EXPONENTIAL–NEMYTSKII NEURAL OPERATORS: LEARNING FIELD EVOLUTION ACROSS STEP SIZES

Anonymous authors
Paper under double-blind review

ABSTRACT

In variable-step operator learning, agreement at observed step sizes does not by itself determine predictions at other temporal resolutions. We propose an operator whose step-size dependence is prescribed by a first-order Duhamel approximation of semilinear dynamics. It combines exact linear propagation with a local nonlinear function of the field and its derivatives, allowing evaluation below the observation interval without supervision densely sampled across step sizes. Under suitable conditions and at fixed spatial resolution, we prove that agreement between models at two step sizes controls their prediction differences throughout a prescribed range. We also give conditions for local identifiability of both components from two step sizes, supporting interpretation of the learned structure. A latent extension applies the same construction to learned fields. We evaluate the base model on two synthetic problems and the extension on two real datasets against several baselines.

1 INTRODUCTION

Many quantities of scientific and practical interest are continuous fields that evolve in space and time: the elevation of a water surface, the concentration of a diffusing species, the liquid-water content of a storm seen by radar (Veillette et al., 2020), or the radiative power of a spreading wildfire (Wooster et al., 2005). The continuous field itself is rarely observed directly. Instead, we typically rely on a finite set of noisy samples taken on a spatial grid at discrete observation times. The task is to learn the underlying spatiotemporal evolution from these samples and to predict the field's future states. Operator learning (Lu et al., 2021; Kovachki et al., 2023) has emerged as a popular approach to this task. Rather than fitting a discrete mapping between fixed-size arrays, it learns the solution operator that advances the field in time as a mapping between function spaces. Because the learned operator acts on the underlying functions rather than their discrete samples, it can be evaluated on spatial grids other than those used during training, and it can be applied to complex systems even when their governing equations remain unknown. The other family of methods are equation learning (Brunton & Kutz, 2019), which focus on recovering governing equations from the data. Our work focuses on operator learning; equation learning is discussed further in Section 2.4.

While operator learning treats space as continuous, many autoregressive models, including FNO-2D (Li et al., 2021), the autoregressive version of DeepONet (Lu et al., 2022) and KNO (Xiong et al., 2024), advance the field at a fixed time interval in both training and prediction. The step size is thereby built into the learned map, which does not directly support prediction at arbitrary temporal intervals. Considerable effort has been devoted to developing methods for variable-step prediction, and we review these approaches in Section 2.4. In this work, we propose a new operator-learning method for variable-step prediction, based on a first-order Duhamel approximation of semilinear dynamics. We combine this construction with multi-step-size training and establish conditions under which prediction maps at two distinct step sizes locally identify the linear and nonlinear components. Our main contributions are as follows.

- *Variable-step prediction with a stability guarantee.* Our model's dependence on the step size is prescribed by its mathematical construction, rather than learned entirely from data. This structure allows us to establish a theoretical stability guarantee for prediction across temporal resolutions, without requiring training data densely sampled across step sizes. We prove that, under stated conditions, close agreement with a reference model at just two

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

distinct training step sizes controls prediction differences throughout a prescribed range of step sizes, including below the observation interval (Theorem 1).

- *Identifiability and interpretability.* When applied directly to the observed fields, our model explicitly learns a linear differential operator and a local nonlinear function of the field and its derivatives. These components can be inspected directly to understand how linear propagation and nonlinear effects contribute to the prediction. This interpretation is ambiguous when different choices of components produce identical predictions. We prove that, under suitable conditions, the linear and nonlinear components of our model are locally identifiable when trained at two distinct step sizes (Theorem 3).

We test the method on two synthetic problems and on two real datasets, and compare it against a variety of baselines from operator learning and equation learning.

2 METHOD

2.1 SETUP AND MODEL DERIVATION

We observe a spatio–temporal field $u(x,t)$, $x \in \Omega \subset \mathbb{R}^{d_\Omega}$, $t \in [0,T]$, only through discrete, possibly noisy samples

$$\big\{\tilde{u}(x_\ell, t_j)\big\}_{\ell=1,\dots,N_x;\; j=1,\dots,N_t}, \tag{1}$$

taken at N_x spatial nodes $x_\ell \in \Omega$ and N_t times $t_j \in [0,T]$. Our goal is to learn the underlying model of the field and use it to make predictions.

We assume that $u(x,t)$ is described by a semilinear evolution equation,

$$\partial_t u = Au + F(u), \qquad u(\cdot,0) = u_0, \tag{2}$$

where A is an unknown linear operator generating a strongly continuous semigroup $(\mathrm{e}^{tA})_{t\geq 0}$ on a Banach space $\mathcal{X}$, and $F : \mathcal{X} \to \mathcal{X}$ is an unknown, locally Lipschitz nonlinear operator. Our aim is to learn both from Eq. (1) alone, in the form of a neural operator that predicts u at later times.

We construct our learning model using the Duhamel representation of Eq. (2), which does not merely split the dynamics into a linear part and a correction, but prescribes how the two combine. Write

$$v(t) := \mathrm{e}^{tA} u_0 \tag{3}$$

for the purely linear evolution of the initial state. A mild solution of Eq. (2) satisfies the Duhamel formula

$$u(t) \;=\; v(t) + \int_0^t \mathrm{e}^{(t-s)A} F\big(u(s)\big)\,\mathrm{d}s, \tag{4}$$

and is the fixed point of $w \mapsto v + \mathcal{N}[w]$, where $\mathcal{N}[w](t) := \int_0^t \mathrm{e}^{(t-s)A} F(w(s))\,\mathrm{d}s$, constructed by Picard iteration from $w^{(0)} := v$. Successive Picard iterates $w^{(k+1)} := v + \mathcal{N}[w^{(k)}]$ differ by $\mathcal{N}[w^{(k)}] - \mathcal{N}[w^{(k-1)}]$, which gains one order in t at each step since F is Lipschitz, so only the first correction $\mathcal{N}[v]$ enters at first order. Keeping it alone, $u \approx v + \mathcal{N}[v]$, and expanding $\mathcal{N}[v]$ near $t = 0$ (substitute $r = t - s$, then use $\mathrm{e}^{rA} \approx I$ and $v(t-r) \approx v(t)$) gives, at $t = \tau$,

$$u(\tau) \;\approx\; v(\tau) \;+\; \tau\, F\big(v(\tau)\big). \tag{5}$$

Over a short horizon the solution is the exactly propagated linear flow plus a single local correction evaluated at the current linear state. Eq. (5) is the model we train to predict u forward in time.

2.2 PARAMETRISATION AND TRAINING

To use Eq. (5) we must learn A and F from data. We parametrise each by a structured class matching its role, rather than by a generic neural network.

Spectral linear propagator. We assume A is a constant-coefficient differential operator of order at most d on a periodic domain, $A = \sum_{|\mathbf{n}|\le d} a_{\mathbf{n}} \partial^{\mathbf{n}}$ with $a_{\mathbf{n}} \in \mathbb{R}$ and $\partial^{\mathbf{n}} = \partial_1^{n_1} \cdots \partial_{d_\Omega}^{n_{d_\Omega}}$. The Fourier transform $\mathcal{F}$ replaces ∂_j by $\mathrm{i}k_j$, so A is the Fourier multiplier with symbol

$$g(k) \;=\; \sum_{|\mathbf{n}|\le d} c_{\mathbf{n}}\,(\mathrm{i}k)^{\mathbf{n}}, \qquad c_{\mathbf{n}} = a_{\mathbf{n}} \in \mathbb{R}, \tag{6}$$

a polynomial in $\mathrm{i}k$ whose coefficients are those of A itself; in one dimension $g(k) = \sum_{n\le d} c_n (\mathrm{i}k)^n$. The linear flow Eq. (3) becomes $v(x,\tau) = \mathcal{F}^{-1}[\mathrm{e}^{\tau g(k)} \widehat{u}_0(k)](x)$ with $\widehat{u}_0 = \mathcal{F}[u_0]$, and the coefficients $\{c_{\mathbf{n}}\}$ are the trainable parameters. The propagator is the *exact* solution of the linear part, so the $O(\tau^2)$ truncation in Eq. (5) is attributable to the nonlinear term alone. In the implementation we truncate the frequency modes to $|k| \le K$, where K is a hyperparameter (values in Appendix A.2.1).

For a bound uniform in k, we enforce $\mathrm{Re}\, g(k) \le \alpha$ for all $k \in \mathbb{R}^{d_\Omega}$, so that $|\mathrm{e}^{\tau g(k)}| \le \mathrm{e}^{\alpha\tau}$: the linear flow is bounded on $L^2(\Omega)$ at every step size and every discretisation. This is imposed by the following construction. Only terms of even total order contribute to $\mathrm{Re}\, g$; the odd-order terms, which carry transport and dispersion, are purely imaginary and are left free, as is the constant term $c_{\mathbf{0}} =: \alpha$. Among the even-order terms, those in which every k_j appears to an even power (k_1^2, $k_1^2 k_2^2$, and so on) are nonnegative everywhere and receive a nonpositive coefficient, so they can only push $\mathrm{Re}\, g$ down; the rest, such as $k_1 k_2$, change sign with direction and are dropped.

Nemytskii operator. For the nonlinear part of Eq. (2) we do not consider nonlocal nonlinearities, and assume instead that F is local of some order $\rho \in \mathbb{N}_0$: the value of $F(u)$ at x depends on u only through $u(x)$ and its spatial derivatives at x up to order ρ (the ρ-jet of u at x). Write

$$\Phi[w](x) \;:=\; \big(w(x),\, \nabla w(x),\, \ldots,\, \nabla^\rho w(x),\, \eta(x)\big) \;\in\; \mathbb{R}^m \tag{7}$$

for the *feature map*, which collects the field, its spatial derivatives up to order ρ, and, optionally, externally given fields $\eta(x)$; its number of entries is m. We then assume

$$F(u)(x) \;=\; f\big(\Phi[u](x)\big), \qquad f : \mathbb{R}^m \to \mathbb{R}, \tag{8}$$

so that F is a generalised Nemytskii (superposition) operator (Li et al., 2025) induced by the finite-dimensional map f. Accordingly, the goal becomes to learn the finite-dimensional function $f : \mathbb{R}^m \to \mathbb{R}$ and we represent it by a multilayer perceptron (MLP) with smooth activations. The derivatives in Φ are evaluated spectrally, $\nabla^j v = \mathcal{F}^{-1}[(\mathrm{i}k)^j \widehat{v}]$.

