is not dynamics. A model can carry exactly the right energy at every scale and still transport a perturbation without amplifying it.

4.3 Stochastic loss

If a deterministic map cannot grow perturbations, a stochastic one might. The stochastic objective stoch replaces the anchor with the energy score,

$$\mathcal{L}_{\mathrm{ES}} \;=\; \mathbb{E}\,\big\|X - y\big\| \;-\; \tfrac{1}{2}\,\mathbb{E}\,\big\|X - X'\big\|, \tag{12}$$

in which X and X' are independent draws from the emulator's predictive distribution. The score is smallest when that distribution is calibrated. The spread it learns is not of the butterfly's kind. Measured on balanced perturbations its divergent fraction is 0.33, which is the baseline's own value, against the truth's 0.68, so the noise it injects is as rotational as the growth the deterministic network already produces. Neither does it carry an amplitude that compounds. Its 12-day balanced growth is a factor of 41.6, and its saturation fidelity, 0.67, stays near the baseline's 0.53 (Figure 2). Varying the amplitude of the injected noise shows that the spread grows only in proportion to the noise supplied, not through any dynamical instability, and its predictability-limit index, $\Pi = 0.44$ against the baseline's 0.01, is the closest any objective here comes to the truth's, supplied by the injected noise rather than by growth. The energy score calibrates the uncertainty of a single forecast step, but it does not install a growing tangent.

4.4 Matching the growth directly

If the objective is perturbation growth, the most direct choice is to match it. The objective dke adds

$$\mathcal{P}_{\mathrm{dke}} \;=\; \sum_{\tau} \big(\log g_\theta(\tau) - \log g_\star(\tau)\big)^2, \qquad g(\tau) \;=\; \frac{\mathrm{DKE}(\tau)}{\mathrm{DKE}(0)}, \tag{13}$$

matching the emulator's perturbation-energy growth g_θ to the truth's at every lead τ. It is the one objective that produces too much growth: it overshoots the truth's 12-day factor by a wide margin, reaching about 44000, and its saturation fidelity is 31.37 where the truth's is one (Figure 2). The overshoot is a numerical instability of the kind described in Section 3, not a butterfly. The dke penalty exploits the aggregate scalar it is trained on by running away, and it destabilizes free running. Its divergent refinement dkeDiv matches only the divergent-wind part of the growth, g_χ, the part in which the physical butterfly lives. It trains stably and, alone among the single objectives, grows a divergent, non-operational structure, a divergent fraction of 0.66 against the truth's 0.68 with an optimal-growth correlation of 0.39. But its amplitude does not follow: a 12-day factor of 4.5, below the baseline's, and a saturation fidelity lower still (Figure 2). Matching the growth scalar yields either an instability or no growth, depending on how tightly the target is drawn, and neither is the butterfly.

4.5 Tangent-matching loss

The most direct attack on the amplitude is to match the tangent itself. Matching the Jacobian of a network to that of a teacher has been used to transfer knowledge between networks [Srinivas and Fleuret, 2018] and, as Sobolev training, to fit derivatives together with values [Czarnecki et al., 2017], and Kang et al. [2026] proposed randomized Jacobian matching to keep emulators of low-dimensional chaotic systems chaotic. A network's own tangent linear and adjoint operators have also been built deliberately, to serve variational assimilation [Hatfield et al., 2021]. Here the teacher is the differentiable truth itself, and the penalty

$$\mathcal{P}_{\mathrm{tan}} \;=\; \sum_{k} \big\|(J_\theta - J_\star)\, v_k\big\|^2 \tag{14}$$

measures the difference between the emulator's six-hour tangent J_θ and the truth's $J_\star$ along a set of probe directions v_k. The choice of v_k is decisive, and the probes must be balanced. Matching the tangent on unbalanced probes, the naive versions jacUnbal and jacUnbalScale,

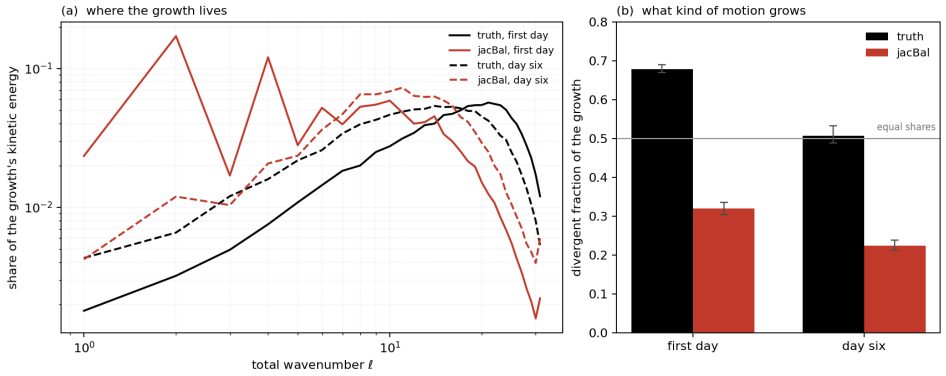

Figure 4: The structure of the growth, for the truth and for the tangent-matching emulator jacBal, from small balanced perturbations (0.001σ, eight members at each of three states). (a) Where the growth lives: the kinetic-energy spectrum of the growth over the first day and at day six, each normalised to unit total so that the panel compares shape and not amplitude. (b) What kind of motion grows: the divergent fraction of the same growth, with the range over the three states. The truth's growth is divergent, 0.68 falling to 0.51; the emulator's is rotational, 0.32 falling to 0.22. The grey line marks equal divergent and rotational shares.

chase the imbalance shock of Section 3 and leave the balanced growth far short, at factors of 31 and 167 against the truth's 20147 (Figure 2). Taking v_k to be the leading singular directions of the balanced tangent instead makes the penalty meaningful. This objective, which we call jacBal, is the only one that substantially restores the amplitude: the 12-day growth at 0.001σ rises from the baseline's 32.8 to 1725, within a factor of twelve of the truth's, the growth becomes amplitude-dependent as the truth's is, and the tangent-amplification score rises to $\sigma_1 = 1.83$ from the baseline's 1.04 (Figure 2). The growth it installs does not saturate as the truth's does, however. Its saturation fidelity is 22 where the truth's is one, so like dke it runs past the saturation level instead of approaching it, and its predictability-limit index is negative. By the two amplitude scores that motivate the literature, the 12-day growth and the tangent gain, the butterfly is largely back.

This restoration comes at a cost with two parts. The first is forecast skill. Varying the weight λ of the tangent penalty trades the deficit against the skill. At the strongest weight the one-step zonal-wind error has risen from 0.473 to 0.862 m s^{-1} and the ten-day error from 3.17 to 9.04 m s^{-1}. The emulator whose amplitude is restored thus has nearly twice the one-step error and nearly three times the ten-day error.

The second part of the cost is structural and more damaging (Figure 4). The three structure scores of Section 3 all report the wrong growth. The optimal-growth correlation is $\rho_{\mathrm{op}} = 0.64$, where the truth's is -0.02. The divergent fraction is $f_{\mathrm{div}} = 0.32$, where the truth's is 0.68. And the spectral cascade runs downscale, $\Delta\bar{\ell} < 0$, where the truth's runs upscale. jacBal has installed optimal growth, the large-scale rotational growth that singular vectors describe, and not the physical butterfly.

The scale-aware version shows that this is the tangent's doing and not the loss's. The variant jacBalScale aims the same penalty at small-scale, divergent probe directions instead. It improves nothing. Its optimal-growth correlation is $\rho_{\mathrm{op}} = 0.88$, further from the truth's than jacBal's 0.64. It barely raises the tangent score, to $\sigma_1 = 1.05$. And its multi-day growth collapses to a factor of 16.5, below even the untreated baseline's, while it pays the largest skill cost of any loss, a ten-day error of 11.3 (Figure 2). The divergent subspace has no genuine linear amplitude to explore: the exact tangent along small-scale divergent directions is near one (Section 5.8), and the large response such a perturbation appears to have at finite amplitude is a gravity-wave adjustment, not growth, so the objective chases a shock and installs nothing real. Tangent matching thus meets a trade-off it cannot escape. Aimed at the large-scale subspace,

it gains amplitude with the wrong, rotational structure. Aimed at the divergent subspace, it finds no linear amplitude to gain. Section 5.8 offers an explanation of both outcomes.

4.6 Combinations

The failures are complementary, which invites combining them. We trained the stochastic and tangent penalties together, stoch+jacBal, and the divergent-growth and tangent penalties together, dkeDiv+jacBal, each with the balanced tangent penalty, the one objective of Section 4.5 that restores amplitude. Such a penalty has to be weighted on its own scale. The balanced probes draw a response whose energy is smaller than that of the unbalanced ones by a factor of 60000, so a weight carried across from the naive version would leave the term with no influence on the training, and we set each weight instead so that the tangent term enters with the share of the gradient it has in jacBal itself.

The stochastic combination gains the largest growth in the family and nothing else. Its 12-day factor at 0.001σ is 8495 against the truth's 20147, where the stochastic objective alone reaches 41.6 and jacBal 1725, and its tangent gain is 2.37, above jacBal's 1.83 (Figure 2). That growth is an instability. Its saturation fidelity is 61, twice dke's own, so the perturbation runs far past the level at which two independent atmospheres differ. Its structure stays rotational, a divergent fraction of 0.33, and its one-step wind error doubles, from 0.426 to 0.839.

