DOES ACCURACY IMPLY CORRECTNESS IN TEMPORAL EVOLUTION FORECASTING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Claims of transferable temporal evolution require response evidence matched to the declared condition changes. Low error on natural trajectories alone does not supply that evidence. Temporal Evolution Forecasting (TEF) expresses this behavioral requirement through declared information interfaces and conditional-response diagnostics. Controlled experiments show why familiar evidence cannot substitute for these tests: longer histories can improve natural accuracy while amplifying response error, an autoregressive model's local advantage can disappear during rollout, and mechanism-matched positive controls can recover the intended response. Across seven real-world benchmarks, forecast revisions and input sensitivity distinguish properties that natural error alone does not describe, without supplying intervention-response ground truth. Authors should therefore state which evolution capability they claim, evaluate natural forecasts alongside responses to the relevant changes, and establish that their diagnostics reward the intended behavior. The six protocols studied here illustrate this requirement; the appropriate evidence depends on the claim and on which paired outcomes the setting can identify.

1 INTRODUCTION

Our position is that claims of transferable temporal evolution require response evidence matched to the declared condition changes. Natural forecast accuracy establishes predictive performance on the observed distribution. A stronger claim, such as reliable response to a new state or operating condition, creates an additional evidential obligation: the forecast must reproduce how the future changes under that condition.

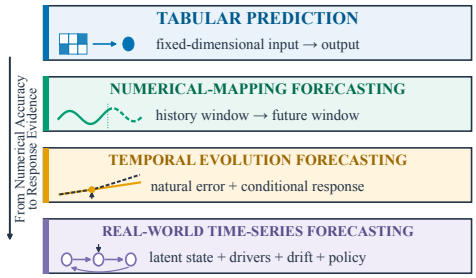

Figure 1: Forecasting tasks and their evidence requirements. TEF adds conditional responses to natural error; real-world forecasting also addresses latent states, drivers, drift, and policy feedback. The hierarchy is conceptual.

Time-series forecasting predicts a future window from a history window and usually evaluates error on natural sequences (Godahewa et al., 2021; Aksu et al., 2024; Zhou et al., 2020). Those sequences arise from an evolving system (Brunton et al., 2015; Chen et al., 2018), but low error can reflect either state evolution or stable correlations in shape, periodicity, and trend. Models without explicit temporal structure can also achieve low benchmark error (Li et al., 2023; Toner & Darlow, 2024). Natural accuracy therefore establishes closeness to an observed future without determining response consistency after conditions change.

Numerical-mapping forecasting fits the correspondence between historical and future values. Periodicity (Lin et al., 2024), trend (Zeng et al., 2023), and tabular correlations (Li et al., 2023; Toner & Darlow, 2024) can support this objective. **Temporal Evolution Forecasting** adds a behavioral requirement: responses should remain consistent with the target system under declared condition changes and information interfaces. Figure 1 organizes these evidence requirements. TEF does not require identifying a unique internal mechanism.

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.

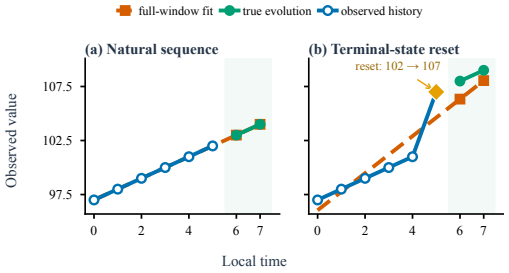

Figure 2: Template fitting is exact on the natural sequence (a) but under-responds to a terminal-state reset (b), predicting $[106.33, 108.05]$ instead of the true future $[108, 109]$.

Figure 2 illustrates the distinction. Information pathways that agree on natural sequences can respond differently after a declared state reset. Whether that reset is relevant depends on the capability being claimed. We develop the position through controllable systems whose paired futures are known. Six diagnostics vary state coverage, parameters, forecast origins, or operating regimes; comparisons among machine-learning models (ML), deep-learning models (DL), and foundation models (FM) test which evidence supports an evolution claim. Seven real-world benchmarks then distinguish natural accuracy, forecast revision, and input sensitivity using the outcomes available from observations.

The argument has three consequences for evaluation:

- ❶ Natural accuracy and conditional response can separate. Paired counterexamples establish why low error on natural futures cannot, by itself, validate forecasts after a condition change.
- ❷ History length, model scale, and autoregressive form cannot substitute for response evidence. History and regularization ablations, rollout comparisons, and cross-model results demonstrate this distinction under the tested interfaces.
- ❸ Evolution claims should specify their information interface, condition changes, and corresponding diagnostics. Mechanism-matched positive controls check that the diagnostics reward the intended response; TEF and the six protocols instantiate this requirement.

2 Related Work

Existing forecasting benchmarks primarily compare natural forecast error under fixed data splits and forecast horizons. Monash extends evaluation across domains, while GIFT-Eval broadens the evaluation of general-purpose models (Godahewa et al., 2021; Aksu et al., 2024). Deep forecasting has progressed from autoregressive recurrence, deep state-space models, and residual bases (Salinas et al., 2020; Rangapuram et al., 2018; Oreshkin et al., 2019) to multi-horizon fusion, hierarchical interpolation (Lim et al., 2021; Challu et al., 2023), direct linear prediction, patching, and cross-variable modeling (Zeng et al., 2023; Nie et al., 2022; Liu et al., 2024). Convolutional interaction, pyramidal attention, frequency-domain memory, and Koopman or frequency-domain MLPs further provide multiscale and non-stationary modeling routes (Liu et al., 2022; 2021; Zhou et al., 2022; Liu et al., 2023; Yi et al., 2023). In addition to TimesFM, Chronos, Moirai, Time-MoE, Chronos-2, and Toto 2.0 (Das et al., 2024; Ansari et al., 2024; Woo et al., 2024; Shi et al., 2025; Ansari et al., 2025; Khwaja et al., 2026), pretraining approaches include language-model transfer and unified multitask methods such as One Fits All, Time-LLM, TEST, and UniTS (Zhou et al., 2023; Jin et al., 2024; Sun et al., 2024; Gao et al., 2024). These studies substantially expand model and data scale, but their central evidence still comes primarily from output error on natural trajectories.

A complementary line of work learns repeatedly applicable state transitions through system identification and continuous-time models (Brunton et al., 2015; Chen et al., 2018), emphasizing structural relationships among state, operating conditions, and evolution rules. We connect these two lines of research. Rather than requiring a predictor to adopt an explicit dynamics architecture, we test whether low natural error supports the behavioral claim of reusable evolution. Paired condition changes distinguish numerical mapping on the natural-history manifold from response consistency after a condition changes, without equating success on finite diagnostics with recovery of a unique equation or causal mechanism (Runge et al., 2019).

3 Numerical-Mapping Forecasting

Structured data forecasting. Let $x_t \in \mathbb{R}^D$ denote the observation of a system at time t, and let L and H denote the lengths of the history and forecast windows, respectively. The corresponding

history and future windows are

$$X_t = [x_{t-L+1}, \ldots, x_t] \in \mathbb{R}^{L\times D}, \qquad Y_t = [x_{t+1}, \ldots, x_{t+H}] \in \mathbb{R}^{H\times D}. \tag{1}$$

A predictor parameterized by θ, $G_\theta : \mathbb{R}^{L\times D} \to \mathbb{R}^{H\times D}$, maps a history window to a future prediction, $\widehat{Y}_t = G_\theta(X_t)$. Let P_{nat} denote the joint distribution of window pairs (X, Y) generated by the system under natural, unmodified conditions. The population objective is to minimize the natural forecasting risk

$$R_{\text{nat}}(G_\theta) = \mathbb{E}_{(X,Y)\sim P_{\text{nat}}}\left[\ell\big(G_\theta(X), Y\big)\right], \tag{2}$$

where ℓ is a forecasting loss and P^X_{nat} is the marginal distribution of histories. Vectorizing X and Y makes this objective multi-output regression: historical observations supply a fixed-dimensional input and the future window supplies the supervised target.

The distinctive structure of time-series data. Unlike arbitrary tabular feature columns, positions within a time-series window carry temporal meaning. Adjacent windows often overlap and are dependent. Predictors can exploit lagged relationships, shapes, trends, periodicity, and cross-variable dependencies without requiring these relationships to remain valid after a condition change.

A numerical objective does not identify a conditional response. The numerical-mapping objective rewards accurate future values without requiring the model to recover the true mechanism internally, so directly modeling observational relationships is often effective. Current algorithms emphasize dependencies among time-series observations (Nie et al., 2022; Hu et al., 2025; Zeng et al., 2023; Wang et al., 2024; Shi et al., 2025); tabular-prediction techniques can also achieve strong forecast accuracy (Toner & Darlow, 2024; Chen & Guestrin, 2016). A model may hard-match shape, trend, and cross-step relationships in a long window and implicitly generate intermediate forecasts. These are valid routes to accurate prediction. The evidential gap appears when natural accuracy is used to support a claim about responses outside the evaluated conditions.

4 Temporal Evolution Forecasting

Numerical accuracy is an adequate criterion when the sole objective is to predict another sample from the same observational distribution. Once forecasting follows a change in state, operating condition, or forecast origin, however, accuracy alone cannot establish that a model has learned a transferable rule. A forecasting rule must also reproduce how the future responds to the declared condition change. Beyond the numerical compatibility measured by Equation (2), we therefore test whether forecast behavior exhibits reusable consistency with temporal evolution.

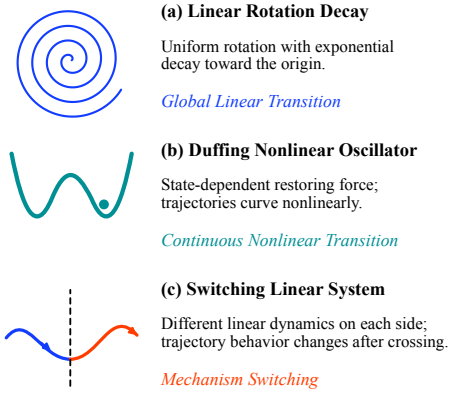

Figure 3: Illustrative mechanisms of the three controlled systems.

4.1 Temporal Evolution Mechanisms

A real process comprises both state evolution and observation. Let s_t be a latent state, ϕ a system condition, and ε_{t+1} and η_t process and observation disturbances, respectively. These disturbances represent state changes that occur at the current time and other changes that are difficult to predict; they need not be mere noise. The process is

$$s_{t+1} = F_\phi(s_t, \varepsilon_{t+1}), \qquad x_t = \mathcal{O}(s_t, \eta_t). \tag{3}$$

The state-space view requires the current state to summarize the historical information needed to predict the future: conditional on s_t, earlier states provide no additional information about the next transition (Kalman, 1960; Baum & Petrie, 1966; Shalizi & Crutchfield, 2001). Under full observability, $x_t = s_t$; under partial observability, a history encoder $z_t = \Psi(X_t)$ must instead form a predictively sufficient current state.

Temporal Evolution Forecasting specifies a behavioral requirement under declared condition changes. We do not attempt to recover every disturbance, latent variable, or unique internal mechanism in Equation (3) from finite observations. Our controlled experiments fix the disturbances and equate observations with states, testing whether forecasts and paired responses agree with repeated application of the known transition. Section 8 further relates this controlled requirement to real-world forecasting. Under these conditions,

$$x_{t+h} = F_\phi^h(s_t), \quad h = 1, \dots, H. \tag{4}$$

The same one-step transition is repeatedly applied to generate the multi-step future. *Temporal Evolution Forecasting* requires forecast behavior to be compatible with this evolution, but it does not require the model to explicitly recover F_ϕ, a unique causal mechanism, or every latent variable. Section 5 operationalizes this behavioral requirement with paired futures.

4.2 Partial Equivalence Between Temporal Evolution and Numerical Mapping

A long history can implicitly represent an evolution mechanism. With a long input, a model can map multiple past values directly to the future. Here, "direct mapping" describes a statistical information pathway; it does not imply that a model cannot generate intermediate states internally.

As illustrated in Figure 2(a), consider an arithmetic-progression system with state (a_t, d_t) and observation $x_t = a_t$, where the increment d_t remains fixed. The true one-step transition is $(a_t, d_t) \mapsto (a_t + d_t, d_t)$. A natural trajectory satisfies $x_t - x_{t-1} = d_t$, so the same future observation can also be written as $\widetilde{x}_{t+1} = 2x_t - x_{t-1}$. A model can use this multi-step observational relation to produce the next value, then continue the rollout with $\widetilde{x}_{t+1}$ as an implicit intermediate state. This recurrence is not the true one-step evolution of the sufficient state (a_t, d_t). Under our behavioral definition, however, it remains a valid TEF predictor whenever its forecasts and paired responses agree with the declared evolution. Determining whether the model internally recovered the true physical mechanism is outside the scope of this work.

Template-matching risk in a long history. A long history may be necessary to recover a sufficient state under partial observability. When the current state is already sufficient, however, earlier observations can become redundant shortcuts. Figure 2(b) shows an explicit long-window mapping for a fixed-length arithmetic sequence: treat every history point as a sample on the same line, fit the line, and forecast by evaluating it at future times.

For a history $X_t = [x_{t-L+1}, \dots, x_t]$ of length L, fit a linear template at local indices $j = 0, \dots, L-1$:

$$\begin{aligned} (\widehat{b}, \widehat{d}) &= \arg\min_{b,d} \sum_{j=0}^{L-1} \left(x_{t-L+1+j} - b - dj\right)^2, \\ \widehat{x}_{t+h}^{\text{fit}} &= \widehat{b} + \widehat{d}(L - 1 + h). \end{aligned} \tag{5}$$

On the natural history $X_t = [97, 98, 99, 100, 101, 102]$ with $L = 6$ and $d_t = 1$, Equation (5) predicts $[103, 104]$ exactly. Now keep the earlier history and increment fixed but reset the terminal state from 102 to 107, giving $X_t^{(c)} = [97, 98, 99, 100, 101, 107]$. The true future is $[108, 109]$, whereas refitting the same template yields $[106.33, 108.05]$. Thus, the predictor responds to the edit only partially: the unchanged earlier history pulls the fit toward the old trajectory. Supplementary Appendix A gives the closed-form coefficients, fractional predictions, and full derivation for both numerical cases.

5 Matching Diagnostics to Evolution Claims

Temporal Evolution Forecasting is a behavioral requirement. Natural and OOD protocols evaluate forecast accuracy, while Restart compares overlapping forecasts from two origins. For state and regime edits, we run the known transition under the original and changed conditions, then compare the true and predicted future changes. These paired responses evaluate the declared capability without identifying the model's internal rule.

The condition change determines the evidence needed. Transfer across state regions calls for observations from those regions; a state-reset claim calls for the corresponding reset response; a regime-selection claim calls for evidence about the regime change. The protocols below instantiate these distinct obligations in controllable systems.

5.1 Temporal Evolution Systems

To make paired responses exact and reproducible, we use three classes of minimal systems with known mechanisms and controllable conditions. For all two-dimensional systems, let $x_t = (q_t, v_t)^\top \in \mathbb{R}^2$. Their transitions are

$$\begin{aligned} \text{Linear: } x_{t+1} &= \rho \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix} x_t, \\ \text{Duffing: } q_{t+1} &= q_t + \Delta t\, v_t, \\ v_{t+1} &= v_t + \Delta t(-k q_t - c v_t - \beta q_t^3), \\ \text{Switching: } x_{t+1} &= \begin{cases} A_+ x_t, & q_t \geq 0, \\ A_- x_t, & q_t < 0. \end{cases} \end{aligned} \tag{6}$$

Here ρ and ω control decay and rotation; Δt is the integration step; k, c, and β are the Duffing stiffness, damping, and cubic-nonlinearity parameters; and A_+ and A_- are the two regime-specific linear maps. Linear tests a globally shared transition, Duffing tests continuous nonlinearity and parameter dependence, and Switching tests discrete mechanism selection. All three systems can be re-simulated exactly, enabling paired ground-truth futures before and after a condition change. Supplementary Appendix B gives their numerical values, support sets, and edit magnitudes.

5.2 Diagnostic Protocols

Table 1 lists six diagnostics, applied where the system supplies the relevant condition change. State and Parameter OOD are reported for Linear and Duffing; Switching instead tests the regime edit. We do not combine the protocols into a single score because they rule out different alternative explanations: extrapolation beyond the observed support does not substitute for state response; a temporally consistent but inaccurate forecast may still deviate from the true future; and an ordinary state edit cannot show that a model switches operating regimes correctly.