Exponential-Nemytskii neural operator. Combining the two components using Eq. (5),

$$\mathcal{S}_\tau[u_0](x) \;=\; \underbrace{\mathcal{F}^{-1}\big[\mathrm{e}^{\tau g}\mathcal{F}u_0\big](x)}_{v(x,\tau)} \;+\; \tau\, f\big(\Phi[v(\cdot,\tau)](x)\big), \tag{9}$$

with trainable objects g and f. We call the resulting model the *Exponential–Nemytskii Neural Operator* (ENNO), reflecting its combination of exponential linear propagation and a learned Nemytskii map acting on local field features. Predictions over horizons longer than a single first-order step are obtained by applying Eq. (9) autoregressively: starting from u_0,

$$u_{j+1} \;=\; \mathcal{S}_\tau[u_j], \qquad j = 0, \ldots, n-1, \qquad \text{so that} \qquad u_n \;=\; \mathcal{S}_\tau^n[u_0] \tag{10}$$

is the prediction at time $t = n\tau$.

Training. In training we use only the one-step model Eq. (9); the multi-step composition Eq. (10) is used only for prediction. A key feature of our approach is that the model is trained with several step sizes τ, which is required for the identifiability of (g, f) (Section 2.3.2). Concretely, fix step sizes $\tau_1 < \cdots < \tau_S$, $S \ge 2$, chosen among the differences of observation times, and for each τ_s let $\mathcal{D}_s$ be the set of observation times t_j for which $t_j + \tau_s$ is also an observation time. For each $t_j \in \mathcal{D}_s$ we advance $\tilde{u}(\cdot, t_j)$ by one step of size τ_s and compare with the observation at $t_j + \tau_s$; averaging the squared error over the grid and over $\mathcal{D}_s$ gives

$$\mathcal{L}_s \;=\; \frac{1}{|\mathcal{D}_s|} \sum_{t_j \in \mathcal{D}_s} \frac{1}{N_x} \sum_{\ell=1}^{N_x} \big|\, \mathcal{S}_{\tau_s}\big[\tilde{u}(\cdot, t_j)\big](x_\ell) - \tilde{u}(x_\ell, t_j + \tau_s)\,\big|^2, \tag{11}$$

and the training objective is their sum, $\mathcal{L} = \sum_{s=1}^{S} \mathcal{L}_s$. Since τ enters $\mathcal{S}_\tau$ explicitly, the same (g, f) is used for every s. With a single step size the constant term of g and the linear part of f cannot be separated (Lemma 2); a second step size removes this ambiguity (Theorem 3(ii)).

Latent fields. An extension of the method handles the case where $u(x,t)$ does not follow the semi-linear model Eq. (2): we map $u(x,t)$ into a latent field, or several latent fields, using an encoding–decoding structure. Namely, we use an encoder $\mathcal{E}$ to map the observed history into the latent fields $w = (w^1, \dots, w^M)$, advance all of them with the same g and f, and use a decoder $\mathcal{R}$ to map the latent state back:

$$w_0 = \mathcal{E}[u_{-T+1}, \dots, u_0], \qquad w_{j+1} = \big(\mathcal{S}_\tau[w_j^1], \dots, \mathcal{S}_\tau[w_j^M]\big), \qquad u_j = \mathcal{R}[w_j]. \tag{12}$$

Here we restrict both $\mathcal{E}$ and $\mathcal{R}$ to be Nemytskii operators, generated by pointwise maps $\mathbb{R}^T \to \mathbb{R}^M$ and $\mathbb{R}^M \to \mathbb{R}$, so every learnable object in the model is generated by a map between finite-dimensional spaces. The encoder and decoder are trained jointly with g and f. Taking $T > 1$ acts as a delay embedding: when u is a partial observation of a higher-dimensional state, a single frame need not determine the subsequent evolution. Only the initial latent state uses the history; the rollout remains Markovian in w. Note that g and f are shared across the latent fields, so the parameter count does not grow with M: the M fields evolve independently under the same dynamics but from different initial channels produced by $\mathcal{E}$, and are recombined nonlinearly by $\mathcal{R}$. The price is interpretability: in the base model g and f describe the mechanisms governing u itself, whereas here they describe the latent fields and no longer correspond directly to the equation for u.

2.3 Theory

2.3.1 Stability across step sizes

We work at fixed spatial resolution K on a periodic box Ω. Let $\mathcal{V}_K$ be the real Fourier space with integer mode indices $|k| \le K$, let P_K be its orthogonal projection, and write $\mathcal{U}_R := \{u \in \mathcal{V}_K : \|u\|_{L^2} \le R\}$. All operator norms below are induced by the L^2 norm. For $\vartheta = (g, f)$, let A have symbol g as in Eq. (6), and set $F_K(v) = P_K[f(\Phi_{\mathrm{st}}[v])]$ and $\mathcal{S}_t^{K,\vartheta} u = \mathrm{e}^{tA}u + tF_K(\mathrm{e}^{tA}u)$. Here f is a spatially homogeneous, locally Lipschitz scalar function, and $\Phi_{\mathrm{st}}[v]$ collects the field and its distinct partial derivatives through order $\rho \ge 0$. For a second pair $\widetilde{\vartheta} = (\widetilde{g}, \widetilde{f})$ from the same class, define $\widetilde{A}$, $\widetilde{F}_K$, and $\mathcal{S}_t^{K,\widetilde{\vartheta}}$ analogously. Let $\mathcal{M}_K$ denote the real-preserving Fourier multipliers on $\mathcal{V}_K$, and define the nonlinear contrast $\mathcal{N}_F(v) := F_K(v) - 2F_K(v/2) + F_K(0)$. We have the following theorem where the proof is given in Appendix A.1.1.

Theorem 1 (Cross-step stability). *Fix $R, \Delta, M > 0$, and suppose $\|A\|_{\mathrm{op}}, \|\widetilde{A}\|_{\mathrm{op}} \le M$. Assume that there exist $0 < r_0 < R$ and $\sigma, \zeta > 0$ such that*

$$\sup_{v \in \mathcal{U}_{r_0}} \|\mathcal{N}_F(\mathrm{e}^H v) - \mathcal{N}_F(v)\|_{L^2} \ge \sigma \|H\|_{\mathrm{op}}, \qquad H \in \mathcal{M}_K, \quad \|H\|_{\mathrm{op}} \le \zeta. \tag{13}$$

Let L_Δ be a Lipschitz constant for F_K on $\mathcal{U}_{\exp(2M\Delta)R}$. There exist $0 < \tau_* \le \Delta$, $0 < r_\Delta < R$, and $C_\Delta > 0$, depending only on $\Omega, K, M, R, \Delta, r_0, \sigma, \zeta, L_\Delta$, such that for any $0 < \tau_0 < \tau_1 \le \tau_*$, setting

$$\varepsilon_{\mathrm{obs}} := \max_{j=0,1} \sup_{u \in \mathcal{U}_R} \frac{\|\mathcal{S}_{\tau_j}^{K,\vartheta} u - \mathcal{S}_{\tau_j}^{K,\widetilde{\vartheta}} u\|_{L^2}}{\tau_j},$$

we have, for every $0 < \delta \le \Delta$,

$$\sup_{u \in \mathcal{U}_{r_\Delta}} \frac{\|\mathcal{S}_\delta^{K,\vartheta} u - \mathcal{S}_\delta^{K,\widetilde{\vartheta}} u\|_{L^2}}{\delta} \le C_\Delta \varepsilon_{\mathrm{obs}} \left(1 + \frac{2 \min_{\tau \in [\tau_0, \tau_1]} |\delta - \tau|}{\tau_1 - \tau_0}\right). \tag{14}$$

The linear dependence on this ratio is sharp in general.

Interpretation. The theorem can be understood as follows. If a model's predictions are close to those of a reference model at two training step sizes, their prediction difference per unit time also has an explicit upper bound at other step sizes. At intermediate step sizes, the same theoretical bound applies as at the training steps. Outside this interval, the bound grows linearly with the extrapolation distance relative to the separation between the training steps. This provides a stability guarantee for using the same model at different temporal resolutions.

2.3.2 IDENTIFIABILITY

We study structural identifiability of (g, f) from the one-step prediction maps. Write $\Phi[\phi] = (\Phi_{\mathrm{st}}[\phi], \eta)$, where Φ_{st} lists each distinct partial derivative through order ρ once, including the field value. Accordingly, write $z = (z_{\mathrm{st}}, z_\eta)$, with z_0 the field value. Let $\eta : \Omega \to \mathbb{R}^{d_\eta}$ be continuous and $E := \eta(\Omega)$. Proofs of the two results below are in Appendix A.1.2.

Lemma 2 (Single-step degeneracy). *Fix $\tau_0 > 0$. For $\gamma \in \mathbb{R}$, set $\varrho_\gamma := \mathrm{e}^{\gamma\tau_0}$ and $\chi_\gamma := (1 - \varrho_\gamma)/(\varrho_\gamma\tau_0)$, and define*

$$g_\gamma := g + \gamma, \qquad f_\gamma(z_{\mathrm{st}}, z_\eta) := f(z_{\mathrm{st}}/\varrho_\gamma, z_\eta) + \chi_\gamma z_0. \tag{15}$$

Then

$$\mathcal{S}_{\tau_0}[\phi; g_\gamma, f_\gamma] = \mathcal{S}_{\tau_0}[\phi; g, f] \tag{16}$$

whenever the maps are defined.

Theorem 3 (Identifiability). *Let (g, f) and (g', f') have symbols in the real-coefficient polynomial class of Eq. (6) and nonlinearities that are C^1 in z_{st} and continuous in z_η. Fix $r > 0$ and $K \geq K_0(d, \rho, d_\Omega)$, with the sufficient threshold K_0 specified in Appendix A.1.2. At every step size τ considered, assume $\tau\,|\operatorname{Im}\widetilde{g}(k)| < \pi/4$ for $\widetilde{g} \in \{g, g'\}$ and all $k \in \mathbb{Z}^{d_\Omega}$ with $|k| \leq K$.*

(i) One step size. *Let $\tau_0 > 0$. Suppose $\mathcal{S}_{\tau_0}[\phi; g, f] = \mathcal{S}_{\tau_0}[\phi; g', f']$ on Ω for every $\phi \in \mathcal{V}_K$ with $\|\phi\|_{L^2} < r$, and $1 + \tau_0\,\partial_{z_0} f(0, \eta(x_0)) \neq 0$ for some $x_0 \in \Omega$. Then, for some $\gamma \in \mathbb{R}$ and $r_* > 0$, $g' = g + \gamma$ and $f' = f_\gamma$ on $B_{r_*}(0) \times E$ with $B_{r_*}(0) \subset \mathbb{R}^{m_{\mathrm{st}}}$, where f_γ is defined by Eq. (15). In particular, $c'_{\mathbf{n}} = c_{\mathbf{n}}$ for $1 \leq |\mathbf{n}| \leq d$.*

(ii) Two step sizes. *Under the hypotheses of* (i), *suppose also $\mathcal{S}_{\tau_1}[\phi; g, f] = \mathcal{S}_{\tau_1}[\phi; g', f']$ on Ω for the same ϕ, where $\tau_1 > 0$, $\tau_1 \neq \tau_0$. If, for some $z_\eta^* \in E$, on no neighbourhood of the origin is $f(\cdot, z_\eta^*)$ of the form $a + bz_0$ for constants $a, b \in \mathbb{R}$, then $\gamma = 0$. Thus $g' = g$, and $f' = f$ on $B_{r_*}(0) \times E$.*

Interpretation. Lemma 2 shows that the linear and nonlinear components cannot be uniquely separated at a single step size, and longer rollouts at that same step size do not resolve this ambiguity. Theorem 3 states that a second distinct step size removes this ambiguity under the stated conditions. Thus, using two distinct step sizes resolves the ambiguity between the linear and nonlinear components, establishing structural identifiability of the model.