The divergent combination is the one that adds. Its 12-day growth rises from the divergent objective's 4.5 to 486 and its tangent gain from 1.08 to 2.14, while the divergent fraction stays at 0.53 and the optimal-growth correlation at 0.31, both on the physical side of the thresholds (Figure 2). Its saturation fidelity is 2.29, the closest to the truth's one of any objective that carries amplitude at all. The two penalties here act on different quantities, which flow grows and how much it grows, and they do not fight: the divergent penalty keeps the structure it supplies alone and the tangent penalty adds the amplitude it supplies alone.

What the composition does not reach is still most of the butterfly. The growth falls short of the truth's by a factor of 41, the energy moves downscale, $\Delta\bar{\ell} = -8.0$, where the truth's moves upscale, and the forecast cost is the largest of any objective that trains stably, a one-step wind error of 1.31 m s^{-1} against the baseline's 0.473. The combination meets two of the six scores, the same two the divergent objective meets alone, and gains a part of the amplitude without meeting a third.

Across the whole family the pattern is uniform (Figure 2). No objective meets more than two of the six scores. The direct growth penalty dke reaches its two on the amplitude side of the table, and the two divergent-growth objectives reach theirs on the structure side. dke matches the truth's amplification, in the 12-day growth and in the tangent gain, and its growth is as uncorrelated with optimal growth as the truth's is, but it reaches these as a numerical instability, which a saturation fidelity of 31 and a one-step wind error of 8.4 m s^{-1} expose. The two divergent-growth objectives, dkeDiv and dkeDiv+jacBal, recover the divergent fraction, 0.66 and 0.53 against the truth's 0.68, the first on perturbations that grow by 4.5 over 12 days and the second on perturbations that grow by 486, where the truth grows by 20147. The divergent structure can therefore be had together with a part of the amplitude, and that is how far the family reaches: no objective grows the truth's structure at the truth's rate, and the one that comes nearest to both pays the largest forecast error of any objective that trains stably. This is a negative result about a large and natural class of objectives, obtained where the teacher is exact and freely available: if the physical butterfly could be trained into an emulator by supervising a six-hour map, it would be trainable here. Why it cannot be is the subject of the next section.

5 Why the emulators lose the butterfly

Section 4 established what the emulators do: they forecast well, and they grow small perturbations far too little, whatever objective they are trained with. This section asks why. We proceed from the simplest observation to the most specific experiment, and for each step we say what it shows and what it does not. The mathematical statements behind the argument are collected in Appendix A, and the section can be read without them.

5.1 Two kinds of accuracy

The quantity this section is about is the emulator's sensitivity: what it does to a small difference between two atmospheres. Displace the initial state by δx and the forecast changes by $M_\theta(x + \delta x) - M_\theta(x)$. As the displacement is made small this becomes $J_\theta(x)\,\delta x$, where J_θ is the derivative of the emulator's map, its tangent. A differentiable twin lets us evaluate J_θ and the truth's $J_\star$ exactly, which is what makes the comparison in this section possible at all. Probing a data-driven model with prescribed disturbances, and reading what it does with them rather than how it scores, is the approach Hakim and Masanam [2024] take to the dynamics such a model has learned.

What can be wrong with an emulator whose forecasts look right? A forecast can be judged in two ways. The first asks whether the predicted state is close to the true one. The second asks whether the prediction changes in the way the truth does when the initial state changes a little. The first is a question about values, the second about slopes. The one-step error of Equation 10 measures the first kind of accuracy, of the forecast. The sensitivity measures the second, of the response to a perturbation, and the two need not go together.

They need not go together because a function can be close to another and have a different slope. Figure 5a makes the point with

$$f(x) \;=\; \sin 2\pi x \;+\; \varepsilon \sin 2\pi k x, \qquad g(x) \;=\; \sin 2\pi x, \tag{15}$$

a map f with a small fast component and the smooth g that a squared-error fit would find. With $\varepsilon = 0.04$ and $k = 25$ the two values never differ by more than ε, four percent, while their slopes differ by up to $2\pi k\varepsilon \approx 6.3$, as large as the slope of g itself. Fitting values to within four percent says nothing about the derivative. The rest of this section asks whether the emulators are in this position, and if so what is playing the part of the fast component.

The emulators behave in the same way. Their forecasts are close in value: the one-step error is about ten percent of the climatological spread for CNN and eleven percent for SFNO. Their sensitivity is not close. Figure 5b puts the two measurements side by side at the same twenty states. The forecast error is small and stays small as the lead grows: in units of the climatological spread its median is 0.099 at six hours, 0.120 at 24 and 0.147 at 48 for CNN, and 0.109, 0.160 and 0.231 for SFNO, and at six hours it lies between 0.088 and 0.125 at every state for both. The error of the exact tangent at those same states is 0.44 at the most favourable and 1.00 at the least, with a median of 0.71. It is also the one of the two that varies with the flow, by more than a factor of two across the states where the forecast error varies by a third. The two architectures track each other to within 0.016 at every state, so what the tangent measures is a property of these emulators rather than of one network.

Figure 5b shows that, for these emulators, forecast accuracy and sensitivity accuracy are different measurements, and that the second fails while the first succeeds. It does not show why.

5.2 Dependence of the sensitivity on the flow

Where does the sensitivity error come from? The atmosphere amplifies different states very differently, and a first clue is whether the emulator notices the difference. We compute the leading six-hour gain exactly, by automatic differentiation, for the truth and for the convolutional

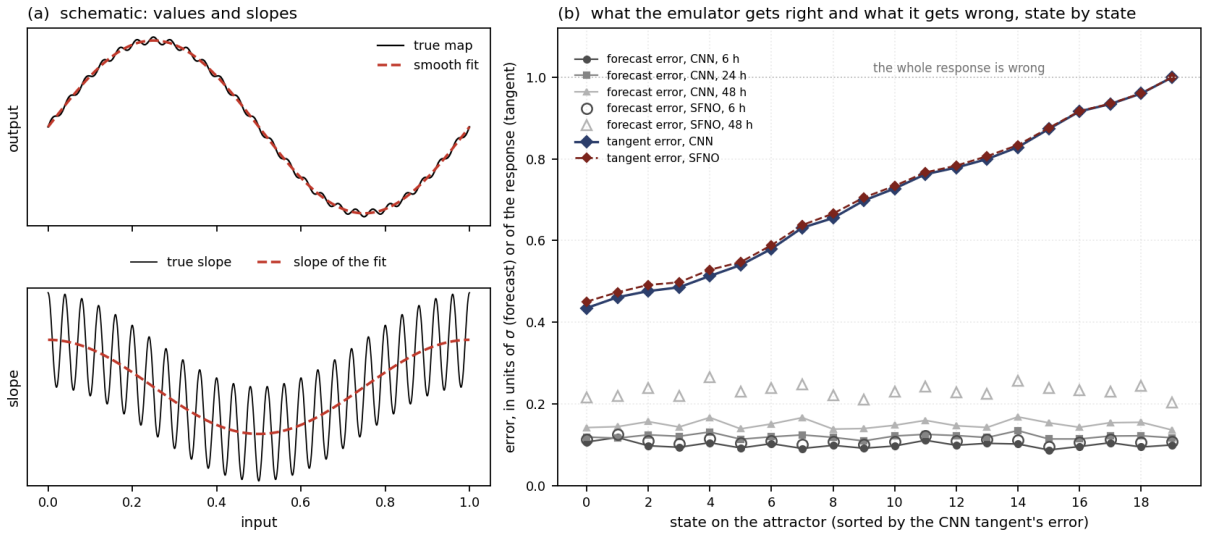

Figure 5: (a) Schematic drawn with synthetic curves: a map with a small, fast component and a smooth fit to it (top), and the slopes of the two (bottom). The values agree, the slopes do not. (b) The same disagreement, measured, state by state, at 20 states spread over the attractor. Grey: each emulator's own forecast error at leads of 6, 24 and 48 hours, in units of the climatological standard deviation, filled for CNN and open for SFNO. Diamonds: the error of the same emulator's exact six-hour tangent, as a fraction of the truth's response to a balanced perturbation, a median over 12 probes per state. States are sorted by the latter for CNN.

emulator CNN at 20 states spread over the attractor (Figure 6). The truth's exact gain ranges from 1.33 at the most quiescent state to 20.0 at the most explosive one, a factor of fifteen, with a median of 2.5. CNN's lies between 1.02 and 1.07 at every state, with a median of 1.04, and SFNO is equally flat, median 1.04 over 1.02 to 1.07.

The one percent of variation the emulator does have is not random: its gain and the truth's are ordered alike over the twenty states, with a rank correlation of +0.40 for CNN and +0.52 for SFNO, visible in the inset of Figure 6. The network notices which flows amplify; it does not amplify them. The emulator does respond to perturbations. Its tangent changes a perturbation by about 30 percent of the perturbation's size, turning and reshaping it, but it leaves its size almost unchanged, with a gain of 0.98 along those directions, against the leading gain of 1.04 along the one that grows fastest. It reshapes perturbations but does not amplify them, and it treats every flow alike.

