

DYNAMICAL PARAMETERS: AN INTERPRETABILITY FRAMEWORK FOR TIME-SERIES FOUNDATION MODELS

Kang Yang Gaofeng Dong Liying Han Mani Srivastava*
 Department of Electrical and Computer Engineering
 University of California, Los Angeles
 {kyang73, gfdong, liying98}@ucla.edu mbs@ucla.edu

ABSTRACT

This work studies a central gap in interpreting time-series foundation models (TSFMs): a dynamical property may be accessible in a hidden state even when the forecast fails to respond correctly as that property changes. We formalize these properties as *Dynamical Parameters*, including trend slope, oscillation frequency, and autoregressive dependence. We compare their *representation accessibility*, measured by recovery from hidden states, with their *forecast response*, measured by agreement with the expected forecast change. Across nine frozen TSFMs and thirteen laws, 42 of 63 model-parameter cells achieve accessibility above 0.95, whereas their median reference-aligned response relative to the conditional reference is only 0.46. To explain this gap, causal geometry compares the hidden-state change required to produce the reference response with the change induced by the parameter intervention. Directly modifying the hidden state recovers the reference response, but the parameter intervention often moves the state in a different direction. These results show that accessible parameter information need not be expressed in forecasts when input changes miss the required hidden-state direction.

1 INTRODUCTION

Classical forecasting connects an observed history to its predicted future through an explicit pipeline. A mathematical model specifies a governing law and its associated *Dynamical Parameters*, such as trend slope, oscillation frequency, and autoregressive dependence. These parameters are estimated from the history and determine how the fitted model extrapolates the future (Kalman, 1960; Harvey, 1989; Durbin & Koopman, 2012; Ljung, 1999). Time-series foundation models (TSFMs), by contrast, produce general-purpose forecasts without explicitly specifying a governing law or its parameters (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024; Shi et al., 2025; Ekambaram et al., 2024). Inspired by classical forecasting, we interpret TSFMs through two properties. *Representation accessibility* measures whether the governing law and its parameters can be recovered from hidden representations. *Forecast response* measures whether changing a parameter produces the corresponding change in the forecast. These properties relate what a TSFM represents to what its forecast expresses without assuming that it follows the classical pipeline internally.

Recent interpretability studies examine hidden representations through probing and forecast behavior through activation steering, concept erasure, and feature ablation (Goswami et al., 2024; Wiliński et al., 2025; Sanyal et al., 2025; Pandey et al., 2025; Pagani et al., 2026; Mishra, 2026; Zhou et al., 2026). Generator-level interventions further measure forecast sensitivity to controlled changes in temporal patterns (Jander et al., 2026). However, successful probing establishes that a parameter is accessible, not that its changes are appropriately reflected in the forecast. Likewise, altering a forecast through intervention establishes influence, not agreement with the direction, magnitude, and temporal shape implied by the changed dynamics. Existing evidence therefore does not determine whether accessible Dynamical Parameters correspond to quantitatively correct forecast responses.

Determining whether accessible parameters are reflected in forecasts requires references for both parameter recovery and forecast change. Because a finite, noisy history may not fully identify a

*Mani Srivastava holds concurrent appointments as a Professor of ECE and CS (joint) at the University of California, Los Angeles.

parameter, we compare probe performance with a law- and nuisance-informed conditional reference. To isolate the forecast consequence of a parameter change, we construct matched histories that differ in one designated parameter but share the remaining parameters, initial conditions, and stochastic innovations. The predictable difference between their futures defines the corresponding reference response. Reference-aligned gain and full-vector fidelity then measure whether the forecast change matches this target in direction, magnitude, and temporal shape. These measurements account for finite-history uncertainty and distinguish response fidelity from mere forecast sensitivity.

When a parameter is accessible but its matched-input forecast response is weak, the gap can arise for two reasons. The selected hidden state may not support the reference response, or the parameter intervention may not induce the hidden-state change required to produce it. We develop *causal geometry* to distinguish these possibilities. The forecast Jacobian yields a small hidden-state edit toward the reference response, which we term a *local causal direction*. A nonlinear intervention then verifies its effect. Comparing this required edit with the hidden-state change induced by the matched parameter intervention distinguishes limited response capacity from a mismatch between the required and induced changes. Shared subspaces summarize structure across the effective edits, while a readout predicts their input-specific *causal coordinates* from the original hidden state.

Across nine frozen TSFMs, including Chronos-2, TimesFM 3, and Moirai (Ansari et al., 2025; Jain & Sen, 2026; Woo et al., 2024), we study ten single-parameter and three multi-parameter laws spanning linear trend, oscillation, and dual tone. Three principal findings emerge:

- **Accessible Parameters Are Not Consistently Expressed in Forecasts.** Across 63 model-parameter cells, 42 achieve accessibility above 0.95, while the median conditional-reference-normalized aligned gain is 0.46. Three models reverse the expected trend-slope response, and full-vector metrics reveal further errors in magnitude and temporal shape.
- **Hidden-State Edits Recover Responses Missed by Input Changes.** Calibrated hidden-state edits recover the reference responses with low full-vector error, but input changes induce hidden-state changes that align weakly with these edits. Coordinates predicted from the original hidden states increase median aligned gain from 0.23 to 0.58.
- **Compact Geometry Persists Under Shuffled Targets.** Low-rank subspaces summarize effective edit directions, and shuffled targets often yield comparable compactness. The shared causal-subspace structure therefore extends beyond the correct parameter-response pairing.

These results separate parameter accessibility from forecast expression and link the gap to misalignment between required and input-induced hidden-state changes. Shuffled targets show comparable compactness, indicating low-rank editability beyond the correct parameter-response pairing.

2 RELATED WORK

Time-Series Foundation Models (TSFMs). TSFMs are pretrained on large-scale time-series corpora and demonstrate strong zero-shot forecasting across diverse domains (Garza et al., 2023; Rasul et al., 2023; Ansari et al., 2024; Das et al., 2024). This work studies whether *Dynamical Parameters* are accessible in their hidden states and expressed through their forecast responses.