2.4 RELATED WORK

We review existing methods that support variable-step prediction. Equation-learning methods can recover continuous-time governing equations from data (Brunton et al., 2016; Rudy et al., 2017; Long et al., 2018). These equations permit flexible time stepping, although prediction still relies on numerical integration of the learned dynamics. Here we focus on operator-learning methods, which learn evolution maps directly. One such approach treats time as a query coordinate, as in DeepONet with a space–time trunk (Lu et al., 2021) and nonlinear coordinate decoders (Seidman et al., 2022). These models directly learn spatiotemporal neural representations and may require substantial paired training data (Wang et al., 2021). The second approach, exemplified by Poseidon (Herde et al., 2024) and OSG-FNO (Chen et al., 2025), takes the step size τ as an input. Such models can in principle be evaluated at any step size, but their dependence on τ is entirely learned through training. Outside the range of observed step sizes—in particular below the finest step the data provide—fitting the available data alone cannot guarantee reliable prediction. A third approach learns a vector field in function space and predicts by numerical integration: TI-DeepONet (Nayak & Goswami, 2026) fits it through a numerical integrator, while CFO (Hou et al., 2026) fits it to estimated time derivatives. The learned dynamics may consequently inherit errors from the numerical integration or derivative estimation used during training.

In our numerical experiments we compare against one representative of each of the three families above: NOMAD (Seidman et al., 2022) for temporal-spatial representation, OSG-FNO (Chen et al., 2025) for step-size learning, and TI-DeepONet (Nayak & Goswami, 2026) for learned vector fields. We further compare against E-WSINDy (Fasel et al., 2022) from the equation-learning literature, KNO (Xiong et al., 2024) as a fixed-step method, and persistence.

3 NUMERICAL EXPERIMENTS

We present four numerical experiments: two synthetic examples and two applications to real-world data. We compare our method with 5 benchmarks discussed in Section 2.4. We use the recommended or default settings from the original papers and implementations whenever available and applicable. Otherwise, we select suitable hyperparameters based on validation performance. The architecture, spectral cutoff, symbol degree and data settings of ENNO in each experiment are listed in Appendix A.2.1. All neural models are trained for at most 3,000 epochs. All baselines are configured to a parameter count comparable to ours, except OSG-FNO, which we could not reduce below three times ours without seriously degrading its training.

3.1 THE KDV EQUATION

The first synthetic example is the Korteweg–de Vries equation $u_t + 6uu_x + u_{xxx} = 0$ on the periodic domain $[0, 10]$. Its symbol $g(k) = \mathrm{i}k^3$ lies in the polynomial class of Eq. (6) once $d \geq 3$, with $c_3 = -1$ and all other coefficients zero, so the learned symbol can be compared with the truth coefficient by coefficient. Training uses 2030 trajectories on an $N_x = 128$ grid, corrupted by Gaussian noise at 5% of each trajectory's own standard deviation; 50 and 200 further trajectories with unseen initial conditions serve for validation and test. The model is trained with two step sizes, $\tau_1 = 10^{-3}$ and $\tau_2 = 2 \times 10^{-3}$. KNO needs ten frames of context, so all methods are scored on the same targets, $u(t_1), u(t_5), u(t_{10})$; the single-frame methods are conditioned on $u(t_0)$ and KNO on the ten frames up to and including $u(t_0)$.

Table 1: KdV: relative L^2 error (%) at three target frames, averaged over the 200 test trajectories, for symbol degrees $d = 1, \ldots, 4$.

	t_1	t_5	t_{10}
$d = 1$	2.895	3.676	4.200
$d = 2$	2.179	2.695	2.989
$d = 3$	**0.313**	**0.721**	**1.115**
$d = 4$	0.319	0.743	1.151

The degree d decides whether the true model can be recovered, so we first ablate it in Table 1, with everything else held fixed. The error falls by a factor of 2.7 from $d = 2$ to $d = 3$ and does not improve clearly at $d = 4$. We use $d = 3$ in what follows. We also ablate the latent-field structure, which brings no improvement here, as expected since the true model lies in our class at this degree (Appendix A.2.2). To test the identifiability of the two components, we show the recovered coefficients in Table 2, where all of them agree well with the truth. Table 3 compares against the baselines, reporting the mean over the 200 test trajectories with its standard error of the mean (SE); ENNO is lowest at all four target times. The $t_{1/2}$ column evaluates each model at half the smallest training step, which none of them has seen: KNO cannot be evaluated there at all, and OSG-FNO and TI-DeepONet are worse at $t_{1/2}$ than at t_1 despite the shorter horizon.

Table 2: KdV: recovered symbol coefficients at $d = 3$ against the truth.

	c_0	c_1	c_2	c_3
truth	0	0	0	-1
recovered	-0.061	-0.081	4.9×10^{-5}	-0.979

3.2 THE ADVECTION-DIFFUSION EQUATION

The second synthetic example is the advection-diffusion equation $u_t = \nu(u_{xx} + u_{yy}) - u(u_x + u_y)$ with $\nu = 0.05$ on the periodic domain $[0, 1]^2$. Its symbol $g(\mathbf{k}) = -\nu|\mathbf{k}|^2$ lies in the polynomial class of Eq. (6) once $d \geq 2$, with $c_{(2,0)} = c_{(0,2)} = \nu$ and all other coefficients zero, so the learned symbol can again be compared with the truth coefficient by coefficient. Training uses 100 trajectories on a 128×128 grid, corrupted by Gaussian noise at 5% of each trajectory's own standard deviation; 10

Table 3: KdV: relative L^2 error (%) at four target times, reported as mean $\pm$ SE over the 200 test trajectories. $t_{1/2}$ is half the smallest training step and is not an observation time; methods that cannot be evaluated below the observation interval are marked "—".

Method	$t_{1/2}$	t_1	t_5	t_{10}
ENNO ($d = 3$)	**0.197** $\pm$ 0.008	**0.313** $\pm$ 0.015	**0.721** $\pm$ 0.032	**1.115** $\pm$ 0.049
TI-DeepONet	2.374 $\pm$ 0.066	1.369 $\pm$ 0.035	2.556 $\pm$ 0.078	3.329 $\pm$ 0.105
E-WSINDy	0.830 $\pm$ 0.061	1.459 $\pm$ 0.079	2.436 $\pm$ 0.124	2.793 $\pm$ 0.143
OSG-FNO	3.264 $\pm$ 0.083	2.045 $\pm$ 0.089	2.008 $\pm$ 0.088	2.112 $\pm$ 0.080
KNO	—	2.049 $\pm$ 0.096	1.824 $\pm$ 0.083	1.912 $\pm$ 0.087
NOMAD	4.800 $\pm$ 0.147	4.815 $\pm$ 0.152	4.804 $\pm$ 0.152	4.832 $\pm$ 0.149

and 25 further trajectories serve for validation and test. The encoder and decoder are identity maps, and the model is trained with two step sizes, $\tau_1 = 10^{-3}$ and $\tau_2 = 2 \times 10^{-3}$. As on KdV, all methods use the last input frame only, except KNO which uses the preceding ten.

The ablation in Table 4 selects $d = 2$, the true order, with no clear improvement at $d = 3$. We first estimate the symbol coefficients: the diffusion coefficient is recovered as 0.0488 against the true 0.05, and the remaining coefficients, whose true values are zero, stay below 0.07 in magnitude (Table 5). Table 6 compares against the baselines at comparable parameter counts. ENNO is lowest at all four target times, with E-WSINDy the closest competitor, unlike on KdV, where its candidate library must cover a third-order dispersive term. At $t_{1/2}$ it remains lowest, while KNO cannot be evaluated there at all. Neither branch–trunk method tracks the dynamics here: the error of NOMAD is already at the scale of the solution at t_1 and barely changes over the horizon, while that of TI-DeepONet grows from about 15% at t_1 to 51% at t_{10}. Since TI-DeepONet is competitive on KdV (second-best at t_1), the failure here is not one of training budget.

Table 4: Advection–diffusion: relative L^2 error (%) at three target frames, averaged over the 25 test trajectories, for symbol degrees $d = 1, 2, 3$.

	t_1	t_5	t_{10}
$d = 1$	2.439	4.011	4.647
$d = 2$	**0.021**	**0.074**	**0.129**
$d = 3$	0.021	0.074	0.128

Table 5: Advection–diffusion: recovered symbol coefficients at $d = 2$ against the truth. The two first-order coefficients are listed as x/y.

	$c_{(0,0)}$	$c_{(1,0)}/c_{(0,1)}$	$c_{(2,0)} = c_{(0,2)}$
truth	0	0/0	0.05
recovered	0.029	0.054/0.061	0.0488

Table 6: Advection–diffusion: relative L^2 error (%) at four target times, reported as mean $\pm$ SE over the 25 test trajectories. $t_{1/2}$ is half the smallest training step and is not an observation time; methods that cannot be evaluated below the observation interval are marked "—".

Method	$t_{1/2}$	t_1	t_5	t_{10}
ENNO ($d = 2$)	**0.011** $\pm$ 0.001	**0.021** $\pm$ 0.002	**0.074** $\pm$ 0.008	**0.129** $\pm$ 0.013
E-WSINDy	0.030 $\pm$ 0.003	0.030 $\pm$ 0.003	0.145 $\pm$ 0.016	0.283 $\pm$ 0.030
KNO	—	0.269 $\pm$ 0.037	0.376 $\pm$ 0.049	0.562 $\pm$ 0.072
OSG-FNO	0.262 $\pm$ 0.037	0.524 $\pm$ 0.060	2.617 $\pm$ 0.303	5.223 $\pm$ 0.607
TI-DeepONet	14.597 $\pm$ 2.122	14.551 $\pm$ 2.068	29.970 $\pm$ 3.330	50.831 $\pm$ 5.843
NOMAD	79.330 $\pm$ 5.491	60.124 $\pm$ 4.971	59.634 $\pm$ 4.909	58.969 $\pm$ 4.898

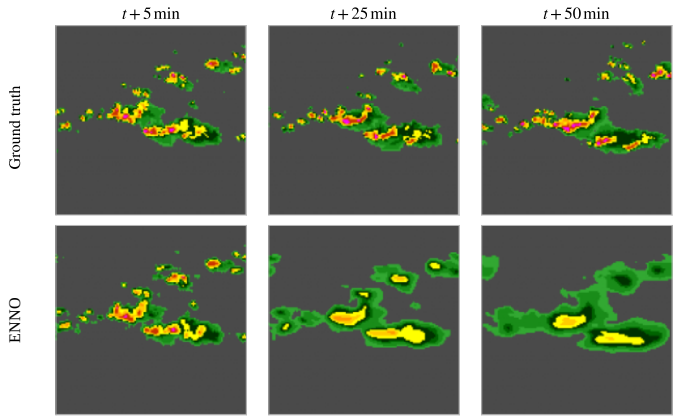


Figure 1: Qualitative SEVIR forecasts for one test event at lead times of 5, 25 and 50 minutes, matching Table 7. Top row: ground truth; bottom row: ENNO's autoregressive predictions.