5.3 Evaluation Metrics

Natural, State OOD, and Parameter OOD forecasts use nMSE, normalized by the per-channel training variance and then averaged across channels:

$$\mathrm{nMSE} = \frac{1}{D} \sum_{d=1}^{D} \frac{(NH)^{-1} \sum_{i,h} (\widehat{Y}_{i,h,d} - Y_{i,h,d})^2}{\max(\sigma^2_{\mathrm{train},d}, \epsilon)}. \tag{7}$$

For temporal restart, the model forecasts from times t and $t+k$ on the same true sequence. Using the same denominator, we compare the two predictions over their $H-k$ overlapping future steps. A terminal-state edit leaves the earlier history unchanged, declares the most recent observation to be a new restart state, and runs the simulator from the original and edited states. The edit generally creates a non-natural history window. It therefore falsifies whether a predictor can restart reliably through the *declared state-reset interface*; it is not an unconditional judgment about mechanism learning within naturally generated sequences.

For N edited pairs under random seed s, define the true future change $\Delta Y_i = Y_i^{(c)} - Y_i$ and predicted change $\Delta \widehat{Y}_i = G_\theta(X_i^{(c)}) - G_\theta(X_i)$. Effect Error is a global ratio over every sample, horizon, and state dimension for that seed:

$$E^{(s)}_{\mathrm{effect}} = \frac{\sum_{i=1}^{N} \|\Delta \widehat{Y}_i^{(s)} - \Delta Y_i^{(s)}\|_F^2}{\sum_{i=1}^{N} \|\Delta Y_i^{(s)}\|_F^2 + \epsilon}, \qquad \epsilon > 0, \tag{8}$$

where $\epsilon = 10^{-12}$. We first compute this global ratio for each seed and then take the arithmetic mean over $s \in \{0, 1, 2\}$. Effect Error is neither an average of IID, OOD, Restart, and Sign Accuracy nor a simple mean of per-sample ratios. $E_{\mathrm{effect}} = 0$ denotes the correct response, while a value near 1 matches the zero-response baseline. Directional accuracy is

$$\mathrm{SignAcc} = \frac{1}{NHD} \sum_{i,h,d} \mathbf{1}\left[\mathrm{sgn}(\Delta \widehat{Y}_{i,h,d}) = \mathrm{sgn}(\Delta Y_{i,h,d})\right]. \tag{9}$$

Table 1: Six complementary protocols and their diagnostic targets.

Protocol	Changed or compared condition and primary diagnostic target
IID	Independently generated sequences from the same mechanism and state coverage; verifies basic fitting ability.
State OOD	Changes initial-state coverage while holding equations and parameters fixed; tests transfer across state regions.
Param. OOD	Changes ϕ while retaining the equation family; tests conditional generalization within that family.
Restart	Forecasts overlapping futures from two origins on the same sequence; tests compositional consistency.
State edit	Restarts the simulator from a new terminal state to generate a paired future; tests the declared state-reset interface.
Regime edit	Moves the terminal state across the Switching threshold; tests operating-regime selection and switching.

The horizon-wise error $E(h)$ recomputes Equation (8) with the Frobenius norms restricted to forecast step h. Finite-difference Jacobians estimate $\partial \widehat{Y}_h / \partial x_t$ with the per-channel central-difference step $\eta_d = \max(10^{-3}\sigma_{x_t,d}, 10^{-5})$. Their relative error is $\|\widehat{J}_h - J_h\|_F / (\|J_h\|_F + \epsilon)$, averaged over samples and seeds. Because a cross-variable response requires an output to read the edited channel, aggregate Effect Error and Sign Accuracy compare only multivariate (MV) models. Channel-independent (CI) models still participate in IID, OOD, and restart evaluation. The supplementary material gives the restart formula and the information-interface lower bound.

6 Empirical Evidence for the Position

The controlled experiments test whether natural accuracy, autoregressive form, and model scale support the response behavior being claimed. Figures 4 and 5 summarize the accuracy–response distinctions.

6.1 Experimental Setup

We use the systems and protocols in Section 5. D2, D4, and D8 have state dimension, history length, and horizon $(D, L, H) = (2, 16, 16), (4, 32, 32), (8, 64, 64)$, respectively: one, two, and four weakly coupled state pairs (q, v). Splits contain independent sequences; Appendix B gives exact parameters, support sets, edits, and sample counts. Trainable models use seeds $0, 1, 2$, with configurations selected solely by full-H-step natural-validation nMSE. Test, OOD, restart, and edit labels never enter selection. Results are three-seed means or means $\pm$ sample standard deviations.

6.2 History Length and Autoregressive Rollout

Figure 4 retrains Window Direct Ridge on fixed D2 Duffing splits, varying history length or regularization under the same terminal-state reset. As L' grows from 1 to 16, IID nMSE falls from 1.01×10^{-3} to 6.72×10^{-6} while Effect Error rises from 0.0162 to 1.46×10^{5}. Raising α from 10^{-8} to 1 reduces the latter to 0.493, still above the $L' = 1$ result. The high condition number documented in the appendix makes the largest error a sign of off-support sensitivity rather than a stable effect-size estimate. Better natural fit did not establish reliable response to the reset.

Table 2 compares parameter-matched Direct-last and One-Step AR checkpoints on the same paired edits. On Duffing, AR has lower response error at step 1 but higher error at step 16; its relative Jacobian error grows faster on all three systems. Linear and Switching show no first-step advantage. Local accuracy therefore cannot replace a test of the composed response.

6.3 Positive Controls and Complementary Diagnostics

A diagnostic must admit the intended response through its declared information interface. Table 3 adds transition information to natural-validation-selected Ridge models: Generic Poly3 uses cubic state features, the Duffing parameter-visible model receives the parameters and matched interactions,

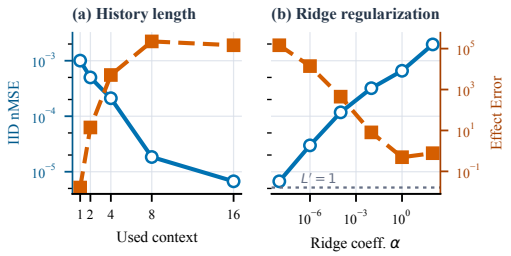

Figure 4: Duffing history and regularization ablations. Longer history lowers IID nMSE but amplifies Effect Error; regularization reduces amplification. Curves show three-seed means ± standard deviations.

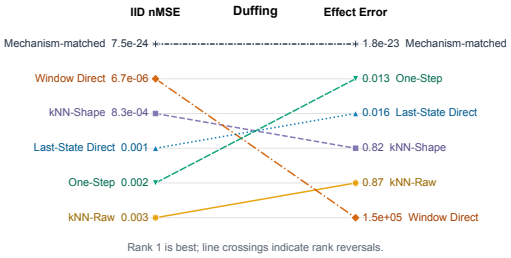

Figure 5: IID nMSE and Effect Error ranks for six models on D2 Duffing. Endpoints give values; crossing lines show ranking reversals. Mechanism-matched is Polynomial One-Step Ridge.

Table 2: D2 Effect and relative Jacobian errors for matched Direct-last and One-Step AR MLPs. Bold marks the lower three-seed mean.

System	Model	$E_1 \downarrow$	$E_{16} \downarrow$	**Jacobian error** $(1 \to 16) \downarrow$
Linear	Direct-last	**0.000272**	**0.00074**	**0.0196** $\to$ **0.0274**
	One-Step AR	0.001895	0.09493	0.0425 $\to$ 0.3123
Duffing	Direct-last	0.000405	**0.01394**	0.0209 $\to$ **0.1117**
	One-Step AR	**0.000187**	0.04453	**0.0129** $\to$ 0.1775
Switching	Direct-last	**0.001761**	**0.02343**	0.0994 $\to$ **0.3496**
	One-Step AR	0.003710	0.11654	**0.0986** $\to$ 0.4233

and Switching True-gate uses the actual branch rule. Generic cubic features help partially; the parameter-visible Duffing and True-gate Switching controls achieve near-zero Effect Error, as does Linear One-Step under the tested Linear conditions. The controls establish that the target response is achievable with the stated information access and representation.

Table 3: D2 16-step Effect Error; bold marks mechanism-aware controls (mean ± s.d.).

System	Model/interface	Effect Error $\downarrow$
Linear	**Linear One-Step**	$\mathbf{3.17 \times 10^{-22} \pm 2.21 \times 10^{-25}}$
Duffing	Linear One-Step	0.0162 ± 0.0004
Duffing	Generic Poly3	0.00567 ± 0.00042
Duffing	**θ-visible Poly3**	$\mathbf{2.18 \times 10^{-19} \pm 2.06 \times 10^{-20}}$
Switching	Linear One-Step	0.2165 ± 0.0030
Switching	Generic Poly3	0.0294 ± 0.0008
Switching	**True-gate Piecewise**	$\mathbf{6.26 \times 10^{-22} \pm 4.51 \times 10^{-23}}$

In Figure 5, Window Direct ranks second by IID nMSE but last by Effect Error, whereas One-Step rises from fifth to second. Table 4 also shows that low OOD or Restart error can coexist with high Effect Error. These diagnostics address different claims. The comparisons span direct, last-state, and one-step Ridge or kNN predictors, deep forecasters, and frozen foundation models. The joint-scale results in Appendix F.2 show that scale, depth, and pretraining do not guarantee conditional responses under the tested interfaces; they do not define a cross-category architecture ranking. Appendix C specifies the configurations and information-access rules, including the joint-state and Duffing-parameter inputs of the main Ridge and kNN baselines.

Table 4: Representative D2 results across natural, OOD, restart, and paired-response protocols. Models reuse configurations selected only by natural validation nMSE; entries average three seeds. "–" marks undefined protocols. Complete results and split details appear in the supplement.

Category	System	Model	IID	State OOD	Param. OOD	Restart	Effect	Sign
ML	Duffing	Window Direct Ridge	6.72×10^{-6}	9.47×10^{-4}	2.39×10^{-4}	1.46×10^{-5}	1.46×10^{5}	0.7997
ML	Duffing	Last-State Direct Ridge	0.0010	0.0096	0.0090	9.18×10^{-4}	0.0162	1.000
ML	Switching	One-Step Ridge	0.4524	–	–	0.2272	0.1866	0.8312
DL	Duffing	iTransformer	7.92×10^{-4}	0.0091	0.0087	9.93×10^{-4}	0.7164	0.8608
DL	Duffing	One-Step MLP	0.0077	0.1086	0.0151	0.0035	0.0638	0.9997
DL	Switching	One-Step MLP	0.0133	–	–	0.0079	0.0228	0.9570
FM	Duffing	Chronos-2	0.0489	0.6760	0.1057	0.0793	1.3088	0.6982
FM	Duffing	Moirai	0.4823	1.2210	0.7364	0.4427	0.7182	0.7048
FM	Switching	Moirai	4.0632	–	–	6.4896	5.4758	0.5056

7 Real-World Forecasting Diagnostics

Real-world observations reveal forecast errors and changes in predictions, but generally not paired futures under interventions. We examine accuracy, forecast revisions, and terminal-input sensitivity as distinct observable properties in seven benchmarks. Figure 6 summarizes the main comparisons.

Setup. We use ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Traffic, and Weather, with chronological splits and training-set standardization. All channels and eligible test origins are included at stride one. Standardizing each channel with its training statistics makes MSE equal to the training-variance-normalized nMSE used above. The history length is $L = 192$ and horizons are $H \in \{96, 192\}$. Six prediction rules cover TrainMean, Persistence, SeasonalNaive, Last-value Direct Ridge, Window Direct Ridge, and Window AR Ridge. Each Ridge model fits a channel-independent map shared across channels. Regularization is selected by full-horizon natural validation MSE, including recursive rollout for AR, over $\lambda \in \{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 100\}$. Appendix O provides the splits, model definitions, and complete results.

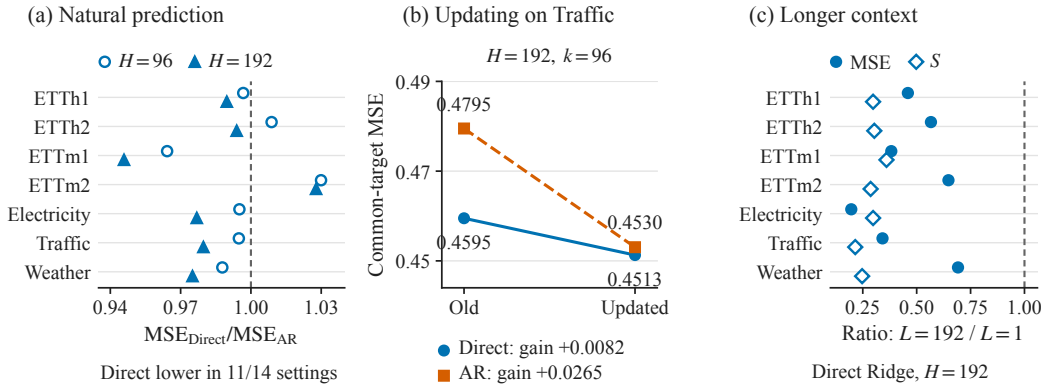

Figure 6: Real-world diagnostics. (a) Direct/AR natural MSE ratios; values below one favor Direct. (b) Traffic forecasts before and after updating, scored on common targets. (c) Ratios of MSE and terminal sensitivity at $L' = 192$ versus 1, selecting regularization separately. Better accuracy, larger update gain, and lower sensitivity are distinct properties.

Accuracy and useful revisions. Window Direct obtains the lowest natural MSE in 11 of the 14 dataset–horizon settings, and Window AR leads in the remaining three. We advance the forecast origin by $k \in \{H/4, H/2\}$ steps and compare both forecasts on their overlapping future. Revision measures their mean squared difference. Update gain is old MSE minus updated MSE on that common target. On Traffic at $H = 192$ and $k = 96$, Direct has lower natural MSE than AR, 0.4583 versus 0.4678, but a smaller update gain, 0.0082 versus 0.0265. Direct still has lower updated MSE, 0.4513 versus 0.4530, because the old forecasts differ in accuracy. Revision, improvement, and final error thus

answer different questions. TrainMean has zero revision because it ignores new observations. Across all rules and offsets, seven of 168 update gains are negative.

Context and sensitivity. For Direct Ridge, terminal sensitivity is $S = (H^{-1} \sum_h \widehat{J}_h^2)^{1/2}$, where $\widehat{J}_h$ is the derivative of lead-h prediction with respect to the final standardized input. Comparing context lengths of one and 192 steps, both natural MSE and S are lower at the longer endpoint on all seven datasets at $H = 192$. On Electricity, MSE falls from 0.8303 to 0.1633 and S from 0.4126 to 0.1229. Reduced sensitivity alone does not establish better prediction: on ETTh2, increasing regularization from 10^{-8} to 100 lowers S from 0.2218 to 0.0036 while raising MSE from 0.4253 to 1.1749. Without paired intervention outcomes or a true system Jacobian, they cannot determine whether its response matches the generating process. The controlled tests supply the paired-future ground truth needed to assess response correctness.

Observed state variation. We also group test origins by the RMS of their final standardized state, using the training 95th percentile as a threshold. Larger observed states do not uniformly imply larger errors. At $H = 192$, Window Direct has MSE 0.3372 on Weather's 88 tail origins versus 0.2233 on the other 10,260. ETTh1 reverses this ordering: MSE is 0.3863 on 216 tail origins versus 0.4234 on the other 2,473. These groups describe observed variation; a training quantile does not establish a change of mechanism or an out-of-support state.

8 Implications and Alternative Views

The two experiment settings specify complementary evidence: observed futures support accuracy and revision diagnostics, whereas paired interventions test response correctness. Appendix N.1 formalizes forecasting with hidden states, drivers, and changing conditions. Identifiable interventions can support average-response tests; when actions alter the process, the conditioning policy must be explicit (Runge et al., 2019). The distinction also applies to scenario forecasts, physics/operator models (Raissi et al., 2019; Lu et al., 2021), and world models (Hafner et al., 2025).

Accuracy is sufficient on the same distribution. Suitable error and calibration measures can address that objective. Response evidence becomes necessary when the claim extends to transferable evolution, mechanism learning, or a specified condition change.

An unnatural state edit may be unfair. A terminal-state edit is relevant only when the claimed interface permits a reset through that observation. Failure challenges reliability under that interface, rather than proving that no dynamic regularities were learned.

No fixed suite covers every evolution claim. The six protocols illustrate distinct obligations in controllable systems. Diagnostics should match the claimed changes and establish ground truth or identification assumptions; a score cannot compensate for an unidentified target.

