MISSPECIFIED PHYSICS CAN HELP AND BIAS EXTRAPOLATION IN HYBRID MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid models that combine physical equations with a data-driven component are expected to extrapolate more reliably than purely data-driven models as the physical component remains active beyond the training data. This study investigates whether a misspecified physical component can still provide useful extrapolative structure beyond the training data and whether its misspecification shapes the extrapolation error of the corresponding hybrid model. In a controlled reaction-diffusion system, this study compares a hybrid model with correct diffusion, hybrid models whose supplied diffusion is misspecified in its directional dependence, its discretization, its magnitude, or by omitting one process, and a purely data-driven model. All models achieve small held-out prediction errors and give similar responses to patterns well represented during training, yet their responses differ when extrapolating to patterns scarcely represented in the training data. For scarcely represented patterns, the hybrid model with correct diffusion closely follows the true response, and moderately misspecified physics gives lower errors than the purely data-driven model, while sufficiently severe misspecification loses that advantage. Hybrid models with misspecified physics carry the specific response errors of their supplied physics into extrapolation, with nearly the same magnitude and response pattern. Intervention analysis attributes the extrapolation error mainly to the misspecified physical component, as replacing the misspecified physics with the correct physics, without retraining the data-driven component, removes most of the error. A misspecified physical component can therefore provide useful extrapolative structure while also imposing a systematic extrapolation bias that follows the response error of the supplied physics.

1 INTRODUCTION

Hybrid models, which combine a physical component derived from knowledge of a system with a data-driven component learned from data, are increasingly used to predict complex natural and engineered systems (Karpatne et al., 2024; Kasilingam et al., 2024; Kochkov et al., 2024; Li & Hu, 2026). These models are expected to provide reliable predictions when applied to conditions not well represented by available observations, referred to broadly as extrapolation. This expectation has been supported by studies showing that hybrid models outperform physics-based and data-driven benchmarks in terms of extrapolability (Sansana et al., 2024; Abbate et al., 2025; Malekjani et al., 2026; Yan et al., 2026). Schweidtmann et al. (2024) review evidence that incorporating fundamental knowledge improves predictions beyond previously tested conditions, while noting that the extrapolation benefit depends on how the hybrid model is constructed and on the range of conditions considered.

In practice, available physical knowledge may be incomplete or imperfect. Some studies report that imperfect physics still improve predictive and extrapolative performance (Pathak et al., 2018; Wan et al., 2018; Gehra & Günthner, 2026). Errors in supplied physical knowledge, however, can bias the hybrid model. A misspecified physical component remains useful while requiring substantially more empirical adjustment from the data-driven component than a well-specified physical component (Overgaard et al., 2026). The extrapolative advantage of a hybrid model shrinks as the physical component becomes less informative, and in some extrapolation cases, hybrid models built on the simplified mechanistic understanding are even less accurate than a purely data-driven model (Duncan & Räth, 2023; Carter et al., 2025; Geremia et al., 2025).

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.

Existing studies have not fully resolved how errors in the physical component shape the corresponding extrapolation error of the trained hybrid model. Specifically, it remains unclear 1) whether a known error in the supplied physical response can be quantitatively traced into the response of the hybrid model after training, or 2) whether the extrapolation error in the hybrid model can be attributed to the physical component. Addressing these questions is also necessary to interpret why the extrapolation advantage of a hybrid model deteriorates as physical misspecification becomes sufficiently severe. The loss of benefit could reflect the systematic error introduced by the misspecified physical component, the empirical adjustment made by the data-driven component to compensate for that error, or their interaction.

To address these questions, this study combines theoretical analysis with controlled experiments in a nonlinear reaction-diffusion system. An exact result for an idealized linear additive hybrid model is first derived to characterize how differences between supplied physical components are reflected in trained hybrid models after finite training. The corresponding behavior is then examined empirically in a FitzHugh-Nagumo reaction-diffusion system with known true dynamics and controlled physical misspecification. The intervention analysis is further applied to isolate the contribution of misspecified physics to extrapolation errors of the hybrid model. The magnitude of the misspecification is varied to investigate how severe the misspecification can become before the hybrid model loses its extrapolation accuracy advantage over the purely data-driven model.

The contributions of this study are as follows:

1) Derives an exact finite-training relation between differences in supplied physical components and differences between trained hybrid models.

2) Shows that, on scarcely represented patterns, hybrid models carry the specific response errors of misspecified physical components with nearly the same magnitude and direction.

3) Attributes most of extrapolation error to the misspecified physical component and demonstrates that sufficiently severe misspecification removes the advantage over a purely data-driven model.

2 Related work

Extrapolation beyond conditions represented in training data. In many scientific applications, the prediction task requires models to be applied under conditions not represented in the training data (Meyer & Pebesma, 2021). For spatial extrapolation, differentiable hydrological models have outperformed purely data-driven models when entire regions were excluded from training (Feng et al., 2023). Temporal extrapolation creates a similar problem when models trained under historical conditions are applied to a changing climate. Hybrid hydrological models trained on a cooler period have shown a smaller decline in accuracy when evaluated under warmer conditions, although the purely data-driven model retained higher overall streamflow accuracy (Bohl et al., 2025). Extrapolation to extreme events has also been evaluated by excluding years containing extreme flows from training (Acuña Espinoza et al., 2025). In atmospheric modeling, a data-driven convection parameterization trained only on a control climate did not reproduce the response to a warmer climate, whereas including examples from the warmer climate enabled that response to be learned (O'Gorman & Dwyer, 2018). These studies show that spatial, temporal, and event-based extrapolation can require predictions under conditions that are absent or weakly represented in the training data. They therefore motivate evaluating data-driven and hybrid models under the types of distribution shifts that arise from the scientific prediction task itself. However, they do not systematically establish how known errors in the supplied physics are reflected in the response of a trained hybrid model under conditions scarcely represented during training.

Hybrid models with imperfect physics and adjustment by the data-driven component. Hybrid models can use physical knowledge to reduce what must be learned from observations, while allowing a data-driven component to account for incomplete or inaccurate physics (Schweidtmann et al., 2024). Imperfect physics can still improve predictive or extrapolative performance (Pathak et al., 2018; Wan et al., 2018; Gehra & Günthner, 2026), but this benefit can decline as the physical component becomes less informative, and sufficiently poor mechanistic information can remove the advantage over a purely data-driven model (Duncan & Räth, 2023; Carter et al., 2025; Geremia et al., 2025). In additive formulations, the data-driven component is trained to account for behavior that the physical component does not reproduce, through corrections to the governing equations or their predictions

(Farchi et al., 2021; Yin et al., 2021; Blakseth et al., 2022; Garg et al., 2022). Other formulations place the data-driven component within the physical model to estimate parameters, represent unresolved processes, or modify incompletely known relationships (Koppa et al., 2022; Giampiccolo et al., 2024; Schweidtmann et al., 2024). Accurate predictions can nevertheless coexist with substantial adjustment for misspecified physics. In hybrid hydrological models, a neural network can vary the parameters of an inadequate conceptual model while maintaining predictive accuracy (Álvarez Chaves et al., 2026). Similarly, comparisons of mechanistic and empirical contributions show that a misspecified physical component can remain useful while requiring greater empirical adjustment than a well-specified physical component (Overgaard et al., 2026). More generally, allowing departures from imposed structural assumptions can improve prediction when those assumptions are imperfect (Finzi et al., 2021; Wang et al., 2022). Previous work has also reduced ambiguity between physical and data-driven contributions through constrained decompositions, regularization, or orthogonal discrepancy formulations (Yin et al., 2021; Takeishi & Kalousis, 2021; Manna et al., 2026). These approaches constrain or diagnose how the two components share the fitted behavior. What remains unclear is whether, after ordinary hybrid training, a known error in the supplied physical component remains quantitatively visible in the response of the complete hybrid model, and whether the resulting extrapolation error can be attributed to that component by changing it while leaving the trained data-driven component unchanged.

Learning from unevenly represented training data. Finite observations do not provide equal information about all variations of a dynamical system. What can be inferred depends on the variation present in the data and on the model properties to be determined (Schoukens & Ljung, 2019; Van Waarde et al., 2020). Prediction error on held-out samples also does not fully characterize whether a learned dynamical model responds correctly to changes in its state; small regression error can coexist with incorrect dynamical behavior (Park et al., 2024). In an idealized linear setting, Floryan (2024) shows slower convergence along directions associated with smaller eigenvalues of the empirical second-moment matrix, while the dynamics in directions absent from the training data cannot be determined from those data alone. Analyses of wide neural networks likewise relate convergence rates to eigenvalues of the neural tangent kernel, under the assumptions of the respective analyses (Jacot et al., 2018; Cao et al., 2021). For the nonuniform sampling distributions they study, Basri et al. (2020) show that learning can be slower in regions with a lower density of training examples. Limited representation, however, does not necessarily prevent generalization. Learning from one example can affect predictions for other inputs: to first order in the parameter update, this influence depends on the similarity between the gradients of the outputs with respect to the model parameters and on the prediction error for the training example (Ren, 2025). For the dynamical systems they study, Norton et al. (2025) show that a data-driven model trained on trajectories from one basin of attraction can reproduce behavior in an unobserved basin. These results suggest that limited representation of a pattern can restrict the information available for learning its response, while adjustments learned from other patterns may still generalize to it. For a hybrid model with imperfect physics, the unresolved question is therefore how errors in the supplied physics and the information available for learning their adjustment jointly shape the response of the hybrid model to patterns scarcely represented during training.

3 METHODS

3.1 LINEAR ANALYSIS

Let $F(x) = Ax$ for $x \in \mathbb{R}^d$ denote the true map, and let $\mathcal{H}_s(x) = (P_s + W_s)x$ be an additive hybrid model approximating F, where $A, P_s, W_s \in \mathbb{R}^{d\times d}$. The supplied physical operator P_s is fixed, and the data-driven component W_s is trained on

$$\mathcal{L}_s(W) = \frac{1}{2n}\sum_{i=1}^{n} \|(P_s + W)x_i - Ax_i\|_2^2. \tag{1}$$

Define the empirical second-moment matrix

$$M = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top. \tag{2}$$

It is symmetric and positive semidefinite; for zero-mean inputs, it coincides with the empirical covariance matrix. Train two models using P_s and P_0 by full-batch gradient descent with the same step size $\eta > 0$, inputs, targets, and initialization $W_{s,0} = W_{0,0}$, without regularization. After t updates, write

$$\Delta P = P_s - P_0, \qquad \Delta\mathcal{H}_t = (P_s + W_{s,t}) - (P_0 + W_{0,t}). \tag{3}$$

Proposition 1. *Under these assumptions,*

$$\boxed{\Delta\mathcal{H}_t = \Delta P(I - \eta M)^t.} \tag{4}$$

The proof is in Appendix A. Equation 4 shows how the difference between the two hybrid models after a finite number of training steps depends on the training data (Floryan, 2024). Let $q \in \mathbb{R}^d$ be an eigenvector of M with eigenvalue λ_q, so that $Mq = \lambda_q q$. Then the difference between the two trained hybrid models when applied to q is

$$\Delta\mathcal{H}_t q = (1 - \eta\lambda_q)^t \Delta P q. \tag{5}$$

For $0 < \eta\lambda_q < 1$, this factor decays faster for larger λ_q. When $\lambda_q = 0$, the supplied difference remains unchanged. If $M \neq 0$, all positive-eigenvalue directions contract for $0 < \eta < 2/\lambda_{\max}(M)$; if $M = 0$, every difference remains unchanged.