Interpretability of TSFMs. Probing and representation analyses recover temporal properties and continuous parameters, including trend, frequency, periodicity, phase, and amplitude, from hidden representations (Goswami et al., 2024; Wiliński et al., 2025; Pandey et al., 2025; Pagani et al., 2026; Chemeris et al., 2026). Complementing this evidence about accessible information, steering, activation transplantation, sparse feature discovery, and component ablation show that internal features can alter forecasts (Wiliński et al., 2025; Sanyal et al., 2025; Mishra, 2026; Oublal et al., 2026; Bao et al., 2026). At the level of the generating process, (Jander et al., 2026) measure scalar forecast responses to controlled parameter changes in two TSFMs across six temporal patterns. These results leave open how parameter accessibility relates to the direction, magnitude, and full-horizon shape of the forecast response when that parameter changes. We characterize this relationship against conditional references across nine TSFMs and thirteen laws. To examine discrepancies between accessibility and response, causal geometry compares natural hidden changes with local edits derived by Jacobian inversion and tests input-specific causal coordinates.

State-Space Models. Classical state-space inference estimates latent states and dynamical parameters from observed histories for forecasting (Kalman, 1960; Harvey, 1989; Hamilton, 1994; Durbin & Koopman, 2012; Ljung, 1999). Koopman forecasters impose spectral or stability constraints on

their learned latent transitions, while Transformers trained on dynamical systems learn in-context estimation and filtering (Forootani et al., 2026; Forootani & Iervolino, 2026; Garg et al., 2022; Akyürek et al., 2023; von Oswald et al., 2023; Bai et al., 2023; Goel & Bartlett, 2024; Du et al., 2023; Xie et al., 2022; Müller et al., 2022). We compare parameter accessibility and controlled forecast response in frozen general-purpose TSFMs, relating them through hidden interventions.

3 METHOD

Given an observed history $x_{1:T}$, a pretrained TSFM produces hidden representations and a forecast. We characterize two properties of *Dynamical Parameters*:

- **Representation Accessibility (§3.2).** This property measures whether the governing law and its parameter values are linearly accessible from the selected hidden representations.
- **Forecast Response (§3.3).** This property measures whether changing a parameter produces the reference forecast change in direction, magnitude, and temporal shape.

We relate these properties through **causal geometry (§3.4)**, which compares the hidden-state change required to produce the reference response with the change induced by the parameter intervention.

3.1 CONTROLLED STATE-SPACE FRAMEWORK

Dynamical Laws and Parameters. Each time series is defined by a dynamical law $m \in \{1, \dots, M\}$ and a parameter vector that shapes its temporal evolution

$$z_m = (z_{m,1}, \dots, z_{m,K_m}) \in \mathcal{Z}_m.$$

The law m specifies the form of the dynamics, while z_m contains the quantitative properties governing the evolution of the series. Each component $z_{m,j}$ is a *Dynamical Parameter*. Every law instantiates a common state-space form in which z_m controls the transition dynamics, input, or readout.

Given m , z_m , and stochastic variation ε , the controlled generator produces a history and future:

$$(x_{1:T}, y_{T+1:T+H}) = G_m(z_m, \varepsilon). \quad (1)$$

The TSFM observes only the history and produces a forecast and layer-wise hidden representations:

$$\hat{y}_{T+1:T+H} = F_\theta(x_{1:T}), \quad h^\ell = F_\theta^\ell(x_{1:T}).$$

The generator exposes the law and parameters that the TSFM can recover only from the history.

Parameter Interventions. To isolate the effect of parameter $z_{m,j}$, two parameter vectors are constructed that differ only in that component:

$$z_{m,j}^B = \mathcal{J}_{m,j}(z_{m,j}^A, \delta), \quad z_{m,j'}^B = z_{m,j'}^A \quad \text{for } j' \neq j.$$

Both sequences are generated using the same stochastic realization ε :

$$(x^A, y^A) = G_m(z_m^A, \varepsilon), \quad (x^B, y^B) = G_m(z_m^B, \varepsilon). \quad (2)$$

Here, x^A and x^B are the matched histories, while y^A and y^B are their futures. Sharing ε removes stochastic variation unrelated to the parameter intervention. The resulting reference response is

$$\Delta y_{m,j}^{\text{ref}} = y^B - y^A. \quad (3)$$

The known law m , Dynamical Parameter vector z_m , and reference response $\Delta y_{m,j}^{\text{ref}}$ provide targets for analyzing representation accessibility, forecast response, and causal geometry.

3.2 REPRESENTATION ACCESSIBILITY

Representation accessibility measures how accurately the dynamical law m and its parameters z_m can be recovered from hidden representations using linear probes. The analysis has three steps.

(I) Probing Laws and Parameters. At each layer ℓ , a linear classifier predicts the dynamical law from h^ℓ , while a linear ridge probe predicts each parameter within its corresponding law:

$$\hat{m}^\ell = \underset{m'}{\operatorname{argmax}} (W_\ell h^\ell + b_\ell)_{m'}, \quad \hat{z}_{m,j}^\ell = w_{\ell,m,j}^\top h^\ell + b_{\ell,m,j}.$$

(II) Dynamical Law Decodability. For each law $m \in \{1, \dots, M\}$, let A_m^* denote an upper reference for identifying m . For balanced M -way classification, $A^0 = 1/M$ is the uniform-guessing accuracy. The layer-wise accuracy and normalized law decodability are

$$A_m^\ell = \Pr(\hat{m}^\ell = m \mid m), \quad D_m(\ell) = \frac{A_m^\ell - A^0}{A_m^* - A^0}. \quad (4)$$

When a reliable Bayes ceiling is unavailable, $A_m^* = 1$ provides a conservative upper reference. A value of $D_m(\ell)$ near 1 indicates that the hidden representation approaches this reference, while a value near 0 indicates chance-level identification.

(III) Parameter Accessibility. For each Dynamical Parameter $z_{m,j}$, we compare the probe correlation $r_{m,j}^\ell$ with a conditional posterior-mean reference:

$$r_{m,j}^\ell = \text{corr}(\hat{z}_{m,j}^\ell, z_{m,j}), \quad r_{m,j}^* = \text{corr}(\mathbb{E}[z_{m,j} \mid x_{1:T}, m, \eta], z_{m,j}), \quad D_{m,j}(\ell) = \frac{r_{m,j}^\ell}{r_{m,j}^*}. \quad (5)$$

Here η collects the scale, offset, noise, phase, and non-designated parameters supplied to the reference observer but not to the TSFM. A value of $D_{m,j}(\ell)$ near 0 indicates no linear association, while a value near 1 means the probe approaches this law- and nuisance-informed reference.

3.3 FORECAST RESPONSE

Forecast response measures whether changing a Dynamical Parameter changes the TSFM forecast in the direction, magnitude, and temporal shape implied by the law. The analysis has three steps.