Figure 6 shows that the emulator's gain is weak and nearly the same at every flow, where the truth's varies by a factor of fifteen over the same twenty states. It does not show what removes the flow dependence.

5.3 One way to lose sensitivity: averaging

Why would a network trained on exact data lose precisely the flow dependence? Consider the training set as a cloud of atmospheres. A network of finite size cannot place an atmosphere in this cloud more precisely than some resolution. Every atmosphere lying within that resolution of a given one therefore receives the same forecast, and the forecast that minimizes the squared error over all of them is the average of their true six-hour changes. States that cannot be told apart are a familiar object in state estimation, where the states of a chaotic system that noisy observations cannot separate are studied as indistinguishable states [Judd and Smith, 2001, 2004]. There it is the observations that fail to separate the states, whereas here it is the emulator that fails, while the truth separates them exactly.

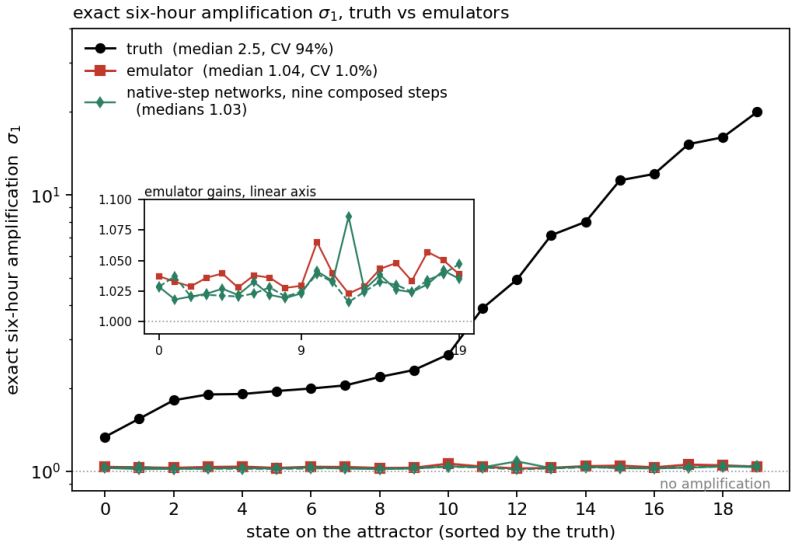

Figure 6: Exact leading six-hour gain σ_1 over 12 balanced perturbation directions at 20 states on the attractor, for the truth (black) and the convolutional emulator CNN (red), with the states sorted by the truth's value and a logarithmic vertical axis. Green: the networks trained at the model's own forty-minute step on both leapfrog levels and tuned through rollouts, whose six-hour gain is the composition of nine of their own steps, at the same states and in the same directions (Section 5.8). The inset shows the three emulator series on a linear axis, where the truth is far above the top.

Two consequences follow at once. The first is that the emulator's response to a small perturbation becomes the average of the responses of all the atmospheres it cannot separate. The second is that this average is both small and blind, small because each of those atmospheres amplifies a perturbation in its own direction so that the responses cancel, and blind because it carries no information about which of them is the actual atmosphere. One mechanism thus gives both symptoms of Section 5.2, weak growth and blindness to the flow. Figure 5a is the same statement in one dimension, where the fit follows the smooth part of a map and loses the slope of its fast component. It is worth noting here that our networks predict an increment and add it to the state,

$$M_\theta(x) \;=\; x + \widehat{\Delta}_\theta(x), \tag{16}$$

so that an increment responding weakly leaves a gain near one rather than near zero. A perturbation that the network does not resolve passes through the first term unchanged.

This explanation turns on four scales, all of them measured in units of σ and all of them measurable in the twin. The first is e, the emulator's one-step error of Equation 10, which is about a tenth of σ. The second is $s^\star$, the displacement of an atmosphere whose six-hour consequence equals e. A difference between two atmospheres smaller than $s^\star$ changes their futures by less than the fit's own error, so the data cannot teach the network to separate them, and $s^\star$ is a lower bound on the resolution. The third is Δ, the spacing of the training atmospheres. Between two of them the network can only interpolate, so the data cannot pin the map more finely than Δ either, and the resolution is at least the larger of $s^\star$ and Δ. The fourth is w, the width over which the truth's own response to a perturbation changes direction, that is, the distance one must move the atmosphere before the butterfly it grows is a different one. If w is smaller than the resolution, the responses that are averaged already disagree with each other, and the average loses the growth. This is an explanation rather than a theory, in the sense that it names the quantities involved and the order in which they must stand. The subsections that follow measure them one by one, ask directly whether the emulator responds

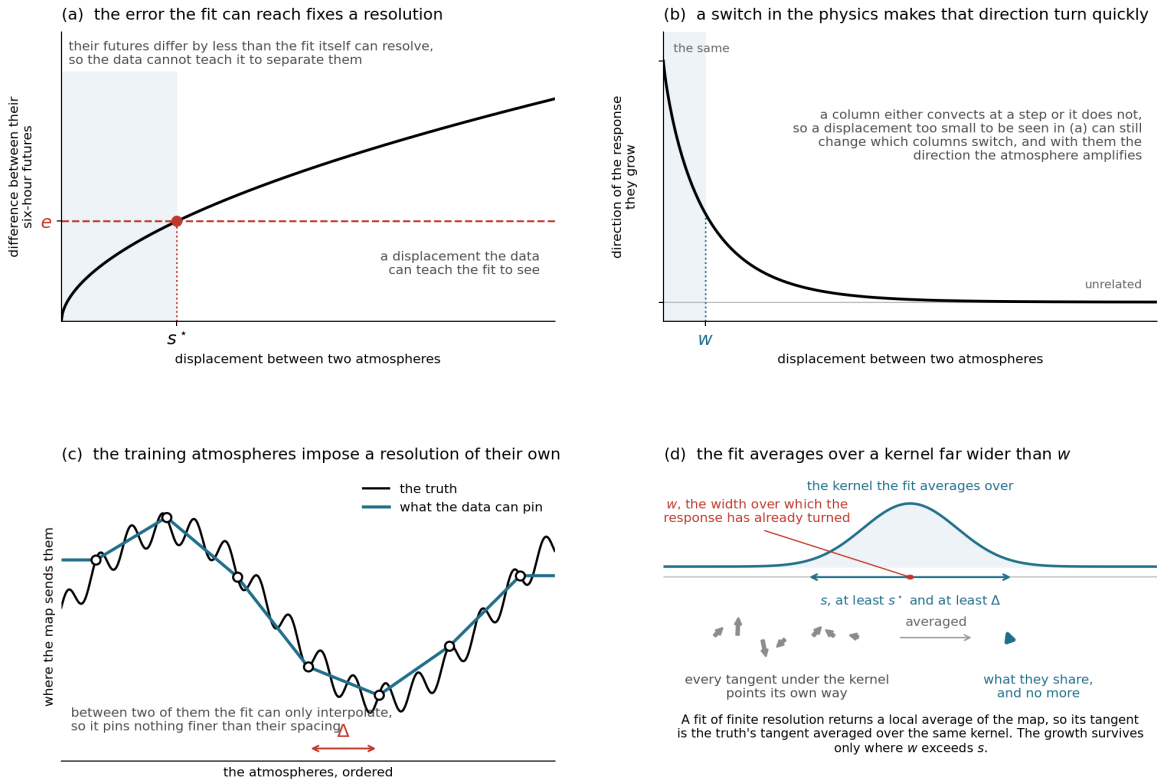

Figure 7: The conceptual explanation for the loss of the butterfly. (a) Two atmospheres whose six-hour futures differ by less than the error e that the fit itself can reach cannot be separated by the data, which fixes a resolution $s^\star$. (b) The direction of the response the two atmospheres grow agrees only over a width w, because a column either convects at a step or it does not, so a displacement far too small to register in (a) can still change which columns switch. (c) The training atmospheres lie a distance Δ apart, and between two of them the fit can only interpolate, so the data pin nothing finer than Δ whatever the capacity. (d) A squared-error fit of finite resolution therefore returns a local average of the map over a kernel whose width s is at least $s^\star$ and at least Δ, and its tangent is the truth's tangent averaged over that kernel. Where w is smaller than s the tangents under the kernel have already turned, and their average keeps only what they share.

like the average, and end with the order the four scales are found to stand in. Figure 7 sets the four out and shows how they decide the question.