9 Conclusion

Transferable evolution claims require evidence about the responses they assert. Our examples show why accuracy, history length, architecture, and scale cannot supply it alone. On real-world benchmarks, forecast revisions and input sensitivity add distinct information to natural error, while paired response correctness requires additional ground truth. Authors should specify the claimed capability, information interface, and condition changes. Evaluations should report natural forecast quality alongside relevant response diagnostics, using positive controls to check the intended behavior. Reviewers should assess whether that evidence supports the claim. The six protocols instantiate this recommendation for controllable systems; the evidential obligation follows the claim, not a fixed list of tests.

AI USE STATEMENT

We used large language models (LLMs) for code generation, experimental data analysis, figure generation, manuscript polishing, and translation. We also used LLMs to assist with the proofs in Appendix and to organize the data presented in the appendices. All other LLM-generated content was reviewed by the human authors. We take full responsibility for all content of this paper.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data: Sparse identification of nonlinear dynamical systems. *arXiv preprint arXiv:1509.03580*, 2015.

Cristian Challu, Kin G Olivares, Boris N Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. Nhits: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 6989–6997, 2023.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*, 2018.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first international conference on machine learning*, 2024.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: A unified multi-task time series model. *arXiv preprint arXiv:2403.00131*, 2024.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

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

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. Timefilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *International Conference on Machine Learning*, pp. 24893–24911. PMLR, 2025.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International conference on learning representations*, volume 2024, pp. 23857–23880, 2024.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, et al. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026.

Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. Revisiting long-term time series forecasting: An investigation on linear mapping. *arXiv preprint arXiv:2305.10721*, 2023.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4):1748–1764, 2021.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Advances in neural information processing systems*, 35:5816–5828, 2022.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*, 2021.

Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. Koopa: Learning non-stationary time series dynamics with koopman predictors. *Advances in neural information processing systems*, 36:12271–12290, 2023.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

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

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31, 2018.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3):1181–1191, 2020.

Cosma Rohilla Shalizi and James P Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of statistical physics*, 104(3):817–879, 2001.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International conference on learning representations*, volume 2025, pp. 34635–34667, 2025.

Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. Test: Text prototype aligned embedding to activate llm's ability for time series. In *International Conference on Learning Representations*, volume 2024, pp. 37854–37881, 2024.

William Toner and Luke Nicholas Darlow. An analysis of linear time series forecasting models. In *Forty-first International Conference on Machine Learning*, 2024.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 38626–38652, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. *Advances in neural information processing systems*, 36:76656–76679, 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *arXiv preprint arXiv:2012.07436*, 2020.

Tian Zhou, Ziqing Ma, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, Rong Jin, et al. Film: Frequency improved legendre memory model for long-term time series forecasting. *Advances in neural information processing systems*, 35:12677–12690, 2022.

Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.

This supplement provides derivations, experimental settings, full hyperparameter ranges, and complete result tables for the main paper. Section and figure references without an appendix letter refer to the main paper.

A Full Derivation of the Linear-Template Example

This section expands the two arithmetic-progression examples summarized in Figure 2 and Section 4.2. For a history $X_t = [x_{t-L+1}, \dots, x_t]$, assign local indices $j = 0, \dots, L-1$ and fit

$$x_{t-L+1+j} \approx b + dj$$

by ordinary least squares. Let

$$\bar{\jmath} = \frac{L-1}{2}, \qquad \bar{x} = \frac{1}{L}\sum_{j=0}^{L-1} x_{t-L+1+j}.$$

The fitted slope and intercept in Equation (5) are

$$\widehat{d} = \frac{\sum_{j=0}^{L-1}(j-\bar{\jmath})(x_{t-L+1+j}-\bar{x})}{\sum_{j=0}^{L-1}(j-\bar{\jmath})^2}, \qquad \widehat{b} = \bar{x} - \widehat{d}\,\bar{\jmath}. \tag{A1}$$

The h-step prediction evaluates the fitted line at local index $L-1+h$:

$$\widehat{x}^{\text{fit}}_{t+h} = \widehat{b} + \widehat{d}(L-1+h).$$

Natural arithmetic sequence. For $L = 6$ and

$$X_t = [97, 98, 99, 100, 101, 102],$$

we have $\bar{\jmath} = 5/2$, $\widehat{d} = 1$, and $\widehat{b} = 97$. Evaluating the fitted line at indices 6 and 7 gives

$$\widehat{Y}^{\text{fit}}_t = [103, 104],$$

which exactly equals the true future of the arithmetic system.

Terminal-state reset. Now keep the first five observations and the system increment fixed but reset the terminal state from 102 to 107:

$$X^{(c)}_t = [97, 98, 99, 100, 101, 107].$$

Here $\bar{x} = 301/3$. The centered-index denominator is $35/2$, the corresponding numerator is 30, and Equation (A1) therefore gives

$$\widehat{d}^{(c)} = \frac{12}{7}, \qquad \widehat{b}^{(c)} = \frac{301}{3} - \frac{12}{7}\frac{5}{2} = \frac{2017}{21}.$$

The complete fractional prediction is

$$\widehat{Y}^{(c),\text{fit}}_t = \left[\frac{2233}{21}, \frac{2269}{21}\right] \approx [106.33, 108.05]. \tag{A2}$$

The valid reset state is $(a_t, d_t) = (107, 1)$, so the system-consistent future is $[108, 109]$. The template responds to the changed terminal value, but the unchanged earlier observations rotate the fitted line toward the old trajectory and yield only a partial response.

Support interpretation. The edited history is not a naturally generated arithmetic sequence: its final increment is 6, whereas every preceding increment is 1. The terminal state $(107, 1)$ is nevertheless valid under the declared state-reset interface. Thus, the comparison tests whether the predictor can restart from a valid new state; it does not claim that the edited full window lies on the natural-history support.

B Dynamical Systems and Data Settings

B.1 Linear Rotation–Decay System

For Linear-D2, Equation (6) uses decay $\rho = 0.98$ and angular increment $\omega = 0.2\pi$. Natural forecast origins have polar angle in $[-\pi, 0)$, whereas State OOD origins use $[0, \pi]$. Linear-D4 uses two two-dimensional blocks with $\rho = (0.98, 0.975)$ and $\omega/\pi = (0.20, 0.14)$. Linear-D8 uses four blocks with $\rho = (0.98, 0.975, 0.97, 0.965)$ and $\omega/\pi = (0.20, 0.14, 0.09, 0.055)$. D4 and D8 add weak cyclic coupling of strength 0.02 between the position coordinates of adjacent blocks. If the spectral radius of the combined matrix reaches 1, the implementation scales the entire matrix back into the stable range.

B.2 Duffing-Type Nonlinear Oscillator

All Duffing scales use Euler step $\Delta t = 0.05$. For D2, train, validation, and IID forecast origins satisfy $|q| \le 0.75$, whereas State OOD origins satisfy $|q| \ge 1.0$. Ordinary state edits change one q or v coordinate by ± 0.5. Duffing-D4 and Duffing-D8 contain two and four state pairs, respectively. Each pair independently samples a tuple (k_i, c_i, β_i), and the position variables receive cyclic discrete-Laplacian coupling of strength 0.03. For both higher-dimensional settings, train, validation, and IID forecast origins satisfy $D_q^{-1} \sum_i |q_i| \le 0.75$, whereas the corrected State OOD split satisfies $D_q^{-1} \sum_i |q_i| \ge 0.90$, where $D_q = D/2$. The two supports are separated by an explicit gap.

B.3 Switching Linear System

The two-dimensional transition matrices are

$$A_+ = \begin{bmatrix} 0.95 & 0.20 \\ -0.10 & 0.90 \end{bmatrix}, \qquad A_- = \begin{bmatrix} 0.80 & 0.45 \\ -0.30 & 0.75 \end{bmatrix}.$$

Switching-D4 and Switching-D8 apply the same A_+ and A_- matrices to two and four two-dimensional state pairs, respectively, and select a global regime from the mean position $\bar{q}_t = D_q^{-1} \sum_i q_{i,t}$. A cross-regime edit moves $\bar{q}_t$ from one side of the threshold to the other.

B.4 Initial-State Support and OOD Splits

Each sequence independently samples its initial state. Linear-D2 draws $x_0 \sim \mathrm{Unif}([-1, 1]^2)$ subject to $\|x_0\|_2 \ge 0.1$. Linear-D4/D8 draw every coordinate from $[-1, 1]$. Each position coordinate of Duffing and Switching is drawn from $[-1.5, 1.5]$, and each velocity coordinate from $[-1, 1]$. The main text states the D2 OOD thresholds. D4/D8 replace $|q|$ with the mean absolute position $D_q^{-1} \sum_i |q_i|$: the natural split is at most 0.75 and State OOD is at least 0.90. Switching has no separate State OOD or Parameter OOD split.

Duffing parameter supports are disjoint by construction:

Table A1: Duffing parameter supports at each scale. For D4/D8, every state pair independently samples one tuple from the support listed in the same row.

Split	k	c	β
Train/validation/IID	$\{0.8, 1.0, 1.2\}$	$\{0.05, 0.10, 0.15\}$	$\{0.3, 0.5, 0.7\}$
Interpolation OOD	$\{0.9, 1.1\}$	$\{0.075, 0.125\}$	$\{0.4, 0.6\}$
Extrapolation OOD	$[1.3, 1.5]$	$[0.17, 0.20]$	$[0.8, 1.0]$

Ordinary D4/D8 interventions uniformly select the affected state pair. For Switching-D2, a cross-regime edit sets the threshold coordinate to ± 0.1. For Switching-D4/D8, we apply the same shift to every position coordinate so that $\bar{q}$ reaches ± 0.1, with the two crossing directions appearing equally often.

B.5 Data Scale

The joint scales and split sizes are: DL and FM models at D2 reuse exactly the NPZ data from the

Table A2: State dimension, history length, forecast horizon, and independent split sizes. Counts are ordered as train/validation/IID/OOD/restart/edit.

Scale	D	L	H	State pairs	Split sizes
D2	2	16	16	1	5000/1000/2000/2000/1000/1000
D4	4	32	32	2	7500/1500/3500/3500/1500/1500
D8	8	64	64	4	10000/2000/5000/5000/2000/2000

lightweight main experiment, and all three model categories at D4 also share the same generated data. The two restart offsets are $k = 4, 8$ for D2, $k = 8, 16$ for D4, and $k = 16, 32$ for D8. Appendix J reports the complete D8 results for general lightweight models.

B.6 Additional Evaluation Details

Section 5 defines the six protocols and metrics. Appendix E gives the paired construction, restart formula, and information-interface lower bound. Raw outputs and aggregate results for channel-independent models remain in the source result files for audit, but do not enter cross-variable response rankings.

C Model Details

C.1 Lightweight Models: Information Scope and Forecast Decomposition

The main text explains the comparative role of each lightweight model. Here we record the implementation based on classical Ridge regression (Hoerl & Kennard, 1970) and the two-dimensional mechanism-aware positive controls.

Persistence copies the last observed state to every future position. kNN-Raw retrieves neighbors from the flattened full-history window, while kNN-Shape performs shape matching after instance-normalizing each window. Window Direct Ridge learns the multi-output map $\mathrm{vec}(X) \to \mathrm{vec}(Y)$. Last-State Direct Ridge uses only the current state x_t to output the complete future. One-Step Ridge learns $x_t \to x_{t+1}$ and rolls it out recurrently. For Duffing, these kNN and Ridge models append the known (k, c, β) to the input features. Polynomial One-Step Ridge additionally constructs explicit features such as kq, cv, and βq^3. Regime Ridge selects a local linear model in the Switching system according to the sign of q_t.

Cross-scale lightweight experiments reuse the original definitions of Persistence, kNN-Raw, kNN-Shape, Window Direct Ridge, Last-State Direct Ridge, and One-Step Ridge. We select only by validation error over

$$k \in \{1, 5, 20\}, \qquad \alpha \in \{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 100\}.$$

Exact kNN performs one query at the maximum k and reuses the same neighbor ordering for all three candidates, which is algebraically equivalent to querying each k separately. Duffing retains known parameter inputs at all three scales. The feature and branch definitions of Polynomial One-Step Ridge and Regime Ridge are specific to two-dimensional states; direct extension to D4/D8 would change the model definition. They are therefore used only as mechanism-matched positive controls at D2 and are excluded from cross-scale curves.

C.2 Deep Models: Architectural Paradigms and Information Interfaces

The architectures follow DLinear, PatchTST, iTransformer (Zeng et al., 2023; Nie et al., 2022; Liu et al., 2024), GRU (Cho et al., 2014), and MLP (Hornik et al., 1989); the main text states their comparative roles. The native Direct MLP and One-Step MLP are not parameter-matched, so parameter-controlled conclusions rely only on the matched group below.

The small deep models are DLinear, PatchTST, iTransformer, GRU-AR, Direct MLP, and One-Step MLP. Each output channel of DLinear and PatchTST cannot read other input channels, so we classify them as channel-independent. iTransformer, GRU-AR, Direct MLP, and One-Step MLP consume the joint state and are classified as multivariate. The trainable parameter counts recorded in the run logs for D2/D4/D8 are: iTransformer $69{,}200/71{,}392/75{,}776$, GRU-AR $8{,}706/9{,}092/9{,}864$, Direct MLP $24{,}864/49{,}536/148{,}224$, and One-Step MLP $17{,}154/17{,}668/18{,}696$. Direct MLP and One-Step MLP are therefore not parameter-matched even when they share the same hidden width and number of layers. The actual learning-rate set is $\{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$ and the batch-size set is $\{512, 1024, 2048\}$; every model is selected by validation nMSE. Four high-learning-rate One-Step MLP trials at D4 produced non-finite losses under AMP and were successfully rerun in FP32 with identical hyperparameters. Twelve of the 45 high-learning-rate One-Step MLP trials on D8 Duffing failed. For every reported seed, the prespecified rule selects among finite configurations.

C.3 Training, Selection, and Point-Forecast Configurations

ML models standardize inputs and targets separately with training-set statistics. Ridge regularization is selected by validation nMSE from

$$\alpha \in \{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 100\}.$$

DL models use per-channel training means and standard deviations. DLinear uses a moving-average kernel of 25; PatchTST uses patch length 8 and stride 4; PatchTST and iTransformer use $d_{\text{model}} = 64$, 2 layers, 4 heads, and FFN width 128; GRU-AR and the two-layer MLPs use hidden width 128 and dropout 0.1. Training uses AdamW, MSE, weight decay 10^{-4}, gradient clipping at 1.0, and no learning-rate schedule, for at most 100 epochs with patience 10. Linear layers use the PyTorch default initialization, while Transformer positional parameters use a truncated normal initialization with standard deviation 0.02.

Parameter-matched MLPs use AdamW, MSE, weight decay 10^{-4}, and per-channel training normalization for at most 120 epochs with patience 12. Learning rates are again selected solely by natural validation nMSE. All six FMs remain frozen and run with batch size 64. Inputs are standardized per sample and per channel and then inverse-transformed. Sampling models generate 20 samples and use the median; TimesFM uses the mean; Chronos-2 uses its official point forecast; and Toto 2.0 uses the 0.5 quantile. We use neither prompting nor fine-tuning.

C.4 D2–D8 Parameter-Matched Mechanism Controls

The parameter-matched experiment in the main text uses three-layer GELU MLPs. The parameter counts for the three interfaces are

$$\begin{aligned}\text{D2} &: 17{,}168/17{,}132/17{,}154,\\ \text{D4} &: 17{,}343/17{,}543/17{,}403,\\ \text{D8} &: 18{,}243/18{,}240/18{,}423.\end{aligned}$$

The remaining training and selection rules are shared as described in the main text.

The hidden widths of the three models are $102/114/128$ at D2, $55/81/127$ at D4, and $17/32/127$ at D8. Widths differ to offset parameter-count differences caused by the three input/output interfaces. Table A3 reports the learning rate selected by natural validation nMSE for seeds $0, 1, 2$ in each system, showing that the parameter-matched conclusions do not depend on an undisclosed single optimization configuration.

C.5 Time-Series Foundation Models: Reproducible Zero-Shot Coverage

Foundation models are included when a public checkpoint is available, the official interface supports forecasting without fine-tuning, and the model can adapt to our history and forecast lengths without using observations outside the input window. TimesFM, Chronos, Time-MoE, and Toto 2.0 provide channel-wise zero-shot controls across distinct public pretrained implementations, but an output channel cannot read other state variables. Chronos-2 and Moirai are included because their native multivariate interfaces accept the joint state in one inference call, making them eligible for cross-variable Effect Error and Sign Accuracy (Das et al., 2024; Ansari et al., 2024; 2025; Woo et al.,

Table A3: Exact hidden widths and per-seed selected learning rates for the parameter-matched MLPs. Each learning-rate triplet is ordered by seed $0/1/2$; DW, DL, and AR denote Direct-window, Direct-last, and One-Step AR, respectively.