For a fixed matrix B, consider $P_a = P_0 + aB$. At a common update count t, $\Delta\mathcal{H}_{a,t} = aB(I - \eta M)^t$. For $a \neq 0$ and $Bq \neq 0$, the projection coefficient is

$$\beta_{a,t}(q) = \frac{\langle \Delta\mathcal{H}_{a,t} q, aBq \rangle}{\|aBq\|_2^2} = \frac{\langle B(I - \eta M)^t q, Bq \rangle}{\|Bq\|_2^2}. \tag{6}$$

Thus, the difference between the responses of the two hybrid models is proportional to a, while the projection coefficient does not depend on a.

Appendix B derives a conditional extension for a nonlinear network under linearization with respect to its parameters at initialization, including a result for directional responses. The following experiment examines whether analogous dependence appears in nonlinear hybrid models.

3.2 Controlled nonlinear experiment

The reference system is the FitzHugh-Nagumo reaction-diffusion system (FitzHugh, 1961; Nagumo et al., 1962; Cebrián-Lacasa et al., 2024),

$$\frac{\partial u}{\partial t} = D_u \nabla^2 u + u - u^3 - k - v, \qquad \frac{\partial v}{\partial t} = D_v \nabla^2 v + u - v, \tag{7}$$

with $D_u = 0.005$, $D_v = 0.01$, and $k = 0.005$. Diffusion derivatives are evaluated spectrally on a periodic 64×64 grid over a square domain of side length 2. Reference trajectories are integrated with fourth-order Runge-Kutta (RK4) method using a solver step of 0.002, with states recorded every 0.01 time units. This discrete system is used as the reference throughout the experiments.

For $x = (u, v)$, let s index the supplied physical specification, with $s = 0$ denoting the correct diffusion. The corresponding hybrid represents the right-hand side as

$$H_s(x) = P_s(x) + C_s(x), \qquad P_0(u, v) = (D_u \nabla^2 u, D_v \nabla^2 v), \tag{8}$$

where the data-driven component C_s is a convolutional network trained separately for each supplied physical component. In this study, as diffusion is the only supplied physical process, C_s learns the reaction dynamics together with any adjustment needed for errors in the supplied diffusion. The supplied diffusion is misspecified in its directional dependence, its discretization, its magnitude, or by omitting diffusion from the v equation (Appendix C). The purely data-driven model predicts the full right-hand side with the same network architecture but no supplied physical component.

Initial fields are independent Gaussian random fields filtered to Fourier radius at most 6 and rescaled to spatial standard deviation 0.5. Independent initial conditions generate 80 training, 24 validation, and 64 test trajectories, each containing 80 recorded states. During training, each model uses the same RK4 solver to predict the state 0.01 time units later, with mean squared error as the loss. Five seeds use paired network initialization and identical minibatch ordering across model conditions. Appendix D provides the network architecture and training details.

3.3 EXTRAPOLATION AND RESPONSE MEASUREMENTS

After training, a held-out state is perturbed along a sine or cosine Fourier pattern q in either u or v, indexed by (m_x, m_y) with $\rho = \sqrt{m_x^2 + m_y^2}$. Patterns are scaled so that adding ϵq changes the corresponding standardized coefficient by ϵ. Representation of each pattern in the training data is measured by the root mean square (RMS) and largest absolute coefficient across the 6,320 training inputs. For the principal $\rho = 6$ patterns, $\epsilon = 0.01$ is below the largest observed coefficient, whereas for the $\rho = 20$ patterns it exceeds the largest observed coefficient by at least two orders of magnitude. They are referred to as well-represented and scarcely represented patterns, respectively. The governing equations, parameters, domain, and numerical resolution remain unchanged. Appendix E provides the normalization, held-out states, and complete pattern sets.

For a differentiable function G, write $DG(x)q$ for its directional derivative at state x along pattern q. For a misspecified physical component P_s and its corresponding trained hybrid model H_s, define

$$p_s(q) = (P_s - P_0)q, \qquad h_s(x, q) = D(H_s - H_0)(x)q. \tag{9}$$

Here, p_s is the difference between the responses of P_s and P_0 to q, and h_s is the corresponding difference between the responses of H_s and H_0. For nonzero denominators, define

$$\beta_s(x, q) = \frac{\langle h_s(x, q), p_s(q)\rangle}{\|p_s(q)\|_2^2}, \tag{10}$$

$$\gamma_s(x, q) = \frac{\langle h_s(x, q), p_s(q)\rangle}{\|h_s(x, q)\|_2 \|p_s(q)\|_2}. \tag{11}$$

A value $\beta_s = 1$ means that the projection of h_s onto p_s matches p_s in magnitude and sign; $\gamma_s = 1$ means that h_s and p_s have the same direction.

Response accuracy is evaluated against the known reference dynamics F. Let Φ_G denote the state obtained by integrating right-hand side G from x for 0.01 time units. The response error is

$$E_G(x, q) = \|D\Phi_G(x)q - D\Phi_F(x)q\|_{\mathrm{RMS}}. \tag{12}$$

Directional derivatives are computed using automatic differentiation. Appendix E provides the aggregation procedure, variation across seeds, zero-difference cases, and finite-difference checks.

3.4 COMPONENT REPLACEMENT AND MISSPECIFICATION SEVERITY

To test whether the extrapolation error is attributable to the physical component, P_s is replaced after training while the trained data-driven component C_s is unchanged:

$$P_s + C_s \;\longrightarrow\; P_0 + C_s, \tag{13}$$

Let Φ_{ab} denote the prediction after 0.01 time units using physical component P_a and data-driven component C_b. The replacement therefore compares Φ_{ss} with Φ_{0s}. Response errors are compared before and after replacement for patterns well represented and scarcely represented during training. Additional component combinations and controls are described in Appendix F.

To test how extrapolation changes with the magnitude of physical misspecification, additional hybrid models are trained with

$$P_\alpha = \alpha P_0 \tag{14}$$

using 12 values of α from 0.1 to 3. Each value of α is evaluated with one hybrid model per seed. For each seed, the same purely data-driven model is used as the baseline for all tested values of α. For the common set $\mathcal{Q}$ of scarcely represented measurements, define

$$B_\alpha = \left(\frac{\mathrm{mean}_{r\in\mathcal{Q}}\, E_{\alpha,r}^2}{\mathrm{mean}_{r\in\mathcal{Q}}\, E_{d,r}^2} \right)^{1/2}, \tag{15}$$

where d denotes the purely data-driven model. $B_\alpha < 1$ indicates lower response error for the hybrid model than for the purely data-driven model. The loss of this advantage is reported between adjacent tested values of α, without estimating a continuous threshold. Appendix G provides the tested values, pattern set, and additional response analysis.

Measurements are first aggregated within each seed and then summarized across the five seeds.

4 RESULTS

4.1 SMALL HELD-OUT ERRORS COEXIST WITH DIFFERENT EXTRAPOLATION RESPONSES

All six model conditions have small relative errors on the standard held-out sets (Figure 1a). Across five training seeds, the mean relative prediction error is 0.0204% for the hybrid model with correct diffusion, ranges from 0.0207% to 0.0336% across the four hybrid models with misspecified physics, and is 0.0448% for the purely data-driven model. Their responses differ on patterns scarcely represented in the training data (Figure 1b). On the common set of ten Fourier index pairs with $\rho \geq 19$, response error is divided within each seed by the response error of the purely data-driven model on the same patterns. The mean ratio across seeds is 0.028 for the hybrid model with correct diffusion and 0.264, 0.247, 0.369, and 0.746 for misspecification in directional dependence, discretization, magnitude, and omission of v diffusion, respectively. Each hybrid model with misspecified physics therefore has lower response error than the purely data-driven model, while the hybrid model with correct diffusion has lower response error than all four. The discretization case illustrates that small

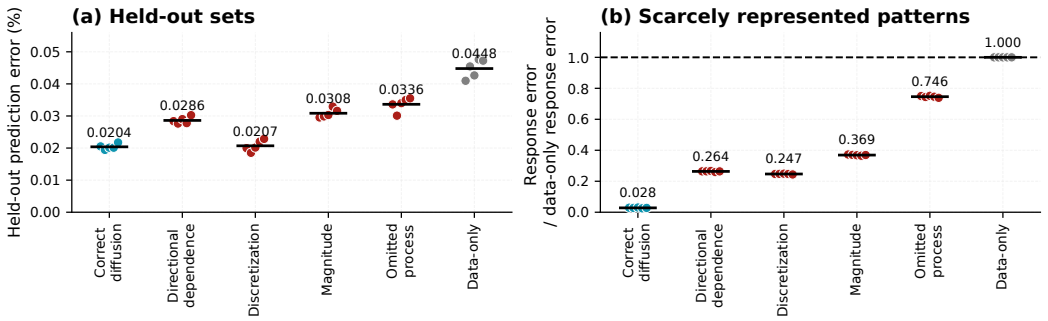

Figure 1: Small held-out prediction errors coexist with different responses during extrapolation. (a) Relative held-out prediction error. (b) Response error on scarcely represented patterns relative to the purely data-driven model. Points show five seeds; vertical marks show their means.

errors on standard held-out sets do not necessarily imply accurate responses to scarcely represented patterns. Its mean held-out prediction error, 0.0207%, is almost identical to the 0.0204% error of the hybrid model with correct diffusion, whereas its response error on scarcely represented patterns is about 8.8 times larger. Its response error is nevertheless only about one quarter of that of the purely data-driven model. Misspecified physics can therefore provide useful extrapolative structure while leaving substantial error in the response of the hybrid model.

4.2 ERRORS IN THE SUPPLIED PHYSICS ARE QUANTITATIVELY TRACED INTO THE RESPONSES OF HYBRID MODELS

Figure 2 compares p_s, the difference between the responses of the misspecified and correct physical components, with h_s, the corresponding difference between the responses of their trained hybrid models. These responses are evaluated before integrating the models forward in time. The analysis uses 173 Fourier index pairs (m_x, m_y) with $19 \leq \rho \leq 24$, with each pair evaluated using both sine and cosine patterns in both u and v. For every tested pattern, a standardized coefficient of 0.01 is at least 116 times the largest absolute coefficient observed in the training data. Thus, all patterns included in this analysis are scarcely represented during training. The mean projection coefficients across the five seeds are 0.980, 0.995, 0.962, and 0.972 for directional dependence, discretization, magnitude, and omission of v diffusion, respectively. Thus, across the four misspecifications, the difference between the responses of the trained hybrid models retains 96.2% to 99.5% of the magnitude and sign of the corresponding difference between the responses of their physical components. The mean cosine similarities are 0.99948, 0.99996, 0.99918, and 0.99961, showing that the differences between responses also have nearly the same direction. For the scarcely represented patterns tested here, the differences between the responses of the trained hybrid models therefore closely match both the magnitude and direction of the known differences between the responses of their physical components. Across individual directions for which the supplied response difference exceeds the

seed-variation resolution criterion, the median $\|h_s - p_s\|/\|p_s\|$ ranges from 0.007 to 0.071 across misspecifications and perturbed variables (Appendix E.4). The correspondence is observed for all four

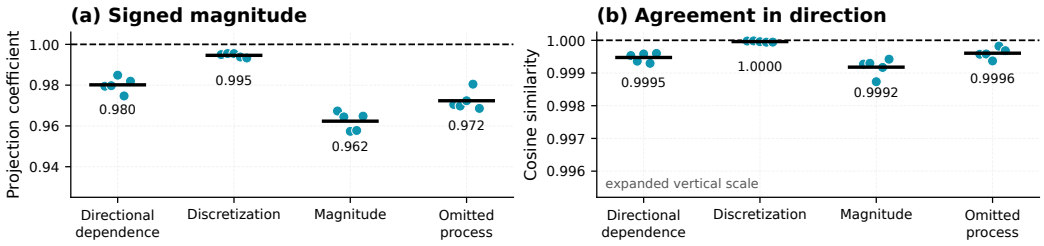

Figure 2: Differences between the responses of trained hybrid models closely follow those between their physical components. (a) Projection coefficient β_s. (b) Cosine similarity γ_s on scarcely represented Fourier patterns with $19 \leq \rho \leq 24$. Points show five seeds; vertical marks show their means.

forms of misspecification. The next experiment tests whether replacing only the supplied physical component changes the response error of the trained hybrid model.