3.3 Radar nowcasting

Radar nowcasting aims to predict the precipitation field 0–2 hours ahead from recent radar observations, filling the gap that numerical weather prediction leaves at these lead times. We take the vertically integrated liquid (VIL) product of SEVIR (Veillette et al., 2020) as the observed field $u(x, t)$, sampled on a uniform 128×128 grid over a square domain $\Omega \subset \mathbb{R}^2$ of 384×384 km^2, with consecutive frames five minutes apart. Given the ten most recent frames, the task is to predict the next ten frames at five-minute intervals, covering lead times of 5–50 minutes. We use 11,159 training events, 200 validation events and 500 test events; the test events come from separate files covering November–December 2019. Our model encodes the ten input frames into 20 latent fields and is trained with two step sizes, $\tau_1 = 5$ and $\tau_2 = 10$ minutes. All methods predict the same final ten frames of each test event.

Table 7: SEVIR test MSE ($\times 10^{-3}$) at three forecast lead times, reported as mean $\pm$ SE over the 500 test events. Bold marks the best result. "Failed" denotes nonfinite predictions during rollout.

Method	5 minutes	25 minutes	50 minutes
ENNO	**1.4** $\pm$ 0.05	**5.4** $\pm$ 0.15	**9.2** $\pm$ 0.25
KNO	1.8 $\pm$ 0.05	6.7 $\pm$ 0.17	10.3 $\pm$ 0.28
OSG-FNO	1.8 $\pm$ 0.04	8.6 $\pm$ 0.15	13.7 $\pm$ 0.26
NOMAD	5.6 $\pm$ 0.14	8.9 $\pm$ 0.24	13.4 $\pm$ 0.40
TI-DeepONet	6.4 $\pm$ 0.16	11.4 $\pm$ 0.29	19.5 $\pm$ 0.50
E-WSINDy	12.5 $\pm$ 0.30	Failed	Failed
Persistence	2.6 $\pm$ 0.09	15.3 $\pm$ 0.42	24.3 $\pm$ 0.66

Our model achieves the lowest MSE at all three lead times in Table 7. The comparison with KNO is particularly informative because both models receive the same ten-frame history. Relative to KNO, our model reduces MSE by about 22%, 19% and 11% at 5, 25 and 50 minutes, respectively. The improvement is therefore present in the first prediction and persists at all three lead times, demonstrating a sustained advantage over the tested forecast horizon. Figure 1 illustrates the rollout on a test event. ENNO tracks the motion and coarse structure of the storm cells throughout, while the high-intensity cores are progressively smoothed.

3.4 Wildfire prediction

Wildfire management relies on knowing where a fire will burn next and how intensely. The goal of this example is to predict the next day's fire radiative power (FRP) field from the FRP observed on the preceding days together with environmental drivers—wind, terrain, vegetation and moisture—

recorded on the most recent day. We use WildfireSpreadTS (Gerard et al., 2023), which records 607 wildfire events across the western United States between 2018 and 2021 as daily multi-channel images at 375 m resolution, one channel holding the observed FRP and the rest holding environmental drivers for the same day. We take $T_{\text{in}} \in \{1, 5\}$ daily FRP fields as the state, transformed by $\log(1 + \text{FRP})$ since FRP spans about four orders of magnitude, and 25 environmental channels from the most recent input day. The latter enter the model as the externally given fields η of Eq. (7), so that f can modulate the predicted intensity pointwise according to local wind, slope and fuel conditions. Our model is trained with two step sizes, $\tau_1 = 1$ and $\tau_2 = 2$ days; evaluation is always at one day ahead. All neural predictors are trained with the same weighted squared loss on log-FRP, giving weight 50 to pixels with positive FRP and weight 1 to the remaining valid pixels: fire pixels make up under 0.15% of the valid pixels, and without this weighting training can favour near-zero predictions. Pixels introduced by padding events to a common grid size are excluded from both training and evaluation.

We use four folds over 2018–2021, one per test year. Each fold trains on two years, selects a checkpoint on a third, and evaluates on the remaining year. Model selection and evaluation use the area under the precision–recall curve (AUC-PR), with predicted log-FRP serving as a continuous score for detecting pixels with positive FRP. Table 8 shows that our model achieves the highest mean AUC-PR at both history lengths. With one input day, it leads in every held-out year, attaining a mean of 0.3210 compared with 0.2895 for KNO. With five input days, our model and OSG-FNO achieve similar means (0.3122 versus 0.3098), each leading in two of the four held-out years; OSG-FNO uses more than ten times as many trainable parameters (8.4M versus 0.75M). Longer histories do not improve our model's mean AUC-PR here, and persistence exceeds three of the trained baselines with one input day.

Table 8: WildfireSpreadTS test AUC-PR, reported as mean $\pm$ sample standard deviation across four held-out years. Bold marks the highest mean. "Unsupported" indicates that the evaluated implementation does not support five-frame inputs.

	AUC-PR	
Method	$T_{\text{in}} = 1$	$T_{\text{in}} = 5$
ENNO	$\mathbf{0.3210} \pm 0.0848$	$\mathbf{0.3122} \pm 0.0835$
KNO	0.2895 ± 0.0867	0.2889 ± 0.0744
OSG-FNO	0.2664 ± 0.0811	0.3098 ± 0.0887
TI-DeepONet	0.1435 ± 0.0373	Unsupported
NOMAD	0.1290 ± 0.0394	Unsupported
E-WSINDy	0.0337 ± 0.0635	0.2533 ± 0.0954
Persistence	0.2473 ± 0.0984	0.2530 ± 0.0972

4 Conclusions

We presented ENNO, an operator-learning method that combines exact linear propagation with a local nonlinear correction and prescribes its step-size dependence through a first-order Duhamel approximation. Under suitable conditions, prediction maps at two distinct step sizes locally identify these components, and agreement between models at those steps controls prediction differences across a prescribed range. Experiments on two synthetic PDEs and two real-world datasets demonstrate strong predictive performance, with a latent extension handling the real-data applications. The synthetic examples also show accurate recovery of the principal differential coefficients. The current guarantees concern the base model at fixed spatial resolution. Extending the analysis to the latent construction and obtaining stability bounds uniform under spatial refinement are left to future work.

AI USE STATEMENT

In this work, we used generative AI tools for tasks that require disclosure as follows. The theoretical results of this paper—the model construction, the statements of Theorem 1, Lemma 2 and Theorem 3, and the proof strategies—were conceived and developed in full by the authors, and complete proofs were written before any AI tool was involved. We then used generative AI tools to check these proofs. The tools identified a number of errors and proposed corrections, some of which we adopted; they also suggested improvements to the arguments, some of which we adopted. Every adopted correction and improvement was verified by the authors before being incorporated, and none of them changed the statements of the results. We disclose this use under both assisting in the writing of proofs and providing critical ingredients for proving mathematical claims.

We have reviewed all AI-assisted work. All AI assistance was carried out under the direction of the authors, and the resulting material was subsequently checked and verified by two authors: every proof, including every AI-suggested correction, was verified line by line independently of the tool's output; AI-edited text was checked against the claims the authors intended to make. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The model architecture, Fourier cutoff, symbol degree, nonlinear-net inputs and the constraint parametrisation used in each experiment are listed in Appendix A.2.1, together with the data-generation procedure for the two synthetic problems and the splits for all four experiments. SEVIR and WildfireSpreadTS are publicly available; the preprocessing applied to each is described in Sections 3.3 and 3.4 and Appendix A.2.1. Code for the model, the baselines and the evaluation pipeline will be released upon acceptance.

REFERENCES

Steven L. Brunton and J. Nathan Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

Junfeng Chen, Kailiang Wu, and Dongbin Xiu. DUE: A deep learning framework and library for modeling unknown equations. *SIAM Review*, 67(4):873–902, 2025.

Urban Fasel, J. Nathan Kutz, Bingni W. Brunton, and Steven L. Brunton. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A*, 478(2260):20210904, 2022.

Sebastian Gerard, Yu Zhao, and Josephine Sullivan. WildfireSpreadTS: A dataset of multi-modal time series for wildfire spread prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 72525–72624, 2024.

Xianglong Hou, Xinquan Huang, and Paris Perdikaris. CFO: Learning continuous-time PDE dynamics via flow-matched neural operators. In *International Conference on Learning Representations*, 2026.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Jingye Li, Alex Bespalov, and Jinglai Li. Nemytskii neural operator: A nonlinear model reduction method for parametrized partial differential equations. *arXiv preprint*, 2025. arXiv:2511.07684.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. PDE-Net: Learning PDEs from data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3208–3216, 2018.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Lu Lu, Xuhui Meng, Shengze Cai, Zhiping Mao, Somdatta Goswami, Zhongqiang Zhang, and George Em Karniadakis. A comprehensive and fair comparison of two neural operators (with practical extensions) based on FAIR data. *Computer Methods in Applied Mechanics and Engineering*, 393:114778, 2022.

Dibyajyoti Nayak and Somdatta Goswami. TI-DeepONet: Learnable time integration for stable long-term extrapolation. *Computer Methods in Applied Mechanics and Engineering*, 456:118960, 2026.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

Jacob Seidman, Georgios Kissas, Paris Perdikaris, and George J. Pappas. NOMAD: Nonlinear manifold decoders for operator learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5601–5613, 2022.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Sifan Wang, Hanwen Wang, and Paris Perdikaris. Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Science Advances*, 7(40):eabi8605, 2021.

M. J. Wooster, G. Roberts, G. L. W. Perry, and Y. J. Kaufman. Retrieval of biomass combustion rates and totals from fire radiative power observations: FRP derivation and calibration relationships between biomass consumption and fire radiative energy release. *Journal of Geophysical Research: Atmospheres*, 110(D24):D24311, 2005.

Wei Xiong, Xiaomeng Huang, Ziyang Zhang, Ruixuan Deng, Pei Sun, and Yang Tian. Koopman neural operator as a mesh-free solver of non-linear partial differential equations. *Journal of Computational Physics*, 513:113194, 2024.