Scale (width DW/DL/AR)	System	DW	DL	AR
D2 (102/114/128)	Linear	.003/.003/.003	.01/.01/.01	.0003/.003/.003
	Duffing	.003/.003/.003	.003/.003/.003	.003/.003/.001
	Switching	.003/.003/.003	.01/.01/.01	.01/.01/.01
D4 (55/81/127)	Linear	.003/.001/.01	.01/.003/.003	.003/.01/.003
	Duffing	.003/.01/.01	.01/.003/.003	.003/.003/.003
	Switching	.01/.01/.01	.01/.01/.01	.01/.01/.01
D8 (17/32/127)	Linear	.001/.003/.001	.01/.01/.01	.01/.001/.01
	Duffing	.003/.003/.003	.01/.01/.01	.003/.003/.003
	Switching	.01/.01/.01	.01/.01/.01	.001/.01/.01

2024; Shi et al., 2025; Khwaja et al., 2026). This selection covers both channel-independent and multivariate information pathways. It is not a systematic ranking of the foundation-model family and does not compare fine-tuning or prompting strategies.

The foundation models and checkpoints are:

TimesFM	`google/timesfm-2.5-200m-pytorch`
Chronos	`amazon/chronos-t5-small`
Chronos-2	`amazon/chronos-2`
Moirai	`Salesforce/moirai-1.0-R-small`
Time-MoE	`Maple728/TimeMoE-50M`
Toto 2.0	`Datadog/Toto-2.0-4m`

Table A4: Audit of model sizes used in the D2 parameter-scale analysis. ML counts fitted Ridge coefficients and intercepts; parenthetical values include the three known Duffing parameters. DL counts trainable parameters in the executed models. FM counts all pretrained parameters, which remain frozen for zero-shot inference in this study.

Category	Model	D2 parameter count	Counting convention
ML	One-Step Ridge	6 (12)	coefficients and intercepts
ML	Window Direct Ridge	1,056 (1,152)	coefficients and intercepts
DL	GRU-AR	8,706	trainable parameters
DL	One-Step MLP	17,154	trainable parameters
DL	iTransformer	69,200	trainable parameters
FM	Moirai	14M	total pretrained parameters
FM	Chronos-2	120M	total pretrained parameters

Every interface uses per-series scaling computed only from the current input sequence. Chronos-2 estimates a separate scale for each channel and then passes the complete variable group to the official `predict_quantiles` interface in one call. Toto 2.0 runs through the official `toto2` interface in an isolated Python environment. Internally, it left-pads a D2 history of length 16 to one length-32 patch by repeating only the leftmost value; this introduces no observations outside the window. The main evaluation process exchanges inputs and outputs with the inference process through temporary NPZ files. Models with incompatible dependencies use isolated environments; isolation does not alter the evaluation protocol or data scaling.

D Support and Scope of the Identification Result

Why the arithmetic state reset lies outside the natural support. In Figure 2, every natural six-step history with increment 1 has five adjacent differences equal to 1. The reset history $[97, 98, 99, 100, 101, 107]$ violates this constraint and therefore does not belong to the support of natural histories. Nevertheless, $(a_t, d_t) = (107, 1)$ is a valid current state of the system and uniquely determines the future $[108, 109]$. The diagnostic therefore tests whether a model can restart from a declared new state; it does not mischaracterize the entire edited window as a natural trajectory.

What the result establishes. The construction in Section 4.2 uses only the fact that natural risk depends on the natural support. It does not depend on a particular model class, a smoothness assumption, or a hard matcher. The template-based hard match in the figure is an explicit witness: natural forecasting, including level-OOD forecasting, can be correct while the response to a state change remains wrong. The result does not claim that every hard match fails in this way.

What the result does not establish. A single failure outside the support cannot identify which shortcut a model uses, and finitely many successful tests cannot recover a unique internal mechanism. Within linear, analytic, or otherwise restricted hypothesis classes, structural assumptions may constrain extensions from the natural support. We therefore report natural forecasting, OOD generalization, temporal restart, terminal-state edits, and cross-regime edits separately and interpret each as scoped behavioral evidence rather than complete mechanism recovery.

E Formal Details of Paired Diagnostics

Section 5 introduces the six protocols and their diagnostic targets. This section records the sample construction, normalization, and eligibility rules required to reproduce the results.

E.1 Ground-Truth Rollouts and Paired Samples

For state z and condition ϕ, denote the H-step ground-truth rollout by

$$\mathcal{F}_\phi^{1:H}(z) = \big[F_\phi(z), F_\phi^2(z), \ldots, F_\phi^H(z)\big].$$

An intervention sample contains a natural history X, its terminal state x_t, an edited history $X^{(c)}$ that differs only at the terminal state, and two futures:

$$Y = \mathcal{F}_\phi^{1:H}(x_t), \qquad Y^{(c)} = \mathcal{F}_\phi^{1:H}(x_t^{(c)}).$$

The rollouts share the condition, horizon, and earlier history and differ only in their initial state. Appendix B specifies the edit magnitudes.

E.2 Aggregation Level of Effect Error

Equation (8) corresponds to `effect_error_global` in the implementation and implicitly weights sample contributions by the energy of the true effect. It differs from `effect_error_sample_mean`, which first computes a per-sample ratio and then averages it. The latter does not appear in the paper tables.

E.3 Normalized Forecast Error and Restart Error

IID, State OOD, and Parameter OOD follow Equation (7). Here we provide the restart metric omitted from the main text.

For restart offset k, forecasts beginning at times t and $t+k$ share $H-k$ future steps. Restart nMSE is

$$E_{\text{restart}} = \frac{1}{D}\sum_{d=1}^{D} \frac{\frac{1}{N(H-k)}\sum_{i=1}^{N}\sum_{h=1}^{H-k}\left(\widehat{Y}_{i,h+k,d}^{(t)} - \widehat{Y}_{i,h,d}^{(t+k)}\right)^2}{\max(\sigma_{\text{train},d}^2, \epsilon)}. \tag{A3}$$

This quantity measures consistency between two forecasts, not the error of either forecast relative to the true future. It must therefore be reported alongside, rather than substituted for, natural and OOD nMSE.

E.4 Eligibility by Variable Access

A channel-independent predictor satisfies

$$\widehat{Y}_{:,:,j} = f_j(X_{:,:,j}). \tag{A4}$$

If an intervention changes only channel k, then $\Delta\widehat{Y}_{:,:,j} = 0$ for every $j \neq k$. When the true system couples variables, the information interface induces the lower bound

$$E_{\text{effect}} \geq \frac{\sum_{j \neq k} \|\Delta Y_{:,:,j}\|_F^2}{\|\Delta Y\|_F^2 + \epsilon}. \tag{A5}$$

We determine evaluation eligibility from this rule before observing any scores, following Section 5.3. Appendix C lists the model groups.

F Complete Main-Experiment Tables

Table A5: Main results on Linear Rotation (mean ± standard deviation over three random seeds). Persistence is only a null baseline; its cross-variable Effect Error is excluded from comparison.

Model	IID nMSE ↓	State OOD nMSE ↓	Effect Error ↓
Window Direct Ridge	$8.68 \times 10^{-27} \pm 9.46 \times 10^{-29}$	$5.96 \times 10^{-26} \pm 1.27 \times 10^{-27}$	0.8669 ± 0.0019
Last-State Direct Ridge	$2.51 \times 10^{-24} \pm 3.50 \times 10^{-26}$	$8.64 \times 10^{-24} \pm 1.33 \times 10^{-25}$	$4.00 \times 10^{-24} \pm 2.31 \times 10^{-27}$
One-Step Ridge	$3.14 \times 10^{-25} \pm 4.95 \times 10^{-26}$	$3.13 \times 10^{-25} \pm 4.76 \times 10^{-26}$	$3.11 \times 10^{-25} \pm 4.95 \times 10^{-26}$
kNN-Shape	$1.30 \times 10^{-7} \pm 2.38 \times 10^{-9}$	0.8073 ± 0.0146	0.9150 ± 0.0012
kNN-Raw	$9.28 \times 10^{-5} \pm 5.67 \times 10^{-6}$	0.5138 ± 0.0085	0.8723 ± 0.0018
Persistence	2.638 ± 0.0072	2.660 ± 0.0394	–
Oracle	0 ± 0	0 ± 0	0 ± 0

Table A6: Main results on Duffing (mean ± standard deviation over three random seeds). Persistence is only a null baseline; its cross-variable Effect Error is excluded from comparison.

Model	IID nMSE ↓	State OOD nMSE ↓	Param. extrapolation ratio ↓	Effect Error ↓
Window Direct Ridge	$6.72 \times 10^{-6} \pm 1.94 \times 10^{-7}$	$9.47 \times 10^{-4} \pm 4.76 \times 10^{-5}$	35.43 ± 5.763	$1.46 \times 10^{5} \pm 8905.1$
Last-State Direct Ridge	$0.0010 \pm 7.08 \times 10^{-5}$	$0.0096 \pm 2.20 \times 10^{-4}$	9.000 ± 0.5215	$0.0162 \pm 3.77 \times 10^{-4}$
One-Step Ridge	$0.0022 \pm 5.75 \times 10^{-5}$	$0.0043 \pm 4.81 \times 10^{-5}$	2.498 ± 0.0670	$0.0134 \pm 2.56 \times 10^{-4}$
Polynomial One-Step	$7.47 \times 10^{-24} \pm 4.20 \times 10^{-24}$	$2.25 \times 10^{-23} \pm 1.58 \times 10^{-23}$	–	$1.84 \times 10^{-23} \pm 9.36 \times 10^{-24}$
kNN-Shape	$8.32 \times 10^{-4} \pm 5.31 \times 10^{-5}$	$0.0698 \pm 1.60 \times 10^{-4}$	13.25 ± 3.906	0.8237 ± 0.0089
kNN-Raw	$0.0027 \pm 5.15 \times 10^{-5}$	0.3701 ± 0.0017	8.623 ± 0.4521	$0.8685 \pm 9.44 \times 10^{-4}$
Persistence	0.5487 ± 0.0039	1.110 ± 0.0077	1.526 ± 0.0295	–
Oracle	0 ± 0	0 ± 0	–	0 ± 0

Table A7: Main results on Switching (mean ± standard deviation over three random seeds). Persistence is only a null baseline; its cross-variable intervention metrics are excluded from comparison.

Model	IID nMSE ↓	Effect Error ↓	Sign Accuracy ↑
kNN-Shape	$9.08 \times 10^{-5} \pm 1.09 \times 10^{-5}$	1.205 ± 0.0080	0.4976 ± 0.0076
kNN-Raw	$0.0015 \pm 9.75 \times 10^{-5}$	0.8347 ± 0.0031	0.7901 ± 0.0060
Window Direct Ridge	$0.0123 \pm 5.43 \times 10^{-4}$	2.910 ± 0.1372	0.8069 ± 0.0029
Last-State Direct Ridge	0.1174 ± 0.0020	0.2216 ± 0.0083	0.8314 ± 0.0017
One-Step Ridge	0.4524 ± 0.0023	0.1866 ± 0.0042	0.8312 ± 0.0025
Regime Ridge	$2.46 \times 10^{-25} \pm 1.93 \times 10^{-26}$	$2.74 \times 10^{-25} \pm 1.68 \times 10^{-26}$	1.000 ± 0
Persistence	4.019 ± 0.0789	–	–
Oracle	0 ± 0	0 ± 0	1.000 ± 0

F.1 Scope of the Complete Results

The tables above retain models and cross-seed standard deviations omitted from the compact comparisons in the main text. Appendix G reports the complete six-protocol matrix, and Appendix E defines evaluation eligibility.

F.2 Supporting Comparison Across Joint Scales

The joint-scale comparison retains the broader model coverage behind the controlled examples. Scale, history length, horizon, sample size, coupling, and some parameter counts change together; these results describe the tested configurations rather than isolate a causal effect of scale.

Table A8: Supporting joint-scale comparison of IID nMSE and Effect Error for representative multivariate ML, DL, and frozen foundation models. Values average seeds 0, 1, and 2; categories are not parameter-matched. Complete system-specific results and numerical-stability qualifications are reported in this appendix.

			D2		D4		D8	
Category	**System**	**Model**	**IID**	**Effect**	**IID**	**Effect**	**IID**	**Effect**
ML	Linear	Window Direct Ridge	8.68×10^{-27}	0.8669	1.44×10^{-27}	0.9363	2.24×10^{-28}	0.9688
ML	Linear	One-Step Ridge	3.14×10^{-25}	3.11×10^{-25}	1.03×10^{-25}	1.04×10^{-25}	4.04×10^{-25}	4.16×10^{-25}
ML	Duffing	Window Direct Ridge	6.72×10^{-6}	1.46×10^{5}	6.44×10^{-4}	1.95×10^{6}	0.0017	6.12×10^{6}
ML	Duffing	One-Step Ridge	0.0022	0.0134	0.0341	0.1225	0.0838	0.2213
ML	Switching	Window Direct Ridge	0.0123	2.9102	0.0090	2.0024	0.0113	0.9446
ML	Switching	One-Step Ridge	0.4524	0.1866	2.5193	0.2947	187.9	0.2906
DL	Linear	iTransformer	1.79×10^{-6}	0.9334	2.55×10^{-6}	0.9793	4.53×10^{-5}	0.9923
DL	Linear	One-Step MLP	0.0029	0.0223	0.0083	0.0897	0.1363	0.3314
DL	Duffing	iTransformer	7.92×10^{-4}	0.7164	0.0021	0.4784	9.67×10^{-4}	0.8059
DL	Duffing	One-Step MLP	0.0077	0.0638	0.0518	0.2010	0.4456	0.6125
DL	Switching	iTransformer	0.0040	0.6467	0.0184	0.6716	0.3681	1.0081
DL	Switching	One-Step MLP	0.0133	0.0228	0.1085	0.1190	1.0095	1.0152
FM	Linear	Chronos-2	0.4490	1.8633	0.1785	1.1482	2.659	5.3431
FM	Linear	Moirai	1.7812	2.1041	3.4256	4.2105	473.4	119.3
FM	Duffing	Chronos-2	0.0489	1.3088	0.1611	12.10	0.5034	17.11
FM	Duffing	Moirai	0.4823	0.7182	1.0643	3.7788	1.7094	560.7
FM	Switching	Chronos-2	10.34	31.77	33.19	29.94	113.6	23.88
FM	Switching	Moirai	4.0632	5.4758	90.92	39.12	4.15×10^{6}	5.82×10^{4}

G Complete Cross-Protocol and Cross-System Results

The endpoint tables below report per-model results for lightweight D2 models, deep D8 models, and zero-shot D8 foundation models. To fit each category on a portrait A4 page, we separate the forecast/consistency metrics (IID, State OOD, Parameter OOD, and Restart) from the paired-response metrics (Effect Error and Sign Accuracy). This split changes only the layout; each pair of tables still contains the same six diagnostics. Parameter OOD uses the extrapolation split, and Restart uses the larger offset ($k = 8$ for D2 and $k = 32$ for D8). The Duffing-D8 State OOD entries for DL and FM come from reruns with the corrected state intervals. Shared column names support only within-category analysis and do not define rankings across categories.

Table A9: Lightweight D2 forecasting and restart results. Values are mean $\pm$ standard deviation over three seeds; Parameter OOD is the Duffing extrapolation split and Restart uses $k = 8$.