In symbols, the explanation reads as follows. A squared-error fit of finite resolution returns not the map but a local average of it,

$$M_\theta(x) \;\approx\; \langle M_\star \rangle_s(x) \;=\; \int K_s(x' - x)\, M_\star(x')\, dx', \tag{17}$$

over a kernel K_s whose width s is the resolution. That width is at least $s^\star$, because atmospheres whose futures differ by less than the error e cannot be separated by the data, and at least Δ, because between training atmospheres the fit interpolates. The derivative of a local average is the local average of the derivative, so the emulator's tangent is the truth's tangent averaged over the same kernel, $J_\theta(x) \approx \langle J_\star \rangle_s(x)$. Whether this average keeps the tangent is decided by w against s. If $w > s$, every tangent under the kernel points the same way, $\langle J_\star \rangle_s \approx J_\star(x)$, and the fit keeps the growth. If $w < s$, the kernel spans tangents whose leading responses point in

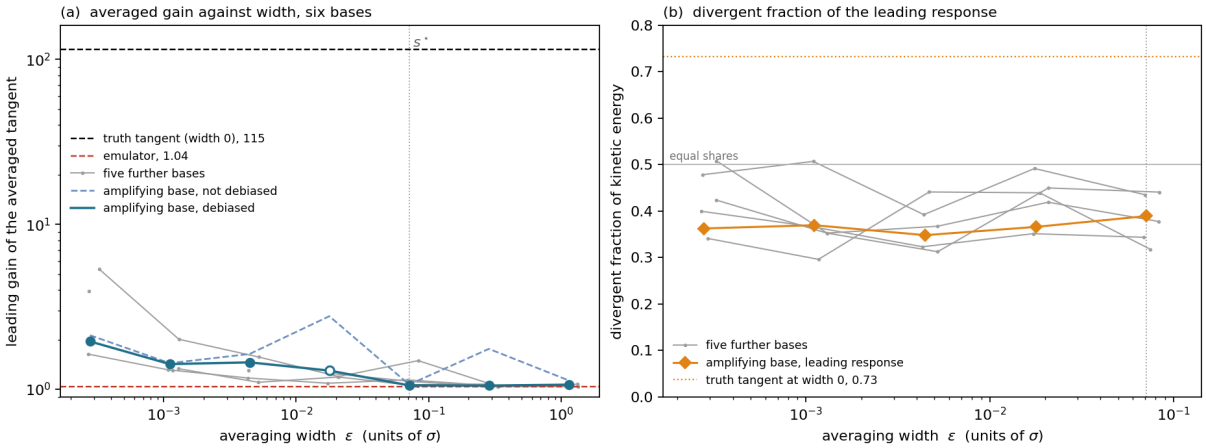

Figure 8: (a) Leading gain of the truth's tangent averaged over nearby atmospheres, against the averaging width ϵ in units of σ. The dotted line marks $s^\star$, the width at which the truth's six-hour response to the displacement equals the emulator's one-step error, at the amplifying state. At that state, whose own gain is 115, points are the leading gain of the mean of 384 displaced atmospheres. Grey lines: five further base states, whose own gains run from 1.3 to 9.0, drawn only at the widths at which they are read. Dashed lines: the truth's gain at width zero at the amplifying state, and the emulator's. (b) Divergent fraction of the leading response of the averaged tangent against the same widths, at the same six states. Dotted line: the truth's tangent at width zero at the amplifying state. Grey line: equal divergent and rotational shares.

unrelated directions. Their mean keeps what they share and cancels the rest, so the gain falls toward that of the shared part, and the flow dependence, which lived in the cancelled part, is lost. The condition for losing the tangent is thus $w < s$ with $s \geq \max(s^\star, \Delta)$.

5.4 What averaging does to the gain

How wide is w? It is measured on the truth itself (Figure 8). At an amplifying state, chosen for its gain of 115, which is stronger than any of the twenty states of Figure 6, we compute the truth's exact tangent at many nearby atmospheres and average these tangents. Each nearby atmosphere is the base state displaced by its difference from a twin trajectory that started three days earlier with a tiny perturbation, rescaled to a chosen width. It is therefore a genuine state of the model, which shares the base state's large scales and differs from it in small-scale detail.

The quantity being measured is the tangent of the averaged map and its response to a perturbation h,

$$\bar{J} \;=\; \mathbb{E}_\eta\big[J_\star(x+\eta)\big], \qquad \bar{r} \;=\; \bar{J}h \;=\; \mathbb{E}_\eta\big[r_\eta\big], \qquad r_\eta \;=\; J_\star(x+\eta)\,h, \tag{18}$$

where the expectation is over the displacement η alone, that is, over which of the atmospheres that the fit cannot separate the state happens to be, and the gain is the size of $\bar{r}$ relative to h, maximized over perturbations. What is reported is therefore the length of an expectation, and the length is a quadratic function of it. Writing $\bar{r}_N$ for the mean of N drawn displacements, $\bar{r}_N = \mathbb{E}\,r + z$ with $\mathbb{E}\,z = 0$, so that

$$\mathbb{E}\|\bar{r}_N\|^2 \;=\; \|\mathbb{E}\,r\|^2 \;+\; \frac{\mathbb{E}\|r - \mathbb{E}\,r\|^2}{N}. \tag{19}$$

The sample mean is unbiased for the mean response, but its length is not unbiased for the length of the mean response.

To remove the bias, we draw 384 atmospheres at each width and split them into two halves of 192, averaging each separately into $\bar{r}_A$ and $\bar{r}_B$, whose fluctuations z_A and z_B are independent

and of mean zero. Then

$$\mathbb{E}\langle \bar{r}_A, \bar{r}_B \rangle \;=\; \|\mathbb{E}\, r\|^2 + \mathbb{E}\langle z_A, z_B \rangle \;=\; \|\mathbb{E}\, r\|^2 \tag{20}$$

exactly, because the mean of the product of two independent fluctuations is zero. Nothing is assumed about the distribution of the responses beyond a finite second moment, and no variance is estimated. The estimator is also a correction of the naive one, since for the mean $\bar{r}$ of all 384,

$$\langle \bar{r}_A, \bar{r}_B \rangle \;=\; \|\bar{r}\|^2 \;-\; \tfrac{1}{4}\|\bar{r}_A - \bar{r}_B\|^2, \tag{21}$$

so the subtracted term measures the bias of Equation 19 directly. We report $\mathbb{E}\langle \bar{r}_A, \bar{r}_B \rangle$ only where it stays below half of $\|\bar{r}\|^2$, and leave it unread otherwise. For the leading gain the same subtraction is applied to every entry of the matrix of responses to the eight perturbations, the gain being its largest singular value over their span.

At the amplifying state the averaged gain falls from 115 to 1.95 at the smallest width, 0.0003σ, to 1.06 at $s^\star$, and to between 1.05 and 1.07 at the two widest displacements, 0.28σ and 1.1σ (Figure 8a). Over six states whose own gains run from 1.3 to 115 it lies between 1.6 and 5.4 at the smallest width, between 1.05 and 1.5 at $s^\star$, and between 1.04 and 1.08 at the two widest displacements, against the emulator's own 1.01 to 1.04, and it no longer follows the states' own gains. The truth's flow dependence, a factor of 90, is replaced by one of a few percent. Averaging also changes the structure of the response. The leading response of the truth's tangent at the base state leans toward divergent flow, with a divergent fraction of 0.73, whereas after averaging it is rotational, 0.39 at $s^\star$ (Figure 8b). Where does the collapse begin? The experiment cannot say, because the collapse has already happened at the smallest width that can be tested. It therefore bounds w from above, at 0.0003σ, rather than locating it.

Why is 0.0003σ the smallest width that can be tested? This limit comes from the arithmetic rather than from the design of the experiment. The model runs in single precision, which carries a field to about seven decimal digits, and its six-hour map amplifies a change at that level by a large factor. A balanced perturbation of $10^{-6}\sigma$, constructed as in Section 3 but at the size of the rounding of the state, grows to between $2 \times 10^{-4}\sigma$ and $3 \times 10^{-4}\sigma$ within a single six-hour step, a factor of about 200. Two atmospheres closer together than this are no longer two atmospheres as far as the six-hour map is concerned. Their futures are separated by the rounding that the integration performs as much as by their displacement, and the tangents of the single-precision program at the two states differ by which rounding-sensitive branches the arithmetic takes rather than by which atmosphere it is given. An average over widths below 0.0003σ would therefore be an average over the arithmetic and not over atmospheres. The smallest width we test is thus the smallest at which the displaced atmospheres are still different states of the model, and the bound on w is set by the precision of the model rather than by the flow.

The moist physics explains why so small a displacement is enough. Convection is a switch, in the sense that a column either convects at a given model step or it does not. Two atmospheres that differ only slightly can therefore differ in which columns switch, and the tangent of a map containing such switches changes abruptly from one atmosphere to the next. We measured how often this happens. Integrating the base atmosphere and 32 displaced ones for six hours, and recording at every column and every step whether the column convects, we find that convection switches on or off in 2.4 percent of the column-steps that convect in either atmosphere at the smallest width, 0.00027σ, and in 32 percent at 0.069σ. Even the smallest displacement that can be applied already flips one column-step in forty. The fast component of the truth's map, the schematic of Figure 5a, is therefore narrower than three ten-thousandths of σ, and it is convection switching that makes it.

5.5 Whether the emulator responds like the average

Figure 8 shows that averaging the truth's own tangent over nearby atmospheres produces a deficit of the kind seen in the emulator. It does not show that the trained network performs this average. We therefore ask the question directly: does the emulator respond to a perturbation the way the average does, or the way the actual atmosphere does? At each of the six base states, for eight balanced perturbations h, we compare three responses, all of them computed exactly: the truth's at the base, $J_\star h$, the truth's averaged over the 384 displaced atmospheres at each width, $\bar{J}h$, and the emulator's, $J_\theta h$. By Equation 16 each response is the perturbation itself carried along plus the response of the increment, and the first part is common to all three. We therefore compare the three with the component along h removed, which leaves the part that the dynamics add to the perturbation.