4.3 Replacing the physical component removes most of the response error during extrapolation

Replacing the misspecified diffusion with the correct diffusion, while leaving the trained data-driven component unchanged, substantially reduces response error on the scarcely represented Fourier patterns with indices $(20, 0)$ and $(0, 20)$ (Figure 3a). Mean response error decreases from 0.0345 to 0.00315 for directional misspecification, from 0.0213 to 0.00260 for discretization misspecification, from 0.0310 to 0.00346 for magnitude misspecification, and from 0.0846 to 0.00199 when v diffusion is omitted. These changes correspond to reductions of approximately 91%, 88%, 89%, and 98%, respectively. For omission of v diffusion, only perturbations in v are included because only they directly change the omitted diffusion response. The replacement increases mean response error on

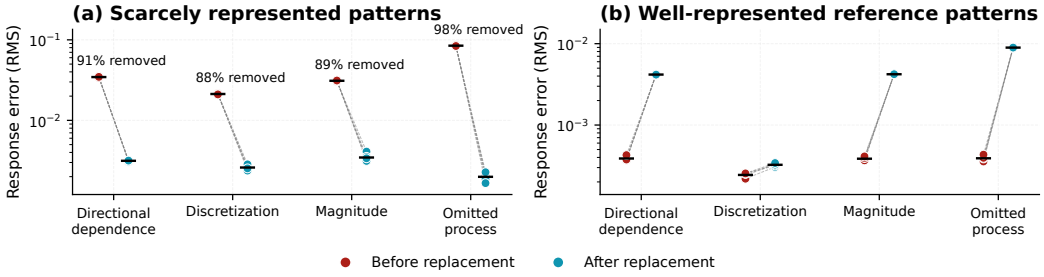

Figure 3: Replacing the misspecified physical component reduces response error on scarcely represented patterns and increases it on well-represented patterns. The trained data-driven component is unchanged. (a) Fourier patterns $(20, 0)$ and $(0, 20)$. (b) Fourier patterns $(6, 0)$ and $(0, 6)$.

the well-represented Fourier patterns with indices $(6, 0)$ and $(0, 6)$ by factors of approximately 10.8, 1.33, 10.9, and 22.9, respectively (Figure 3b). On these well-represented patterns, the data-driven component has adapted to the supplied misspecified physics. In the original hybrid model, the response of the data-driven component compensates much of the response error introduced by the misspecified physical component. Replacing only the physical component removes the error being compensated while leaving the data-driven component unchanged, causing the response error to increase. Replacement also reduces response error on the off-axis Fourier patterns with indices $(10, 8)$ and $(8, 10)$. On the principal $\rho = 20$ patterns, the median factorial interaction magnitude ranges from 0.10% to 1.82% of the observed difference between the responses of the misspecified and correct hybrid models across the four misspecifications (Appendix F). The intervention therefore attributes most of the response error on the tested scarcely represented patterns to the misspecified physical

component, conditional on the trained data-driven component being held fixed. The next analysis examines how the magnitude of that misspecification affects whether the hybrid model retains an advantage over the purely data-driven model.

4.4 Severe misspecification removes the advantage over the purely data-driven model

To examine how the severity of physical misspecification affects extrapolation, the supplied diffusion is multiplied by α, where $\alpha = 1$ corresponds to the correct diffusion. For the scarcely represented Fourier patterns used in this experiment, the five-seed mean ratio of the hybrid model's response error to that of the purely data-driven model increases from 0.032 at $\alpha = 1$ to 0.386 at $\alpha = 1.5$, 0.713 at $\alpha = 2$, and 0.991 at $\alpha = 2.5$ (Figure 4a). The ratio then increases to 1.113 at $\alpha = 2.75$ and 1.233 at $\alpha = 3$. All five seeds have ratios below 1 at $\alpha = 2.5$, whereas all five have ratios above 1 at $\alpha = 2.75$. Thus, between these two tested values of α, the hybrid model changes from having lower response error than the purely data-driven model to having higher response error. When the supplied

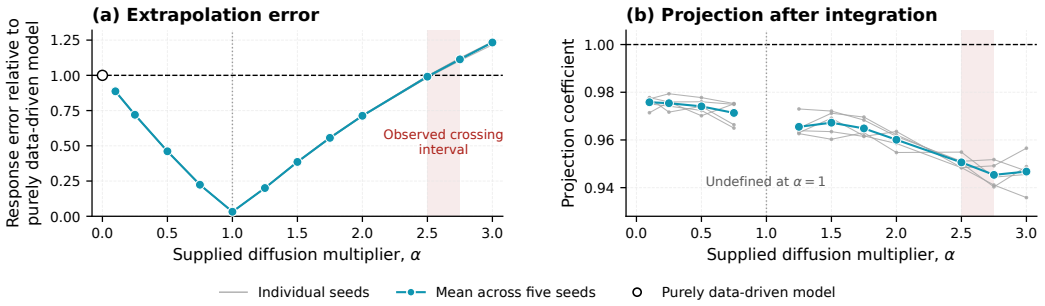

Figure 4: **Severe misspecification removes the extrapolation advantage while the projection remains close to one.** (a) Response error relative to the purely data-driven model. (b) Projection coefficient after integration over 0.01 time units. Thin lines show five seeds and thick lines their means; shading marks $2.5 \leq \alpha \leq 2.75$. The projection is undefined at $\alpha = 1$.

diffusion is reduced below its correct magnitude, the hybrid model still has a mean error ratio of 0.887 at $\alpha = 0.1$, even though the supplied diffusion is only one tenth of the correct magnitude. Across the tested values of α, the extrapolation advantage weakens as the supplied diffusion moves farther from its correct magnitude. A diffusion-only RK4 calculation closely reproduces the overall severity trend but not the exact ordering near the crossing, showing that attenuation by the supplied diffusion explains most of the variation while differences between the trained models remain consequential near the observed crossing (Appendix G). After integration over 0.01 time units, Figure 4b tests whether differences between the responses of separately trained hybrid models continue to follow the response change caused by the supplied physical component. For each value of α, the difference between the response of the corresponding hybrid model and that of the hybrid model with correct diffusion is compared with the response change produced by replacing its supplied diffusion with the correct diffusion while keeping its trained data-driven component unchanged. The mean projection coefficients range from 0.945 to 0.976, with mean cosine similarities above 0.999 (Table 7 in the Appendix). Thus, even after the hybrid model loses its extrapolation advantage, the differences between the responses of the trained hybrid models remain close in magnitude and direction to the changes caused by the supplied physical component.

5 Discussion

A misspecified physical component can improve extrapolation relative to a purely data-driven model while remaining the main source of the remaining response error. Previous studies show that the benefit of hybrid modeling depends on the quality of the supplied physics (Geremia et al., 2025) and that accurate predictions can conceal substantial adjustment by a data-driven component (Álvarez Chaves et al., 2026). The present study establishes a more specific connection in a controlled system: on scarcely represented patterns, differences in responses between hybrid models with

misspecified and correct diffusion closely follow differences in responses between their supplied physical components, and replacing the misspecified component removes most of the error against the true response. The same replacement increases error on well-represented patterns, consistent with the data-driven component having adapted to the original misspecified physical component. This attribution therefore applies to the tested patterns with the trained data-driven component held fixed.

The linear analysis specifies how finite training can preserve dependence on the supplied physics. Under the stated assumptions and conditions on the learning rate, for a perturbation along an eigenvector of the empirical second-moment matrix, the difference between hybrid models with misspecified and correct physics contracts at a rate determined by the corresponding eigenvalue. If that eigenvalue is zero, the difference remains unchanged. This extends the dependence of learning on variation in the training data (Floryan, 2024) to additive hybrid models with different physical components. Under parameter linearization, the difference between directional responses at a new state depends on the directional derivative of the neural tangent kernel that relates that state to the training inputs (Appendix B). These analyses motivate the observed dependence on the supplied physics but do not predict how much of the response difference remains across different forms of misspecification or where the hybrid model loses its advantage over the purely data-driven model (Appendix H).

The severity experiment further separates continued dependence on the supplied physics from comparative accuracy. As the supplied diffusion is increased above its correct magnitude, the hybrid model loses its advantage over the purely data-driven model while the projection coefficient remains close to one. The hybrid model therefore loses its accuracy advantage even while the difference between its response and that of the hybrid model with correct physics continues to follow the response change caused by the diffusion misspecification. This is qualitatively consistent with the linear analysis, although the exact independence of the projection from misspecification magnitude in that analysis applies at a common number of training updates, whereas the experiment uses checkpoints selected using validation loss and a different projection defined after integration. A diffusion-only calculation further shows that attenuation of the tested Fourier modes explains most of the severity curve. The trained-model result is therefore that this physical trend remains visible after learning, while smaller differences between the trained models affect the ordering near the observed crossing. The crossing itself remains specific to the tested diffusion formulation, pattern set, and comparison with the purely data-driven model.

For Scientific AI, these findings motivate examining which physical assumptions continue to influence predictions under scientifically relevant changes, including new regions, warmer climates, and extreme events (Feng et al., 2023; Bohl et al., 2025; Acuña Espinoza et al., 2025). Fourier patterns provide a controlled construction here; changes derived from independent scenario simulations, with the relevant physical constraints respected, provide one possible starting point in other systems. Hybrid models can be trained with two plausible physical formulations using common data and training settings. For the same chosen changes, one can then test whether differences in responses between these hybrid models follow the magnitude and pattern of differences in responses between their physical components. Without an accurate reference, these comparisons reveal dependence on the physical assumptions under examination; identifying the correct formulation and detecting errors shared by all candidates still require independent evidence. Where the components can be replaced consistently, holding the trained data-driven component fixed provides a further test of how the choice of physical component affects the response.

The present evidence concerns one reaction-diffusion system, errors in linear diffusion, one network architecture, and one training procedure. The five training seeds use the same trajectory data, so variation across independently generated datasets is not measured. The response measurements concern controlled state changes at fixed governing equations and, for the accuracy comparisons, one integration interval. Extension to sustained changes in forcing or long-term scientific predictions requires further evaluation. The replacement intervention measures the effect of changing the physical component on the complete trained hybrid model, including interactions between the physical and data-driven components. Other ways of coupling physics and learning require interventions designed for their specific model structure.

AI USE STATEMENT

In this work, the author used generative AI tools to assist with editing the manuscript and improving its readability. All AI-assisted edits were reviewed and refined by the author through an iterative, sentence-level process involving repeated interaction, revision, and careful evaluation. The author made the final decisions regarding all wording and content and takes full responsibility for the final work, including all text, claims, and other materials produced with the assistance of generative AI.

Generative AI was not used to conduct the literature review or to develop or implement the research code, except for limited assistance with debugging. The scientific analysis, methodology, literature review, and computational work were independently conducted and evaluated by the author.

REPRODUCIBILITY STATEMENT

The Appendix provides all details needed to reproduce the experiments, including the construction of the training, validation, and test data and the procedures used for the extrapolation and intervention analyses. Source code and supplementary materials needed to reproduce the reported results are provided anonymously with the submission.

REFERENCES

Joseph A. Abbate, Emiliano Fable, Giovanni Tardini, Brian A. Grierson, Alexei Y. Pankin, and Egemen Kolemen. Validating physics-based (astra/transp), data-driven (d3d+ aug), and physics+ data hybrid models for quantitatively accurate yet generalizable guidance for iter operators. In *67th Annual Meeting of the APS Division of Plasma Physics*. APS, 2025. URL https://meetings-archive.aps.org/dpp/2025/bo04/3/.