System	Model	IID nMSE	State OOD	Param. OOD	Restart
Linear	Persistence	2.6378 ± 0.0072	2.6603 ± 0.0394	–	1.6913 ± 0.0605
	kNN-Raw	$9.28 \times 10^{-5} \pm 5.67 \times 10^{-6}$	0.5138 ± 0.0085	–	0.1176 ± 0.0142
	kNN-Shape	$1.30 \times 10^{-7} \pm 2.38 \times 10^{-9}$	0.8073 ± 0.0146	–	0.1588 ± 0.0204
	Window Direct Ridge	$8.68 \times 10^{-27} \pm 9.46 \times 10^{-29}$	$5.96 \times 10^{-26} \pm 1.27 \times 10^{-27}$	–	$1.41 \times 10^{-26} \pm 6.69 \times 10^{-28}$
	Last-State Direct Ridge	$2.51 \times 10^{-24} \pm 3.50 \times 10^{-26}$	$8.64 \times 10^{-24} \pm 1.33 \times 10^{-25}$	–	$2.10 \times 10^{-24} \pm 3.18 \times 10^{-26}$
	One-Step Ridge	$3.14 \times 10^{-25} \pm 4.95 \times 10^{-26}$	$3.13 \times 10^{-25} \pm 4.76 \times 10^{-26}$	–	$2.14 \times 10^{-25} \pm 2.94 \times 10^{-26}$
	Oracle	0 ± 0	0 ± 0	–	0 ± 0
Duffing	Persistence	0.5487 ± 0.0039	1.1103 ± 0.0077	0.8373 ± 0.0136	0.3837 ± 0.0081
	kNN-Raw	$0.0027 \pm 5.15 \times 10^{-5}$	0.3701 ± 0.0017	0.0231 ± 0.0011	0.0488 ± 0.0040
	kNN-Shape	$8.32 \times 10^{-4} \pm 5.31 \times 10^{-5}$	$0.0698 \pm 1.60 \times 10^{-4}$	0.0109 ± 0.0026	$0.0018 \pm 1.61 \times 10^{-4}$
	Window Direct Ridge	$6.72 \times 10^{-6} \pm 1.94 \times 10^{-7}$	$9.47 \times 10^{-4} \pm 4.76 \times 10^{-5}$	$2.39 \times 10^{-4} \pm 4.37 \times 10^{-5}$	$1.46 \times 10^{-5} \pm 1.36 \times 10^{-6}$
	Last-State Direct Ridge	$0.0010 \pm 7.08 \times 10^{-5}$	$0.0096 \pm 2.20 \times 10^{-4}$	$0.0090 \pm 3.43 \times 10^{-4}$	$9.18 \times 10^{-4} \pm 2.57 \times 10^{-5}$
	One-Step Ridge	$0.0022 \pm 5.75 \times 10^{-5}$	$0.0043 \pm 4.81 \times 10^{-5}$	$0.0056 \pm 2.47 \times 10^{-4}$	$0.0021 \pm 4.45 \times 10^{-5}$
	Polynomial One-Step	$7.47 \times 10^{-24} \pm 4.20 \times 10^{-24}$	$2.25 \times 10^{-23} \pm 1.58 \times 10^{-23}$	$3.38 \times 10^{-23} \pm 1.31 \times 10^{-23}$	$5.37 \times 10^{-24} \pm 2.70 \times 10^{-24}$
	Oracle	0 ± 0	0 ± 0	0 ± 0	0 ± 0
Switching	Persistence	4.019 ± 0.0789	–	–	5.0473 ± 0.2907
	kNN-Raw	$0.0015 \pm 9.75 \times 10^{-5}$	–	–	$0.0020 \pm 3.37 \times 10^{-4}$
	kNN-Shape	$9.08 \times 10^{-5} \pm 1.09 \times 10^{-5}$	–	–	$8.44 \times 10^{-5} \pm 2.83 \times 10^{-5}$
	Window Direct Ridge	$0.0123 \pm 5.43 \times 10^{-4}$	–	–	$0.0112 \pm 4.71 \times 10^{-4}$
	Last-State Direct Ridge	0.1174 ± 0.0020	–	–	0.1260 ± 0.0079
	One-Step Ridge	0.4524 ± 0.0023	–	–	0.2272 ± 0.0019
	Regime Ridge	$2.46 \times 10^{-25} \pm 1.93 \times 10^{-26}$	–	–	$1.60 \times 10^{-25} \pm 7.95 \times 10^{-27}$
	Oracle	0 ± 0	–	–	0 ± 0

Table A10: Lightweight D2 paired-response results for the same models as Table A9. Switching uses the cross-regime intervention.

System	Model (interface)	Effect Error ↓	Sign Accuracy ↑
Linear	Persistence (null)	*n/a*	*n/a*
	kNN-Raw (MV)	0.8723 ± 0.0018	0.9826 ± 0.0015
	kNN-Shape (MV)	0.9150 ± 0.0012	0.6739 ± 0.0067
	Window Direct Ridge (MV)	0.8669 ± 0.0019	1 ± 0
	Last-State Direct Ridge (MV)	$4.00 \times 10^{-24} \pm 2.31 \times 10^{-27}$	1 ± 0
	One-Step Ridge (MV)	$3.11 \times 10^{-25} \pm 4.95 \times 10^{-26}$	1 ± 0
	Oracle	0 ± 0	1 ± 0
Duffing	Persistence (null)	*n/a*	*n/a*
	kNN-Raw (MV)	$0.8685 \pm 9.44 \times 10^{-4}$	0.9093 ± 0.0028
	kNN-Shape (MV)	0.8237 ± 0.0089	0.7405 ± 0.0080
	Window Direct Ridge (MV)	$1.46 \times 10^{5} \pm 8.91 \times 10^{3}$	0.7997 ± 0.0049
	Last-State Direct Ridge (MV)	$0.0162 \pm 3.77 \times 10^{-4}$	$1 \pm 1.80 \times 10^{-5}$
	One-Step Ridge (MV)	$0.0134 \pm 2.56 \times 10^{-4}$	$1 \pm 1.80 \times 10^{-5}$
	Polynomial One-Step (MV)	$1.84 \times 10^{-23} \pm 9.36 \times 10^{-24}$	1 ± 0
	Oracle	0 ± 0	1 ± 0
Switching	Persistence (null)	*n/a*	*n/a*
	kNN-Raw (MV)	0.8347 ± 0.0031	0.7901 ± 0.0060
	kNN-Shape (MV)	1.2048 ± 0.0080	0.4976 ± 0.0076
	Window Direct Ridge (MV)	2.9102 ± 0.1372	0.8069 ± 0.0029
	Last-State Direct Ridge (MV)	0.2216 ± 0.0083	0.8314 ± 0.0017
	One-Step Ridge (MV)	0.1866 ± 0.0042	0.8312 ± 0.0025
	Regime Ridge (MV)	$2.74 \times 10^{-25} \pm 1.68 \times 10^{-26}$	1 ± 0
	Oracle	0 ± 0	1 ± 0

n/a marks interfaces excluded from cross-variable intervention comparisons.

Table A11: DL D8 forecasting and restart results. Values are mean ± standard deviation over three seeds; Parameter OOD is the Duffing extrapolation split and Restart uses $k = 32$.

System	Model	IID nMSE	State OOD	Param. OOD	Restart
Linear-D8	DLinear	$8.71 \times 10^{-5} \pm 9.62 \times 10^{-6}$	$3.61 \times 10^{-4} \pm 5.27 \times 10^{-5}$	–	$1.45 \times 10^{-4} \pm 2.52 \times 10^{-5}$
	PatchTST	$4.46 \times 10^{-4} \pm 7.80 \times 10^{-5}$	$5.55 \times 10^{-4} \pm 1.20 \times 10^{-4}$	–	$0.0012 \pm 2.44 \times 10^{-4}$
	iTransformer	$4.53 \times 10^{-5} \pm 2.17 \times 10^{-6}$	$8.56 \times 10^{-5} \pm 1.04 \times 10^{-5}$	–	$6.25 \times 10^{-4} \pm 2.87 \times 10^{-4}$
	GRU-AR	0.0139 ± 0.0030	0.0344 ± 0.0027	–	0.0133 ± 0.0014
	Direct MLP	$0.0019 \pm 3.75 \times 10^{-4}$	0.0048 ± 0.0018	–	$0.0018 \pm 7.77 \times 10^{-4}$
	One-Step MLP	0.1363 ± 0.1184	0.1464 ± 0.1169	–	0.0161 ± 0.0021
Duffing-D8	DLinear	$0.0149 \pm 4.03 \times 10^{-4}$	$0.0158 \pm 3.53 \times 10^{-4}$	0.1054 ± 0.0050	$0.0350 \pm 2.90 \times 10^{-4}$
	PatchTST	$0.0019 \pm 1.18 \times 10^{-4}$	$0.0025 \pm 1.70 \times 10^{-4}$	0.0180 ± 0.0027	$0.0061 \pm 1.49 \times 10^{-4}$
	iTransformer	$9.67 \times 10^{-4} \pm 6.98 \times 10^{-5}$	$0.0017 \pm 3.25 \times 10^{-4}$	$0.0196 \pm 7.73 \times 10^{-4}$	$0.0032 \pm 2.77 \times 10^{-4}$
	GRU-AR	0.0119 ± 0.0015	0.0177 ± 0.0024	0.0448 ± 0.0045	$0.0129 \pm 9.94 \times 10^{-4}$
	Direct MLP	$0.0115 \pm 6.28 \times 10^{-4}$	0.0210 ± 0.0011	0.0640 ± 0.0053	0.0206 ± 0.0021
	One-Step MLP	0.4456 ± 0.1005	0.5978 ± 0.1730	0.4744 ± 0.1097	0.0689 ± 0.0204
Switching-D8	DLinear	0.0858 ± 0.0048	–	–	0.0548 ± 0.0039
	PatchTST	0.9478 ± 0.0515	–	–	1.7336 ± 0.9887
	iTransformer	0.3681 ± 0.0565	–	–	0.8995 ± 0.3874
	GRU-AR	0.2380 ± 0.0890	–	–	0.0090 ± 0.0042
	Direct MLP	0.0847 ± 0.0099	–	–	0.0118 ± 0.0029
	One-Step MLP	1.0095 ± 0.0028	–	–	$7.37 \times 10^{-6} \pm 2.32 \times 10^{-6}$

Table A12: DL D8 paired-response results for the same models as Table A11. Switching uses the cross-regime intervention.

System	Model (interface)	Effect Error ↓	Sign Accuracy ↑
Linear-D8	DLinear (CI)	*n/a*	*n/a*
	PatchTST (CI)	*n/a*	*n/a*
	iTransformer (MV)	0.9923 ± 0.0064	0.5058 ± 0.0049
	GRU-AR (MV)	0.4830 ± 0.0244	0.6398 ± 0.0156
	Direct MLP (MV)	0.9908 ± 0.0038	0.5308 ± 0.0213
	One-Step MLP (MV)	0.3314 ± 0.0425	0.6680 ± 0.0087
Duffing-D8	DLinear (CI)	*n/a*	*n/a*
	PatchTST (CI)	*n/a*	*n/a*
	iTransformer (MV)	0.8059 ± 0.0049	0.6637 ± 0.0129
	GRU-AR (MV)	0.6990 ± 0.0028	0.6288 ± 0.0295
	Direct MLP (MV)	0.7754 ± 0.0190	$0.5873 \pm 8.90 \times 10^{-4}$
	One-Step MLP (MV)	0.6125 ± 0.0666	0.6094 ± 0.0156
Switching-D8	DLinear (CI)	*n/a*	*n/a*
	PatchTST (CI)	*n/a*	*n/a*
	iTransformer (MV)	1.0081 ± 0.0096	0.5181 ± 0.0177
	GRU-AR (MV)	1.0731 ± 0.0358	0.5880 ± 0.0645
	Direct MLP (MV)	1.0015 ± 0.0027	0.4852 ± 0.0359
	One-Step MLP (MV)	1.0152 ± 0.0319	0.1512 ± 0.1164

CI: channel-independent; MV: multivariate. *n/a* marks interfaces excluded from cross-variable intervention comparisons.

Table A13: Zero-shot FM D8 forecasting and restart results. Values are mean ± standard deviation over three seeds; Parameter OOD is the Duffing extrapolation split and Restart uses $k = 32$.

System	Model	IID nMSE	State OOD	Param. OOD	Restart
Linear-D8	TimesFM	5.8131 ± 0.0119	5.7806 ± 0.0242	–	8.2198 ± 0.0332
	Chronos	140.7 ± 1.4116	139.1 ± 1.7289	–	203.9 ± 3.6873
	Time-MoE	50.27 ± 0.3557	50.33 ± 0.4355	–	61.37 ± 0.3939
	Toto 2.0	1.0686 ± 0.0051	1.1913 ± 0.0137	–	0.7524 ± 0.0035
	Chronos-2	2.659 ± 0.0093	2.2603 ± 0.0158	–	2.997 ± 0.0085
	Moirai	473.4 ± 1.9061	471.1 ± 6.0646	–	775.3 ± 8.4705
Duffing-D8	TimesFM	0.4686 ± 0.0021	0.6199 ± 0.0039	0.3100 ± 0.0012	0.7122 ± 0.0104
	Chronos	0.4110 ± 0.0041	0.7369 ± 0.0011	1.2128 ± 0.0186	2.773 ± 0.1100
	Time-MoE	0.6633 ± 0.0015	$1.1197 \pm 5.43 \times 10^{-4}$	0.4118 ± 0.0026	1.1788 ± 0.0109
	Toto 2.0	0.4782 ± 0.0068	0.8116 ± 0.0042	0.5196 ± 0.0033	1.0442 ± 0.0093
	Chronos-2	0.5034 ± 0.0016	1.1739 ± 0.0036	0.3425 ± 0.0012	1.1047 ± 0.0194
	Moirai	1.7094 ± 0.0278	2.7694 ± 0.0133	9.333 ± 1.2266	17.90 ± 0.7706
Switching-D8	TimesFM	$4.90 \times 10^3 \pm 163.5$	–	–	$5.29 \times 10^3 \pm 145.7$
	Chronos	$2.92 \times 10^4 \pm 556.7$	–	–	$4.31 \times 10^4 \pm 770.2$
	Time-MoE	$4.83 \times 10^3 \pm 43.79$	–	–	$7.32 \times 10^3 \pm 81.78$
	Toto 2.0	50.29 ± 0.4836	–	–	85.04 ± 0.9992
	Chronos-2	113.6 ± 0.5410	–	–	175.4 ± 2.463
	Moirai	$4.15 \times 10^6 \pm 1.89 \times 10^6$	–	–	$3.08 \times 10^6 \pm 1.08 \times 10^6$

Table A14: Zero-shot FM D8 paired-response results for the same models as Table A13. Switching uses the cross-regime intervention.

System	Model (interface)	Effect Error ↓	Sign Accuracy ↑
Linear-D8	TimesFM (CI)	*n/a*	*n/a*
	Chronos (CI)	*n/a*	*n/a*
	Time-MoE (CI)	*n/a*	*n/a*
	Toto 2.0 (CI)	*n/a*	*n/a*
	Chronos-2 (MV)	5.3431 ± 0.1544	$0.5101 \pm 5.51 \times 10^{-4}$
	Moirai (MV)	119.3 ± 5.1562	$0.5005 \pm 1.61 \times 10^{-4}$
Duffing-D8	TimesFM (CI)	*n/a*	*n/a*
	Chronos (CI)	*n/a*	*n/a*
	Time-MoE (CI)	*n/a*	*n/a*
	Toto 2.0 (CI)	*n/a*	*n/a*
	Chronos-2 (MV)	17.11 ± 0.4704	$0.4392 \pm 3.95 \times 10^{-4}$
	Moirai (MV)	560.7 ± 195.5	$0.5069 \pm 1.85 \times 10^{-4}$
Switching-D8	TimesFM (CI)	*n/a*	*n/a*
	Chronos (CI)	*n/a*	*n/a*
	Time-MoE (CI)	*n/a*	*n/a*
	Toto 2.0 (CI)	*n/a*	*n/a*
	Chronos-2 (MV)	23.88 ± 0.1855	$0.5376 \pm 5.10 \times 10^{-4}$
	Moirai (MV)	$5.82 \times 10^4 \pm 8.42 \times 10^4$	$0.5126 \pm 7.97 \times 10^{-4}$

CI: channel-independent; MV: multivariate. *n/a* marks interfaces excluded from cross-variable intervention comparisons.

H Complete D2–D8 Scaling Results

Table A15 reports IID nMSE and Effect Error for seven shared models across three systems and three scales along the joint scaling path. Because the path changes several experimental factors simultaneously, we discuss only the joint scale sensitivity of a fixed model.

H.1 Detailed Analysis of Joint Scaling

Forecast–response separation persists with scale, but its direction is system-dependent. For Linear, the IID nMSE of Window Direct Ridge remains near machine precision at every scale, while Effect Error rises from 0.8669 to 0.9363 and 0.9688; One-Step Ridge remains numerically exact under both metrics. For Duffing, Window Direct IID nMSE increases from 6.72×10^{-6} to 6.44×10^{-4} and 0.0017, while Effect Error simultaneously rises from 1.46×10^5 to 1.95×10^6 and 6.12×10^6. One-Step Ridge Effect Error also increases from 0.0134 to 0.1225 and 0.2213, but remains several orders of magnitude below Window Direct. Switching exhibits the opposite trend: Window Direct Effect Error decreases from 2.910 to 2.002 and 0.945. The IID nMSE of One-Step Ridge rises from 0.452 to 2.519 and 187.9, while its Effect Error remains between 0.187 and 0.295. These same-direction, opposite-direction, and approximately independent changes show that system properties, rather than scale alone, determine how the two errors evolve.