(I) Matched Forecast Response. For the matched pair in Equation 2, the TSFM forecast response is

$$\Delta \hat{y}_{m,j} = F_\theta(x^B) - F_\theta(x^A).$$

(II) Response Agreement. To compare the forecast and reference responses while allowing a constant level offset, define the centered horizon response

$$\mathcal{C}(a) = a - \frac{1}{H}(\mathbf{1}^\top a)\mathbf{1}, \quad a \in \mathbb{R}^H. \quad (6)$$

For N matched pairs, the per-pair response coefficient and its mean are

$$u_{i,m,j} = \frac{\langle \mathcal{C}(\Delta \hat{y}_{i,m,j}), \mathcal{C}(\Delta y_{i,m,j}^{\text{ref}}) \rangle}{\|\mathcal{C}(\Delta y_{i,m,j}^{\text{ref}})\|_2^2}, \quad U_{m,j} = \frac{1}{N} \sum_{i=1}^N u_{i,m,j}. \quad (7)$$

Here, $u_{i,m,j}$ is the least-squares slope of the forecast response on the reference response, with a free level offset. The mean $U_{m,j}$ near 1 indicates a reference-aligned component of the expected size, near 0 indicates little aligned response, and a negative value indicates an opposing component.

FULL-VECTOR FIDELITY AND AGGREGATION. The response coefficient $U_{m,j}$ measures only the component aligned with the reference and therefore does not establish fidelity of the full forecast-change trajectory. For each matched pair i , let q_i denote the realized forecast change and r_i the corresponding reference change. Define $\tilde{q}_i = \mathcal{C}(q_i)$, $\tilde{r}_i = \mathcal{C}(r_i)$, $E = \sum_i \|\tilde{r}_i\|_2^2$, and the pairwise aligned gain $g_i = \langle \tilde{q}_i, \tilde{r}_i \rangle / \|\tilde{r}_i\|_2^2$. We report four pooled metrics:

$$g = \frac{\sum_i \langle \tilde{q}_i, \tilde{r}_i \rangle}{E}, \quad s = \frac{\sum_i \langle \tilde{q}_i, \tilde{r}_i \rangle}{\sqrt{(\sum_i \|\tilde{q}_i\|_2^2) E}}, \quad e = \sqrt{\frac{\sum_i \|\tilde{q}_i - \tilde{r}_i\|_2^2}{E}}, \quad e_\perp = \sqrt{\frac{\sum_i \|\tilde{q}_i - g_i \tilde{r}_i\|_2^2}{E}}. \quad (8)$$

Here, g measures the reference-aligned component, s measures directional agreement, e measures full centered-trajectory error, and $e_\perp$ measures the component orthogonal to each pair's reference. These quantities satisfy the pooled decomposition $e^2 = \sum_i \|\tilde{r}_i\|_2^2 (g_i - 1)^2 / E + e_\perp^2$; thus, total error reflects both gain mismatch and orthogonal trajectory error. For the natural response, $g_i = u_{i,m,j}$, so $U_{m,j} = N^{-1} \sum_i g_i$ weights pairs equally, whereas pooled g weights each pair by its reference energy. We report both because a few large reference changes can dominate the pooled metric.

(III) Conditional-Reference-Normalized Response. Finite histories and stochastic variation can prevent even an optimal history-based forecaster from reproducing the realized reference response. We therefore define the law- and nuisance-informed conditional forecast

$$F_m^*(x_{1:T}; \eta) = \mathbb{E}[y_{T+1:T+H} \mid x_{1:T}, m, \eta], \quad \Delta y_{m,j}^* = F_m^*(x^B; \eta) - F_m^*(x^A; \eta),$$

where η collects the documented nuisance variables supplied to the conditional forecast. Substituting $\Delta y_{i,m,j}^*$ for $\Delta \hat{y}_{i,m,j}$ in Equation 7 gives the conditional-reference coefficient $U_{m,j}^*$. Across intervention magnitudes $\mathcal{M}_{m,j}$, the normalized forecast response is

$$\bar{U}_{m,j} = \frac{\sum_{\delta \in \mathcal{M}_{m,j}} U_{m,j}(\delta)}{\sum_{\delta \in \mathcal{M}_{m,j}} U_{m,j}^*(\delta)}. \quad (9)$$

A value of $\bar{U}_{m,j}$ near 1 means the model matches the conditional reference’s aggregate aligned component across magnitudes; near 0 means little alignment, and negative values mean an opposing response. Because this signed normalization conditions on m and η , it is a comparison, not a history-only upper bound: $\bar{U}_{m,j}$ can exceed one and does not establish full-vector fidelity.

3.4 CAUSAL GEOMETRY

Causal geometry characterizes how hidden-state interventions affect the frozen downstream computation and whether their effective directions share structure across inputs. Here, “causal” refers only to the effect of modifying the hidden state at a specified layer; it neither identifies causal variables in the generating process nor establishes that the model naturally uses the resulting directions. For a fixed law m , parameter $z_{m,j}$, and intervention magnitude δ , we suppress dependence on m , j , and δ below.

(I) Local Causal Directions. For matched pair i at layer ℓ , let $h_i^\ell = F_\theta^\ell(x_i^A)$ denote the representation of x_i^A , and let $F_\theta^{>\ell}$ denote the remaining computation to the forecast. The intervention modifies the selected hidden-state positions while leaving the downstream computation fixed. Linearizing the forecast around h_i^ℓ , regularized Jacobian inversion yields

$$J_i^\ell = \left. \frac{\partial F_\theta^{>\ell}(h)}{\partial h} \right|_{h=h_i^\ell}, \quad v_i^\ell = \underset{v}{\operatorname{argmin}} \left\| J_i^\ell v - \Delta y_i^{\text{ref}} \right\|_2^2 + \lambda \|v\|_2^2. \quad (10)$$

Here, J_i^ℓ maps hidden-state perturbations to first-order forecast changes, and v_i^ℓ is the regularized, input-specific *local causal direction*. We evaluate this direction in the nonlinear TSFM and summarize agreement across N matched pairs. For intervention strength $\alpha > 0$, the realized response per unit intervention and pooled aligned gain are

$$\Delta \hat{y}_i^\ell(v; \alpha) = \frac{F_\theta^{>\ell}(h_i^\ell + \alpha v) - F_\theta^{>\ell}(h_i^\ell)}{\alpha}, \quad U^{\text{pool}} = \frac{\sum_i \langle \mathcal{C}(\Delta \hat{y}_i^\ell(v_i; \alpha)), \mathcal{C}(\Delta y_i^{\text{ref}}) \rangle}{\sum_i \|\mathcal{C}(\Delta y_i^{\text{ref}})\|_2^2}. \quad (11)$$

The inversion targets the complete reference response. The primary aligned-gain score removes constant level offsets using $\mathcal{C}$ in Equation 6, whereas raw and level diagnostics retain them. A value of U^{pool} near 1 indicates the correct pooled aligned component, near 0 indicates little aligned response, and a negative value indicates an opposing component. Setting $v_i = v_i^\ell$ defines $U_{\text{local}}(\ell)$, which measures whether the Jacobian-derived directions remain effective in the nonlinear model.