Eduardo Acuña Espinoza, Ralf Loritz, Frederik Kratzert, Daniel Klotz, Martin Gauch, Manuel Álvarez Chaves, and Uwe Ehret. Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events. *Hydrology and Earth System Sciences*, 29(5):1277–1294, 2025. URL https://doi.org/10.5194/hess-29-1277-2025.

Manuel Álvarez Chaves, Eduardo Acuña Espinoza, Uwe Ehret, and Anneli Guthke. When physics gets in the way: an entropy-based evaluation of conceptual constraints in hybrid hydrological models. *Hydrology and Earth System Sciences*, 30(3):629–658, 2026. URL https://hess.copernicus.org/articles/30/629/2026/.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 685–694. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/basri20a.html.

Sindre Stenen Blakseth, Adil Rasheed, Trond Kvamsdal, and Omer San. Deep neural network enabled corrective source term approach to hybrid analysis and modeling. *Neural Networks*, 146:181–199, 2022. URL https://doi.org/10.1016/j.neunet.2021.11.021.

Jan P Bohl, Raul R Wood, Corinna Frank, Paul C Astagneau, Jonas Peters, and Manuela I Brunner. Hybrid models generalize better to warmer climate conditions than process-based and purely data-driven models. *EGUsphere*, 2025:1–45, 2025. URL https://doi.org/10.5194/hess-30-4667-2026.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2205–2211. International Joint Conferences on Artificial Intelligence Organization, 8 2021. URL https://doi.org/10.24963/ijcai.2021/304. Main Track.

Allan Carter, Syed Imtiaz, and Greg Naterer. Imperfect physics-guided neural networks. *Chemical Engineering Science*, 305:121153, 2025. URL https://doi.org/10.1016/j.ces.2024.121153.

Daniel Cebrián-Lacasa, Pedro Parra-Rivas, Daniel Ruiz-Reynés, and Lendert Gelens. Six decades of the fitzhugh-nagumo model: A guide through its spatio-temporal dynamics and influence across disciplines. *Physics Reports*, 1096:1–39, 2024. URL https://doi.org/10.1016/j.physrep.2024.09.014.

Dennis Duncan and Christoph Räth. Optimizing the combination of data-driven and model-based elements in hybrid reservoir computing. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(10), 2023. URL https://doi.org/10.1063/5.0164013.

Alban Farchi, Patrick Laloyaux, Massimo Bonavita, and Marc Bocquet. Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, 147(739):3067–3084, 2021. URL https://doi.org/10.1002/qj.4116.

Dapeng Feng, Hylke Beck, Kathryn Lawson, and Chaopeng Shen. The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. *Hydrology and Earth System Sciences*, 27(12):2357–2373, 2023. URL https://doi.org/10.5194/hess-27-2357-2023.

Marc Finzi, Gregory Benton, and Andrew G Wilson. Residual pathway priors for soft equivariance constraints. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30037–30049. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fc394e9935fbd62c8aedc372464e1965-Paper.pdf.

Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, 1(6):445–466, 1961. URL https://doi.org/10.1016/S0006-3495(61)86902-6.

Daniel Floryan. On instabilities in neural network-based physics simulators. *arXiv preprint arXiv:2406.13101*, 2024. URL https://arxiv.org/abs/2406.13101.

Shailesh Garg, Souvik Chakraborty, and Budhaditya Hazra. Physics-integrated hybrid framework for model form error identification in nonlinear dynamical systems. *Mechanical Systems and Signal Processing*, 173:109039, 2022. URL https://doi.org/10.1016/j.ymssp.2022.109039.

Tobias Gehra and Michael Günthner. A hybrid modeling approach combining machine learning and physical phenomenological methods to predict highly transient engine emissions. *Automotive and Engine Technology*, 11(1):6, 2026. URL https://doi.org/10.1007/s41104-026-00168-y.

Margherita Geremia, Tiziana Marella, Elaheh Ardalani, Samira Beyramysoltan, Sayantan Chattoraj, Pierantonio Facco, Massimiliano Barolo, and Fabrizio Bezzo. On the impact of mechanistic model quality and data availability in hybrid model development. *Computers & Chemical Engineering*, pp. 109536, 2025. URL https://doi.org/10.1016/j.compchemeng.2025.109536.

Stefano Giampiccolo, Federico Reali, Anna Fochesato, Giovanni Iacca, and Luca Marchetti. Robust parameter estimation and identifiability analysis with hybrid neural ordinary differential equations in computational biology. *NPJ Systems Biology and Applications*, 10(1):139, 2024. URL https://doi.org/10.1038/s41540-024-00460-3.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Anuj Karpatne, Xiaowei Jia, and Vipin Kumar. Knowledge-guided machine learning: Current trends and future prospects. *arXiv preprint arXiv:2403.15989*, 2024. URL https://arxiv.org/html/2403.15989.

Sathish Kasilingam, Ruoyu Yang, Shubhendu Kumar Singh, Mojtaba A Farahani, Rahul Rai, and Thorsten Wuest. Physics-based and data-driven hybrid modeling in manufacturing: a review. *Production & Manufacturing Research*, 12(1):2305358, 2024. URL https://doi.org/10.1080/21693277.2024.2305358.

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*, 632:1060–1066, 2024. doi: 10.1038/s41586-024-07744-y. URL https://doi.org/10.1038/s41586-024-07744-y.

Akash Koppa, Dominik Rains, Petra Hulsman, Rafael Poyatos, and Diego G Miralles. A deep learning-based hybrid model of global terrestrial evaporation. *Nature communications*, 13(1):1912, 2022. URL https://doi.org/10.1038/s41467-022-29543-7.

Jiating Li and Qiao Hu. Beyond data-driven: Integrating scientific knowledge and machine learning for crop prediction. *European Journal of Agronomy*, 180:128261, 2026. URL https://doi.org/10.1016/j.eja.2026.128261.

Narjes Malekjani, Abdolreza Kharaghani, and Evangelos Tsotsas. Physics-informed vs. data-driven neural networks for n-heptane droplet evaporation. *Chemical Engineering Research and Design*, 2026. URL https://doi.org/10.1016/j.cherd.2026.03.016.

Swapnil Manna, Timothy J. Rogers, and Lawrence Bull. Orthogonal discrepancy kernels for learning with partial physics. *arXiv preprint arXiv:2606.21199*, 2026. URL https://arxiv.org/abs/2606.21199.

Hanna Meyer and Edzer Pebesma. Predicting into unknown space? estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633, 2021. URL https://doi.org/10.1111/2041-210X.13650.

Jin-ichi Nagumo, Suguru Arimoto, and Shuji Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, 1962. URL https://doi.org/10.1109/JRPROC.1962.288235.

Declan A Norton, Yuanzhao Zhang, and Michelle Girvan. Learning beyond experience: Generalizing to unseen state space with reservoir computing. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 35(10), 2025. URL https://doi.org/10.1063/5.0283421.

Paul A O'Gorman and John G Dwyer. Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, 10(10):2548–2563, 2018. URL https://doi.org/10.1029/2018MS001351.

Tobias Overgaard, Maria-Ona Bertran, Bo Friis Nielsen, Pierantonio Facco, and Massimiliano Barolo. Beyond black-box models: Measuring physics-driven and data-driven contributions in hybrid process models. *Digital Chemical Engineering*, pp. 100307, 2026. URL https://doi.org/10.1016/j.dche.2026.100307.

Jeongjin Park, Nicole Tianjiao Yang, and Nisha Chandramoorthy. When are dynamical systems learned from time series data statistically accurate? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 43975–44008. Curran Associates, Inc., 2024. doi: 10.52202/079017-1396. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4dc57702c987e1e72f0dd2921edb5ded-Paper-Conference.pdf.

Jaideep Pathak, Alexander Wikner, Rebeckah Fussell, Sarthak Chandra, Brian R Hunt, Michelle Girvan, and Edward Ott. Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model. *Chaos: An interdisciplinary journal of nonlinear science*, 28(4), 2018. URL https://doi.org/10.1063/1.5028373.

Yi Ren. Learning dynamics of deep learning–force analysis of deep neural networks. *arXiv preprint arXiv:2509.19554*, 2025. URL https://arxiv.org/abs/2509.19554.

Joel Sansana, Ricardo Rendall, Ivan Castillo, Leo Chiang, and Marco S Reis. Hybrid modeling for improved extrapolation and transfer learning in the chemical processing industry. *Chemical Engineering Science*, 300:120568, 2024. URL https://doi.org/10.1016/j.ces.2024.120568.

Johan Schoukens and Lennart Ljung. Nonlinear system identification: A user-oriented road map. *IEEE Control Systems Magazine*, 39(6):28–99, 2019. URL https://doi.org/10.1109/MCS.2019.2938121.

Artur M Schweidtmann, Dongda Zhang, and Moritz Von Stosch. A review and perspective on hybrid modeling methodologies. *Digital Chemical Engineering*, 10:100136, 2024. URL https://doi.org/10.1111/2041-210X.13650.

Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14809–14821. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7ca57a9f85a19a6e4b9a248c1daca185-Paper.pdf.

Henk J Van Waarde, Jaap Eising, Harry L Trentelman, and M Kanat Camlibel. Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65 (11):4753–4768, 2020. URL https://doi.org/10.1109/TAC.2020.2966717.

Zhong Yi Wan, Pantelis Vlachas, Petros Koumoutsakos, and Themistoklis Sapsis. Data-assisted reduced-order modeling of extreme events in complex dynamical systems. *PloS one*, 13(5): e0197704, 2018. URL https://doi.org/10.1371/journal.pone.0197704.

Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23078–23091. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wang22aa.html.

Zhaoyang Yan, Yingxi Xie, Zhen Zhang, Feilong Liu, Shu Yang, Di Xing, and Longsheng Lu. Robust extrapolation of sparse particle image velocimetry measurements via a mamba-integrated physics-informed neural network framework. *Engineering Applications of Artificial Intelligence*, 182:115925, 2026. URL https://doi.org/10.1016/j.engappai.2026.115925.

Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel De Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, 2021. URL https://doi.org/10.1088/1742-5468/ac3ae5.

A PROOF OF THE LINEAR HYBRID RESULT

For the loss in Eq. 1,

$$\nabla_W \mathcal{L}_s(W) = \frac{1}{n}\sum_{i=1}^{n}((P_s + W - A)x_i)x_i^\top = (P_s + W - A)M. \tag{16}$$

Full-batch gradient descent gives

$$W_{s,t+1} = W_{s,t} - \eta(P_s + W_{s,t} - A)M, \tag{17}$$

$$W_{0,t+1} = W_{0,t} - \eta(P_0 + W_{0,t} - A)M. \tag{18}$$

Writing $\Delta W_t = W_{s,t} - W_{0,t}$ and subtracting these updates yields

$$\Delta W_{t+1} = \Delta W_t - \eta(\Delta P + \Delta W_t)M. \tag{19}$$

Since $\Delta\mathcal{H}_t = \Delta P + \Delta W_t$,

$$\Delta\mathcal{H}_{t+1} = \Delta\mathcal{H}_t(I - \eta M). \tag{20}$$

Identical initialization implies $\Delta\mathcal{H}_0 = \Delta P$, so induction proves

$$\Delta\mathcal{H}_t = \Delta P(I - \eta M)^t. \tag{21}$$

In fact, the targets need not be linear: replacing Ax_i by arbitrary shared targets y_i leaves Eq. 19 unchanged because the target terms cancel. The essential restriction here is the unrestricted linear data-driven component and its specified training update.

If $Mq = \lambda_q q$, then $(I - \eta M)^t q = (1 - \eta\lambda_q)^t q$. For $0 < \eta\lambda_q < 1$, the multiplier decreases monotonically to zero. For $0 < \eta\lambda_q < 2$, its absolute value contracts, although its sign can alternate when $\eta\lambda_q > 1$. For $\lambda_q = 0$, it is one for every t. Thus, when $M \neq 0$, $0 < \eta < 2/\lambda_{\max}(M)$ is sufficient for contraction in all positive-eigenvalue directions. When $M = 0$, every update difference remains unchanged and no division by $\lambda_{\max}(M)$ is needed.