The response advantage of AR models is not stable across scale. Across Linear, Duffing, and Switching, the Effect Error of One-Step MLP changes as $0.0223 \to 0.0897 \to 0.331$, $0.0638 \to 0.201 \to 0.613$, and $0.0228 \to 0.119 \to 1.015$, respectively. GRU-AR follows the same direction, rising from 0.131, 0.486, and 0.608 to D8 values of 0.483, 0.699, and 1.073. As a non-AR control, iTransformer stays near the no-response reference on Linear ($0.933 \to 0.979 \to 0.992$), but changes non-monotonically on Duffing ($0.716 \to 0.478 \to 0.806$); on Switching-D8, all three deep models have Effect Error near 1. Lightweight One-Step Ridge further bounds this conclusion: it remains numerically exact for Linear; its Duffing Effect Error grows from 0.0134 to 0.2213; and its Switching Effect Error stays near 0.2–0.3 even as IID nMSE reaches 187.9. Benefits of recurrence therefore depend jointly on the system, scale, and representational capacity.

Foundation models exhibit stronger model–system interactions. The IID nMSE/Effect Error of Chronos-2 changes from $0.449/1.863 \to 0.179/1.148 \to 2.659/5.343$ on Linear, from $0.0489/1.309$ to $0.503/17.11$ on Duffing, and from $10.34/31.77 \to 33.19/29.94 \to 113.6/23.88$ on Switching. Moirai diverges more sharply: Linear rises from $1.781/2.104$ to $473.4/119.3$, Duffing from $0.482/0.718$ to $1.709/560.7$, and Switching from $4.063/5.476$ to $4.15 \times 10^6/5.82 \times 10^4$.

Neither pretraining nor a native multivariate interface yields a scale law consistent across systems; the direction and magnitude of change for the same foundation model can differ completely across mechanisms.

H.2 Detailed Analysis of Parameter Scale and Performance

Changes from ML to DL are jointly determined by structure and system. For Linear, One-Step Ridge contains only 6 fitted coefficients, yet both errors are near machine precision. For the three deep models with 8,706–69,200 parameters, IID nMSE ranges from 10^{-6} to 10^{-3}, while Effect Error ranges from 0.0223 for One-Step MLP to 0.9334 for iTransformer. For Duffing, One-Step Ridge yields IID nMSE/Effect Error of $0.0022/0.0134$; the three deep models fall in ranges of 7.92×10^{-4}–0.0077 and 0.0638–0.7164, respectively. Larger deep models improve some natural forecasts but do not uniformly improve response. For Switching, the roughly 17.2K-parameter One-Step MLP attains the lowest Effect Error of 0.0228, compared with 0.6467 for the larger iTransformer and 0.1866 for the smaller One-Step Ridge. Even among ML and DL models trained on the same experimental data, parameter count does not determine the relative ranking of the two metrics.

The transition from DL to FM is not monotonic. The two FMs increase scale to 14M and 120M parameters, yet their IID nMSE/Effect Error remains $1.781/2.104$ and $0.449/1.863$ on Linear, $0.482/0.718$ and $0.0489/1.309$ on Duffing, and $4.063/5.476$ and $10.34/31.77$ on Switching. Relative to the smaller Moirai, Chronos-2 improves both metrics only on Linear, improves natural forecasting alone on Duffing, and degrades both metrics on Switching. Together with the ML and DL results, these values reveal strong model–system interaction rather than a performance envelope explained by parameter count alone. Table A4 defines the counting convention. Because FMs are evaluated zero-shot, this comparison must not be interpreted as a controlled ablation of pretraining scale.

Table A15: Complete results for seven multivariate models along the joint D2–D8 scale path (mean $\pm$ standard deviation over three random seeds). The path changes state dimension, horizon, sample count, system coupling, and some parameter counts simultaneously; values track only changes of a fixed model.

System	Category	Model	Scale	IID nMSE ↓	Effect Error ↓
Linear	ML	Window Direct	D2	$8.68 \times 10^{-27} \pm 9.46 \times 10^{-29}$	0.8669 ± 0.0019
Linear	ML	Window Direct	D4	$1.44 \times 10^{-27} \pm 1.92 \times 10^{-29}$	$0.9363 \pm 8.09 \times 10^{-4}$
Linear	ML	Window Direct	D8	$2.24 \times 10^{-28} \pm 2.29 \times 10^{-30}$	$0.9688 \pm 5.56 \times 10^{-4}$
Linear	ML	One-Step Ridge	D2	$3.14 \times 10^{-25} \pm 4.95 \times 10^{-26}$	$3.11 \times 10^{-25} \pm 4.95 \times 10^{-26}$
Linear	ML	One-Step Ridge	D4	$1.03 \times 10^{-25} \pm 4.97 \times 10^{-26}$	$1.04 \times 10^{-25} \pm 5.09 \times 10^{-26}$
Linear	ML	One-Step Ridge	D8	$4.04 \times 10^{-25} \pm 1.23 \times 10^{-25}$	$4.16 \times 10^{-25} \pm 1.49 \times 10^{-25}$
Linear	DL	iTransformer	D2	$1.79 \times 10^{-6} \pm 6.06 \times 10^{-7}$	0.9334 ± 0.0339
Linear	DL	iTransformer	D4	$2.55 \times 10^{-6} \pm 2.29 \times 10^{-7}$	0.9793 ± 0.0157
Linear	DL	iTransformer	D8	$4.53 \times 10^{-5} \pm 2.17 \times 10^{-6}$	0.9923 ± 0.0064
Linear	DL	GRU-AR	D2	$5.51 \times 10^{-4} \pm 1.37 \times 10^{-4}$	0.1315 ± 0.0115
Linear	DL	GRU-AR	D4	0.0038 ± 0.0011	0.1402 ± 0.0149
Linear	DL	GRU-AR	D8	0.0139 ± 0.003	0.483 ± 0.0244
Linear	DL	One-Step MLP	D2	$0.0029 \pm 4.66 \times 10^{-4}$	0.0223 ± 0.0041
Linear	DL	One-Step MLP	D4	0.0083 ± 0.001	0.0897 ± 0.0027
Linear	DL	One-Step MLP	D8	0.1363 ± 0.1184	0.3314 ± 0.0425
Linear	FM	Chronos-2	D2	0.449 ± 0.0048	1.8633 ± 0.0326
Linear	FM	Chronos-2	D4	0.1785 ± 0.0023	1.1482 ± 0.0095
Linear	FM	Chronos-2	D8	2.659 ± 0.0093	5.3431 ± 0.1544
Linear	FM	Moirai	D2	1.7812 ± 0.0146	2.1041 ± 0.0279
Linear	FM	Moirai	D4	3.4256 ± 0.0378	4.2105 ± 0.0432
Linear	FM	Moirai	D8	473.4 ± 1.9061	119.3 ± 5.1562
Duffing	ML	Window Direct	D2	$6.72 \times 10^{-6} \pm 1.94 \times 10^{-7}$	$1.46 \times 10^{5} \pm 8.91 \times 10^{3}$
Duffing	ML	Window Direct	D4	$6.44 \times 10^{-4} \pm 1.92 \times 10^{-5}$	$1.95 \times 10^{6} \pm 5.93 \times 10^{4}$
Duffing	ML	Window Direct	D8	$0.0017 \pm 1.47 \times 10^{-5}$	$6.12 \times 10^{6} \pm 1.3 \times 10^{5}$
Duffing	ML	One-Step Ridge	D2	$0.0022 \pm 5.75 \times 10^{-5}$	$0.0134 \pm 2.56 \times 10^{-4}$
Duffing	ML	One-Step Ridge	D4	$0.0341 \pm 7.04 \times 10^{-4}$	0.1225 ± 0.0076
Duffing	ML	One-Step Ridge	D8	0.0838 ± 0.0028	$0.2213 \pm 8.18 \times 10^{-4}$
Duffing	DL	iTransformer	D2	$7.92 \times 10^{-4} \pm 4.39 \times 10^{-5}$	0.7164 ± 0.1938
Duffing	DL	iTransformer	D4	$0.0021 \pm 9.57 \times 10^{-5}$	0.4784 ± 0.0434
Duffing	DL	iTransformer	D8	$9.67 \times 10^{-4} \pm 6.98 \times 10^{-5}$	0.8059 ± 0.0049
Duffing	DL	GRU-AR	D2	$0.0012 \pm 1.01 \times 10^{-4}$	0.4857 ± 0.0372
Duffing	DL	GRU-AR	D4	0.0063 ± 0.0013	0.642 ± 0.0195

Continued on next page

Table A15 (continued)

System	Category	Model	Scale	IID nMSE ↓	Effect Error ↓
Duffing	DL	GRU-AR	D8	0.0119 ± 0.0015	0.699 ± 0.0028
Duffing	DL	One-Step MLP	D2	$0.0077 \pm 7.08 \times 10^{-4}$	0.0638 ± 0.0053
Duffing	DL	One-Step MLP	D4	0.0518 ± 0.0018	0.201 ± 0.0063
Duffing	DL	One-Step MLP	D8	0.4456 ± 0.1005	0.6125 ± 0.0666
Duffing	FM	Chronos-2	D2	0.0489 ± 0.0011	1.3088 ± 0.0267
Duffing	FM	Chronos-2	D4	0.1611 ± 0.0012	12.10 ± 0.1811
Duffing	FM	Chronos-2	D8	0.5034 ± 0.0016	17.11 ± 0.4704
Duffing	FM	Moirai	D2	0.4823 ± 0.0061	0.7182 ± 0.0132
Duffing	FM	Moirai	D4	1.0643 ± 0.0105	3.7788 ± 0.145
Duffing	FM	Moirai	D8	1.7094 ± 0.0278	560.7 ± 195.5
Switching	ML	Window Direct	D2	$0.0123 \pm 5.43 \times 10^{-4}$	2.9102 ± 0.1372
Switching	ML	Window Direct	D4	$0.009 \pm 1.97 \times 10^{-4}$	2.0024 ± 0.4315
Switching	ML	Window Direct	D8	$0.0113 \pm 3.7 \times 10^{-4}$	0.9446 ± 0.8393
Switching	ML	One-Step Ridge	D2	0.4524 ± 0.0023	0.1866 ± 0.0042
Switching	ML	One-Step Ridge	D4	2.5193 ± 0.0681	0.2947 ± 0.0019
Switching	ML	One-Step Ridge	D8	187.9 ± 1.4497	$0.2906 \pm 7.06 \times 10^{-4}$
Switching	DL	iTransformer	D2	$0.004 \pm 1.37 \times 10^{-4}$	0.6467 ± 0.0147
Switching	DL	iTransformer	D4	$0.0184 \pm 6.91 \times 10^{-4}$	0.6716 ± 0.058
Switching	DL	iTransformer	D8	0.3681 ± 0.0565	1.0081 ± 0.0096
Switching	DL	GRU-AR	D2	$0.0041 \pm 2.58 \times 10^{-4}$	0.6082 ± 0.0545
Switching	DL	GRU-AR	D4	$0.0102 \pm 5.43 \times 10^{-4}$	0.8661 ± 0.0788
Switching	DL	GRU-AR	D8	0.238 ± 0.089	1.0731 ± 0.0358
Switching	DL	One-Step MLP	D2	0.0133 ± 0.0012	0.0228 ± 0.0017
Switching	DL	One-Step MLP	D4	0.1085 ± 0.0033	0.119 ± 0.0135
Switching	DL	One-Step MLP	D8	1.0095 ± 0.0028	1.0152 ± 0.0319
Switching	FM	Chronos-2	D2	10.34 ± 0.4551	31.77 ± 0.6231
Switching	FM	Chronos-2	D4	33.19 ± 0.6457	29.94 ± 0.9326
Switching	FM	Chronos-2	D8	113.6 ± 0.541	23.88 ± 0.1855
Switching	FM	Moirai	D2	4.0632 ± 0.0496	5.4758 ± 0.0091
Switching	FM	Moirai	D4	90.92 ± 3.6779	39.12 ± 0.4239
Switching	FM	Moirai	D8	$4.15 \times 10^6 \pm 1.89 \times 10^6$	$5.82 \times 10^4 \pm 8.42 \times 10^4$

I Numerical-Stability Qualifications for the Main Scaling Table

For D2 Duffing, the Window Direct Ridge training design matrix has condition number $(1.19 \pm 0.21) \times 10^{17}$. Its very large Effect Error should therefore be interpreted as sensitivity of a highly collinear extension beyond the natural-history support, not as a stable effect-size estimate.

For D8 Switching with Moirai, the three seed-level Effect Error values are 1.58×10^4, 1.55×10^5, and 3.52×10^3. Their arithmetic mean is dominated by a highly skewed seed distribution. The main table reports the prespecified mean aggregation, while these seed-level values expose its sensitivity.

J Scale-Matched Experiments for Lightweight Models

To test whether the lightweight-model phenomenon is caused only by experimental scale, we transfer six general lightweight models to the D8 data and forecasting setting used in the main text and report the larger restart offset $k = 32$. This experiment matches data scale only; it does not match parameter count, optimization, or information interface. It therefore evaluates scale robustness within the lightweight category.

The separation between natural forecasting and conditional response persists after matching the data scale. On Linear-D8, Window Direct Ridge attains an IID nMSE of 2.24×10^{-28} but an Effect Error of 0.9688, whereas both errors of Last-State Direct Ridge and One-Step Ridge remain near machine precision. On Duffing-D8, Window Direct Ridge reaches an IID nMSE of 0.0017 and an Effect Error of 6.12×10^6. Last-State Direct Ridge and One-Step Ridge reduce Effect Error to 0.2354 and 0.2213, respectively, while their IID nMSE increases to 0.0714 and 0.0838. Switching-D8 produces yet another ordering: Window Direct Ridge has an IID nMSE of 0.0113 and an Effect Error of 0.9446 ± 0.8393, whereas One-Step Ridge has an IID nMSE of 187.86 and an Effect Error of 0.2906. Increasing D, L, H, and the sample count therefore does not unify natural fit and dynamical response under a single error metric, nor does it produce one universally dominant model across systems.

The long-window linear design also exposes numerical limitations. The condition numbers of the Window Direct Ridge design matrices are approximately 10^{17}–10^{18} for Linear-D8 and Duffing-D8,

and one Duffing seed is diagnosed as rank-deficient; all forecasts and aggregate metrics remain finite for all three seeds. An SVD-solver check on Duffing-D8 seed 0 still selects $\alpha = 10^{-8}$ on validation: IID nMSE changes from 0.0017046 to 0.0017045, and Effect Error from 6.118×10^6 to 6.113×10^6, both by less than 0.1%. The magnitude in the table is therefore not caused solely by the default solver, but its absolute value should still be interpreted as the sensitivity of a highly collinear model beyond naturally generated sequences, rather than as a stable effect-size estimate. Duffing retains known parameter inputs, and the two-dimensional Polynomial One-Step Ridge and Regime Ridge are not forcibly extended. These boundaries, together with only three random seeds, limit extrapolation across models.

Table A16: Complete lightweight-model results at the scale of the deep experiments ($D = 8, L = H = 64$; training, validation, IID, OOD, restart, and intervention counts are $10000, 2000, 5000, 5000, 2000, 2000$; mean $\pm$ standard deviation over three random seeds). Parameter OOD uses Duffing extrapolation, Restart uses $k = 32$, and Effect on Switching is the cross-regime intervention.