(II) Shared Causal Subspace. Although the local causal directions are input-specific, they may concentrate in a compact subspace shared across inputs. The subspace is estimated from a discovery bank $\mathcal{D}_{\text{disc}}$ and evaluated on a disjoint bank $\mathcal{D}_{\text{eval}}$. At layer ℓ , singular-value decomposition gives

$$V_\ell = [v_i^\ell]_{i \in \mathcal{D}_{\text{disc}}} = Q_\ell \Sigma_\ell R_\ell^\top, \quad \mathcal{S}_{\ell,k} = \operatorname{span}(Q_{\ell,k}), \quad \tilde{v}_i^{\ell,k} = Q_{\ell,k} Q_{\ell,k}^\top v_i^\ell, \quad (12)$$

where $Q_{\ell,k}$ contains the leading k left singular vectors. Substituting the projected directions $\tilde{v}_i^{\ell,k}$ into Equation 11 on $\mathcal{D}_{\text{eval}}$ defines $U_{\text{projected}}(\ell, k)$.

For each valid layer, the smallest rank retaining 90% of the local reference-aligned gain is

$$k_{90}(\ell) = \min \left\{ k \in \mathcal{K}_\ell : \frac{U_{\text{projected}}(\ell, k)}{U_{\text{local}}(\ell)} \geq 0.9 \right\}.$$

A small $k_{90}(\ell)$ indicates a low-rank shared causal subspace relative to the intervened state dimension. If no evaluated rank reaches the threshold, $k_{90}(\ell)$ is reported as a lower bound. However, $\tilde{v}_i^{\ell,k}$ still requires the reference-derived local direction v_i^ℓ , so this analysis establishes subspace compactness but not whether its input-specific coordinates are accessible from the hidden representation.

(III) Causal Coordinate Accessibility. The final step asks whether the input-specific coordinates of the local causal direction can be inferred from the unperturbed representation h_i^ℓ . The local direction

is projected into the shared subspace, its coordinates are predicted from h_i^ℓ , and the predicted direction is mapped back into the hidden space:

$$c_i^{\ell,k} = Q_{\ell,k}^\top v_i^\ell, \quad \hat{c}_i^{\ell,k} = P_{\ell,k} h_i^\ell + b_{\ell,k}^c, \quad \hat{v}_i^{\ell,k} = Q_{\ell,k} \hat{c}_i^{\ell,k}. \quad (13)$$

Here, $P_{\ell,k}$ and $b_{\ell,k}^c$ are the ridge-readout coefficients in the original hidden coordinates and incorporate discovery-bank standardization. The readout alone is fitted on $\mathcal{D}_{\text{disc}}$, while the TSFM remains frozen. On $\mathcal{D}_{\text{eval}}$, the predicted direction $\hat{v}_i^{\ell,k}$ requires neither J_i^ℓ nor Δy_i^{ref} . Using these directions in Equation 11 defines the deployed response score $U_{\text{deployed}}(\ell, k)$.

A fixed-direction control uses the evaluation-bank mean direction, rescaled to match each input’s oracle norm. Its response score is $U_{\text{fixed}}(\ell)$. Because the control receives the oracle norm, comparison with $U_{\text{deployed}}(\ell, k)$ isolates the benefit of input-specific coordinates from perturbation magnitude. High projected response at low rank indicates a compact shared subspace, while high deployed response shows that its input-specific coordinates are accessible from unperturbed representations.

4 EVALUATION

Models. We evaluate Chronos-2 (C2) (Ansari et al., 2025), TimesFM 2.5 (F2) (Das et al., 2024), TimesFM 3 (F3) (Jain & Sen, 2026), Toto-Open-Base 1.0 (TO) (Cohen et al., 2025), Moirai (MO) (Woo et al., 2024), Sundial (SU) (Liu et al., 2025b), Time-MoE (ME) (Shi et al., 2025), TTM-R2 (TT) (Ekambaram et al., 2024), and Timer-XL (TX) (Liu et al., 2025a). We use $H = 16$ and $T = 256$ by default. The panel spans diverse architectures, scales, contexts, and training objectives, as detailed in §A.1.

Laws. The analysis includes thirteen laws, each instantiating a common state-space form whose parameters determine the dynamics, input, or readout. Ten are single-parameter laws: linear trend (TR), oscillation (OS), mean reversion (MR), random walk with drift (RW), exponential relaxation (RX), saturating growth (SG), level shift (LS), damped oscillation (DO), seasonal autoregression (SA), and transient pulse (TP). Three are multi-parameter laws requiring joint identification: dual tone (DT; low tone TL, high tone TH, and amplitude ratio TW), linear state space (SS; modal frequency SC and eigenvalue modulus SR), and forced diffusion (FD; drive frequency WC and diffusion time WD), yielding seven parameters in total. Generators, parameter ranges, and constraints appear in §A.2.

Evaluation Protocol. All TSFM weights remain frozen. Probes are fitted on training datasets, with layers and penalties selected on validation datasets and evaluated once on independent test datasets. Each standard forecast-response dataset contains 32,768 matched pairs per intervention magnitude. Larger datasets are used for high-variance settings to improve stability across runs. Causal geometry uses separate fitting, validation, and test datasets. Implementation details for representation accessibility, forecast response, and causal geometry appear in §A.3, §A.4, and §A.5, respectively.

4.1 DYNAMICAL LAW IDENTIFICATION

Dynamical-law identification tests whether TSFM representations distinguish the thirteen laws. Balanced training, validation, and test banks treat each multi-parameter law as one class with jointly varying parameters. A multinomial logistic probe is fitted at each layer, with layer and penalty selected on validation data. For $D_m(\ell)$, we set chance accuracy to $A_0 = 1/13$ and the reference to $A_m^* = 1$, yielding normalized above-chance accuracy. Results appear in Figure 1.

Dynamical Laws Are Broadly Decodable.

Every model distinguishes the thirteen laws far above chance. Mean law decodability ranges from 0.757 to 0.971 across models, with a mean of 0.929. Thus, law identity remains broadly accessible even when parameters vary jointly, although this does not establish a unique learned mechanism.