For $P_a = P_0 + aB$, with B fixed,

$$\Delta\mathcal{H}_{a,t} = aB(I - \eta M)^t. \tag{22}$$

Consequently, for $a \neq 0$ and $Bq \neq 0$,

$$\frac{\langle \Delta\mathcal{H}_{a,t}q, aBq\rangle}{\|aBq\|_2^2} = \frac{a^2\langle B(I - \eta M)^t q, Bq\rangle}{a^2\|Bq\|_2^2}, \tag{23}$$

which proves Eq. 6 at a fixed update count. This does not assert that the projection equals one, lies between zero and one, or remains invariant when different conditions are evaluated at different update counts.

B NONLINEAR DATA-DRIVEN COMPONENTS AND DIRECTIONAL RESPONSES

We distinguish exact identities for differentiable hybrid models from a conditional training result obtained by linearizing the network in its parameters. The latter is related to the neural tangent kernel formulation (Jacot et al., 2018). It allows nonlinear dependence on the input and nonlinear supplied physical maps, but does not assume that the finite convolutional networks in the experiment follow this linearization.

B.1 AN EXACT RESPONSE IDENTITY

For any differentiable P_s, P_0, C_s, C_0, set

$$p_s(x, q) = D(P_s - P_0)(x)q, \qquad c_s(x, q) = D(C_s - C_0)(x)q. \tag{24}$$

Differentiating $H_s - H_0 = (P_s - P_0) + (C_s - C_0)$ gives

$$h_s(x, q) = p_s(x, q) + c_s(x, q). \tag{25}$$

For the linear diffusion components in the experiment, $p_s(x, q) = (P_s - P_0)q$ is independent of the reference field x. For nonlinear physical maps, it generally depends on x. The identity alone does

not show that c_s is small: the data-driven components could retain, cancel, or enlarge the supplied difference.

Nor is a difference between two hybrid responses automatically an error against truth. For any differentiable reference map F,

$$D(H_s - F)(x)q = h_s(x, q) + D(H_0 - F)(x)q. \tag{26}$$

The discrepancy between the hybrid difference and the true response error is exactly the response error of the correct hybrid. The same identity holds for the corresponding saved-step maps, without assuming that those maps decompose additively into physical and data-driven parts.

B.2 A NETWORK LINEARIZED AT COMMON INITIALIZATION

Let $C_\theta : \mathbb{R}^d \to \mathbb{R}^d$ have parameters $\theta \in \mathbb{R}^p$ and common initialization θ^{init}. Define the parameter Jacobian, the matrix of output derivatives with respect to parameters, by

$$J_0(x) = \partial_\theta C_{\theta^{\text{init}}}(x) \in \mathbb{R}^{d\times p}. \tag{27}$$

The linearized network is a separate mathematical model,

$$C_\theta^{\text{lin}}(x) = C_{\theta^{\text{init}}}(x) + J_0(x)(\theta - \theta^{\text{init}}). \tag{28}$$

Its dependence on x can be nonlinear; its dependence on θ is affine, and J_0 is fixed during training. For fixed physical maps P_s and P_0, train $H_s^{\text{lin}}(x;\theta) = P_s(x) + C_\theta^{\text{lin}}(x)$ by full-batch gradient descent with fixed step size $\eta > 0$, no regularization, and common inputs and targets:

$$\mathcal{L}_s^{\text{lin}}(\theta) = \frac{1}{2n}\sum_{i=1}^{n} \|H_s^{\text{lin}}(x_i;\theta) - y_i\|_2^2. \tag{29}$$

No linearity assumption is imposed on the target map producing y_i or on P_s and P_0.

Stack the Jacobians and physical differences as

$$\mathbf{J} = \begin{pmatrix} J_0(x_1) \\ \vdots \\ J_0(x_n) \end{pmatrix}, \qquad \Delta\mathbf{p} = \begin{pmatrix} P_s(x_1) - P_0(x_1) \\ \vdots \\ P_s(x_n) - P_0(x_n) \end{pmatrix}. \tag{30}$$

Define the empirical neural tangent kernel matrix

$$\mathbf{\Theta} = \frac{1}{n}\mathbf{J}\mathbf{J}^\top \in \mathbb{R}^{nd\times nd}. \tag{31}$$

It records inner products of parameter derivatives at the training inputs and is symmetric and positive semidefinite. Let $\Delta\mathbf{h}_t$ stack the differences between the two linearized hybrid outputs at those inputs after t updates. These are output values, not the directional responses $h_s(x, q)$.

Proposition 2. *For the linearized models above, identical initialization gives*

$$\boxed{\Delta\mathbf{h}_t = (I_{nd} - \eta\mathbf{\Theta})^t \Delta\mathbf{p}.} \tag{32}$$

Proof. Let $\mathbf{h}_{s,t}$ stack the linearized hybrid outputs, let $\mathbf{y}$ stack the targets, and write $\boldsymbol{\varepsilon}_{s,t} = \mathbf{h}_{s,t} - \mathbf{y}$. The gradient and parameter update are

$$\nabla_\theta \mathcal{L}_s^{\text{lin}} = \frac{1}{n}\mathbf{J}^\top \boldsymbol{\varepsilon}_{s,t}, \qquad \theta_{s,t+1} - \theta_{s,t} = -\frac{\eta}{n}\mathbf{J}^\top \boldsymbol{\varepsilon}_{s,t}. \tag{33}$$

The physical outputs and the initial network outputs are fixed, so

$$\mathbf{h}_{s,t+1} - \mathbf{h}_{s,t} = \mathbf{J}(\theta_{s,t+1} - \theta_{s,t}) \tag{34}$$

$$= -\frac{\eta}{n}\mathbf{J}\mathbf{J}^\top \boldsymbol{\varepsilon}_{s,t} = -\eta\mathbf{\Theta}\boldsymbol{\varepsilon}_{s,t}. \tag{35}$$

The same matrix applies to the model using P_0. Subtracting the two updates cancels the targets and gives $\Delta\mathbf{h}_{t+1} = (I_{nd} - \eta\mathbf{\Theta})\Delta\mathbf{h}_t$. Since $\Delta\mathbf{h}_0 = \Delta\mathbf{p}$, induction proves the result. □

For an eigenvector $\mathbf{v}$ with $\mathbf{\Theta}\mathbf{v} = \mu\mathbf{v}$,

$$\mathbf{v}^\top \Delta\mathbf{h}_t = (1 - \eta\mu)^t \mathbf{v}^\top \Delta\mathbf{p}. \tag{36}$$

For $\mathbf{\Theta} \neq 0$, positive-eigenvalue components contract when $0 < \eta < 2/\lambda_{\max}(\mathbf{\Theta})$; components in its null space remain unchanged. If $\mathbf{\Theta} = 0$, the entire difference remains unchanged. These eigenvectors belong to the space of stacked training outputs. They are not spatial Fourier patterns or eigenvectors of the input second-moment matrix.

B.3 Outputs and directional responses at a new state

Write

$$A_\Theta = I_{nd} - \eta\mathbf{\Theta}, \qquad S_t = \sum_{k=0}^{t-1} A_\Theta^k, \qquad S_0 = 0, \tag{37}$$

and define

$$\mathbf{K}(x) = \frac{1}{n} J_0(x)\mathbf{J}^\top \in \mathbb{R}^{d\times nd}. \tag{38}$$

Summing the parameter-update differences gives

$$\theta_{s,t} - \theta_{0,t} = -\frac{\eta}{n}\mathbf{J}^\top S_t \Delta\mathbf{p}. \tag{39}$$

Substitution into Eq. 28 yields the exact output difference within this model:

$$H_s^{\mathrm{lin}}(x;\theta_{s,t}) - H_0^{\mathrm{lin}}(x;\theta_{0,t}) = \Delta P(x) - \eta\mathbf{K}(x)S_t\Delta\mathbf{p}, \tag{40}$$

where $\Delta P(x) = P_s(x) - P_0(x)$. In particular, the sum ends at $t-1$ and its prefactor is η; the factor $1/n$ is already included in $\mathbf{K}$.

To connect this result to the measured responses, assume that P_s, P_0, and J_0 are differentiable in x near the reference state. Define

$$\mathbf{K}_q(x) = D_x\mathbf{K}(x)q = \frac{1}{n}(D_x J_0(x)q)\mathbf{J}^\top. \tag{41}$$

The training inputs, parameters after t updates, and $S_t\Delta\mathbf{p}$ are held fixed when the test state is perturbed. Differentiating Eq. 40 therefore proves

$$\boxed{h_{s,t}^{\mathrm{lin}}(x,q) = p_s(x,q) - \eta\mathbf{K}_q(x)S_t\Delta\mathbf{p}.} \tag{42}$$

Here $h_{s,t}^{\mathrm{lin}} = D(H_s^{\mathrm{lin}} - H_0^{\mathrm{lin}})(x)q$. The relevant quantity for directional responses is $\mathbf{K}_q$, not $\mathbf{K}$ alone. A small function value does not in general imply a small derivative.

For $p = p_s(x,q) \neq 0$, define

$$\delta_t(x,q) = \frac{\eta\|\mathbf{K}_q(x)S_t\Delta\mathbf{p}\|_2}{\|p\|_2}. \tag{43}$$

Then Eq. 42 gives

$$\frac{\|h_{s,t}^{\mathrm{lin}} - p\|_2}{\|p\|_2} = \delta_t, \qquad |\beta_{s,t}^{\mathrm{lin}} - 1| \le \delta_t, \qquad r_{\perp,s,t}^{\mathrm{lin}} \le \delta_t. \tag{44}$$

The projection bound follows from Cauchy-Schwarz; the perpendicular bound follows because orthogonal projection cannot increase a Euclidean norm. For $\delta_t < 1$, the triangle inequality also gives $\gamma_{s,t}^{\mathrm{lin}} \ge (1-\delta_t)/(1+\delta_t)$. Thus a small correction relative to the supplied response difference is sufficient for both its magnitude and direction to remain close.

For completeness, let $\|A\|_{\mathrm{op}} = \sup_{\|v\|_2=1}\|Av\|_2$ denote the operator norm. When $\|A_\Theta\|_{\mathrm{op}} \le 1$,

$$\delta_t \le \frac{\eta\|\mathbf{K}_q(x)\|_{\mathrm{op}}\|S_t\Delta\mathbf{p}\|_2}{\|p\|_2} \le \frac{\eta t\|\mathbf{K}_q(x)\|_{\mathrm{op}}\|\Delta\mathbf{p}\|_2}{\|p\|_2}. \tag{45}$$

These are conditional bounds, not measurements made in this study. The training Fourier-coefficient statistics do not establish that their right-hand sides are small.

B.4 Scaling and the linear special case

For $P_a(x) = P_0(x) + aB(x)$, let $\mathbf{b}$ stack $B(x_i)$. Within the same frozen linearization, for a common update count t,

$$\Delta\mathbf{h}_{a,t} = aA_\Theta^t\mathbf{b}, \tag{46}$$

$$H_a^{\mathrm{lin}}(x) - H_0^{\mathrm{lin}}(x) = a\big(B(x) - \eta\mathbf{K}(x)S_t\mathbf{b}\big), \tag{47}$$

$$h_{a,t}^{\mathrm{lin}}(x,q) = a\big(DB(x)q - \eta\mathbf{K}_q(x)S_t\mathbf{b}\big). \tag{48}$$

For $a \neq 0$ and $DB(x)q \neq 0$, projecting the last expression onto $aDB(x)q$ and dividing by $\|aDB(x)q\|_2^2$ cancels a^2. The normalized projection is therefore independent of scale. This cancellation requires no small-kernel assumption and does not assert that the projection is close to one. The same statement holds for output differences when their supplied comparison vector is nonzero.