System	Model (interface)	IID nMSE ↓	State OOD ↓	Param. OOD ↓	Restart ↓	Effect ↓	Sign ↑
Linear-D8	Persistence (null)	6.0248 ± 0.0193	6.0292 ± 0.0287	–	5.6666 ± 0.0353	*n/a*	*n/a*
	kNN-Raw (MV)	$0.0354 \pm 3.94 \times 10^{-5}$	$0.2022 \pm 4.32 \times 10^{-4}$	–	0.0718 ± 0.0046	$0.9794 \pm 1.45 \times 10^{-4}$	0.6039 ± 0.0012
	kNN-Shape (MV)	$0.0605 \pm 8.10 \times 10^{-4}$	0.3416 ± 0.0083	–	0.1553 ± 0.0037	$0.9906 \pm 5.54 \times 10^{-4}$	0.5182 ± 0.0015
	Window Direct Ridge (MV)	$2.24 \times 10^{-28} \pm 2.29 \times 10^{-30}$	$3.89 \times 10^{-28} \pm 3.07 \times 10^{-30}$	–	$8.21 \times 10^{-29} \pm 1.63 \times 10^{-30}$	$0.9688 \pm 5.56 \times 10^{-4}$	0.7062 ± 0.0033
	Last-State Direct Ridge (MV)	$9.45 \times 10^{-25} \pm 4.18 \times 10^{-27}$	$1.31 \times 10^{-24} \pm 2.42 \times 10^{-27}$	–	$1.73 \times 10^{-25} \pm 5.03 \times 10^{-27}$	$1.07 \times 10^{-24} \pm 2.50 \times 10^{-27}$	1 ± 0
	One-Step Ridge (MV)	$4.04 \times 10^{-25} \pm 1.23 \times 10^{-25}$	$4.03 \times 10^{-25} \pm 1.22 \times 10^{-25}$	–	$2.17 \times 10^{-25} \pm 7.26 \times 10^{-26}$	$4.16 \times 10^{-25} \pm 1.49 \times 10^{-25}$	1 ± 0
Duffing-D8	Persistence (null)	2.4255 ± 0.0239	3.7844 ± 0.0117	2.3238 ± 0.0076	3.1212 ± 0.0261	*n/a*	*n/a*
	kNN-Raw (MV)	0.1445 ± 0.0035	0.2851 ± 0.0033	0.4479 ± 0.0089	0.1329 ± 0.0015	0.9917 ± 0.0023	0.591 ± 0.0028
	kNN-Shape (MV)	0.3822 ± 0.0153	0.3634 ± 0.0064	1.3752 ± 0.0202	0.4182 ± 0.0084	0.9966 ± 0.0034	0.5078 ± 0.0011
	Window Direct Ridge (MV)	$0.0017 \pm 1.47 \times 10^{-5}$	$0.0023 \pm 2.78 \times 10^{-5}$	0.0771 ± 0.0064	$0.0383 \pm 4.58 \times 10^{-4}$	$6.12 \times 10^6 \pm 1.30 \times 10^5$	0.6253 ± 0.0235
	Last-State Direct Ridge (MV)	0.0714 ± 0.0025	$0.0585 \pm 7.65 \times 10^{-4}$	0.2822 ± 0.0051	0.0845 ± 0.0013	0.2354 ± 0.0024	0.8239 ± 0.0074
	One-Step Ridge (MV)	0.0838 ± 0.0028	$0.074 \pm 6.22 \times 10^{-4}$	0.2907 ± 0.0035	$0.0816 \pm 9.57 \times 10^{-4}$	$0.2213 \pm 8.18 \times 10^{-4}$	0.8073 ± 0.0099
Switching-D8	Persistence (null)	11.36 ± 0.0513	–	–	10.5995 ± 0.0277	*n/a*	*n/a*
	kNN-Raw (MV)	0.0926 ± 0.0016	–	–	0.0527 ± 0.007	$0.9916 \pm 1.91 \times 10^{-4}$	0.6406 ± 0.0228
	kNN-Shape (MV)	144.782 ± 4.204	–	–	138.5552 ± 2.368	$1.0026 \pm 2.50 \times 10^{-4}$	0.5112 ± 0.0011
	Window Direct Ridge (MV)	$0.0113 \pm 3.70 \times 10^{-4}$	–	–	$2.69 \times 10^{-4} \pm 8.59 \times 10^{-6}$	0.9446 ± 0.8393	0.7141 ± 0.0989
	Last-State Direct Ridge (MV)	0.1771 ± 0.0024	–	–	0.0383 ± 0.0011	$0.3373 \pm 9.43 \times 10^{-4}$	$0.8446 \pm 1.04 \times 10^{-4}$
	One-Step Ridge (MV)	187.8566 ± 1.4497	–	–	27.0647 ± 0.175	$0.2906 \pm 7.06 \times 10^{-4}$	0.8416 ± 0.0052

"–" means that the protocol is undefined for the system; *n/a* means that Persistence is a no-response null baseline and is excluded from cross-variable intervention rankings. Duffing retains the known-parameter input used in the lightweight main experiment. Polynomial One-Step Ridge and Regime Ridge depend on two-dimensional features or branch definitions and are excluded from this generic scale comparison.

K Complete Ablation Tables

The following tables report the complete numerical results corresponding to Figure 4, together with weight norms and design-matrix condition numbers. Section 6 describes the experimental design.

Table A17: Context-length ablation for Window Direct Ridge on Duffing.

History used	IID nMSE ↓	Effect Error ↓	Weight Frobenius norm	Condition number
1	$0.0010 \pm 7.08 \times 10^{-5}$	$0.0162 \pm 3.77 \times 10^{-4}$	5.615	1.636
2	$5.00 \times 10^{-4} \pm 3.22 \times 10^{-5}$	14.00 ± 0.1412	32.51	3.69×10^{15}
4	$2.11 \times 10^{-4} \pm 2.88 \times 10^{-5}$	5128.9 ± 1273.8	1382.9	4.02×10^{16}
8	$1.84 \times 10^{-5} \pm 1.72 \times 10^{-6}$	$2.25 \times 10^5 \pm 9876.1$	1.10×10^4	7.29×10^{16}
16	$6.72 \times 10^{-6} \pm 1.94 \times 10^{-7}$	$1.46 \times 10^5 \pm 8905.1$	6607.1	1.19×10^{17}

Table A18: Ridge-regularization ablation for Window Direct Ridge on Duffing.

α	IID nMSE ↓	Effect Error ↓	Weight norm	Maximum weight
10^{-8}	$6.72 \times 10^{-6} \pm 1.94 \times 10^{-7}$	$1.46 \times 10^5 \pm 8905.1$	6607.1	2231.3
10^{-6}	$2.98 \times 10^{-5} \pm 2.95 \times 10^{-6}$	$1.40 \times 10^4 \pm 384.5$	1780.9	419.5
10^{-4}	$1.18 \times 10^{-4} \pm 8.05 \times 10^{-6}$	448.2 ± 15.50	218.0	57.65
10^{-2}	$3.20 \times 10^{-4} \pm 2.14 \times 10^{-5}$	8.043 ± 0.1980	29.36	6.715
1	$6.59 \times 10^{-4} \pm 4.19 \times 10^{-5}$	0.4928 ± 0.0083	4.547	0.7999
100	$0.0020 \pm 1.25 \times 10^{-4}$	$0.7732 \pm 8.53 \times 10^{-4}$	1.774	0.1583

L Deep-Model and Foundation-Model Results

Table A19: Deep-model results grouped by interface at $D = 8, L = 64, H = 64$ (hyperparameters selected on validation data; mean $\pm$ standard deviation over three random seeds). "–" marks inapplicable cross-variable intervention metrics.

System	Interface	Model	IID nMSE ↓	Effect Error ↓	Sign Accuracy ↑
Linear-D8	Channel-independent	DLinear	$8.71 \times 10^{-5} \pm 9.62 \times 10^{-6}$	–	–
Linear-D8	Channel-independent	PatchTST	$4.46 \times 10^{-4} \pm 7.8 \times 10^{-5}$	–	–
Linear-D8	Multivariate	iTransformer	$4.53 \times 10^{-5} \pm 2.17 \times 10^{-6}$	0.9923 ± 0.0064	0.5058 ± 0.0049
Linear-D8	Multivariate	GRU-AR	0.0139 ± 0.003	0.483 ± 0.0244	0.6398 ± 0.0156
Linear-D8	Multivariate	Direct MLP	$0.0019 \pm 3.75 \times 10^{-4}$	0.9908 ± 0.0038	0.5308 ± 0.0213
Linear-D8	Multivariate	One-Step MLP	0.1363 ± 0.1184	0.3314 ± 0.0425	0.668 ± 0.0087
Duffing-D8	Channel-independent	DLinear	$0.0149 \pm 4.03 \times 10^{-4}$	–	–
Duffing-D8	Channel-independent	PatchTST	$0.0019 \pm 1.18 \times 10^{-4}$	–	–
Duffing-D8	Multivariate	iTransformer	$9.67 \times 10^{-4} \pm 6.98 \times 10^{-5}$	0.8059 ± 0.0049	0.6637 ± 0.0129
Duffing-D8	Multivariate	GRU-AR	0.0119 ± 0.0015	0.699 ± 0.0028	0.6288 ± 0.0295
Duffing-D8	Multivariate	Direct MLP	$0.0115 \pm 6.28 \times 10^{-4}$	0.7754 ± 0.019	$0.5873 \pm 8.9 \times 10^{-4}$
Duffing-D8	Multivariate	One-Step MLP	0.4456 ± 0.1005	0.6125 ± 0.0666	0.6094 ± 0.0156
Switching-D8	Channel-independent	DLinear	0.0858 ± 0.0048	–	–
Switching-D8	Channel-independent	PatchTST	0.9478 ± 0.0515	–	–
Switching-D8	Multivariate	iTransformer	0.3681 ± 0.0565	1.0081 ± 0.0096	0.5181 ± 0.0177
Switching-D8	Multivariate	GRU-AR	0.238 ± 0.089	1.0731 ± 0.0358	0.588 ± 0.0645
Switching-D8	Multivariate	Direct MLP	0.0847 ± 0.0099	1.0015 ± 0.0027	0.4852 ± 0.0359
Switching-D8	Multivariate	One-Step MLP	1.0095 ± 0.0028	1.0152 ± 0.0319	0.1512 ± 0.1164

Table A20: State OOD rerun on Duffing-D8 after correcting the split. Mean position magnitude at the forecast origin is at most 0.75 for IID and at least 0.90 for State OOD. Validation-selected configurations are retained and fully retrained. Every column comes from the same corrected rerun (mean $\pm$ standard deviation over three random seeds).

Interface	Model	IID nMSE ↓	State OOD nMSE ↓	OOD/IID ↓
Channel-independent	DLinear	$0.0152 \pm 3.92 \times 10^{-4}$	$0.0158 \pm 3.53 \times 10^{-4}$	1.0445 ± 0.0476
	PatchTST	$0.002 \pm 4.36 \times 10^{-5}$	$0.0025 \pm 1.7 \times 10^{-4}$	1.2691 ± 0.0581
Multivariate	iTransformer	$0.0013 \pm 1.92 \times 10^{-4}$	$0.0017 \pm 3.25 \times 10^{-4}$	1.3278 ± 0.0573
	GRU-AR	0.0115 ± 0.0021	0.0177 ± 0.0024	1.5576 ± 0.0998
	Direct MLP	$0.0182 \pm 3.76 \times 10^{-4}$	0.021 ± 0.0011	1.1546 ± 0.084
	One-Step MLP	0.3825 ± 0.1114	0.5978 ± 0.173	1.5635 ± 0.0032

Table A21: Zero-shot foundation-model results grouped by interface at $D = 8, L = 64, H = 64$ (mean $\pm$ standard deviation over three random seeds). "–" marks inapplicable cross-variable intervention metrics.

System	Interface	FM	IID nMSE ↓	Effect Error ↓	Sign Accuracy ↑
Linear-D8	Channel-independent	TimesFM	5.8131 ± 0.0119	–	–
Linear-D8	Channel-independent	Chronos	140.7 ± 1.4116	–	–
Linear-D8	Channel-independent	Time-MoE	50.27 ± 0.3557	–	–
Linear-D8	Channel-independent	Toto 2.0 (4M)	1.0686 ± 0.0051	–	–
Linear-D8	Multivariate	Chronos-2	2.659 ± 0.0093	5.3431 ± 0.1544	$0.5101 \pm 5.51 \times 10^{-4}$
Linear-D8	Multivariate	Moirai	473.4 ± 1.9061	119.3 ± 5.1562	$0.5005 \pm 1.61 \times 10^{-4}$
Duffing-D8	Channel-independent	TimesFM	0.4686 ± 0.0021	–	–
Duffing-D8	Channel-independent	Chronos	0.411 ± 0.0041	–	–
Duffing-D8	Channel-independent	Time-MoE	0.6633 ± 0.0015	–	–
Duffing-D8	Channel-independent	Toto 2.0 (4M)	0.4782 ± 0.0068	–	–
Duffing-D8	Multivariate	Chronos-2	0.5034 ± 0.0016	17.11 ± 0.4704	$0.4392 \pm 3.95 \times 10^{-4}$
Duffing-D8	Multivariate	Moirai	1.7094 ± 0.0278	560.7 ± 195.5	$0.5069 \pm 1.85 \times 10^{-4}$
Switching-D8	Channel-independent	TimesFM	4899.2 ± 163.5	–	–
Switching-D8	Channel-independent	Chronos	$2.92 \times 10^{4} \pm 556.7$	–	–
Switching-D8	Channel-independent	Time-MoE	4834.3 ± 43.79	–	–
Switching-D8	Channel-independent	Toto 2.0 (4M)	50.29 ± 0.4836	–	–
Switching-D8	Multivariate	Chronos-2	113.6 ± 0.541	23.88 ± 0.1855	$0.5376 \pm 5.1 \times 10^{-4}$
Switching-D8	Multivariate	Moirai	$4.15 \times 10^{6} \pm 1.89 \times 10^{6}$	$5.82 \times 10^{4} \pm 8.42 \times 10^{4}$	$0.5126 \pm 7.97 \times 10^{-4}$

M QUALITATIVE FOUNDATION-MODEL FORECASTS

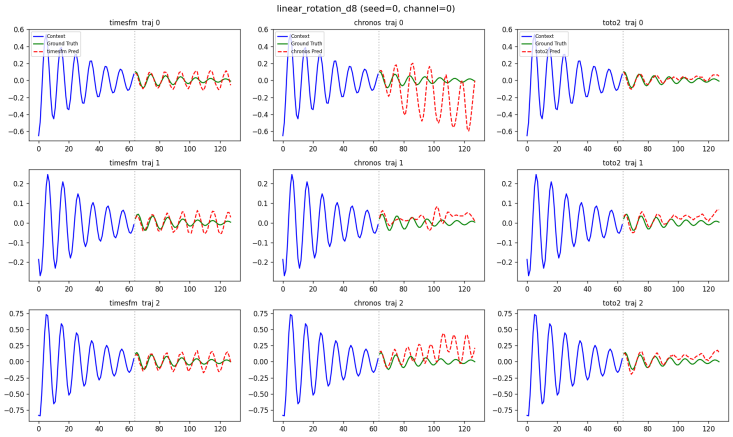

Figure A1: Example zero-shot FM forecasts on Linear-D8. Blue denotes the input history, green the true future, and dashed red the forecast future.

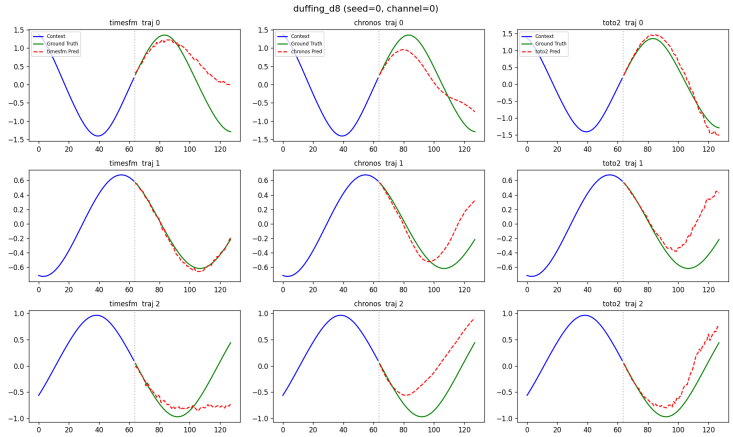

Figure A2: Example zero-shot FM forecasts on Duffing-D8.

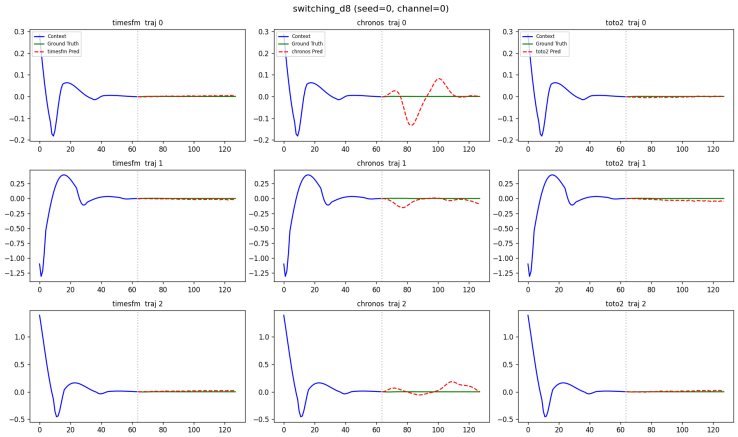

Figure A3: Example zero-shot FM forecasts on Switching-D8.