Two angles are involved here and they answer different questions. Write a response as the part along the perturbation and the part across it,

$$r \;=\; \frac{\langle r, h\rangle}{\|h\|^2}\, h \;+\; r_\perp, \qquad \cos\angle(r,h) \;=\; \frac{\langle r, h\rangle}{\|r\|\,\|h\|}, \qquad \|r_\perp\| \;=\; \|r\|\,\sin\angle(r,h), \tag{22}$$

so that $r_\perp$ is what the dynamics add and $\angle(r,h)$ is how far the response has turned away from the perturbation it was given. The first question is whether the emulator and the average add the same thing, which is the cosine between the two added parts,

$$\cos\angle\big(r_{\theta\perp},\, \bar{r}_\perp\big), \tag{23}$$

and the second is how far each has turned, which sets $\angle(r_\theta, h)$ against $\angle(\bar{r}, h)$. The two are independent of one another. Responses that have turned by the same amount may add structure pointing in quite different directions, and responses whose added structure points the same way may have turned by different amounts. The last equality of Equation 22 ties the second question to the sizes: among responses of equal length, the one that has turned further adds the larger part, in the ratio of the sines. Figure 9a shows the cosine of Equation 23 against the width, with the cosine against the truth at the base as the reference line, and Figure 9b the ratio $\|\bar{r}_\perp\|/\|r_{\theta\perp}\|$ of the two added parts.

At the amplifying state the answer is unambiguous. The emulator's response there is nearly orthogonal to the truth's own response at that atmosphere, with a cosine of 0.01, and it points along the averaged response, with a cosine of 0.55 to 0.74 at every width from $s^\star$ to $16s^\star$, which is about one σ. In other words, the emulator responds like an average of nearby atmospheres and not like the atmosphere it was given. At the other five states, whose own gains are closer to the emulator's, the cosine against the truth at the base is 0.1 to 0.5, while the cosine against the average is 0.5 to 0.8 at the two widest displacements, where the sampling of the average permits at most 0.6 to 1.0. A cosine of 0.5 to 0.8 is not close to one, and it is not against one that it should be read. These responses are vectors of 152,064 components, one for every variable at every point of the grid, and in a space of dimension n that large, two directions drawn at random are almost perpendicular: their cosine has mean zero and standard deviation $n^{-1/2}$, which is about three thousandths here, so a cosine of 0.5 lies some two hundred standard deviations from chance.

Two features of these curves deserve attention. The first is that above a quarter of $s^\star$ the cosine stops rising with the width, because the averaged response itself hardly changes direction over that range once w has been passed. The measurement therefore identifies the direction of the emulator's response with that of an average, but it does not identify the width of the average.

The second feature is that the two responses have the same size and differ in how far they turn the perturbation. Taken whole, the averaged response and the emulator's agree to within five percent at a displacement of one σ. What differs is the share of the response that is turned

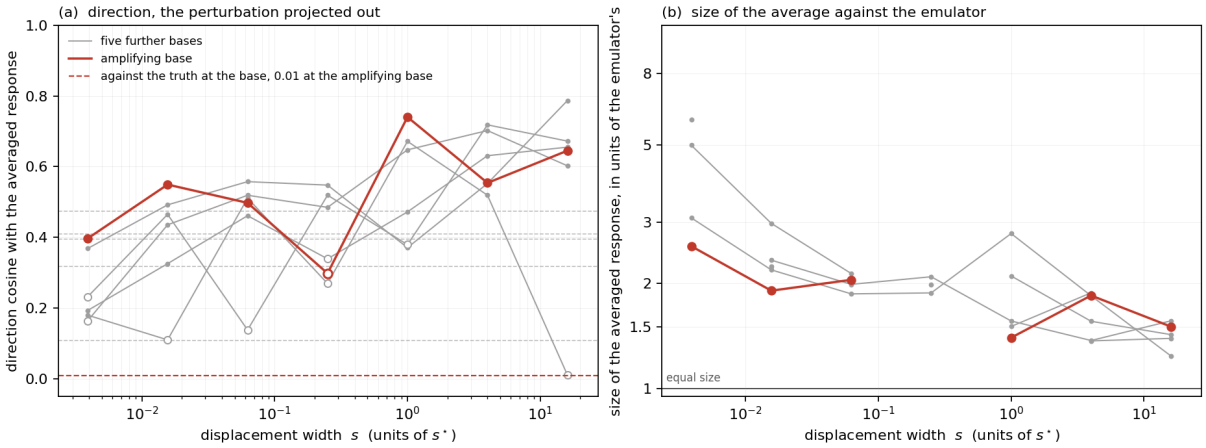

Figure 9: Does the emulator respond like the average? (a) Direction cosine between the spherical emulator's exact response to a balanced perturbation and the truth's response averaged over 384 atmospheres displaced from the base state by the width on the horizontal axis, with the component along the perturbation removed from both responses. Medians over eight perturbations at each of the six base states of Figure 8, the amplifying state in red and the five others in grey. Dashed lines: the cosine between the emulator's response and the truth's own response at the base. Open symbols: the widths at which the two halves of the ensemble agree too poorly for the average to be matched. (b) The size of the averaged response as a multiple of the emulator's, the component along the perturbation again removed from both, at the same states and widths, with those poorly sampled widths omitted. The line marks equal size.

away from the perturbation: the emulator's response keeps a cosine of 0.93 to 0.95 with the perturbation it was given, where the average keeps 0.86 to 0.90, so the part the dynamics add is larger in the average at every width from $s^\star$ to one σ (Figure 9b). These two cosines answer the two questions of Equation 22 and are not to be read against each other. The 0.5 to 0.8 above is the agreement between the parts the two models add, while 0.94 against 0.88 is how far each whole response has turned from the perturbation. By the last equality of Equation 22, responses of the same length that turn by those two amounts differ by a factor of 1.4 in the part they add, which is the measured range. Averaging over genuine atmospheres therefore reproduces the direction of the emulator's response, its loss of flow dependence and its size. What it does not reproduce is how little the network turns the perturbation, and that remainder is not explained here.

5.6 The signature an average leaves in the error

An average also leaves a signature in the error, which we check last. The error of an average is uncorrelated with its own prediction, band by band in wavenumber, so that the fraction R^2 of the true increment's second moment that the emulator reproduces and the fraction Q^2 left in its error add up to one,

$$R(\ell)^2 + Q(\ell)^2 \;=\; 1. \tag{24}$$

Writing P_ℓ for the projection onto total wavenumber band ℓ, and $\Delta_\theta = M_\theta(x) - x$ and $\Delta_\star = M_\star(x) - x$ for the predicted and the true increment over one step,

$$R(\ell) \;=\; \frac{\big(\mathbb{E}\|P_\ell \Delta_\theta\|^2\big)^{1/2}}{\big(\mathbb{E}\|P_\ell \Delta_\star\|^2\big)^{1/2}}, \qquad Q(\ell) \;=\; \frac{\big(\mathbb{E}\|P_\ell (\Delta_\theta - \Delta_\star)\|^2\big)^{1/2}}{\big(\mathbb{E}\|P_\ell \Delta_\star\|^2\big)^{1/2}}, \tag{25}$$

these are the size of what the emulator predicts and the size of its error, each as a fraction of the size of what the truth does in that band. Equation 24 is Pythagoras: it holds exactly when

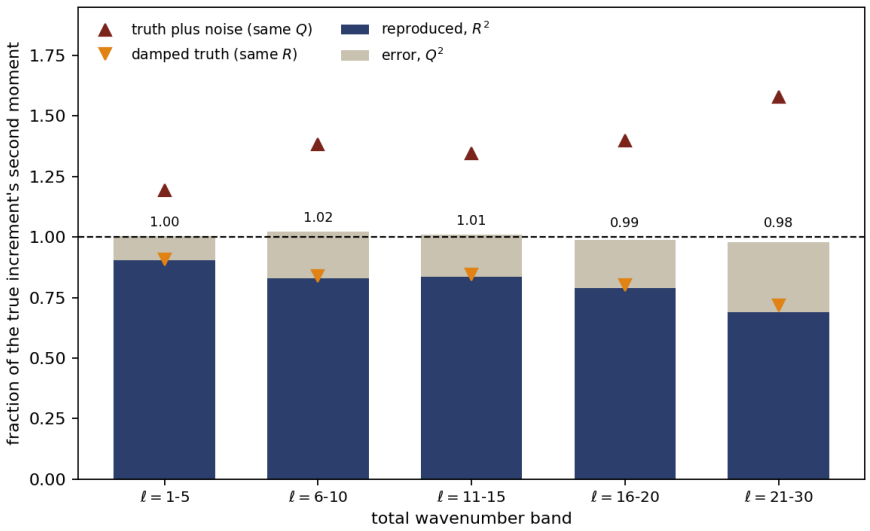

Figure 10: In each total-wavenumber band, the fraction R^2 of the true increment's second moment that the spherical emulator reproduces (dark) and the fraction Q^2 that lies in its error (light), stacked, with R^2+Q^2 printed above each bar. A conditional mean satisfies $R^2+Q^2=1$. Triangles: the same quantity for the truth plus independent noise (pointing up), and for a damped copy of the truth with the measured band gains (pointing down).

the error is uncorrelated with the prediction, which is the defining property of a conditional mean. Both quantities are averages over 256 states of one emulator, so that R and Q belong to a single emulator and a single band. Figure 10 shows the measurement for the spherical emulator. Measured in the same way for each of the six emulators of Section 5.8 separately, with two architectures and three constructions, the sum departs from one by at most 0.013 to 0.027 in any band. The two simplest alternatives fail this test. An emulator that were the truth plus independent noise would give a sum between 1.19 at the largest scales and 1.58 at the smallest, and a damped copy of the truth between 0.91 at the largest scales and 0.72 at the smallest. It is worth emphasizing that the signature is necessary and not sufficient, since ordinary least-squares regression leaves it whatever the underlying mechanism (Appendix A gives an example). This is why it comes last, after the direct comparison of Figure 9.