A Appendix

A.1 Proofs

A.1.1 Proofs of Section 2.3.1

Throughout this appendix, write

$$X := \mathcal{V}_K, \qquad F := F_K, \qquad \widetilde{F} := \widetilde{F}_K, \qquad S_t := \mathcal{S}_t^{K,\vartheta}, \qquad \widetilde{S}_t := \mathcal{S}_t^{K,\widetilde{\vartheta}}.$$

For vectors, $\|\cdot\|$ denotes the $L^2(\Omega)$ norm; for linear operators, it denotes the induced operator norm. Since X is finite-dimensional, the locally Lipschitz maps F and $\widetilde{F}$ are Lipschitz on every bounded ball. All Fourier multipliers considered below preserve real-valued fields and commute with one another.

Proof of the estimate in Theorem 1. Set

$$B := A - \widetilde{A}, \qquad h := \tau_1 - \tau_0, \qquad \varepsilon := \varepsilon_{\mathrm{obs}}.$$

Then $\|B\| \le 2M$. We may choose

$$\tau_* := \min\left\{\Delta, \frac{\zeta}{2M}, \frac{\log(R/r_0)}{3M}\right\}, \qquad r_\Delta := r_0 \mathrm{e}^{-M\Delta}. \tag{17}$$

We first control the difference between the linear operators. For $t > 0$ and $v \in X$, define

$$Z_t(v) := \frac{S_t(\mathrm{e}^{-t\widetilde{A}}v) - \widetilde{S}_t(\mathrm{e}^{-t\widetilde{A}}v)}{t}. \tag{18}$$

Commutativity of A and $\widetilde{A}$ gives the exact identity

$$Z_t(v) = \frac{\mathrm{e}^{tB} - I}{t} v + F(\mathrm{e}^{tB}v) - \widetilde{F}(v). \tag{19}$$

Consequently,

$$Z_t(v) - 2Z_t(v/2) + Z_t(0) = \mathcal{N}_F(\mathrm{e}^{tB}v) - \mathcal{N}_{\widetilde{F}}(v), \tag{20}$$

where $\mathcal{N}_{\widetilde{F}}(v) := \widetilde{F}(v) - 2\widetilde{F}(v/2) + \widetilde{F}(0)$.

Fix $w \in \mathcal{U}_{r_0}$ and set $v = \mathrm{e}^{-\tau_0 B}w$. For $j = 0, 1$,

$$\|\mathrm{e}^{-\tau_j \widetilde{A}}v\| \le \mathrm{e}^{M\tau_j + 2M\tau_0} r_0 \le \mathrm{e}^{3M\tau_*} r_0 \le R.$$

The same bound holds with v replaced by $v/2$ or 0. The definition of ε therefore implies

$$\|Z_{\tau_j}(v)\|,\ \|Z_{\tau_j}(v/2)\|,\ \|Z_{\tau_j}(0)\| \le \varepsilon, \qquad j = 0, 1.$$

Subtracting Eq. (20) at the two observed step sizes eliminates $\mathcal{N}_{\widetilde{F}}(v)$ and yields

$$\|\mathcal{N}_F(\mathrm{e}^{hB}w) - \mathcal{N}_F(w)\| \le 8\varepsilon.$$

Since $hB \in \mathcal{M}_K$ and $\|hB\| \le 2M\tau_* \le \zeta$, Eq. (13) gives

$$\|B\| \le \frac{8\varepsilon}{\sigma h}. \tag{21}$$

To control the variation with the step size, define

$$Q_t(B) := \int_0^1 B\mathrm{e}^{stB}\, ds, \qquad 0 \le t \le \Delta.$$

For $t > 0$, this equals $(\mathrm{e}^{tB} - I)/t$. For $s, t \in [0, \Delta]$, the exponential integral identity gives

$$\|\mathrm{e}^{tB} - \mathrm{e}^{sB}\| \le \|B\|\mathrm{e}^{2M\Delta}|t - s|, \tag{22}$$

$$\|Q_t(B) - Q_s(B)\| \le \frac{1}{2}\|B\|^2 \mathrm{e}^{2M\Delta}|t - s| \le M\|B\|\mathrm{e}^{2M\Delta}|t - s|. \tag{23}$$

For $v \in \mathcal{U}_{r_0}$ and $0 \le t \le \Delta$,

$$\|\mathrm{e}^{tB}v\| \le \mathrm{e}^{2M\Delta} r_0 \le \mathrm{e}^{2M\Delta} R.$$

Thus the Lipschitz bound L_Δ applies to the nonlinear terms in Eq. (19). It follows that, for $s, t > 0$,

$$\|Z_t(v) - Z_s(v)\| \le D_\Delta \|B\| |t - s|, \qquad D_\Delta := r_0(M + L_\Delta)\mathrm{e}^{2M\Delta}. \tag{24}$$

The term $\widetilde{F}(v)$ cancels in this difference.

Now fix $0 < \delta \le \Delta$ and $u \in \mathcal{U}_{r_\Delta}$. Set $v = \mathrm{e}^{\delta \widetilde{A}}u$. By Eq. (17),

$$\|v\| \le \mathrm{e}^{M\Delta} r_\Delta = r_0.$$

Moreover,

$$\|\mathrm{e}^{-\tau_j \widetilde{A}}v\| \le \mathrm{e}^{M\tau_*} r_0 \le R, \qquad j = 0, 1,$$

so $\|Z_{\tau_j}(v)\| \le \varepsilon$. Choose τ_j to be the nearer of the two observed step sizes. Equations (21) and (24) imply

$$\begin{aligned}\frac{\|S_\delta u - \widetilde{S}_\delta u\|}{\delta} &= \|Z_\delta(v)\| \\ &\le \varepsilon + D_\Delta \|B\| |\delta - \tau_j| \\ &\le \varepsilon \left(1 + \frac{8D_\Delta}{\sigma}\frac{|\delta - \tau_j|}{h}\right). \end{aligned} \tag{25}$$

Write

$$d(\delta) := \min_{\tau \in [\tau_0, \tau_1]} |\delta - \tau|.$$

The distance to the nearer endpoint satisfies

$$|\delta - \tau_j| \le \frac{h}{2} + d(\delta).$$

Therefore,

$$1 + \frac{8D_\Delta}{\sigma}\frac{|\delta - \tau_j|}{h} \le \left(1 + \frac{4D_\Delta}{\sigma}\right)\left(1 + \frac{2d(\delta)}{h}\right).$$

Taking the supremum over $u \in \mathcal{U}_{r_\Delta}$ proves Eq. (14), with

$$C_\Delta := 1 + \frac{4D_\Delta}{\sigma}.$$

All constants have the dependence stated in the theorem. □

We next verify the nonlinear separation condition. The following lemma includes the scalar condition without derivative features.

Lemma 4 (Derivative-free nonlinearities)**.** *Suppose $F(v) = P_K[f(v)]$, where $f : \mathbb{R} \to \mathbb{R}$ is locally Lipschitz. Let*

$$\mathcal{C}_f(s) := f(s) - 2f(s/2) + f(0).$$

Assume that, for some $\nu > 0$ and $I = [s_-, s_+]$ with $0 < s_- < s_+ < R/|\Omega|^{1/2}$,

$$|\mathcal{C}_f(s) - \mathcal{C}_f(t)| \ge \nu|s - t|, \qquad s, t \in I.$$

Then Eq. (13) holds for some $0 < r_0 < R$ and $\sigma, \zeta > 0$. These constants can be chosen depending only on Ω, K, R, I, ν.

Proof. Since $\mathcal{C}_f$ is continuous and injective on I, it is strictly monotone there. Hence there exists $\epsilon_f \in \{-1, 1\}$ such that

$$\epsilon_f\big(\mathcal{C}_f(s) - \mathcal{C}_f(t)\big)(s - t) \ge \nu|s - t|^2, \qquad s, t \in I. \tag{26}$$

Let

$$c := \frac{s_- + s_+}{2}, \qquad d_I := \frac{s_+ - s_-}{2}.$$

Finite dimensionality gives a constant $C_K > 0$ such that $\|v\|_{L^\infty} \le C_K\|v\|$ for $v \in X$. Choose $\eta > 0$ sufficiently small that

$$r_0 := c|\Omega|^{1/2} + \eta < R, \qquad C_K\eta \le d_I/4.$$

Next choose $0 < \zeta \le 1/2$ such that

$$C_K(\mathrm{e}^\zeta - 1)r_0 \le d_I/4.$$

Fix $H \in \mathcal{M}_K$ with $\|H\| \le \zeta$. For every real $q \in X$ with $\|q\| = 1$, the fields

$$v_\pm := c\mathbf{1} \pm \eta q$$

belong to $\mathcal{U}_{r_0}$ and take values in I. Also,

$$\|\mathrm{e}^H v_\pm - v_\pm\|_{L^\infty} \le C_K(\mathrm{e}^\zeta - 1)r_0 \le d_I/4,$$

so $\mathrm{e}^{H}v_{\pm}$ take values in I.

For either of these fields, put $w = (\mathrm{e}^{H} - I)v$. Because $w \in X$ and $\mathcal{N}_F(v) = P_K[\mathcal{C}_f(v)]$, orthogonality of P_K and Eq. (26) imply

$$\epsilon_f \langle \mathcal{N}_F(\mathrm{e}^{H}v) - \mathcal{N}_F(v), w\rangle = \epsilon_f \int_\Omega \big(\mathcal{C}_f(\mathrm{e}^{H}v) - \mathcal{C}_f(v)\big)(\mathrm{e}^{H}v - v)\,dx \geq \nu\|w\|^2.$$

Consequently,

$$\|\mathcal{N}_F(\mathrm{e}^{H}v) - \mathcal{N}_F(v)\| \geq \nu\|(\mathrm{e}^{H} - I)v\|.$$

Let $T := \mathrm{e}^{H} - I$. Choose a real unit vector $q \in X$ attaining $\|Tq\| = \|T\|$. Since $Tv_+ - Tv_- = 2\eta Tq$,

$$\max\{\|Tv_+\|, \|Tv_-\|\} \geq \eta\|T\|.$$

Finally, $\|H\| \leq 1/2$ gives

$$\|\mathrm{e}^{H} - I - H\| \leq \frac{1}{2}\mathrm{e}^{1/2}\|H\|^2 \leq \frac{1}{2}\|H\|,$$

and hence $\|T\| \geq \|H\|/2$. Combining these estimates yields

$$\sup_{v\in\mathcal{U}_{r_0}} \|\mathcal{N}_F(\mathrm{e}^{H}v) - \mathcal{N}_F(v)\| \geq \frac{\nu\eta}{2}\|H\|.$$

Thus Eq. (13) holds with $\sigma = \nu\eta/2$. □

The next lemma covers the quadratic transport nonlinearities in the KdV and advection–diffusion examples.

Lemma 5 (Quadratic transport nonlinearities). *Let $\Omega = \prod_{j=1}^{d_\Omega}(\mathbb{R}/\ell_j\mathbb{Z})$, with the Euclidean Fourier cutoff $|k| \leq K$ and $K \geq 2$. Suppose*

$$F(v) = -P_K[v(\mathbf{a}\cdot\nabla v)], \qquad \mathbf{a} \in \mathbb{R}^{d_\Omega} \setminus \{0\}.$$

Then, for every $r_0 > 0$, there exist $\sigma, \zeta > 0$ such that Eq. (13) holds.