N Additional Scope and Coverage Considerations

Section 8 discusses the evidence required by real-world evolution claims. We give the corresponding forecasting target here, followed by the scope of the controlled comparisons.

N.1 Real-World Forecasting Target

Real systems additionally contain known controls u_t, unobserved drivers w_t, and time-varying operating conditions ϕ_t:

$$\begin{aligned} s_{t+1} &= F_{\phi_t}(s_t, u_t, w_t, \varepsilon_{t+1}), \\ \phi_{t+1} &= Q(\phi_t, w_t, \xi_{t+1}), \\ x_t &= \mathcal{O}_t(s_t, \eta_t). \end{aligned} \tag{A6}$$

The predictor must therefore infer latent state and operating condition while separating known inputs from shocks and observation error.

Accordingly, real-world forecasting estimates a conditional distribution under a declared information set and future scenario. Let $\mathcal{I}_t$ contain the observations, temporal information, known exogenous variables, and metadata available at time t, and let $c_{t:t+H-1}$ specify the horizon scenario. The target is

$$\mathcal{P}_{t,H}(\,\cdot\,; c) = \Pr(x_{t+1:t+H} \in \,\cdot\, \mid \mathcal{I}_t,\, c_{t:t+H-1})\,. \tag{A7}$$

The point forecast $G_\theta(X_t)$ is a decision quantity such as a conditional mean or quantile. Changing the information set, policy, or scenario changes the target.

N.2 Scope of the Controlled Comparisons

Joint scaling and cross-category comparisons. The D2–D8 path changes D, L, H, sample count, system coupling, and some model parameter counts simultaneously. It neither isolates a single factor nor matches optimization budgets or information interfaces. Lightweight Duffing models can read known system parameters, whereas DL and FM models cannot; changing the FM evaluation sample count also leaves their zero-shot pretraining data unchanged. Cross-scale results therefore describe the joint sensitivity of a fixed model and do not constitute a cross-category SOTA ranking of ML, DL, and FM models. The D2 parameter-scale figure is likewise a descriptive comparison across training paradigms, not a controlled scaling law.

Numerical stability, model coverage, and forecast horizon. The design matrix of D8 long-window Window Direct Ridge is highly collinear. Although an SVD check preserves the order of magnitude, its enormous Effect Error should be interpreted as sensitivity beyond the natural support, not as a stable effect size. The native Direct MLP and One-Step MLP are unmatched in input interface and parameter count. The main text therefore uses D2–D8 parameter-matched controls to identify history-interface and AR/Direct differences, and localizes rollout compounding with horizon-wise error and Jacobian analyses at D2 only. Jacobian, spectral, coverage, gate, and paired-response-loss analyses remain incomplete at D4/D8. The matched One-Step AR on D8 Switching exhibits extremely large but finite divergence under intervention rollouts, showing that natural-validation selection does not guarantee response stability. Some high-learning-rate Duffing-D8 One-Step MLP configurations also produce non-finite loss. We evaluate only two foundation models with native multivariate interfaces and do not provide controlled ablations of normalization, fine-tuning, or prompting, so the study does not represent the complete foundation-model family. Finally, our experiments cover only $H = 16, 32, 64$ and three representative mechanisms. Longer horizons, stochastic differential equations, delayed or chaotic systems, and open systems with control inputs remain for future work. The current experiments provide an auditable diagnostic starting point, not a complete characterization of every time-series capability.

O OBSERVATIONAL DIAGNOSTICS ON SEVEN REAL-WORLD BENCHMARKS

Section 7 presents the central observational findings. This appendix supplies the complete data splits, prediction rules, accuracy tables, restart results, context and regularization ablations, and observed-state-group analysis.

O.1 DATA, MODELS, AND SELECTION

We use ETTh1, ETTh2, ETTm1, ETTm2, Electricity, Traffic, and Weather, with the chronological splits in Table A22. The ETT splits follow the fixed benchmark convention of Zhou et al. (2020). The remaining datasets use the standard 70% training, 10% validation, and 20% test division, with integer boundaries. Every channel and every eligible forecast origin is included, with stride one. Each channel is standardized using its training mean and population standard deviation. MSE therefore equals the training-variance-normalized nMSE defined in the main text, and MAE uses the same standardized coordinates.

Table A22: Real-world data settings. Split sizes count observations; daily periods use the nominal sampling interval.

Dataset	Channels	Step (min)	Train	Validation	Test	Daily period
ETTh1	7	60	8,640	2,880	2,880	24
ETTh2	7	60	8,640	2,880	2,880	24
ETTm1	7	15	34,560	11,520	11,520	96
ETTm2	7	15	34,560	11,520	11,520	96
Electricity	321	60	18,412	2,632	5,260	24
Traffic	862	60	12,280	1,756	3,508	24
Weather	21	10	36,887	5,270	10,539	144

The main history length is $L = 192$, and the forecast horizons are $H \in \{96, 192\}$. All lengths denote sampling steps. Weather retains the original row order, including one repeated timestamp and one 100-minute interval. Validation and test inputs may use observations preceding their split boundary, while each complete target lies inside its own split. The ETT rows beyond the standard test boundary are unused.

We compare six fully specified prediction rules. TrainMean returns the training mean of each channel. Persistence repeats its last observation. SeasonalNaive repeats the latest nominal daily cycle, using the periods in Table A22. Last-value Direct Ridge maps the last observation to the complete future. Window Direct Ridge maps the 192-step history to the complete future. Window AR Ridge learns a one-step map from the same history and recursively feeds back its own predictions.

The three Ridge models use a shared channel-independent interface. Each training row contains one channel's standardized history at one origin. Pooling these rows fits a single coefficient matrix and intercept shared across channels. This retains every channel without estimating cross-channel coefficients. For N pooled rows, the objective is $N^{-1}\|XW + \mathbf{1}b^\top - Y\|_F^2 + \lambda\|W\|_F^2$, with an unpenalized intercept. We select $\lambda \in \{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 100\}$ separately for each dataset, horizon, and model using full-horizon natural validation MSE. For AR, selection uses the complete recursive forecast. The equivalent sum-loss Ridge coefficient is $\alpha = N\lambda$.

All direct models use the same training origins, beginning after 192 observations and ending at the last complete target. The one-step model uses every complete one-step training target from the same initial origin. The models remain fixed after validation selection. Window Direct has $193H$ fitted coefficients and intercepts, Last-value Direct has $2H$, and Window AR has 193. Their comparisons therefore concern these prediction rules and information interfaces, rather than parameter-matched architectures.

O.2 NATURAL FORECAST ACCURACY

Table A23 reports the complete natural-test results. Window Direct has the lowest MSE in 11 of the 14 dataset–horizon settings, while Window AR leads in the other three. Thus recurrence does not

produce a general accuracy advantage in this comparison. The ranking also depends on the loss: at $H = 192$, SeasonalNaive has the lowest MAE on ETTh2 and ETTm2, although a window Ridge model has the lowest MSE. Accurate natural forecasting is well within the reach of these linear rules, but this table alone does not assess their response to a change in the generating process.

Table A23: Complete natural-test MSE / MAE. Values are averaged across all forecast origins, leads, and standardized channels. Bold marks the lowest value for each metric within a row. Last, Direct, and AR denote the three shared channel-independent Ridge rules.

Dataset	H	Train mean	Persistence	Seasonal Naive	Last-value Direct	Window Direct	Window AR
ETTh1	96	1.1099 / 0.7960	1.2944 / 0.7132	0.5122 / 0.4333	0.9049 / 0.6566	**0.3769 / 0.3928**	0.3781 / 0.3937
ETTh1	192	1.1111 / 0.7980	1.3249 / 0.7331	0.5808 / 0.4692	0.9156 / 0.6713	**0.4204 / 0.4187**	0.4248 / 0.4226
ETTh2	96	3.1560 / 1.3623	0.4317 / 0.4216	0.3905 / 0.3802	0.5745 / 0.5493	0.3148 / 0.3739	**0.3120 / 0.3717**
ETTh2	192	3.1628 / 1.3612	0.5337 / 0.4725	0.4819 / **0.4285**	0.7508 / 0.6296	**0.4251** / 0.4434	0.4277 / 0.4447
ETTm1	96	1.1053 / 0.7913	1.2139 / 0.6648	0.4227 / 0.3865	0.8686 / 0.6214	**0.3060 / 0.3445**	0.3173 / 0.3567
ETTm1	192	1.1045 / 0.7913	1.2612 / 0.6904	0.4631 / 0.4057	0.8914 / 0.6388	**0.3408 / 0.3662**	0.3603 / 0.3828
ETTm2	96	3.1313 / 1.3589	0.2665 / 0.3278	0.2631 / 0.3005	0.3050 / 0.3944	0.1868 / 0.2850	**0.1813 / 0.2800**
ETTm2	192	3.1359 / 1.3593	0.3400 / 0.3707	0.3206 / **0.3366**	0.4115 / 0.4611	0.2664 / 0.3448	**0.2592** / 0.3394
Electricity	96	1.0125 / 0.8298	1.5878 / 0.9455	0.3211 / 0.3258	0.8299 / 0.7172	**0.1515 / 0.2468**	0.1523 / 0.2471
Electricity	192	1.0119 / 0.8300	1.5962 / 0.9507	0.3043 / 0.3237	0.8303 / 0.7184	**0.1633 / 0.2584**	0.1672 / 0.2613
Traffic	96	1.4641 / 0.7992	2.7145 / 1.0772	1.2171 / 0.4975	1.3381 / 0.7280	**0.4501 / 0.3023**	0.4524 / 0.3034
Traffic	192	1.4742 / 0.8011	2.7471 / 1.0851	1.0925 / 0.4580	1.3421 / 0.7274	**0.4583 / 0.3055**	0.4678 / 0.3107
Weather	96	0.6366 / 0.6072	0.2591 / 0.2542	0.3167 / 0.2880	0.3023 / 0.3770	**0.1844 / 0.2439**	0.1867 / 0.2463
Weather	192	0.6374 / 0.6076	0.3092 / 0.2917	0.3426 / 0.3049	0.3242 / 0.3928	**0.2243 / 0.2813**	0.2300 / 0.2894

O.3 Forecast Revisions After New Observations

For each horizon, we evaluate restart offsets $k = H/4$ and $k = H/2$. Both complete forecasts must lie inside the test split. The old prediction at leads $k + 1, \ldots, H$ and the updated prediction at leads $1, \ldots, H - k$ refer to the same realized future. Their mean squared difference is the Restart quantity defined in the main text. Here it measures forecast revision after the origin moves forward. We additionally measure the old and updated MSE on that common target and define update gain as old MSE minus updated MSE.

Table A24 illustrates the distinction at $H = 192$ and $k = 96$. On Traffic, Window Direct has lower natural MSE than Window AR, 0.4583 versus 0.4678, but a smaller update gain, 0.0082 versus 0.0265. The old forecasts differ in accuracy, so the larger gain does not establish a better updated forecast: Direct still has lower updated MSE, 0.4513 versus 0.4530. Revision magnitude, improvement from updating, and final prediction error answer different questions.

Smaller revision is not intrinsically preferable. TrainMean has zero revision at every offset because it ignores new observations. Updating can also increase error: seven of the 168 model–dataset–horizon–offset combinations have negative gain. These comprise SeasonalNaive on Electricity at both offsets for $H = 192$, and five Traffic cases at that horizon: SeasonalNaive and Last-value Direct at both offsets, and Persistence at $k = 96$. The largest negative gain is SeasonalNaive on Traffic at $k = 96$, where old MSE 0.9434 increases to 1.2444, yielding a gain of -0.3010. These are revisions under newly observed data. The moving window also discards its oldest observations, so the two input information sets are not nested.

O.4 History, Regularization, and Input Sensitivity

We retrain Direct Ridge using terminal histories $L' \in \{1, 2, 4, 8, 16, 24, 96, 192\}$, keeping the training and evaluation origins fixed and selecting regularization separately by validation. We also evaluate all six regularization values at $L = 192$. Table A25 reports the endpoints at $H = 192$. The full context and regularization results are retained in the accompanying artifact.

For the resulting multistep linear map, let $\widehat{J}_h = W_{L',h}$ denote the derivative of the forecast at lead h with respect to its own channel's final standardized input. We summarize terminal sensitivity by

Table A24: Forecast revisions at $H = 192$ and $k = 96$. Revision is the mean squared difference between overlapping forecasts. Old and updated MSE use the same future targets; positive update gain means lower error after updating.

Dataset	Window model	Revision	Old MSE	Updated MSE	Update gain
ETTh1	Direct	0.1238	0.4642	0.3793	0.0849
ETTh1	AR	0.1234	0.4739	0.3808	0.0931
ETTh2	Direct	0.1643	0.5300	0.3221	0.2080
ETTh2	AR	0.1672	0.5366	0.3213	0.2153
ETTm1	Direct	0.0955	0.3742	0.3072	0.0670
ETTm1	AR	0.0916	0.4020	0.3179	0.0841
ETTm2	Direct	0.1425	0.3469	0.1887	0.1582
ETTm2	AR	0.1350	0.3381	0.1829	0.1552
Electricity	Direct	0.0386	0.1751	0.1486	0.0265
Electricity	AR	0.0321	0.1820	0.1494	0.0325
Traffic	Direct	0.0675	0.4595	0.4513	0.0082
Traffic	AR	0.0483	0.4795	0.4530	0.0265
Weather	Direct	0.0701	0.2643	0.1852	0.0791
Weather	AR	0.0693	0.2735	0.1876	0.0858

$S = \sqrt{H^{-1} \sum_h \widehat{J}_h^2}$. Editing that input by ± 0.5 training standard deviations changes its predicted future by a mean square of $0.25S^2$. The input-edit response and Jacobian are therefore two expressions of the same model sensitivity, rather than independent evidence of a correct physical response. Cross-channel derivatives are zero by construction.

Table A25: Direct Ridge ablation endpoints at $H = 192$. Each cell lists values in the order shown in its header. Context comparisons select regularization separately on validation; regularization comparisons hold $L = 192$ fixed. S measures model terminal sensitivity.

Dataset	MSE $L' = 1,\ 192$	S $L' = 1,\ 192$	MSE $\lambda = 10^{-8},\ 100$	S $\lambda = 10^{-8},\ 100$
ETTh1	0.9156 / 0.4204	0.5306 / 0.1580	0.4204 / 0.7705	0.1580 / 0.0030
ETTh2	0.7508 / 0.4251	0.6526 / 0.1981	0.4253 / 1.1749	0.2218 / 0.0036
ETTm1	0.8914 / 0.3408	0.6063 / 0.2183	0.3408 / 0.7127	0.2188 / 0.0032
ETTm2	0.4115 / 0.2664	0.7951 / 0.2277	0.2665 / 0.8977	0.2831 / 0.0037
Electricity	0.8303 / 0.1633	0.4126 / 0.1229	0.1633 / 0.6381	0.1231 / 0.0029
Traffic	1.3421 / 0.4583	0.3444 / 0.0741	0.4584 / 1.1333	0.0788 / 0.0025
Weather	0.3242 / 0.2243	0.5021 / 0.1240	0.2243 / 0.3688	0.1240 / 0.0031

Increasing context from one to 192 steps lowers both natural MSE and terminal sensitivity on all seven datasets at $H = 192$. For Electricity, MSE falls from 0.8303 to 0.1633 while sensitivity falls from 0.4126 to 0.1229. Longer histories redistribute the linear forecast's dependence across observations; reduced reliance on the final value can accompany better natural prediction. Stronger regularization also reduces terminal sensitivity, but its accuracy effect differs. On ETTh2, changing λ from 10^{-8} to 100 reduces sensitivity from 0.2218 to 0.0036 while increasing MSE from 0.4253 to 1.1749. A small input response is consequently not an accuracy criterion.

The selected 192-step Direct Ridge designs have estimated centered-design condition numbers between 34.30 and 135.53 across the two horizons. Weight norms and maximum absolute coefficients are recorded with the ablations. These quantities describe the fitted maps and their numerical conditioning. They provide no target value for the sensitivity of the real process.