5.7 What the scales decide

Three measurements of the average are now in hand, and what remains is to check that the four scales themselves stand in the order the explanation requires. They do, and with room to spare. The three scales $e, s^\star, \Delta$ stand far above w. The one-step error e is 0.11σ for the spherical emulator, and the width $s^\star$ is 0.07σ at the amplifying state, lying between 0.069σ and 0.083σ over the six states of Figure 8. The training atmospheres lie 0.71σ apart in the same units, ten times $s^\star$. This spacing is measured as the distance from each of the 29200 states of the free run to its nearest neighbour that is not its neighbour in time since those measure the time step rather than the cloud. The order that the explanation requires is therefore met by a wide margin, $w < 0.0003\sigma \ll s^\star = 0.07\sigma < \Delta$. The atmospheres that a squared-error fit of this accuracy cannot separate differ by more than two hundred times the width over which the truth's response has already changed direction.

Three measurements therefore agree. First, averaging the truth's tangent over atmospheres closer together than the fit can separate reproduces the deficit's kind, a weak gain of 1.04 to 1.08 above $s^\star$ across six states whose own gains run from 1.3 to 115, which no longer follows the flow and is rotational in structure. Second, the emulator's own response points along that average and away from the atmosphere's own response, most sharply where the atmosphere

amplifies most. Third, the emulator's error carries the signature that an average leaves. This is what makes unresolved detail the leading explanation.

What has not been shown should be stated just as plainly. The averaged tangent is a measuring device and not a model of the network (Appendix A). The width of the average that the network performs is not identified, but only bounded from below by $s^\star$ and Δ. And the network turns a perturbation less than any average of genuine atmospheres we formed, so that the explanation accounts for the direction of its response, for its blindness to the flow and for its gain, but not for how much of its response lies along the perturbation.

What would it take to resolve the detail? Two ways are conceivable, one through capacity and one through data, and neither of them is short. The universal approximation theorem [Cybenko, 1989, Hornik et al., 1989] states that a network of sufficient size can approximate any continuous map on a bounded set to any accuracy, and its extension to derivatives [Hornik et al., 1990] states the same of the tangent, so that nothing in principle forbids an error e small enough for $s^\star$ to fall below w. The theorem is, however, an existence statement that names no size, and the size it leaves open is large. Bringing $s^\star$ below w requires a one-step error smaller than the truth's six-hour response to a displacement of w, which is a small fraction of the present 0.11σ, a factor of 200 if the response grew in proportion to the displacement. Our two architectures, built differently, reached one-step errors within a tenth of each other.

The second way is more demanding still. The nearest-neighbour spacing of n atmospheres on an attractor of dimension d shrinks only as $n^{-1/d}$, which is why 29200 atmospheres leave a nearest neighbour only a third closer than a random one. To bring Δ from 0.71σ to 0.0003σ, a factor of 2400, needs 2400^d times as many atmospheres, and even if d were ten that is 10^{34} times twenty years of six-hourly states. Capacity cannot substitute for data here, because atmospheres at spacing Δ pin nothing finer than Δ. What the four widths say is that no affordable increase in capacity or data moves the resolution to where the truth's response is already decided.

5.8 Other explanations

Six other explanations of the weak, flow-independent gain come naturally to mind, and each predicts something the measurements contradict.

Damping. The simplest explanation of weak growth is that the emulator is a blurred version of the truth. Suppose it were exactly that: suppose it made the truth's own forecast change and then shrank it, scale by scale, keeping at least 0.83 of it, which is the most shrinkage we measure in any band. Such a model would still inherit the truth's sensitivity to the weather, because it would still be built from the truth's own response. It would amplify by at least 16.4 at the most explosive of our twenty states and by at least 1.9 at the median one (Appendix A derives the bound). The emulator instead amplifies by 1.04 at every state, between 1.02 and 1.07. It is therefore not a shrunken copy of the truth; it is a different map. The spectral objective of Section 4.2 is the direct test of this explanation, since it puts back exactly the energy that shrinking removes, and it made the growth smaller rather than larger.

Accuracy and capacity. A more accurate network should carry a better tangent. Within the accuracy we can reach, it does not: the convolutional emulator's one-step error is nine percent lower than SFNO's, and its exact gain is the same, 1.04. This is what the widths of Section 5.3 lead one to expect. What decides whether the tangent survives is not a nine percent difference in the error but whether $s^\star$ falls below w, which asks for a factor of about 200, and short of that factor a more accurate average is still an average.

Construction. On SPEEDY, a spherical network trained as a tendency advanced by a Runge-Kutta step, and a map applied twice per step, reach the same validation loss as the direct map,

satisfy Equation 24 equally well, and have the same gain as the direct map. Changing how the map is built did not change what the network resolves.

Information hidden by the time stepping. SPEEDY's state is a pair of leapfrog time levels, one native step apart. The truth advances each step from both, while the emulator sees only the leading one (Section 2.2). The two levels differ by one native step of evolution, so the level the emulator does not see could, in principle, carry information that the network then averages over. To test this, we trained SFNO, with the same recipe, on both levels. Its one-step error is lower by only 4.5 percent, and its gain is 1.027 at the median state and varies by 0.8 percent across the 20 states. The weak, flow-independent gain survives when nothing is hidden by the time stepping.

Coarse-graining of the training data in time. Hassanzadeh et al. [2026] attribute the missing growth to the coarse-graining of the training data, and show in a Lorenz system and in Pangu-Weather that a shorter step brings faster, amplitude-dependent growth and lower forecast skill together. The twin can remove that coarse-graining entirely. Trained at the model's own forty-minute step on both leapfrog levels, on the same number of pairs, the spherical network fits its increment as well as the six-hour network fits its own, a normalised validation loss of 0.22 against 0.21, and it reproduces what they report: growth where the truth's is, an optimal-growth correlation of -0.15 against the truth's 0.00, at the truth's scale, a centroid of 17.1 against 17.4, and divergent early and late, 0.74 and 0.49 against 0.68 and 0.51. It is still not a butterfly, and no longer a forecast model. The growth does not move upscale, the centroid going to 17.3 where the truth's runs to 14.5; it reaches a quarter of saturation in three to seven days against the truth's 12 to 20, and then overshoots to 2000 times the climatological level where the truth sits at 0.98; its six-hour error is one and a half times the six-hour network's, with little skill left by one day; and its composed gain varies across the twenty flows by 27 to 62 percent without following the truth's. Tuned through rollouts as the baseline is, the same network becomes the six-hour emulator's twin in skill, in gain and in structure: within three percent at six hours, a composed gain of 1.03 with about one percent of spread across the flows (the green series of Figure 6), an optimal-growth correlation of $+0.92$, a divergent share of 0.32 and no cascade. A shorter step therefore admits small-scale divergent growth of the right kind and in the right places, but not a butterfly, and training the network to forecast removes it. Two checks bound this: at 12 days the one-step network has already left the attractor, its unperturbed error standing at 8 to 10 times the climatological spread, so what grows there is the model and not the perturbation. And trained on reversed pairs it backcasts within 6 to 14 percent of its forward error to 48 hours, so the growth is not an artefact of a one-sided time stencil.

Growth along other directions. The missing growth is not available elsewhere in the six-hour tangent either. Along small-scale, purely divergent directions, isolated from balanced perturbations by a high-pass filter and a projection onto the divergent wind, the truth's exact gain is 1.2, with a median of 1.1 and a maximum of 3.5 over four states and eight directions, and the emulator's is 0.85. A perturbation of finite size along the same directions responds strongly, but that response is a gravity-wave adjustment that vanishes in the linear limit rather than a growth that a tangent loss could learn.

5.9 Whether this is a property of one twin

The twin's emulators are trained on one model at one modest resolution, so we measured the same quantity on an operational emulator that no part of this study shaped. GraphCast at one degree, differentiated exactly, has a six-hour gain over balanced differences of reanalysis states

of 1.37 in the median over twenty dates through a year, with a spread of five percent across those dates. GraphCast also carries vertical velocity and precipitation, which the twin's state does not, so we measure it as the twin's emulators are measured, with the same balanced perturbation and both norms restricted to the variables the twin carries, the winds, temperature, humidity and surface pressure: the gain is 1.16, with a spread of 2 percent. We draw one thing from this and not another. The spread is the point: in those variables an operational emulator's exact tangent varies across flows by two percent, where the twin's emulator varies by one percent and the twin's truth over the same kind of probe by 94 percent. The level is not comparable with the twin's truth and we do not compare it: there is no tangent of the real atmosphere to set beside it, a gain of 1.16 over six hours is a doubling time of 28 hours and one of 1.37 is 13 hours, neither of them outside the error-doubling times reported for the atmosphere at these scales, so the level by itself shows no deficit, and the perturbations are differences of analyses, which are atmospheres but are not states of GraphCast's own attractor. That an operational emulator underestimates the growth of small perturbations is the measurement of Selz and Craig [2023] and Selz and Craig [2026]. What has not been reported before, and what the exact tangent adds, is how little that growth depends on the weather it is measured in.