Difficulty Concentrates in Related Dynamics.

Forced diffusion (FD) has the lowest average decodability of 0.842 and is the weakest law for seven models. Random walk with drift (RW) averages 0.851 and is weakest for Moirai (MO) and TTM-R2 (TT). Other laws average at least 0.877, indicating that errors concentrate in a small subset of dynamical families.

	C2	F2	F3	TO	MO	SU	ME	TT	TX	TR	OS	MR	RW	RX	SG	LS	DO	SA	TP	DT	SS	FD
C2	0.97	0.93	0.97	0.94	0.98	0.98	1.00	0.99	0.99	0.95	0.98	0.98	0.91									
F2	0.97	0.95	0.97	0.94	0.98	0.98	1.00	0.99	0.99	0.96	0.98	0.98	0.93									
F3	0.94	0.94	0.96	0.92	0.98	0.99	1.00	0.99	0.92	0.95	0.98	0.98	0.91									
TO	0.96	0.91	0.96	0.92	0.97	0.98	1.00	0.99	0.99	0.95	0.98	0.97	0.89									
MO	0.92	0.86	0.89	0.74	0.97	0.98	0.97	0.96	0.98	0.90	0.88	0.91	0.84									
SU	0.96	0.91	0.96	0.92	0.97	0.99	1.00	0.99	0.99	0.95	0.98	0.98	0.88									
ME	0.93	0.87	0.95	0.87	0.97	0.98	0.98	0.98	0.99	0.94	0.96	0.96	0.83									
TT	0.86	0.65	0.65	0.51	0.92	0.95	0.83	0.84	0.92	0.83	0.66	0.66	0.56									
TX	0.96	0.87	0.94	0.90	0.97	0.98	0.99	0.99	0.98	0.95	0.94	0.94	0.83									

Figure 1: Thirteen-law decodability scores $D_m(\ell)$ across the nine pretrained TSFMs.

Differences Are Model- and Law-Specific. Seven models average between 0.939 and 0.971, whereas Moirai (MO) averages 0.908 and TTM-R2 (TT) averages 0.757. TTM-R2 (TT) scores 0.51 on random walk with drift (RW) and 0.56 on forced diffusion (FD). They are its weakest laws. TimesFM 3 (F3) scores 0.92 on seasonal autoregression (SA), compared with 0.99 for TimesFM 2.5 (F2), showing that aggregate scores can obscure law-specific differences. Complete layer-wise profiles and confusion analyses for all models and thirteen laws appear in §B.1.

4.2 SINGLE-PARAMETER LAWS

We study ten single-parameter laws. Accessibility covers eight, excluding random walk with drift (RW) and level shift (LS) because their required references are unavailable. Forecast response and causal geometry cover seven, additionally excluding exponential relaxation (RX) because its response reference is unstable. The common comparison spans seven laws and nine models, totaling 63 cells.

4.2.1 REPRESENTATION ACCESSIBILITY

At each layer, a ridge probe predicts $z_{m,j}$, and its correlation relative to the law- and nuisance-informed reference $r_{m,j}^*$ defines $D_{m,j}(\ell)$. The layer and ridge penalty are selected on validation data and fixed for test evaluation. The main comparison uses the forecast-response generative regime.

Dynamical Parameters Are Broadly Accessible.

In Figure 2, median accessibility across 63 common-regime model-law cells is 0.982, and 42 cells exceed 0.95. Fifteen of 16 scores below 0.90 occur for oscillation (OS) or seasonal autoregression (SA). The exception is TimesFM 3 (F3) on linear trend (TR), while the other five laws are highly accessible in nearly every model across diverse architectures and scales.

Accessibility Reflects Both Preservation and Transformation. On the original eight-law probe banks, linear trend (TR), saturating growth (SG), and exponential relaxation (RX) match near-ceiling raw-input controls, indicating preserved linear access. Mean reversion (MR), damped oscillation (DO), and transient pulse (TP) are highly accessible despite chance-level raw linear controls, showing TSFM processing makes them linearly accessible.

Long-Range Parameters Differentiate Models. Common-regime accessibility ranges from 0.49 to 0.80 for oscillation (OS) and from 0.63 to 0.93 for seasonal autoregression (SA). Layer profiles show early access to trend, whereas frequency has a near-zero raw-linear baseline and peaks beyond the network midpoint in seven models. Exponential relaxation (RX) ranges from 0.82 for TimesFM 3 (F3) to 0.97–0.99 for the other models.

Most scored parameters therefore approach the conditional reference. Across 19 model-law cells, MLP probes improve median accessibility by 0.046 for oscillation (OS) and 0.029 for seasonal autoregression (SA), but do not recover the TimesFM 3 (F3) trend result. Controls, per-law ranges, and layer-wise profiles for all models and laws appear in §B.2.

4.2.2 FORECAST RESPONSE

Using the matched histories of §3.1, we compare each TSFM forecast change with the reference change induced by the designated parameter. For each law, $\bar{U}_{m,j}$ aggregates the evaluated intervention magnitudes and normalizes the reference-aligned response by the conditional reference.

Forecast Responses Are Often Weak or Reversed.

In Figure 3, the median normalized aligned gain across 63 model-parameter cells is 0.46, and only 15 cells lie within 20% of the conditional reference. Six responses remain negative across all intervention magnitudes. No model reaches 0.8 on oscillation (OS), seasonal autoregression (SA), or transient pulse (TP), whose maxima are 0.64, 0.71, and 0.62.

Accessibility and Forecast Response Diverge. Among 42 cells with accessibility above 0.95, 20 have

C2	1.00	0.80	0.97	0.99	1.00	0.93	0.99
F2	1.00	0.77	0.97	0.99	0.99	0.93	0.99
F3	0.37	0.76	0.97	0.99	0.99	0.93	0.99
TO	1.00	0.69	0.97	0.99	0.99	0.88	0.99
MO	1.00	0.59	0.96	0.99	0.98	0.63	0.98
SU	1.00	0.72	0.97	0.99	0.99	0.89	0.99
ME	1.00	0.68	0.97	0.99	0.99	0.86	0.99
TT	1.00	0.49	0.90	0.99	0.93	0.73	0.98
TX	1.00	0.62	0.96	0.99	0.99	0.84	0.98
	TR	OS	MR	SG	DO	SA	TP

Figure 2: Representation accessibility $D_{m,j}$.