Proof. Write $\mathcal{D} := \mathbf{a}\cdot\nabla$. Since F is homogeneous of degree two,

$$\mathcal{N}_F(v) = \frac{1}{2}F(v) = -\frac{1}{4}P_K\mathcal{D}(v^2).$$

Its derivative satisfies

$$D\mathcal{N}_F(v)[Hv] = -\frac{1}{2}P_K\mathcal{D}(vHv). \tag{27}$$

We first show that

$$D\mathcal{N}_F(v)[Hv] = 0 \quad \text{for all real } v \in X \qquad \Longrightarrow \qquad H = 0.$$

Suppose the left-hand side holds. Polarization of Eq. (27) gives

$$P_K\mathcal{D}(uHv + vHu) = 0, \qquad u, v \in X.$$

This bilinear identity extends to the complexification of X. Let $b(k)$ denote the Fourier symbol of H, and set $\alpha_j := 2\pi a_j/\ell_j$. Testing the identity on the Fourier modes indexed by p and q gives

$$\big(\boldsymbol{\alpha}\cdot(p+q)\big)\big(b(p) + b(q)\big) = 0, \qquad |p|, |q|, |p+q| \leq K. \tag{28}$$

Choose i with $\alpha_i \neq 0$. Taking $p = q = e_i$ is allowed because $K \geq 2$, and gives $b(e_i) = 0$. Taking $p = 0$, $q = e_i$ then gives $b(0) = 0$. For every retained mode k with $\boldsymbol{\alpha}\cdot k \neq 0$, taking $p = 0$, $q = k$ gives

$$b(k) = 0.$$

It remains to consider a nonzero retained mode k satisfying $\boldsymbol{\alpha} \cdot k = 0$. If $\alpha_i k_i \neq 0$ for some i, choose $p = -\operatorname{sgn}(k_i) e_i$. Then

$$\boldsymbol{\alpha} \cdot p \neq 0, \qquad |k+p|^2 = |k|^2 - 2|k_i| + 1 \leq |k|^2.$$

Thus p and $k+p$ are retained and have nonzero $\boldsymbol{\alpha}$-frequency. Equation (28) with $q = k$ therefore implies $b(k) = 0$.

Otherwise, $\alpha_i k_i = 0$ for every i. Choose i, j with $\alpha_i \neq 0$ and $k_j \neq 0$. Then $k_i = 0$, $\alpha_j = 0$, and $i \neq j$. Set $p = e_i - \operatorname{sgn}(k_j) e_j$. Now

$$|p| = \sqrt{2} \leq K, \qquad \boldsymbol{\alpha} \cdot p = \alpha_i \neq 0, \qquad |k+p|^2 = |k|^2 - 2|k_j| + 2 \leq |k|^2.$$

The same use of Eq. (28) gives $b(k) = 0$. Hence $b(k) = 0$ for every retained mode, proving $H = 0$.

By finite dimensionality and compactness, this injectivity implies

$$\kappa_K := \min_{\substack{H \in \mathcal{M}_K \\ \|H\|=1}} \sup_{\|v\| \leq 1} \|D\mathcal{N}_F(v)[Hv]\| > 0.$$

Homogeneity in v and H therefore gives

$$\sup_{\|v\| \leq r_0} \|D\mathcal{N}_F(v)[Hv]\| \geq \kappa_K r_0^2 \|H\|. \tag{29}$$

To pass to finite perturbations, introduce the continuous symmetric bilinear map

$$\mathcal{B}(v,w) := -\frac{1}{4} P_K \mathcal{D}(vw), \qquad \mathcal{N}_F(v) = \mathcal{B}(v,v).$$

Choose $C_{\mathcal{B}} > 0$ such that $\|\mathcal{B}(v,w)\| \leq C_{\mathcal{B}} \|v\| \|w\|$. For $\|H\| \leq 1$, set $T = \mathrm{e}^H - I$. The bounds

$$\|T\| \leq \mathrm{e}\|H\|, \qquad \|T - H\| \leq \frac{\mathrm{e}}{2} \|H\|^2$$

and the quadratic identity give

$$\begin{aligned} &\mathcal{N}_F(\mathrm{e}^H v) - \mathcal{N}_F(v) - D\mathcal{N}_F(v)[Hv] \\ &\qquad = 2\mathcal{B}(v, (T-H)v) + \mathcal{B}(Tv, Tv). \end{aligned}$$

Consequently,

$$\|\mathcal{N}_F(\mathrm{e}^H v) - \mathcal{N}_F(v) - D\mathcal{N}_F(v)[Hv]\| \leq C_0 \|v\|^2 \|H\|^2, \qquad C_0 := C_{\mathcal{B}}(\mathrm{e} + \mathrm{e}^2).$$

Together with Eq. (29), this yields

$$\sup_{\|v\| \leq r_0} \|\mathcal{N}_F(\mathrm{e}^H v) - \mathcal{N}_F(v)\| \geq r_0^2 (\kappa_K - C_0 \|H\|) \|H\|.$$

Taking

$$\zeta := \min\{1, \kappa_K / (2C_0)\}, \qquad \sigma := \kappa_K r_0^2 / 2$$

proves the claim. □

For the KdV equation

$$u_t + a u u_x + b u_{xxx} = 0, \qquad a \neq 0,$$

take

$$A = -b\partial_x^3, \qquad F(v) = -a P_K[v v_x].$$

For the two-dimensional equation

$$u_t = \nu(u_{xx} + u_{yy}) - u(u_x + u_y), \qquad \nu = 0.05,$$

on the periodic domain $[0,1]^2$, take

$$A = \nu(\partial_x^2 + \partial_y^2), \qquad F(v) = -P_K[v(v_x + v_y)].$$

In both cases, Lemma 5 verifies Eq. (13) for $K \geq 2$. The linear operators are bounded on X, and the projected quadratic nonlinearities are locally Lipschitz. Thus both equations provide reference

models satisfying the hypotheses of Theorem 1. The constants may depend on the fixed spatial resolution K.

Finally, we establish the sharpness assertion in Theorem 1. We show that the linear growth in

$$\frac{\min_{\tau\in[\tau_0,\tau_1]}|\delta-\tau|}{\tau_1-\tau_0}$$

cannot, in general, be replaced by sublinear growth with constants having the dependence stated in the theorem.

Take the reference model

$$A=0, \qquad F(v)=Q(v):=P_K[v^2].$$

For $f(s)=s^2$, the scalar contrast is $\mathcal{C}_f(s)=s^2/2$. On every positive interval $I=[s_-,s_+]$,

$$|\mathcal{C}_f(s)-\mathcal{C}_f(t)|=\frac{s+t}{2}|s-t|\geq s_-|s-t|.$$

Lemma 4 therefore verifies the nonlinear separation condition.

Fix the reference model, the parameters in the theorem, and the resulting constants τ_* and r_Δ. Choose

$$0<a<\tau_*/3, \qquad 0<|\gamma|\leq M,$$

and define a fixed comparison model by

$$\widetilde{A}=\gamma I, \qquad \widetilde{F}(v)=\mathrm{e}^{-2\gamma a}Q(v)+\chi v, \qquad \chi:=\frac{\mathrm{e}^{-\gamma a}-1}{a}.$$

This is a model of the same form, with a locally Lipschitz scalar nonlinearity. A direct calculation gives

$$D_t(v):=\frac{\widetilde{S}_t v-S_t v}{t}=\alpha(t)v+\beta(t)Q(v), \tag{30}$$

where

$$\alpha(t):=\frac{(1+t\chi)\mathrm{e}^{\gamma t}-1}{t}, \qquad \beta(t):=\mathrm{e}^{2\gamma(t-a)}-1.$$

Both coefficients vanish at $t=a$, so the two models agree exactly at that step size.

Now take observed steps

$$\tau_0=a, \qquad \tau_1=a+h, \qquad 0<h<a/2,$$

and let $h\downarrow 0$, keeping both models fixed. The functions α and β are continuously differentiable on $[a,3a/2]$, and Q is bounded on $\mathcal{U}_R$. Consequently, for a constant C_1 independent of h,

$$\varepsilon_{\mathrm{obs}}=\sup_{v\in\mathcal{U}_R}\|D_{a+h}(v)\|\leq C_1 h. \tag{31}$$

Consider the fixed inference step $\delta=2a\leq\Delta$. Since $Q(-v)=Q(v)$,

$$\frac{D_\delta(v)+D_\delta(-v)}{2}=\beta(\delta)Q(v).$$

Testing the constant fields $v=\pm r_\Delta/|\Omega|^{1/2}$ yields

$$\sup_{v\in\mathcal{U}_{r_\Delta}}\|D_\delta(v)\|\geq|\mathrm{e}^{2\gamma a}-1|\frac{r_\Delta^2}{|\Omega|^{1/2}}=:c_1>0. \tag{32}$$

On the other hand,

$$q_h:=\frac{\min_{\tau\in[a,a+h]}|2a-\tau|}{h}=\frac{a-h}{h}\longrightarrow\infty.$$

If the amplification factor in the theorem could be replaced by $1+\psi(q_h)$ with $\psi(q)=o(q)$, then Eq. (31) would make its right-hand side tend to zero:

$$C\varepsilon_{\mathrm{obs}}(1+\psi(q_h))\leq CC_1h(1+\psi(q_h))\longrightarrow 0.$$

This contradicts Eq. (32). All model parameters and permitted constant dependencies remain fixed as $h\downarrow 0$.

The same argument with the fixed inference step $\delta=a/2$ establishes the corresponding sharpness for extrapolation below the observed interval. Thus the linear growth order is optimal in general; no optimality of the numerical coefficient 2 is asserted.

A.1.2 PROOFS OF SECTION 2.3.2

Write $n := d_\Omega$ and $J_x\phi := \Phi_{\mathrm{st}}[\phi](x)$. Throughout this section, $\mathcal{S}_t$ denotes the original one-step map

$$\mathcal{S}_t[\phi; g, f](x) = (\mathrm{e}^{tA}\phi)(x) + tf\big(J_x\mathrm{e}^{tA}\phi, \eta(x)\big). \tag{33}$$

In particular, there is no projection of the final nonlinear output in this definition. We write the Fourier modes as $\mathrm{e}^{\mathrm{i}k\cdot x}$ on the normalised torus. On a periodic box with side lengths ℓ_j, replace k_j by $2\pi k_j/\ell_j$ in the differential symbols. This invertible rescaling preserves the polynomial degrees and the arguments below; the bandwidth bounds refer to integer mode indices.