When $C_\theta(x) = Wx$, the parameter linearization is exact. Let $X = (x_1, \ldots, x_n)$, $M = XX^\top/n$, and $K_X = X^\top X/n$. The (i, j) block of $\mathbf{\Theta}$ is $(x_i^\top x_j/n)I_d$. Since $MX = XK_X$, induction yields

$$(I - \eta M)^t X = X(I - \eta K_X)^t. \tag{49}$$

Consequently, the matrix expression in Proposition 1 and the stacked-output expression in Proposition 2 describe the same updates in this case.

B.5 What changes when the network Jacobian changes

An exact remainder formula identifies the additional terms for a genuinely nonlinear parameter update. Consider the actual network C_θ, assumed continuously differentiable in its parameters, trained by full-batch gradient descent on the instantaneous squared-error loss with the same initialization and targets. For condition s, let $\mathbf{J}_{s,t}$ be its stacked parameter Jacobian at $\theta_{s,t}$, and define

$$\overline{\mathbf{J}}_{s,t} = \int_0^1 \mathbf{J}\big(\theta_{s,t} + u(\theta_{s,t+1} - \theta_{s,t})\big)\, du, \qquad \mathbf{\Psi}_{s,t} = \frac{1}{n}\overline{\mathbf{J}}_{s,t}\mathbf{J}_{s,t}^\top. \tag{50}$$

Here $\mathbf{J}(\theta)$ denotes the stacked Jacobian evaluated at parameter vector θ. The fundamental theorem of calculus and the gradient update give, exactly,

$$\boldsymbol{\varepsilon}_{s,t+1} = (I - \eta\mathbf{\Psi}_{s,t})\boldsymbol{\varepsilon}_{s,t}. \tag{51}$$

Let $\mathbf{\Theta}$ be the common initial matrix in Eq. 31 and set

$$\mathbf{d}_t = (\mathbf{\Psi}_{s,t} - \mathbf{\Theta})\boldsymbol{\varepsilon}_{s,t} - (\mathbf{\Psi}_{0,t} - \mathbf{\Theta})\boldsymbol{\varepsilon}_{0,t}. \tag{52}$$

Subtracting the two exact updates and iterating proves

$$\Delta\mathbf{h}_t = A_\Theta^t \Delta\mathbf{p} - \eta \sum_{k=0}^{t-1} A_\Theta^{t-1-k}\mathbf{d}_k. \tag{53}$$

In particular, if $\|A_\Theta\|_{\mathrm{op}} \leq 1$,

$$\|\Delta\mathbf{h}_t - A_\Theta^t \Delta\mathbf{p}\|_2 \leq \eta \sum_{k=0}^{t-1} \|\mathbf{d}_k\|_2. \tag{54}$$

This is an exact identity and conditional bound for nonlinear full-batch gradient descent. It reduces to Proposition 2 when the Jacobians remain fixed. Without bounds on the remainder, it does not imply that the frozen formula approximates the trained network. It concerns outputs on the training inputs; it does not supply an unmeasured bound on extrapolated directional derivatives.

B.6 Scope of the nonlinear analysis

The experiment uses finite convolutional networks, mini-batch AdamW, weight decay, gradient clipping, and validation-selected checkpoints, rather than the full-batch updates analyzed above. It also fits an integrated saved-step map. Even at identical neural initialization, the parameter Jacobian of $\Phi_{P_s + C_\theta}$ generally depends on P_s, because the intermediate Runge-Kutta states depend on that operator. Thus the common-Jacobian subtraction used in Proposition 2 cannot be assumed for this training objective. Different validation-selected update counts also fall outside the common-t scaling result. The exact additive response identity applies to the instantaneous model, while the observed tracing, component-replacement effects, and severity dependence remain empirical findings.

C SUPPLIED PHYSICAL COMPONENTS AND OPERATOR CHECKS

The correct physical component is $P_0(u, v) = (D_u \nabla^2 u, D_v \nabla^2 v)$. For $j \in \{u, v\}$, let x_j be the corresponding field and D_j its diffusion coefficient. The four misspecifications are

$$P_{\mathrm{dir},j}(x) = D_j(1.5\,\partial_{xx} + 0.5\,\partial_{yy})x_j, \tag{55}$$

$$P_{\mathrm{fd},j}(x)_{r,c} = \frac{D_j}{h^2}\big(x_{j,r+1,c} + x_{j,r-1,c} + x_{j,r,c+1} + x_{j,r,c-1} - 4x_{j,r,c}\big), \tag{56}$$

$$P_{\mathrm{mag}}(x) = 1.5P_0(x), \tag{57}$$

$$P_{\mathrm{omit}}(u, v) = (D_u \nabla^2 u, 0), \tag{58}$$

where $h = 2/64$ and the stencil indices are periodic. These change directional dependence, the numerical discretization of the Laplacian, the magnitude of both diffusion coefficients, and the presence of v diffusion, respectively.

C.1 FOURIER SYMBOLS AND INTEGRATION CHECKS

For $\kappa_x = 2\pi m_x/L$ and $\kappa_y = 2\pi m_y/L$, the spectral symbols are

$$\widehat{P}_{0,j} = -D_j(\kappa_x^2 + \kappa_y^2), \tag{59}$$

$$\widehat{P}_{\mathrm{dir},j} = -D_j(1.5\kappa_x^2 + 0.5\kappa_y^2), \tag{60}$$

$$\widehat{P}_{\mathrm{mag},j} = -1.5D_j(\kappa_x^2 + \kappa_y^2), \tag{61}$$

$$\widehat{\Delta}_5 = \frac{2}{h^2}\big(\cos(\kappa_x h) + \cos(\kappa_y h) - 2\big). \tag{62}$$

The omitted-process symbol is unchanged for u and zero for v. For the supplied components alone, every nonconstant mode has a negative symbol except the neutral modes of omitted v diffusion. Thus these misspecifications introduce no growing continuous-time diffusion mode.

The implementation also checks all modes on the discrete 64×64 Fourier grid. For a symbol λ, the saved-interval Runge-Kutta gain is

$$g(\lambda) = \left(1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}\right)^5, \qquad z = 0.002\lambda. \tag{63}$$

Table 1 reports maxima over nonconstant modes. The constant mode, excluded from these maxima, has gain one for every diffusion operator. Neither the continuous-time symbols nor these numerical checks guarantee stability of the complete nonlinear hybrid.

Table 1: **Integration check for the supplied physical components.** Largest absolute saved-interval gain over nonconstant modes on the full discrete Fourier grid.

Condition	Variable	Largest absolute gain
Directional dependence	u	0.999753
Directional dependence	v	0.999507
Discretization	u	0.999507
Discretization	v	0.999014
Magnitude	u	0.999260
Magnitude	v	0.998521
Omitted process	u	0.999507
Omitted process	v	1.000000

D DATA GENERATION, NETWORK, AND TRAINING

The reference system uses $L = 2$, $D_u = 0.005$, $D_v = 0.01$, $k = 0.005$, a periodic 64×64 grid, solver step 0.002, and recorded interval 0.01. Initial u and v fields are independent Gaussian random fields, with Fourier coefficients set to zero outside $\sqrt{m_x^2 + m_y^2} \le 6$ and each field rescaled to spatial

standard deviation 0.5. The 80 training, 24 validation, and 64 test trajectories use independent initial conditions generated by this procedure. Each trajectory contains 80 recorded states, yielding 6,320, 1,896, and 5,056 consecutive state pairs.

D.1 Data-driven component

All conditions use the same convolutional architecture with 52,738 trainable parameters. A 1×1 convolution maps the two input variables to 32 hidden channels. Five blocks apply 3×3 circular convolutions with dilations $1, 2, 4, 8, 16$. Each spatial convolution is added to a parallel 1×1 convolution of the same block input, followed by a GELU activation. The output head is a 1×1 convolution, GELU, and a final 1×1 convolution producing two fields.

Inputs are centered and scaled separately for u and v using the mean and standard deviation of the training inputs. The network output for each variable is rescaled using the training-set standard deviation of $(x_{t+1} - x_t)/0.01$. Perturbations are normalized using the training standard deviations, while response errors are measured on the original u and v scale.

D.2 Optimization, precision, and checkpoints

Each 0.01 interval is integrated with five RK4 steps of size 0.002, using the sum of the supplied diffusion and network output as the model right-hand side. Training minimizes mean squared error between the predicted and reference states at the end of the interval. The purely data-driven model uses the same solver, with the network alone providing the right-hand side.

For each of five seeds, all model conditions use the same initial network parameters. For each seed and epoch, the same random ordering of training samples is used across model conditions. Models are trained for 80 epochs with AdamW, a constant learning rate of 2×10^{-3}, weight decay 10^{-6}, and an effective batch size of 256. Gradient accumulation uses microbatches of 128, and the gradient norm is clipped at 1. Validation loss is evaluated every two epochs, and the checkpoint with the lowest validation loss is used for analysis.

Training uses CUDA automatic mixed precision, with bfloat16 where supported or float16 with gradient scaling otherwise. Model performance evaluation and directional derivatives use float32 with automatic mixed precision disabled, while inner products, squared norms, and aggregate statistics use float64. Inner products, squared norms, and aggregate statistics use float64.

E Spatial patterns, training representation, and response protocol

E.1 Pattern construction and normalization

For an integer pair (m_x, m_y) and $\varphi \in \{\cos, \sin\}$, define

$$w^{\varphi}_{m_x,m_y}(r, c) = \varphi\left(\frac{2\pi(m_x c + m_y r)}{64}\right). \tag{64}$$

Place this wave in one variable and zero in the other. Let D_σ scale each variable by its training standard deviation. Scale q so that $\|D_\sigma^{-1} q\|_{\mathrm{RMS}} = 1$, with the RMS taken over both variables and all grid points. There are $N = 2 \times 64^2$ states, and $\|v\|_{\mathrm{RMS}} = \|v\|_2/\sqrt{N}$.

For $z_i = D_\sigma^{-1}(x_i - \bar{x})$ and $\bar{q} = D_\sigma^{-1} q$, define

$$a_i(q) = \frac{1}{N} \bar{q}^\top z_i. \tag{65}$$

Since $\bar{q}^\top \bar{q} = N$, adding ϵq changes this coefficient by exactly ϵ in exact arithmetic. The training statistics are

$$s_{\mathrm{RMS}}(q) = \left(\frac{1}{n}\sum_{i=1}^{n} a_i(q)^2\right)^{1/2}, \qquad s_{\max}(q) = \max_i |a_i(q)|. \tag{66}$$

For $M_z = n^{-1}\sum_i z_i z_i^\top$,

$$s_{\mathrm{RMS}}(q)^2 = \frac{1}{N^2}\bar{q}^\top M_z \bar{q}. \tag{67}$$

This is a quadratic form of the empirical second moment, not an eigenvalue in general. Even if $\bar{q}/\sqrt{N}$ is an eigenvector, its eigenvalue is $N s_{\mathrm{RMS}}^2$ under this normalization. A single coefficient statistic does not specify variation across all joint state directions.

Table 2 compares the perturbation amplitude with observed training coefficients. An amplitude below a training maximum does not by itself guarantee that both perturbed endpoints lie inside that range, since the reference state can have a nonzero coefficient. These statistics also do not establish that perturbed states occur on natural trajectories.

Table 2: **Examples of training representation.** Ranges cover the two coordinate-axis orientations and sine/cosine phases. The last column compares perturbation amplitude with the largest observed absolute coefficient.