C2	0.91	0.64	0.79	1.10	1.06	0.71	0.62
F2	0.92	0.46	0.82	1.97	0.96	0.52	0.28
F3	1.01	0.46	0.85	2.33	0.95	0.57	0.33
TO	0.90	0.35	0.78	1.53	1.12	0.47	0.42
MO	0.12	0.03	0.24	0.60	0.04	-0.01	0.06
SU	0.11	0.22	0.82	0.03	1.15	0.38	0.36
ME	-0.45	0.17	0.80	-0.71	1.10	0.54	0.33
TT	-0.69	0.15	0.33	-1.02	0.35	0.09	0.15
TX	-0.82	0.05	0.74	0.43	0.46	0.23	0.20
	TR	OS	MR	SG	DO	SA	TP

Figure 3: Forecast response score $\bar{U}_{m,j}$.

forecast response below 0.5, and five respond oppositely. Linear trend (TR) provides an example: eight models have accessibility between 0.998 and 0.999, including three with negative responses. Near-reference accessibility does not determine the magnitude or direction of the forecast response.

Accessibility Explains Relative but Not Absolute Variation. Across models, accessibility and response have Spearman correlations of 0.92 for oscillation (OS) and 0.82 for seasonal autoregression (SA), the two laws with the largest accessibility variation. Accessibility therefore helps distinguish models when parameter accessibility is limited, but even the strongest responses remain below the conditional reference. Across all cells, the median centered cosine is 0.54 and the median centered relative error is 1.06, revealing additional disagreement in response magnitude and temporal shape. Dose-specific and full-vector analyses appear in §B.3.

4.2.3 CAUSAL GEOMETRY

Causal geometry explains the accessibility–response gap by comparing the hidden-state edit required for the expected response with the change induced by the parameter intervention. We test whether the state supports the response, then compare required and input-induced changes.

Direct Hidden-State Edits Recover Reference Responses. All 63 cells have a validation-selected layer where the Jacobian-derived edit passes the nonlinear response criterion. At the calibrated step, median aligned gain is 1.00 and centered error is 0.06. Projection onto the shared subspace retains median gain 0.93 with centered error 0.18. Weak matched-input responses therefore coexist with effective edits that produce the reference response across models and dynamical laws alike.

Effective Edits Share Compact Structure. Figure 4 shows a median rank of $k_{90} = 16$ for retaining 90% of local aligned gain. Targets shuffled across inputs within the same dynamical law have the same median rank and tie the correct targets in 51 of 63 cells. The correct targets improve both rank and rank-curve area in only six cells. Compactness therefore reflects shared causal-subspace structure but is not consistently tied to the correct input–response pairing.

C2	4	16	8	2	8	32	64
F2	1	8	4	64	2	16	32
F3	1	8	16	1	2	256	4
TO	2	16	16	16	256	256	64
MO	128	16	8	4	8	32	32
SU	4	8	4	4	16	32	8
ME	≥1024	≥1024	1024	1024	512	≥1024	1024
TT	256	16	16	64	64	32	8
TX	256	256	128	256	2	256	4
	TR	OS	MR	SG	DO	SA	TP

Figure 4: Shared causal-subspace rank k_{90} .

Predicted Edits Partially Recover the Response. A readout predicts each input’s edit from the original state without the corresponding test target or Jacobian. Figure 5 shows that it outperforms the fixed-direction control in 56 of 63 cells, with median aligned gains of 0.58 and 0.23. Its centered error remains 0.68, versus 0.18 for the projected edit. The original hidden state therefore predicts part of the required edit.

Input Changes Miss the Required Directions. Input-induced hidden-state changes have median cosine 0.028 with the required local edits. Only 1.1% of their energy lies in the shared subspace, while 99.6% lies in the downstream Jacobian’s null space. Their median linearized aligned gain is 0.28, indicating a weak reference-aligned effect on the forecast. Together, these results show that input changes follow a different hidden-state path from the effective edits producing the reference response across the panel.

C2	0.93	0.73	0.56	0.95	0.92	0.61	0.76
F2	0.94	0.58	0.49	0.85	0.92	0.52	0.46
F3	1.03	0.69	0.49	0.92	0.93	0.50	0.56
TO	1.00	0.51	0.54	0.92	0.72	0.41	0.46
MO	0.94	0.40	0.49	0.92	0.83	0.35	0.51
SU	0.94	0.72	0.49	0.92	0.89	0.66	0.44
ME	0.52	0.25	0.12	0.51	0.77	0.24	0.27
TT	0.69	0.27	0.43	0.87	0.81	0.44	0.65
TX	0.90	0.21	0.36	0.57	0.91	0.32	0.47
	TR	OS	MR	SG	DO	SA	TP

Figure 5: Deployed pooled aligned gain.

The reference response can be produced from the selected state, but the parameter intervention often fails to induce the required hidden-state direction. Complete causal-geometry results appear in §B.4.

4.3 MULTI-PARAMETER LAWS

For seven parameters across three multi-parameter laws, each intervention changes one parameter while holding the others and the stochastic realization fixed. We compare representation accessibility, matched-input forecast response, and deployed aligned gain. Figure 6 summarizes the three measurements.

C2	0.99	0.99	0.99	0.90	0.99	1.00	0.89
F2	0.99	0.99	1.00	0.91	0.99	1.00	0.85
F3	0.99	0.99	1.00	0.91	0.99	1.00	0.85
TO	0.97	0.98	0.99	0.87	0.99	1.00	0.85
MO	0.86	0.94	0.94	0.77	0.98	0.98	0.81
SU	0.99	0.99	0.99	0.90	0.99	1.00	0.88
ME	0.96	0.97	0.98	0.85	0.99	1.00	0.88
TT	0.91	0.93	0.96	0.61	0.91	0.98	0.62
TX	0.96	0.98	0.99	0.83	0.99	1.00	0.80
	TL	TH	TW	SR	SC	WC	WD

C2	0.97	0.99	1.10	1.02	0.83	0.56	0.96
F2	0.97	0.75	0.99	0.86	0.75	0.40	0.84
F3	1.00	0.88	1.07	0.94	0.81	0.48	0.93
TO	0.82	0.81	1.18	0.98	0.79	0.32	0.78
MO	0.04	-0.00	0.06	0.07	0.07	0.01	0.10
SU	0.96	0.89	1.15	0.85	0.70	0.43	0.76
ME	0.89	0.90	1.11	0.90	0.77	0.40	0.84
TT	0.16	-0.16	0.20	0.17	0.19	-0.04	0.30
TX	0.12	0.12	0.35	0.28	0.26	0.05	0.51
	TL	TH	TW	SR	SC	WC	WD

C2	0.86	0.87	0.93	0.59	0.59	0.73	0.80
F2	0.96	0.95	0.94	0.54	0.61	0.68	0.77
F3	0.98	0.97	0.90	0.49	0.63	0.77	0.84
TO	0.75	0.81	0.86	0.49	0.53	0.50	0.74
MO	0.65	0.42	0.54	0.29	0.32	0.19	0.51
SU	0.87	0.89	0.91	0.49	0.50	0.66	0.77
ME	0.67	0.76	0.77	0.17	0.28	0.38	0.61
TT	0.69	0.47	0.55	0.20	0.31	0.38	0.54
TX	0.73	0.40	0.44	0.26	0.44	0.16	0.61
	TL	TH	TW	SR	SC	WC	WD

(a) Representation accessibility.