Proof of Lemma 2. Let $v = \mathrm{e}^{\tau_0 A}\phi$. Since $g_\gamma = g + \gamma$, the corresponding linear operator is $A_\gamma = A + \gamma I$, and hence

$$\mathrm{e}^{\tau_0 A_\gamma}\phi = \varrho_\gamma v, \qquad J_x(\varrho_\gamma v) = \varrho_\gamma J_x v.$$

Substitution into Eq. (33) gives

$$\begin{aligned}\mathcal{S}_{\tau_0}[\phi; g_\gamma, f_\gamma](x) &= \varrho_\gamma v(x) + \tau_0 f(J_x v, \eta(x)) + \tau_0\chi_\gamma\varrho_\gamma v(x)\\ &= v(x) + \tau_0 f(J_x v, \eta(x)),\end{aligned}$$

where the last equality uses $\varrho_\gamma(1 + \tau_0\chi_\gamma) = 1$. This proves Eq. (16). Repeated composition also gives equality of all rollouts for which both iterates are defined. □

For the identifiability theorem, the following explicit choice of bandwidth suffices:

$$\begin{aligned}D &:= \max\{1, d\},\\ N_* &:= \max\left\{(2\rho+1)D + 2\rho + 1,\ \rho + (\rho+2)(D+2\rho+1)\right\},\\ T_* &:= \left\lceil \frac{N_* - 1}{2} \right\rceil,\\ K_0 &:= \left\lceil \sqrt{n}\,\max\{\rho, DT_*\} \right\rceil.\end{aligned} \tag{34}$$

We first establish the two properties supplied by this threshold.

Lemma 6 (Uniform realisation of local features)**.** *If $K \geq \rho\sqrt{n}$, then $J_x : \mathcal{V}_K \to \mathbb{R}^{m_{\mathrm{st}}}$ is surjective for every $x \in \Omega$. Moreover, there are linear right inverses $\mathcal{R}_x : \mathbb{R}^{m_{\mathrm{st}}} \to \mathcal{V}_K$ satisfying*

$$J_x\mathcal{R}_x = I, \qquad C_J := \sup_{x\in\Omega}\|\mathcal{R}_x\| < \infty,$$

where the operator norms use the Euclidean norm on the feature space and the L^2 norm on $\mathcal{V}_K$.

Proof. Suppose a real linear functional annihilates the range of J_x. Writing its coefficients as a_α, this means

$$\sum_{|\alpha|\leq\rho} a_\alpha\,\partial^\alpha\phi(x) = 0, \qquad \phi \in \mathcal{V}_K.$$

Complexifying this real linear identity and testing Fourier modes gives

$$\sum_{|\alpha|\leq\rho} a_\alpha(\mathrm{i}k)^\alpha = 0, \qquad k \in \mathbb{Z}^n, \quad |k| \leq K.$$

The retained frequencies contain the tensor grid $\{0, \dots, \rho\}^n$. A polynomial of degree at most ρ in each coordinate that vanishes on this grid is identically zero, by repeated univariate interpolation. Thus all a_α vanish, proving surjectivity.

Choose a linear right inverse at one point. Translating its output produces right inverses at all other points, because differentiation commutes with translations. Translations preserve the L^2 norm, so these right inverses have the same operator norm. □

The next lemma supplies the finite-frequency rigidity needed to identify the differential symbol.

Lemma 7 (Finite-frequency rigidity)**.** *Let H, P, Q be complex polynomials in n real variables, with degrees at most d, ρ, ρ, respectively, and suppose $P(0) \neq 0$. Assume that, for every $k \in \mathbb{Z}^n$ with $|k| \leq K$,*

$$\mathrm{e}^{H(k)} P(k) = Q(k), \qquad |\operatorname{Im} H(k)| < \pi/2.$$

If $K \geq K_0$ with K_0 given by Eq. (34), then H is constant.

Proof. We first prove a univariate statement. Let h, p, q be complex polynomials of degrees at most D, ρ, ρ, with p and q nonzero. Suppose that at least N_* distinct real points t_j satisfy

$$\mathrm{e}^{h(t_j)} p(t_j) = q(t_j), \qquad |\operatorname{Im} h(t_j)| < \pi/2. \tag{35}$$

We show that h is constant.

Write $h(t) = a(t) + \mathrm{i}b(t)$ for real t, where a and b are real polynomials. Taking squared moduli in Eq. (35) gives zeros of

$$R(t) := \mathrm{e}^{2a(t)} |p(t)|^2 - |q(t)|^2.$$

Here $|p(t)|^2$ and $|q(t)|^2$ are real polynomials in t of degrees at most $m := 2\rho$.

Suppose first that a is nonconstant. Differentiating R exactly $m+1$ times eliminates the second polynomial and gives

$$R^{(m+1)}(t) = \mathrm{e}^{2a(t)} W(t), \qquad \deg W \leq m + (m+1)(D-1).$$

The polynomial W is nonzero: at each differentiation, the polynomial factor is transformed by $U \mapsto U' + 2a'U$, which maps every nonzero polynomial to a nonzero polynomial when a' is a nonzero polynomial. Repeated Rolle's theorem therefore bounds the number of distinct real zeros of R by

$$m + (m+1)(D-1) + (m+1) = m + (m+1)D.$$

This contradicts the definition of N_*. Hence $a(t) = a_0$ is constant. The polynomial $|q(t)|^2 - \mathrm{e}^{2a_0}|p(t)|^2$ then has more zeros than its degree, so

$$|q(t)|^2 = \mathrm{e}^{2a_0} |p(t)|^2 \qquad \text{for all real } t. \tag{36}$$

Cancel the common polynomial factors in q/p and write the result as U/V, with U and V coprime. For a polynomial W, let $W^{\#}(z) := \overline{W(\overline{z})}$. Equation (36) implies the polynomial identity

$$UU^{\#} = \mathrm{e}^{2a_0} VV^{\#}.$$

In particular, $\deg U = \deg V$. Coprimality implies that U divides $V^{\#}$, and therefore

$$U = CV^{\#}, \qquad |C| = \mathrm{e}^{a_0}.$$

The polynomial V has no real roots, since any such root would also be a root of U. Consequently, the reduced rational function admits a smooth argument on the real line:

$$\mathrm{e}^{-a_0} \frac{U(t)}{V(t)} = \mathrm{e}^{\mathrm{i}\theta(t)}.$$

If the roots of V are $z_1, \ldots, z_\ell$, counted with multiplicity, then $\ell \leq \rho$, $\operatorname{Im} z_j \neq 0$, and

$$\theta'(t) = -2 \sum_{j=1}^{\ell} \frac{\operatorname{Im} z_j}{(t - \operatorname{Re} z_j)^2 + (\operatorname{Im} z_j)^2}. \tag{37}$$

In particular,

$$\int_{\mathbb{R}} |\theta'(t)|\, dt \leq 2\pi\rho.$$

At most ρ of the sample points are zeros of p. At each remaining point, Eq. (35) gives

$$\theta(t_j) - b(t_j) \in 2\pi\mathbb{Z}.$$

The range of θ has width at most $2\pi\rho$, while $b(t_j) \in (-\pi/2, \pi/2)$. Thus at most $\rho + 2$ distinct integers can occur in the preceding identity. By the definition of N_*, one integer s occurs at least $D + 2\rho + 1$ times. The function

$$\theta(t) - b(t) - 2\pi s$$

therefore has at least that many distinct zeros.

By Eq. (37), $\theta' - b'$ is a rational function with a strictly positive real denominator of degree at most 2ρ and a numerator of degree at most $D + 2\rho - 1$. Rolle's theorem supplies at least $D + 2\rho$ distinct zeros of this numerator, so it vanishes identically. Hence $b' = \theta'$ on $\mathbb{R}$. Since $\theta'(t) \to 0$ as $|t| \to \infty$, the polynomial b' must be zero. Both a and b are constant, proving the univariate statement.

We now return to the multivariate polynomials. The identity at $k = 0$ implies $Q(0) \neq 0$. For each nonzero direction $v \in \{0, \dots, D\}^n$, restrict the polynomials to

$$h_v(t) := H(tv), \qquad p_v(t) := P(tv), \qquad q_v(t) := Q(tv).$$

The polynomials p_v and q_v are nonzero because their values at zero are nonzero. For every integer t with $|t| \le T_*$,

$$|tv| \le T_* D\sqrt{n} \le K.$$

Thus these restrictions satisfy the univariate hypotheses at $2T_* + 1 \ge N_*$ distinct points. It follows that $H(tv)$ is constant in t, and in particular $H(v) = H(0)$.

Therefore $H - H(0)$ vanishes on the tensor grid $\{0, \dots, D\}^n$. Its degree in each coordinate is at most D, so repeated univariate interpolation gives $H \equiv H(0)$. □

Proof of Theorem 3. We first prove part (i). Choose x_0 as in its nondegeneracy assumption and set $\eta_0 := \eta(x_0)$. Differentiate the equality of the prediction maps at the zero input and evaluate at x_0. For every real $\psi \in \mathcal{V}_K$, this gives

$$\begin{aligned}(\mathrm{e}^{\tau_0 A}\psi)(x_0) + \tau_0 \sum_{|\alpha|\le\rho} \partial_{z_\alpha} f(0,\eta_0)\, \partial^\alpha (\mathrm{e}^{\tau_0 A}\psi)(x_0) \\ = (\mathrm{e}^{\tau_0 A'}\psi)(x_0) + \tau_0 \sum_{|\alpha|\le\rho} \partial_{z_\alpha} f'(0,\eta_0)\, \partial^\alpha (\mathrm{e}^{\tau_0 A'}\psi)(x_0).\end{aligned}$$

Define the polynomials

$$P(k) := 1 + \tau_0 \sum_{|\alpha|\le\rho} \partial_{z_\alpha} f(0,\eta_0)(\mathrm{i}k)^\alpha, \qquad Q(k) := 1 + \tau_0 \sum_{|\alpha|\le\rho} \partial_{z_\alpha} f'(0,\eta_0)(\mathrm{i}k)^\alpha.$$

Complexifying the linear identity and testing Fourier modes yields

$$\mathrm{e}^{\tau_0 g(k)} P(k) = \mathrm{e}^{\tau_0 g'(k)} Q(k), \qquad |k| \le K. \tag{38}$$

The nondegeneracy assumption states precisely that

$$P(0) = 1 + \tau_0 \partial_{z_0} f(0,\eta_0) \neq 0.$$

Apply Lemma 7 to

$$H(k) := \tau_0 \big(g(k) - g'(k)\big).$$

Indeed, Eq. (38) gives $\mathrm{e}^{H(k)} P(k) = Q(k)$, and the two anti-aliasing bounds imply

$$|\operatorname{Im} H(k)| \le \tau_0 |\operatorname{Im} g(k)| + \tau_0 |\operatorname{Im} g'(k)| < \pi/2.$$

Hence $g - g'$ is constant. Since both differential symbols have real coefficients, their values at zero are real. Thus, for some $\gamma \in \mathbb{R}$,

$$g' = g + \gamma, \qquad A' = A + \gamma I.$$

Equality of the positive-order polynomial coefficients follows immediately.