6 Conclusion

Machine-learning emulators forecast the weather well but do not grow small perturbations the way the atmosphere does. In this study, we have asked why, in a setting where the question can be answered cleanly: a differentiable twin of SPEEDY, in which the emulator is trained on exact data from the model it imitates and the true tangent linear operator is known exactly. We have shown the following.

First, the deficit is large once perturbations are constructed correctly. An unbalanced perturbation excites a gravity-wave shock that is easily mistaken for growth, grid-point noise being amplified 115-fold in one model step, whereas a balanced perturbation, a scaled difference of two states on the attractor, grows steadily from the start. Measured with balanced perturbations, the emulator's growth of small perturbations over 12 days falls short of the truth's by a factor of about six hundred. Its exact six-hour gain is 1.04 and nearly the same at every flow, where the truth's ranges from 1.3 to 20.

Second, no training objective we tested restores the physical butterfly. A spectral loss restores the right energy at every scale and grows less than the baseline. A stochastic head adds spread in proportion to the noise it injects. Matching the growth curve either runs away into a numerical instability or grows far too little. Tangent matching comes closest on the 12-day growth, although it overshoots saturation as the direct growth penalty does, and the growth it installs is rotational, large scale, and cascades downscale. It is the optimal growth that singular vectors describe rather than the divergent, small-scale butterfly, and it nearly doubles the one-step error.

Third, forecast accuracy does not imply sensitivity accuracy, and for these emulators the two differ sharply. The one-step error is about a tenth of the climatological spread, whereas the error of its exact tangent is 0.44 to 1.00 of the truth's response, with a median of 0.71, and its gain is nearly the same at every flow. The evidence points to one explanation: a network cannot distinguish atmospheres that share their large scales and differ in unresolved detail, the prediction that minimizes the squared error for all of them is the average of their futures, and the growth of such an average is weak because each of these atmospheres amplifies a perturbation in its own direction. Averaging the truth's own tangent over nearby atmospheres reproduces a deficit of this kind at every width tested, from three ten-thousandths of the climatological spread: across six base states whose own gains run from 1.3 to 115, the averaged gain falls to between 1.04 and 1.08 above $s^\star$, within a few percent of the emulator's 1.01 to 1.04, no longer follows the flow, and turns rotational. The emulators carry the band-wise signature an average

leaves, to within three percent in every band across two architectures and three constructions of the map. Asked directly, and with the perturbation itself removed from both responses, the emulator's response points along the averaged response rather than along the response of the atmosphere it was given, with a cosine of 0.5 to 0.8 at the wider displacements at every state the ensemble is able to read, and of 0.01 against the truth's own response at the most amplifying state. Its size is reproduced, and what is not is the share of the response the network leaves along the perturbation: the average turns the perturbation a quarter to a half more than the emulator does at a displacement of one climatological spread, so that averaging accounts for the direction of the deficit, for its blindness to the flow and for its size, but not for that share. Beyond the twin, GraphCast's exact six-hour gain in the same variables varies by two percent across twenty dates of reanalysis, against the twin emulator's one percent and 94 percent for the twin's truth. The same averaging also explains why tangent matching installs rotational growth, since a squared-error loss on the tangent is pulled toward the average of the true tangents, and that average is rotational.

It is worth emphasizing what this averaging is and what it is not. It is not an averaging over errors in the training data, the explanation proposed by Selz and Craig [2026]. Our data are exact, and the butterfly is lost all the same. Nor does noise in the training targets create it, since noise that is independent of the state leaves the average unchanged. The averaging is over what the network does not resolve. Better training data, even a nature run, did not restore the butterfly here, and the averaging picture explains why a more accurate average would not either: it remains an average.

These conclusions have limits that should be kept in view. They are established for one model at one modest resolution, T30 with eight levels. The averaged tangent is a measuring device, not a model of the trained network (Appendix A), and the measurements do not prove that the networks are exactly averages, nor that supervising derivatives cannot, in principle, recover the response. What they establish is that none of the objectives, architectures and constructions we tried escaped the deficit, and that averaging over unresolved detail accounts for its kind: it leaves a gain of 1.04 to 1.08 where the truth's runs from 1.3 to 115, and the remainder is not in the gain but in the direction, since the network turns a perturbation a quarter to a half less than the average does. The truth's tangent is rough on a scale narrower than three ten-thousandths of the climatological spread, so its gains describe infinitesimal perturbations rather than any perturbation of finite size. And the explanation is not a theorem. A network able to resolve the detail that controls the growth, or an atmosphere whose growth did not depend on that detail, would keep its amplification. This should be verified for more models and at higher resolution, where the unresolved detail is smaller but the growth it controls is faster.

What would restore the physical butterfly remains open. Our results point to three directions: resolving more of the detail that controls the growth, representing the unresolved state probabilistically so that it is sampled rather than averaged, and supervising the evolution of perturbations over several days rather than their six-hour linear response. The first of these need not proceed atmosphere by atmosphere, and this is the one encouraging thing the widths say. What the atmospheres a network cannot separate have in common is the rule that governs them: convection is switched by the same function of the local column in every one of them, and a network whose form can express that function could learn it from data far coarser than the width over which the truth's response changes direction, as a rule rather than atmosphere by atmosphere. Whether a squared-error fit would find that rule is a different question from whether a network could hold it, since the loss rewards the average that the rule's consequences form, and we have not tested it. The second and third directions are the Mori-Zwanzig reading that Hassanzadeh et al. [2026] give of the same deficit, in which what a deterministic one-step map discards is a memory term that a stochastic or a multi-step objective can restore. We will address these problems in the near future.

Open Research Section

The data supporting the findings of this study were generated using the numerical code developed by the authors. The source code is publicly available at: https://github.com/leducvn/dajax.

Conflict of Interest disclosure

The authors declare there are no conflicts of interest for this manuscript.

Acknowledgments

This work was supported by the JST Moonshot R&D program (Grant JMPJMS2281). Le Duc was also supported by the Grants-in-Aid for Scientific Research (KAKENHI) project (Grant 24K07131: A robust ensemble Kalman filter to innovate short-range severe weather prediction).

A The statements behind Section 5

This appendix collects four statements that Section 5 uses in words. Throughout, X is a state drawn from the training distribution, $\Delta = M_\star(X) - X$ its true six-hour increment, $J_\star = DM_\star$, and $Z = \varphi(X)$ the description of the state that the emulator effectively resolves.

A1. What mean squared error returns. A residual emulator $M_\theta(x) = x + \widehat{\Delta}(\varphi(x))$ minimizes $\mathbb{E}\|\widehat{\Delta}(Z) - \Delta\|^2$. Over functions of Z the minimizer is $\widehat{\Delta}^\star(Z) = \mathbb{E}[\Delta \mid Z]$, the average of the increments of all states that share the description Z. Its error $e = \Delta - \widehat{\Delta}^\star(Z)$ satisfies $\mathbb{E}[e \mid Z] = 0$, hence $\mathbb{E}[e^\mathsf{T} g(Z)] = 0$ for every function g. Taking for g the projection of $\widehat{\Delta}^\star$ onto a band of total wavenumber gives Equation 24, an identity of second moments averaged over the data rather than a statement about individual states. The identity is necessary but not sufficient. Any least-squares fit within a class that contains rescaled copies of its own prediction satisfies it. For example, let X be uniform on $[-1, 1]$ and $\Delta = X + X^2 - \frac{1}{3}$. The best linear predictor of Δ from X is X itself, its error $X^2 - \frac{1}{3}$ is uncorrelated with X, and the identity holds, although the conditional mean of Δ given X is Δ and not X. For a trained network, the reduction to some description Z is therefore an effective hypothesis, and Equation 24 is a test that the hypothesis passes, not a proof of it.

A2. The tangent of an average and the average of tangents. Write the states sharing a description z as $T(z, U)$, with U labelling the unresolved detail. Then the tangent of the average is $\mathbb{E}_U[(J_\star(T(z,U)) - I)\, D_z T(z,U)]$, which carries a transport factor $D_z T$ describing how the family of states deforms as z changes. The average used in Section 5.4, $\bar{J} = \mathbb{E}_\eta[J_\star(x + \eta)]$ over a fixed family of displacements η, drops that factor: it is the tangent of the smoothed map $\mathbb{E}_\eta[M_\star(x + \eta)]$, a different object, but one we can compute. A simple example separates the two. For $X = (Z, U)$ with $M_\star(z, u) = u$ and $\mathbb{E}[U \mid Z] = 0$, the conditional mean is identically zero and so is its tangent, whereas $\mathbb{E}[J_\star] = (0 \;\; 1)$ is not. What the two objects share is the structure that matters here: each is an average of responses, whose size is set by how well the responses agree in direction. Section 5.4 measures that structure on the smoothed map and claims nothing more.