(b) Matched-input response.

(c) Deployed gain.

Figure 6: Multi-parameter accessibility, matched-input response, and deployed gain across joint laws.

Joint Parameters Remain Accessible. Median accessibility is 0.977. For linear state space, modal frequency (SC) exceeds eigenvalue modulus (SR) by 0.08–0.30 in every model. Adding a matched high tone (TH) reduces low-tone (TL) accessibility by at most 0.033 in six models. Joint components therefore produce limited interference, but accessibility still varies across parameters of the same law.

Accessibility and Forecast Response Diverge. Drive frequency (WC) is accessible at 0.98–1.00, yet no matched-input response exceeds 0.56. Diffusion time (WD) is less accessible, but six models attain responses between 0.76 and 0.96. This reversal within forced diffusion further separates accessibility from forecast response under jointly varying parameter settings.

Predicted Edits Partially Recover the Response. Deployed directions outperform the fixed control in 56 of 63 cells across all three multi-parameter laws. Median centered errors are 0.93, 0.14, and 0.61 for matched-input, projected, and deployed responses. TTM-R2 (TT), Timer-XL (TX), and Moirai (MO) attain deployed gains between 0.16 and 0.73 despite weak or opposing matched-input responses. Thus, accessible parameters can support the reference response without being fully expressed by input changes or predicted edits. Further analyses appear in §B.5.

4.4 CONTROLLED CHANGES ON REAL-SERIES BACKGROUNDS

The controlled-law experiments isolate parameter effects but use synthetic backgrounds. To test whether the accessibility–response separation persists on real data, centered trend or sinusoid components are injected into ETTh1 Zhou et al. (2021) windows. Matched pairs share the background and differ only in the injected component. Injection strength is the component root-mean-square (RMS) amplitude relative to context standard deviation, at 0.25, 0.5, and 1.0. The analytic future difference defines the reference response. All nine TSFMs are evaluated on 2,048 test windows per setting. No-injection and shuffled-future controls yield median pooled gains of 0.000 and −0.000.

Results. Across the evaluation, median probe correlation remains high at 0.94, while pooled matched-input gain is −0.50 for linear trend and 0.36 for sinusoid. As shown in Figure 7, ETTh1 exhibits a separation between representation accessibility and forecast response across three injection strengths, with linear trend gain remaining negative from −1.40 at strength 0.25 to −0.62 at strength 1.0, while sinusoid gain increases from 0.09 to 0.67. The response depends on the background: at strength 1.0, linear trend gain reaches 0.89 on Electricity and 0.71 on Weather, while Weather sinusoid gain reaches 0.77. Thus, the accessibility–response separation persists on real-series backgrounds, although its magnitude and sign vary with the background and injection strength. These gains use the injected continuation rather than a conditional reference for the background. The full comparison across ETTh1, Electricity, and Weather is reported in §B.6.

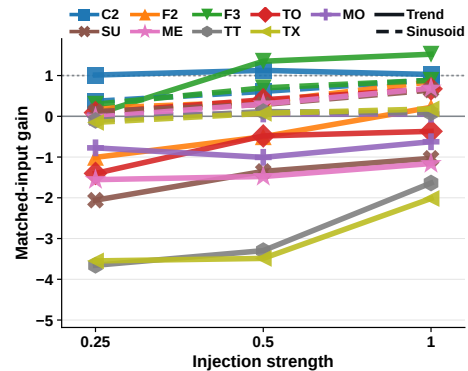


Figure 7: Matched-input responses on ETTh1.

Additional Analyses. The appendix reports mechanistic diagnostics in §B.7, reference sensitivity in §B.8, intervention side effects in §B.9, computational cost in §B.10, and limitations in §B.11. Together, these analyses characterize the robustness and scope of the main findings.

5 CONCLUSION

Across nine TSFMs and thirteen controlled laws, *Dynamical Parameters* are typically linearly accessible even when matched-input forecast responses are weak or reversed. Projected hidden-state edits recover the reference responses with low error, showing that the selected hidden states can support forecast changes not elicited by the corresponding input interventions. Input-induced changes align weakly with the required directions, while deployed edits constructed from predicted coordinates recover these responses only partially. The analysis therefore distinguishes what a TSFM represents, what its forecast expresses, and what its hidden states can produce under intervention.

ACKNOWLEDGMENTS

The research reported in this paper was sponsored in part by the DEVCOM Army Research Laboratory (award # W911NF1720196), the National Science Foundation (award # 2325956), and the National Institutes of Health (award # 1P41EB028242). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

REFERENCES

- Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What Learning Algorithm Is in-Context Learning? Investigations with Linear Models. In *International Conference on Learning Representations (ICLR)*, 2023.
- A. J. Ångström. New Method of Determining the Thermal Conductibility of Bodies. *The London Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 25(166):130–142, 1863.
- Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*, 2024.
- Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From Univariate to Universal Forecasting. *arXiv preprint arXiv:2510.15821*, 2025.
- Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as Statisticians: Provable in-Context Learning with in-Context Algorithm Selection. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- Anthony Bao, Venkata Hasith Vattikuti, Jeffrey Lai, and William Gilpin. Universal Redundancies in Time Series Foundation Models. *arXiv preprint arXiv:2602.01605*, 2026.
- H. S. Carslaw and J. C. Jaeger. *Conduction of Heat in Solids*. Clarendon Press, Oxford, 2nd edition, 1959.
- Alexander Chemeris, Ming Jin, and Randall Balestriero. Aionoscope: Debugging Latent-State Accessibility in Time-Series Representations. *arXiv preprint arXiv:2607.00956*, 2026.
- Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean Ogier du Terrail, Anna-Monica Toon, Kan Wang, Stephan Xie, Zongzhe Xu, Viktoriya Zhukova, David Asker, Ameet Talwalkar, and Othmane Abou-Amal. This Time Is Different: An Observability Perspective on Time Series Foundation Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.
- Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A Decoder-Only Foundation Model for Time-Series Forecasting. In *International Conference on Machine Learning (ICML)*, 2024.