Patterns	Variable	s_{RMS} range	$0.01/s_{\max}$ range
$\rho = 6$ axes	u	0.0428-0.0509	0.044-0.058
$\rho = 6$ axes	v	0.0429-0.0516	0.027-0.036
$\rho = 20$ axes	u	8.29×10^{-6}-9.89×10^{-6}	182-315
$\rho = 20$ axes	v	3.31×10^{-7}-3.96×10^{-7}	4.64×10^{3}-7.93×10^{3}

E.2 Reference states and pattern sets

The principal response analyses use 12 reference fields: test trajectories $0, 1, 2, 3$ at saved indices $0, 16, 48$. Both sine and cosine phases and both perturbed variables are included unless stated otherwise.

The broad instantaneous tracing set contains every nonzero integer pair with

$$0 \le m_x, m_y \le 24, \qquad m_x^2 + m_y^2 \le 24^2. \tag{68}$$

It contains 472 integer pairs and includes both well and scarcely represented patterns. The component-replacement analysis uses

$$(6,0),\ (0,6),\ (10,8),\ (8,10),\ (20,0),\ (0,20),\ (12,12). \tag{69}$$

The severity comparison uses

$$(20,0),\ (0,20),\ (16,12),\ (12,16),\ (17,15),\ (15,17), \tag{70}$$

with $(6,0)$ and $(0,6)$ as well-represented controls. A separate common comparison across misspecification forms uses the patterns with $\rho \ge 19$ in its specified set:

$$\begin{gathered}(14,14),\ (16,12),\ (12,16),\ (17,15),\ (15,17),\\ (20,0),\ (0,20),\ (20,20),\ (24,0),\ (0,24).\end{gathered} \tag{71}$$

This common comparison and the severity comparison use different pattern sets; their ratios are reported separately.

E.3 Instantaneous tracing and aggregation

For one seed, index repeated measurements by r, and stack p_r and h_r from Eq. 9 into $\mathbf{p}$ and $\mathbf{h}$. The broad-set summaries are

$$\beta = \frac{\langle \mathbf{h}, \mathbf{p}\rangle}{\|\mathbf{p}\|_2^2}, \tag{72}$$

$$\gamma = \frac{\langle \mathbf{h}, \mathbf{p}\rangle}{\|\mathbf{h}\|_2 \|\mathbf{p}\|_2}, \tag{73}$$

$$r_\perp = \frac{\|\mathbf{h} - \beta\mathbf{p}\|_2}{\|\mathbf{p}\|_2}. \tag{74}$$

When $p_r \neq 0$, the aggregate projection equals the average of individual projections weighted by $\|p_r\|_2^2$. It is not their unweighted mean and does not establish agreement for every pattern. Null physical differences contribute no projection numerator or denominator, but their hybrid differences can contribute to $\|\mathbf{h}\|_2^2$. These summaries include the full set of tested patterns.

The relative residual satisfies

$$\frac{\|\mathbf{h} - \mathbf{p}\|_2}{\|\mathbf{p}\|_2} = \sqrt{(\beta - 1)^2 + r_\perp^2}. \tag{75}$$

It measures departure from equality between the two response differences. It is not a numerical residual of the exact additive identity $h = p + c$. All reported summaries first aggregate within each seed, then summarize the five seeds.

Table 3: **Instantaneous tracing across the broad pattern set.** Values are five-seed means of concatenated-vector summaries. This set includes both well and scarcely represented patterns; larger supplied differences carry more weight in the projection.

Misspecification	Projection	Cosine similarity	Perpendicular residual
Directional dependence	0.971	0.9975	0.068
Discretization	0.993	0.9999	0.016
Magnitude	0.943	0.9966	0.078
Omitted process	0.958	0.9980	0.060

E.4 Variation across directions in the Figure 2 set

Figure 2 uses the narrower set with $19 \leq \rho \leq 24$: 173 spatial indices, two phases, two perturbed variables, and 12 reference fields per seed. The released instantaneous response records give, for every repeated measurement, the three vector moments $p_r^\top p_r$, $h_r^\top h_r$, and $h_r^\top p_r$. Summing these moments within each seed reproduces the Figure 2 projections 0.980, 0.995, 0.962, and 0.972 as five-seed means in the order used in the main text. The corresponding cosine similarities are 0.999475, 0.999961, 0.999179, and 0.999606.

To examine the variation concealed by a pooled projection, for each seed, spatial index, and perturbed variable, aggregate both phases and all 12 reference fields before calculating

$$d_{s,m,j} = \left(\frac{\sum_r \|h_r - p_r\|_2^2}{\sum_r \|p_r\|_2^2} \right)^{1/2}, \tag{76}$$

where r ranges over the repeated measurements for that seed, index, and variable. The vector norms include both output variables. A direction is included in normalized summaries only if the supplied difference is nonzero and exceeds five times the projected seed background at every included reference field and phase, using the resolution diagnostic defined below. This conservative filter excludes three spatial indices per seed and variable for directional dependence; the other affected variables retain all 173 indices per seed. Table 4 gives descriptive quantiles over the retained directions and five seeds, with the input variable identified explicitly. These directions and reference fields are repeated measurements, so the quantiles are not confidence intervals from independent experiments.

The median normalized residual is at most 0.071 in every affected variable, while the 90th percentile reaches 0.105 for u in the magnitude case and individual directions can depart further. The agreement in Figure 2 therefore applies broadly among resolved directions, with discernible heterogeneity; it is not an equal-error statement for each direction. For omitted v diffusion, perturbing u produces $p = 0$ exactly but a nonzero difference between the trained hybrid responses, whose median RMS is 0.1399. No normalized transmission is defined for that physical null.

Figure 5 groups each affected variable's supplied differences into quartiles within the misspecification and variable. It shows the median and the 10th-90th percentile range of d in each quartile. The lowest-magnitude quartile generally has the larger relative residual; the pooled correspondence is therefore not produced solely by high relative agreement in the strongest supplied-difference directions. This descriptive comparison does not isolate the effects of physical magnitude, spatial frequency, and training representation, and normalization by $\|p\|$ itself affects the trend.

Table 4: **Variation across directions in the Figure 2 set.** Each row summarizes $d_{s,m,j}$ across resolved spatial directions and five seeds after phases and reference fields are pooled within a direction. The last row has zero supplied difference, so it reports the unnormalized RMS hybrid response difference $\|h\|_{\mathrm{RMS}}$ instead of d.

Misspecification	Perturbed variable	Entries	Median	90th percentile
Directional dependence	u	850	0.0442	0.0947
	v	850	0.0277	0.0582
Discretization	u	865	0.0135	0.0224
	v	865	0.0074	0.0114
Magnitude	u	865	0.0707	0.1047
	v	865	0.0368	0.0617
Omitted v diffusion	v	865	0.0309	0.0586
Omitted v diffusion, physical null	u	865	0.1399	0.3158

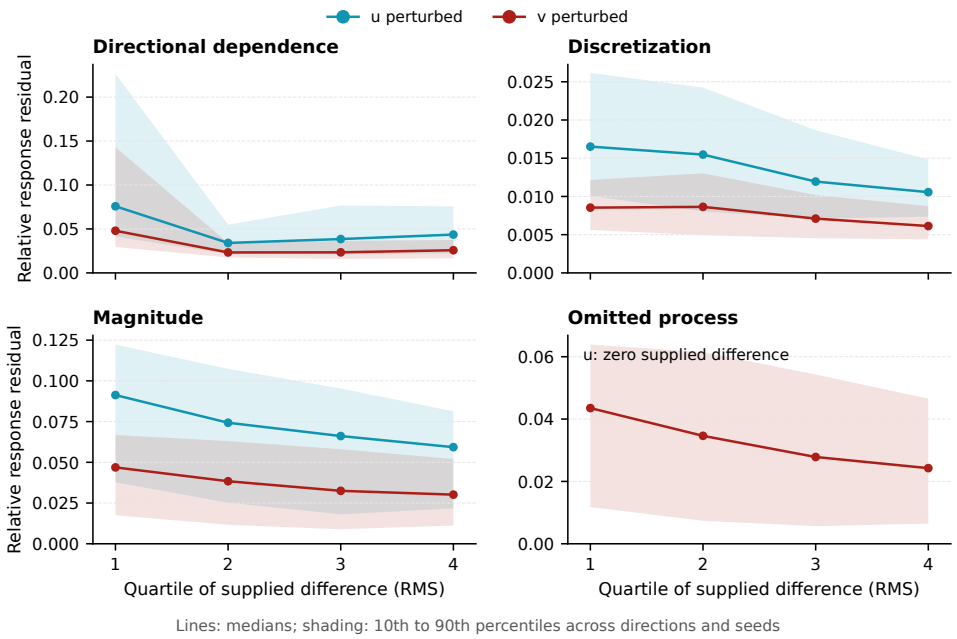

Figure 5: **Variation in tracing across the Figure 2 directions.** Spatial directions with nonzero, resolved supplied differences are ranked by their physical RMS within each misspecification and perturbed variable. Lines show median relative residuals; shading spans the 10th to 90th percentiles across directions and five training seeds. In the omitted-process case, u perturbations have zero supplied difference and are reported in Table 4 using unnormalized RMS. Quartiles are specific to each case and variable, so their physical RMS ranges need not match across panels.

E.5 Variation across seeds, zero-difference cases, and derivative checks

For each condition, pattern, and reference field, the instantaneous analysis computes the standard deviation across the five seeds of the hybrid response projected onto the unit supplied-difference direction, with RMS scaling. The larger standard deviation from the correct and misspecified conditions is the background. A supplied difference is flagged as resolved when its RMS exceeds five times that background. This is a diagnostic based on projected seed variability, not a test using independent reference fields or all pairwise vector distances. Normalized projections and cosines are undefined when their denominators are zero; such cases are interpreted through unnormalized responses. Numerical denominator safeguards in the code do not turn them into scientifically defined normalized quantities.

Selected automatic derivatives are checked against

$$DG(x)q \approx \frac{G(x+\epsilon q) - G(x-\epsilon q)}{2\epsilon}, \qquad \epsilon \in \{0.01, 0.03\}. \tag{77}$$

The tracing checks use seed 0, test trajectory 0 at saved index 16, cosine patterns $(6,0), (10,8), (20,0), (12,12)$, both variables, and all four misspecifications. Instantaneous and saved-step differences are checked separately. The direct replacement checks use seed 0 at the same reference field, patterns $(6,0), (20,0), (0,20)$, both variables, and all four misspecifications. The severity checks use seed 0, multipliers $0.5, 1.5, 3$, and patterns $(6,0), (20,0)$ in both variables. These selected numerical checks do not constitute validation over all five seeds.

F Additional component-replacement analyses

For the saved-step map Φ_{ab} using $P_a + C_b$, define

$$R_{ab}(x,q) = D\Phi_{ab}(x)q, \qquad R_F(x,q) = D\Phi_F(x)q. \tag{78}$$

The four responses $R_{00}, R_{s0}, R_{0s}, R_{ss}$ hold the state, pattern, and seed fixed. The principal comparison replaces R_{ss} by R_{0s}, retaining C_s. The interaction between changing both components is

$$I_s(x,q) = R_{ss} - R_{s0} - R_{0s} + R_{00}. \tag{79}$$

Equivalently, it is the difference between the physical replacement effects measured with C_s and with C_0. It is recorded separately rather than set to zero.

On the two $\rho = 20$ axis patterns, the median of $\|I_s\|_{\mathrm{RMS}}/\|R_{ss} - R_{00}\|_{\mathrm{RMS}}$ over the recorded seed, phase, variable, and reference-field measurements is 0.72%, 0.10%, 1.02%, and 1.82% for directional dependence, discretization, magnitude, and omitted v diffusion, respectively. The omitted-process calculation includes only perturbations of v. These small observed ratios indicate that the physical replacement effect changes little between the two retained corrections for these specific patterns; they do not make the decomposition unique or eliminate interactions at other states.