It remains to identify the nonlinear function on a common feature neighbourhood. Let $\mathcal{R}_x$ and C_J be as in Lemma 6. For any step size t at which the prediction maps agree on the stated input ball, define

$$r_t := \frac{r}{2C_J \|\mathrm{e}^{-tA'}\|_{\mathcal{V}_K \to \mathcal{V}_K}}.$$

This is positive because the exponential is invertible on the finite-dimensional space $\mathcal{V}_K$. For $|z_{\mathrm{st}}| < r_t$, take

$$\phi := \mathrm{e}^{-tA'}\mathcal{R}_x z_{\mathrm{st}}.$$

Then $\|\phi\|_{L^2} < r$, and

$$J_x \mathrm{e}^{tA'}\phi = z_{\mathrm{st}}, \qquad J_x \mathrm{e}^{tA}\phi = \mathrm{e}^{-\gamma t} z_{\mathrm{st}}.$$

Equality of the two prediction maps at x therefore gives

$$f'(z_{\mathrm{st}}, \eta(x)) = f(\mathrm{e}^{-\gamma t} z_{\mathrm{st}}, \eta(x)) + \frac{\mathrm{e}^{-\gamma t} - 1}{t}\, z_0, \qquad |z_{\mathrm{st}}| < r_t. \tag{39}$$

The radius r_t is independent of x. Since $E = \eta(\Omega)$, this identity holds for every $z_\eta \in E$. Taking $t = \tau_0$ and $r_* = r_{\tau_0}$ proves $f' = f_\gamma$ on $B_{r_*}(0) \times E$, completing part (i).

For part (ii), put

$$a_j := \mathrm{e}^{-\gamma \tau_j}, \qquad \chi_j := \frac{a_j - 1}{\tau_j}, \qquad j = 0, 1.$$

Equation (39) applies at both observed step sizes. Thus, on a common neighbourhood of the origin and for every $z_\eta \in E$,

$$f(a_0 z_{\mathrm{st}}, z_\eta) + \chi_0 z_0 = f(a_1 z_{\mathrm{st}}, z_\eta) + \chi_1 z_0. \tag{40}$$

Suppose $\gamma \neq 0$, and choose z_η^* as in the non-affinity assumption. Since $\tau_0 \neq \tau_1$,

$$\lambda := \frac{a_0}{a_1} > 0, \qquad \lambda \neq 1.$$

Setting $w = a_1 z_{\mathrm{st}}$ in Eq. (40) gives

$$f(\lambda w, z_\eta^*) - f(w, z_\eta^*) = b\, w_0, \qquad b := \frac{\chi_1 - \chi_0}{a_1},$$

for all w sufficiently close to zero. Define

$$H_*(w) := f(w, z_\eta^*) - \frac{b}{\lambda - 1} w_0.$$

Then $H_*(\lambda w) = H_*(w)$ locally. Writing $\mu := \min\{\lambda, \lambda^{-1}\} \in (0, 1)$, this implies, on a sufficiently small ball,

$$H_*(\mu w) = H_*(w).$$

Iteration and continuity at zero yield

$$H_*(w) = H_*(\mu^m w) \longrightarrow H_*(0) \qquad (m \to \infty).$$

Consequently,

$$f(w, z_\eta^*) = f(0, z_\eta^*) + \frac{b}{\lambda - 1} w_0$$

on a neighbourhood of the origin. This contradicts the assumed non-affinity. Therefore $\gamma = 0$. Part (i) now gives $g' = g$ and $f' = f$ on its original feature domain $B_{r_*}(0) \times E$. □

A.2 Numerical Experiments

A.2.1 Main ENNO hyperparameters

Tables 9 and 10 list, for each of the four experiments, the architecture of ENNO and the data it was trained and evaluated on. Throughout, K is the Euclidean cutoff on integer mode indices, $|k| \leq K$, as in Section 2.3.1; d is the degree of the symbol g in Eq. (6); ρ is the order of the jet entering the nonlinear map f in Eq. (8); and v denotes the linearly propagated field to which f is applied in Eq. (9). The coefficients of the even monomials in $\mathrm{Re}\, g$ (Section 2.2) are parametrised as $-\mathrm{softplus}(\theta)$, so that $\mathrm{Re}\, g \leq c_0$ holds by construction; the odd-order coefficients and c_0 are unconstrained. One consequence is that a constrained coefficient whose true value is zero cannot reach zero exactly and settles at the floor of the parametrisation; this is the case for c_2 in Table 2. The unconstrained coefficients are initialised from $\mathcal{N}(0, 0.01^2)$. The nonlinear map f and, where they are not the identity, the encoder E and the decoder R are multilayer perceptrons with Tanh

activations; the output layer of f is zero-initialised, so that training starts from the purely linear flow.

ENNO is trained with Adam for at most 3,000 epochs (the cap applied to all neural models in Section 3), and its checkpoint is selected by early stopping on the validation set (patience 15 for the synthetic problems, evaluated on the noised validation trajectories, and 200 for the real data; for WildfireSpreadTS the criterion is validation AUC-PR, Section 3.4). All reported results come from a single training seed per configuration (per fold for WildfireSpreadTS). Relative L^2 errors on the synthetic problems are computed against the clean, noise-free test trajectories; the targets at $t_{1/2}$ come from the same solver runs at the finer output interval (Table 10).

Table 9: Architecture of ENNO in the four experiments.

	KdV	Advection–diffusion	SEVIR (VIL)	WildfireSpreadTS
Latent channels M	1	1	20	4 ($T_{\text{in}} = 1$) / 20 ($T_{\text{in}} = 5$)
History length (frames)	1	1	10	$T_{\text{in}} \in \{1, 5\}$
Encoder E / decoder R	identity (learnable variants in Table 11)	identity	pointwise MLPs $\mathbb{R}^{10} \to \mathbb{R}^{20}$ and $\mathbb{R}^{20} \to \mathbb{R}$, 5 hidden layers $\times$ 256	same design, $\mathbb{R}^{T_{\text{in}}} \to \mathbb{R}^{M}$ and $\mathbb{R}^{M} \to \mathbb{R}$, 5 hidden layers $\times$ 256
Cutoff K ($\lvert k\rvert \le K$)[a]	64	64	64	32
Symbol degree d	3 (ablation over 1–4, Table 1)	2 (ablation over 1–3, Table 4)	6	6
Constant term c_0	trained	trained	fixed at 0	fixed at 0
Jet order ρ	2	2	1	1
Inputs of f	v, v_x, v_{xx}	v, v_x, v_y, v_{xx}, v_{yy}, v_{xy}	v, v_x, v_y	v, v_x, v_y and 25 environmental channels η
Size of f	4 hidden layers $\times$ 128	4 $\times$ 128	4 $\times$ 256	4 $\times$ 256

[a]Euclidean cutoff on integer mode indices, matching V_K in Section 2.3.1. On the 128-point and 128 $\times$ 128 grids, $K = 64$ is the Nyquist index (in two dimensions the corners of the square spectrum beyond radius 64 are dropped); for WildfireSpreadTS, $K = 32$ is a genuine low-pass.

A.2.2 Ablation results of the KdV example

The encoder and decoder of Eq. (12) are identity maps in the configuration reported above, so the model acts directly on u. Table 11 makes them learnable, with M latent fields, at each symbol degree. Where the degree is below the true order, the learnable variant with $M \ge 4$ is lower at every target time, and the gain saturates between $M = 4$ and $M = 16$: at $d = 1$ it reaches 2.449% at t_{10} against 4.189% for the identity. At $d \ge 3$, where the true symbol is already in the class, the pattern splits: the identity is lower at t_1 and t_5, by a factor of 2.7 at t_1, while the learnable variant grows more slowly and ends slightly lower at t_{10}. We report the identity configuration, which is the more accurate of the two over most of the horizon.

Table 10: Data used in the four experiments.

	KdV	Advection–diffusion	SEVIR	WildfireSpreadTS
Domain and grid	$[0, 10]$ periodic, 128 points	$[0, 1]^2$ periodic, 128×128	$384 \times 384\,\mathrm{km}^2$; VIL on 384^2 pixels, average-pooled to 128^2	daily fire maps at 375 m, padded to 360×320
Source	Fourier pseudospectral in space, DOP853 in time (rtol 10^{-8}); 10,001 frames over $T = 1$, every 10th kept	explicit Euler, $\Delta t = 10^{-5}$, central differences in space; 5,001 frames over $T = 0.5$, every 10th kept	real radar, 49 frames per event	real observations, 2018–2021
Initial conditions	3 sine modes (amplitude $\mathcal{U}(-1, 1)/m^2$ for mode m) plus 1 or 2 solitons (amplitude $\mathcal{U}(1.5, 3.5)$)	sums of 1–3 Gaussian bumps (amplitude $\mathcal{U}(0.5, 1.5)$, $\sigma \sim \mathcal{U}(0.05, 0.25)$) and low-mode Fourier fields	—	—
Observation interval	10^{-3}	10^{-3}	5 min	1 day
Training step sizes τ_1, τ_2	$10^{-3},\ 2 \times 10^{-3}$	$10^{-3},\ 2 \times 10^{-3}$	5, 10 min	1, 2 days
Targets at $t_{1/2}$	every 5th frame of the same runs	every 5th frame of the same runs	—	—
Train / val / test	2,030 / 50 / 200 trajectories	100 / 10 / 25 trajectories	11,159 / 200 / 500 events	four leave-one-year-out folds
Noise (train and val, all methods)	i.i.d. Gaussian, 5% of each trajectory's std, fixed seed; test trajectories clean	same	none	none

Table 11: KdV: encoder/decoder ablation. Relative L^2 error (%) at three target frames. "Identity" is the parameter-free configuration reported in the main text; the other rows make them learnable, with M latent fields. Best in each block in bold.

Variant	t_1	t_5	t_{10}
$d = 1$, identity	2.898	3.683	4.189
$d = 1$, $M = 1$	3.069	2.629	3.276
$d = 1$, $M = 4$	**1.866**	2.317	2.468
$d = 1$, $M = 16$	1.877	**2.289**	**2.449**
$d = 2$, identity	2.181	2.700	2.994
$d = 2$, $M = 1$	1.990	2.454	2.856
$d = 2$, $M = 4$	**1.532**	**2.233**	**2.529**
$d = 2$, $M = 16$	1.535	2.235	**2.529**
$d = 3$, identity	**0.313**	**0.721**	1.115
$d = 3$, $M = 1$	0.852	0.915	**0.994**
$d = 3$, $M = 4$	0.846	0.939	1.028
$d = 3$, $M = 16$	0.842	0.935	1.014
$d = 4$, identity	**0.318**	**0.738**	1.152
$d = 4$, $M = 1$	0.846	0.920	**1.007**
$d = 4$, $M = 4$	0.840	0.968	1.086
$d = 4$, $M = 16$	0.832	0.963	1.098