A3. The tangent of a damped truth. A damped copy of the truth, $M_G(x) = x + G\,(M_\star(x) - x)$, with G multiplying each wavenumber band by a gain between $g_{\min}$ and one, has

error $(I - G)\,\delta y$, so that $R = g$, $Q = 1 - g$ and $R^2 + Q^2 = 1 - 2g(1 - g) < 1$ in each band. Its tangent is $J_G = (I - G) + GJ_\star$. For the truth's leading direction v, with $\|J_\star v\| = \sigma_1^\star$,

$$\sigma_1(J_G) \;\geq\; \|GJ_\star v\| - \|(I - G)\,v\| \;\geq\; g_{\min}\,\sigma_1^\star - (1 - g_{\min}), \tag{26}$$

which assumes that G is symmetric in the norm used, as it is for band projectors in the area-weighted norm on the sphere and approximately in our grid norm.

A4. What a tangent penalty returns. Treat the tangent that the network can represent as a function $A(\varphi(x))$ of the resolved description, ignoring that it must be the derivative of the learned map, and draw the probe directions v independently of the state with $\mathbb{E}[vv^\mathsf{T}] = I$. Then the penalty of Equation 14 becomes $\mathbb{E}\|A(\varphi(X)) - J_\star(X)\|_F^2$, whose minimizer is $A^\star(z) = \mathbb{E}[J_\star(X) \mid \varphi(X) = z]$, the average of the true tangents over the states that share a description. For probes that depend on the state, as jacBal's leading singular directions do, the minimizer is a weighted version of the same average, $A^\star(z)\,\mathbb{E}[P \mid z] = \mathbb{E}[J_\star P \mid z]$ with $P = \sum_k v_k v_k^\mathsf{T}$. This is a statement about the regression target, not about what a network can realize, and it underlies the explanation of Section 5.3, not a limit on derivative supervision.

References

ICON. https://gitlab.dkrz.de/icon/icon-model, April 2026.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06185-3.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, May 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09005-y.

Massimo Bonavita. On Some Limitations of Current Machine Learning Weather Prediction Models. *Geophysical Research Letters*, 51(12):e2023GL107377, 2024. ISSN 1944-8007. doi: 10.1029/2023GL107377.

Massimo Bonavita and Alan J. Geer. Forecast verification using information and noise. *Quarterly Journal of the Royal Meteorological Society*, 152(776):e70109, 2026. ISSN 1477-870X. doi: 10.1002/qj.70109.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical Fourier Neural Operators: Learning Stable Dynamics on the Sphere. In *International Conference on Machine Learning*, pages 2806–2823. PMLR, July 2023.

Jiakai Chen, Joel Oskarsson, Simon Driscoll, and Sebastian Schemm. Butterfly Effect and the Kinetic Energy Cascade in Probabilistic Machine Learning Weather Prediction Models, September 2026.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, December 1989. ISSN 1435-568X. doi: 10.1007/BF02551274.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev Training for Neural Networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Dale R. Durran and Mark Gingrich. Atmospheric Predictability: Why Butterflies Are Not of Practical Importance. *Journal of the Atmospheric Sciences*, 71(7):2476–2488, July 2014. ISSN 0022-4928, 1520-0469. doi: 10.1175/JAS-D-14-0007.1.

Maximilian Gelbrecht, Alistair White, Sebastian Bathiany, and Niklas Boers. Differentiable programming for Earth system modeling. *Geoscientific Model Development*, 16(11):3123–3135, June 2023. ISSN 1991-959X. doi: 10.5194/gmd-16-3123-2023.

Gregory J. Hakim and Sanjit Masanam. Dynamical Tests of a Deep Learning Weather Prediction Model. *Artificial Intelligence for the Earth Systems*, 3(3), July 2024. ISSN 2769-7525. doi: 10.1175/AIES-D-23-0090.1.

Pedram Hassanzadeh, Weidong Li, Y. Qiang Sun, Jiangdi Wang, Alexander Wikner, Justin Finkel, and Jonathan Q. Weare. Missing the Butterfly and Predicting the Past: Features or Bugs of Accurate AI Weather Models?, August 2026.

Sam Hatfield, Matthew Chantry, Peter Dueben, Philippe Lopez, Alan Geer, and Tim Palmer. Building Tangent-Linear and Adjoint Models for Data Assimilation With Neural Networks. *Journal of Advances in Modeling Earth Systems*, 13(9):e2021MS002521, 2021. ISSN 1942-2466. doi: 10.1029/2021MS002521.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, January 1989. ISSN 0893-6080. doi: 10.1016/0893-6080(89)90020-8.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, January 1990. ISSN 0893-6080. doi: 10.1016/0893-6080(90)90005-6.

Kevin Judd and Leonard Smith. Indistinguishable states: I. Perfect model scenario. *Physica D: Nonlinear Phenomena*, 151(2):125–141, May 2001. ISSN 0167-2789. doi: 10.1016/S0167-2789(01)00225-1.

Kevin Judd and Leonard A. Smith. Indistinguishable states II: The imperfect model scenario. *Physica D: Nonlinear Phenomena*, 196(3):224–242, September 2004. ISSN 0167-2789. doi: 10.1016/j.physd.2004.03.020.

Shinhoo Kang, Hai V. Nguyen, and Tan Bui-Thanh. Second-order consistency for learning chaotic dynamics via randomized Jacobian matching. https://arxiv.org/abs/2606.01596v2, June 2026.

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

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, December 2023. doi: 10.1126/science.adi2336.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. AIFS – ECMWF's data-driven forecasting system, August 2024.

M. Leutbecher and T. N. Palmer. Ensemble forecasting. *Journal of Computational Physics*, 227(7):3515–3539, March 2008. ISSN 0021-9991. doi: 10.1016/j.jcp.2007.02.014.

E. N. Lorenz. Atmospheric predictability experiments with a large numerical model. *Tellus*, 34 (6):505–513, 1982. ISSN 2153-3490. doi: 10.1111/j.2153-3490.1982.tb01839.x.

Edward N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, March 1963. ISSN 0022-4928, 1520-0469. doi: 10.1175/1520-0469(1963)020¡0130:DNF¿2.0.CO;2.

Edward N. Lorenz. The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3):289–307, 1969. ISSN 2153-3490. doi: 10.1111/j.2153-3490.1969.tb00444.x.

F. Molteni. Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments. *Climate Dynamics*, 20(2): 175–191, January 2003. ISSN 0930-7575, 1432-0894. doi: 10.1007/s00382-002-0268-2.

T N Palmer, A Döring, and G Seregin. The real butterfly effect. *Nonlinearity*, 27(9):R123, August 2014. ISSN 0951-7715. doi: 10.1088/0951-7715/27/9/R123.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators, February 2022.

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. doi: 10.1029/2023MS004019.

R. Rotunno and C. Snyder. A Generalization of Lorenz's Model for the Predictability of Flows with Many Scales of Motion. *Journal of the Atmospheric Sciences*, 65(3):1063–1076, March 2008. ISSN 0022-4928, 1520-0469. doi: 10.1175/2007JAS2449.1.

T. Selz and G. C. Craig. Can Artificial Intelligence-Based Weather Prediction Models Simulate the Butterfly Effect? *Geophysical Research Letters*, 50(20):e2023GL105747, 2023. ISSN 1944-8007. doi: 10.1029/2023GL105747.

T. Selz and G. C. Craig. Can AI-Based Weather Prediction Models Simulate the Butterfly Effect? The Role of Architecture and Implementation. *Journal of Geophysical Research: Machine Learning and Computation*, 3(3):e2025JH001180, 2026. ISSN 2993-5210. doi: 10.1029/2025JH001180.

Suraj Srinivas and Francois Fleuret. Knowledge Transfer with Jacobian Matching. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4723–4731. PMLR, July 2018.

P. Trent Vonich and Gregory J. Hakim. Predictability Limit of the 2021 Pacific Northwest Heatwave From Deep-Learning Sensitivity Analysis. *Geophysical Research Letters*, 51(19): e2024GL110651, 2024. ISSN 1944-8007. doi: 10.1029/2024GL110651.

P. Trent Vonich and Gregory J. Hakim. Atmospheric Predictability beyond 30 Days with Machine Learning. *Artificial Intelligence for the Earth Systems*, 5(3), July 2026. ISSN 2769-7525. doi: 10.1175/AIES-D-26-0009.1.

Günther Zängl, Daniel Reinert, Pilar Rípodas, and Michael Baldauf. The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, 141(687): 563–579, 2015. ISSN 1477-870X. doi: 10.1002/qj.2378.

Fuqing Zhang, Y. Qiang Sun, Linus Magnusson, Roberto Buizza, Shian-Jiann Lin, Jan-Huey Chen, and Kerry Emanuel. What Is the Predictability Limit of Midlatitude Weather? *Journal of the Atmospheric Sciences*, 76(4):1077–1091, April 2019. ISSN 0022-4928, 1520-0469. doi: 10.1175/JAS-D-18-0269.1.