For one repeated measurement,

$$e_{\mathrm{before}} = \|R_{ss} - R_F\|_{\mathrm{RMS}}, \qquad e_{\mathrm{after}} = \|R_{0s} - R_F\|_{\mathrm{RMS}}. \tag{80}$$

For each seed, average these RMS errors over the selected patterns, phases, and reference fields, obtaining $\bar{e}_{\mathrm{before}}$ and $\bar{e}_{\mathrm{after}}$. The corresponding fraction removed is

$$f_{\mathrm{removed}} = 1 - \frac{\bar{e}_{\mathrm{after}}}{\bar{e}_{\mathrm{before}}}, \tag{81}$$

when the denominator is nonzero. Negative values mean that replacement increases error. This intervention summary averages RMS errors, whereas the severity ratio in Eq. 15 takes a square root after averaging squared errors.

Table 5: **Response error before and after physical replacement.** The scarcely represented comparison uses the $\rho = 20$ coordinate-axis patterns, both phases, and both perturbed variables for the first three cases; omitted v diffusion uses only v perturbations. Error columns are means of the five seed-level mean RMS errors. Percentages are calculated from these unrounded column means, rather than averaged per-record percentages.

Misspecification	Before replacement	After replacement	Error removed
Directional dependence	0.03449	0.00315	90.9%
Discretization	0.02126	0.00260	87.8%
Magnitude	0.03102	0.00346	88.9%
Omitted process	0.08463	0.00199	97.6%

The same saved-step response records contain off-axis perturbations and a diagonal control. Table 6 reports, for each pattern, the mean across five seeds of the fraction removed after first averaging the before and after RMS errors over the two phases and 12 reference fields within each seed. Both input

variables are included, except that the omitted-process case uses only v perturbations. The orientation-specific off-axis fractions are substantially smaller than those for the $\rho = 20$ axes, consistent with the different physical responses and training representation of these patterns. For directional dependence, the $(12, 12)$ symbol equals the correct diffusion symbol, and the replacement has essentially no effect on response error.

Table 6: **Fraction of saved-step response error removed by physical replacement on additional patterns.** Values are percentages, averaged over the five seed-level fractions. A zero value for directional dependence at $(12, 12)$ is the physical null control.

Misspecification	$(10, 8)$	$(8, 10)$	$(12, 12)$
Directional dependence	40.8%	55.9%	0.0%
Discretization	38.0%	38.6%	62.1%
Magnitude	63.8%	63.3%	86.8%
Omitted v diffusion	90.2%	90.7%	96.7%

For the well-represented $\rho = 6$ axis patterns, the ratios of mean error after to before replacement are approximately 10.8, 1.33, 10.9, and 22.9 in the same case order. This is consistent with the data-driven component having adjusted to the supplied physics during training. Replacement does not uniformly improve an already trained hybrid.

F.1 An intervention that preserves the unperturbed prediction

Let

$$E_s(z) = \Phi_{ss}(z) - \Phi_{0s}(z) \tag{82}$$

be the conditional map difference with C_s fixed. For reference state x and pattern q, project using the standardized coordinates:

$$\pi_q(z; x) = x + \frac{\langle D_\sigma^{-1}(z - x), D_\sigma^{-1} q\rangle}{\|D_\sigma^{-1} q\|_2^2} q. \tag{83}$$

Define

$$\widetilde{\Phi}_s(z) = \Phi_{ss}(z) - \big(E_s(\pi_q(z; x)) - E_s(x)\big). \tag{84}$$

At $z = x$, $\widetilde{\Phi}_s(x) = \Phi_{ss}(x)$ exactly in real arithmetic; the implementation records the numerical residual. Moreover, for any scalar ϵ for which the maps are defined,

$$\pi_q(x + \epsilon q; x) = x + \epsilon q, \qquad \widetilde{\Phi}_s(x + \epsilon q) = \Phi_{0s}(x + \epsilon q) + E_s(x). \tag{85}$$

It follows that

$$D\widetilde{\Phi}_s(x)q = R_{0s}(x, q). \tag{86}$$

Its centered finite response along this same line also equals that of direct replacement, since the constant $E_s(x)$ cancels. This equality is a consequence of the construction, not independent evidence for a second mechanism. The control shows how the tested response can be changed while preserving the unperturbed prediction.

The implemented opposite-direction control has response

$$2R_{ss} - R_{0s} = R_{0s} + 2(R_{ss} - R_{0s}). \tag{87}$$

It adds the conditional physical difference once more instead of removing it. It does not reverse the sign of the supplied physical operator or of its contribution relative to R_{0s}. These controls use all five seeds, trajectories $0, 1, 2, 3$ at index 16, cosine patterns $(20, 0), (0, 20)$, both variables, and amplitudes $0.01, 0.03$.

G Severity analysis

The family $P_\alpha = \alpha P_0$ uses the multipliers in Eq. 14; the purely data-driven model is trained separately. For each misspecified α, define

$$R_\alpha = D\Phi_{P_\alpha + C_\alpha}(x)q, \tag{88}$$

$$R_1 = D\Phi_{P_0 + C_1}(x)q, \tag{89}$$

$$R_{\alpha\to 1} = D\Phi_{P_0 + C_\alpha}(x)q. \tag{90}$$

Within this subsection, C_α denotes the network trained with P_α; therefore, C_1 is the network trained with the correct physical component $P_1 = P_0$. For each repeated measurement, set

$$h_\alpha = R_\alpha - R_1, \qquad p_\alpha^{\mathrm{map}} = R_\alpha - R_{\alpha\to 1}, \tag{91}$$

and, for a nonzero denominator,

$$\beta_\alpha^{\mathrm{map}} = \frac{\langle h_\alpha, p_\alpha^{\mathrm{map}}\rangle}{\|p_\alpha^{\mathrm{map}}\|_2^2}. \tag{92}$$

The denominator describes replacing the physical component with the trained C_α held fixed. It depends on that network and the reference state through integration; it is not the operator-only instantaneous difference $(P_\alpha - P_0)q$.

In this analysis, the seed summary of $\beta_\alpha^{\mathrm{map}}$ is the unweighted mean of the per-record coefficients over the selected patterns, phases, variables, and reference fields. Cosine similarities are averaged in the same way. This differs from the concatenated-vector summaries of the broad instantaneous tracing analysis. Five-seed mean projections range from approximately 0.976 at $\alpha = 0.1$ to 0.945 at $\alpha = 2.75$ and 0.947 at $\alpha = 3$. The corresponding mean cosines exceed 0.9992 for every positive misspecified multiplier. The relation is close to proportional over this experiment, but not invariant.

For accuracy, set $E_{\alpha,r} = \|R_{\alpha,r} - R_{F,r}\|_{\mathrm{RMS}}$ and define $E_{d,r}$ similarly for the purely data-driven model. Within each seed,

$$B_\alpha = \sqrt{\frac{\mathrm{mean}_{r\in\mathcal{Q}}\, E_{\alpha,r}^2}{\mathrm{mean}_{r\in\mathcal{Q}}\, E_{d,r}^2}}, \tag{93}$$

which is Eq. 15, not a mean of individual error ratios. The denominator is evaluated for the same seed and measurements.

An operator-only diagnostic makes explicit how much of this severity pattern follows from diffusion attenuation. For spatial index $m = (m_x, m_y)$ and physical variable j, the supplied diffusion map has the pure-mode gain

$$g_\alpha(m,j) = \left[R_4\left(-0.002\,\alpha D_j \pi^2 (m_x^2 + m_y^2)\right)\right]^5, \qquad R_4(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}. \tag{94}$$

There are five RK4 substeps over one saved interval, and the periodic domain has length two. Let σ_j be the training standard deviation of the corresponding physical state variable. Approximating the purely data-driven response on these patterns by the identity, and omitting reaction and learned responses, gives

$$\widetilde{B}_\alpha = \left[\frac{\sum_{m,j} \sigma_j^2 \{g_\alpha(m,j) - g_1(m,j)\}^2}{\sum_{m,j} \sigma_j^2 \{1 - g_1(m,j)\}^2}\right]^{1/2}, \tag{95}$$

where the sum uses the six spatial indices listed in Appendix E.2. This deterministic diagnostic uses the measured state standard deviations $\sigma_u = 0.4394786$ and $\sigma_v = 0.2687279$ and contains no fitted parameter. The resulting curve is close to the trained-model curve at the tested misspecified settings, with the largest absolute difference outside $\alpha = 1$ approximately 0.018 at $\alpha = 2.75$. At $\alpha = 1$ the physical numerator is identically zero, whereas the correct hybrid still has a response error ratio of 0.032. At $\alpha = 2.5$ the approximation gives 1.004 while the trained hybrid gives 0.991: these lie on opposite sides of the data-only ratio of one. Thus the approximation captures the dominant severity trend but does not locate the observed ordering change between the tested settings. The diagnostic describes the dominant physical attenuation in this specific set; it does not by itself explain the learned correction or establish a general severity threshold.

At $\alpha = 2.5$, all five seed ratios are below one; at $\alpha = 2.75$, all are above one. These adjacent tested settings mark the observed change in ordering. No unique or continuous threshold between them is inferred, because the models are trained separately and intermediate settings are not evaluated.

H Limits of training representation as a quantitative predictor

The theory does not identify a single Fourier-coefficient statistic with the complete information available to a nonlinear network. A separate analysis specified four criteria for predicting the retained hybrid difference and the benefit relative to the purely data-driven model from measured training

Table 7: **Severity on the specified scarcely represented patterns.** Entries for B_α are five-seed means. The diffusion-only approximation $\widetilde{B}_\alpha$ is deterministic and uses Eq. 95. Accuracy uses a ratio of root mean squared response errors; the projection uses a within-seed mean of per-record saved-step coefficients. The projection is undefined at the correct multiplier because its supplied comparison difference is zero.

α	B_α	$\widetilde{B}_\alpha$	Mean $\beta_\alpha^{\mathrm{map}}$
0.10	0.887	0.884	0.976
0.25	0.720	0.717	0.975
0.50	0.461	0.457	0.974
0.75	0.224	0.219	0.971
1.00	0.032	0.000	-
1.25	0.200	0.202	0.966
1.50	0.386	0.388	0.967
1.75	0.556	0.560	0.965
2.00	0.713	0.720	0.960
2.50	0.991	1.004	0.951
2.75	1.113	1.131	0.945
3.00	1.233	1.249	0.947

representation. Its implementation records predictions before loading the comparison outcomes within that notebook. However, the saved timing audit shows that these predictions postdate the existing severity outcome table from the preceding experiment. The calculations therefore do not constitute a prospective preregistered test of those severity outcomes.

None of the four declared overall criteria passed:

1. The required strong monotonic association was not obtained for every primary misspecification. The weakest Spearman rank correlation was approximately -0.398, against a required value at most -0.8.
2. Fitting on the other misspecification forms and predicting the excluded form gave a worst mean absolute error of approximately 1.053 on scarcely represented patterns, exceeding the limit 0.1.
3. Predicting which conditions would outperform the purely data-driven model gave three incorrect classifications relative to an error ratio of one.
4. The predicted upper crossing, $\alpha \approx 2.259$, lay outside the adjacent tested settings 2.5 and 2.75 where the observed ordering changes.

These negative results are retained. They show that the tested use of representation did not reliably predict the retained physical difference across misspecification forms or the measured loss of extrapolation advantage. They do not establish that every possible predictor using training data must fail.

I Reporting conventions

The five seeds are independent training replicates, with paired comparisons across conditions within each seed. Patterns, phases, reference fields, perturbed variables, and field elements are repeated measurements. They are aggregated within a seed before cross-seed summaries, and the figures display all five seed values. Seed variation is conditional on the fixed trajectory datasets; variation across independently regenerated datasets is not estimated.

Principal checkpoints are selected using validation loss only; the final epoch is retained separately for checkpoint comparisons. Test trajectories and response measurements are not used for this selection. Instantaneous response differences, true saved-step response errors, and conditional saved-step replacement effects are defined separately throughout. The former admit an exact additive identity, while the integrated component effects can interact.