-
- Zhe Du, Haldun Balim, Samet Oymak, and Necmiye Ozay. Can Transformers Learn Optimal Filtering for Unknown Systems? *IEEE Control Systems Letters*, 7:3525–3530, 2023.
- James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2nd edition, 2012.
- Vijay Ekambaram, Arindam Jati, Pankaj Dayama, Sumanta Mukherjee, Nam H. Nguyen, Wesley M. Gifford, Chandra Reddy, and Jayant Kalagnanam. Tiny Time Mixers (TTMs): Fast Pre-Trained Models for Enhanced Zero/Few-Shot Forecasting of Multivariate Time Series. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
- Ali Forootani and Raffaele Iervolino. Learnable Koopman-Enhanced Transformer-Based Time Series Forecasting with Spectral Control. *arXiv preprint arXiv:2602.02592*, 2026.
- Ali Forootani, Mohammad Khosravi, and Masoud Barati. DeepKoopFormer: A Koopman Enhanced Transformer Based Architecture for Time Series Forecasting. *npj Artificial Intelligence*, 2(35), 2026.
- Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. TimeGPT-1. *arXiv preprint arXiv:2310.03589*, 2023.
- Gautam Goel and Peter Bartlett. Can a Transformer Represent a Kalman Filter? In *Learning for Dynamics and Control Conference (LADC)*, 2024.
- Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A Family of Open Time-Series Foundation Models. In *International Conference on Machine Learning (ICML)*, 2024.
- James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
- Andrew C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1989.
- ITU-T. Technical Features of Push-Button Telephone Sets. Recommendation Q.23, International Telecommunication Union, 1988.
- Ayush Jain and Rajat Sen. TimesFM-3: A Zero-Shot Foundation Model for Multivariate Forecasting. Google Research Blog, August 2026. URL <https://research.google/blog/timesfm-3-a-zero-shot-foundation-model-for-multivariate-forecasting/>.
- Mathis Jander, Wouter van Heeswijk, and Martijn Mes. Causal Analysis for Time Series Foundation Models. *arXiv preprint arXiv:2608.24303*, 2026.
- Rudolf Emil Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.
- Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-Context Transformers for Unified Time Series Forecasting. In *International Conference on Learning Representations (ICLR)*, 2025a.
- Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A Family of Highly Capable Time Series Foundation Models. In *International Conference on Machine Learning (ICML)*, 2025b.
- Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, Upper Saddle River, NJ, 2nd edition, 1999.
- Anurag Mishra. Dissecting Chronos: Sparse Autoencoders Reveal Causal Feature Hierarchies in Time Series Foundation Models. *arXiv preprint arXiv:2603.10071*, 2026.

-
- Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers Can Do Bayesian Inference. In *International Conference on Learning Representations (ICLR)*, 2022.
- Khalid Oublal, Quentin Bouniot, Qi Gan, Stephan Cl  men  on, and Zeynep Akata. TimeSAE: Causal Sparse Decoding for Faithful Explanations of Black-Box Time Series Models. In *International Conference on Machine Learning (ICML)*, 2026.
- Alessandro Pagani, Marco Cominelli, Liying Han, Gaofeng Dong, Sergio Benini, Francesco Gringoli, Mattia Savardi, Mani B. Srivastava, Trevor Bihl, Erik P. Blasch, Daniel O. Bringham, Kara Combs, Lance M. Kaplan, and Federico Cerutti. Preliminary Insights in Chronos Frequency Data Understanding and Reconstruction. In *International Conference on Information Fusion (FUSION)*, 2026.
- Atharva Pandey, Abhilash Neog, and Gautam Jajoo. On the Internal Semantics of Time-Series Foundation Models. *arXiv preprint arXiv:2511.15324*, 2025.
- Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Bilo  , Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. *arXiv preprint arXiv:2310.08278*, 2023.
- Debdeep Sanyal, Aaryan Nagpal, Dhruv Kumar, Murari Mandal, and Saurabh Deshpande. time2time: Causal Intervention in Hidden States to Simulate Rare Events in Time Series Foundation Models. *arXiv preprint arXiv:2509.05801*, 2025.
- 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 (ICLR)*, 2025.
- Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, Jo  o Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers Learn in-Context by Gradient Descent. In *International Conference on Machine Learning (ICML)*, 2023.
- Micha  l Wili  ski, Mononito Goswami, Willa Potosnak, Nina   ukowska, and Artur Dubrawski. Exploring Representations and Interventions in Time Series Foundation Models. In *International Conference on Machine Learning (ICML)*, 2025.
- Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified Training of Universal Time Series Forecasting Transformers. In *International Conference on Machine Learning (ICML)*, 2024.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of in-Context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations (ICLR)*, 2022.
- 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. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.
- Xinxing Zhou, Qingren Yao, Yiji Zhao, Chenghao Liu, Flora Salim, Xiaojie Yuan, Yanlong Wen, and Ming Jin. EIDOS: Latent-Space Predictive Learning for Time Series Foundation Models. *arXiv preprint arXiv:2602.14024*, 2026.

APPENDIX CONTENTS

A	Implementation Details	14
A.1	Models and Forecast Outputs	14
A.2	Controlled Dynamical Laws	14
A.3	Representation Accessibility Implementation	18
A.4	Forecast Response Implementation	20
A.5	Causal Geometry Implementation	22
B	Detailed Dynamical-Parameter Results	25
B.1	Dynamical-Law Identification	25
B.2	Single-Parameter Representation Accessibility	27
B.3	Single-Parameter Forecast Response	29
B.4	Single-Parameter Causal Geometry	33
B.5	Multi-Parameter Laws	43
B.6	Additional Results on Real-Series Backgrounds	50
B.7	Cross-Measure and Mechanistic Diagnostics	52
B.8	Conditional-Reference Sensitivity	57
B.9	Intervention Side Effects and Output Semantics	59
B.10	Computational Cost and Operational Boundary	60
B.11	Limitations and Future Work	60

Table 1: Model checkpoints and configurations used throughout the evaluation.

Model	Checkpoint	Params	L	d	T	Architecture
Chronos-2 (Ansari et al., 2025)	amazon/chronos-2	119.5M	12	768	256	Patched encoder
TimesFM 2.5 (Das et al., 2024)	google/timesfm-2.5-200m-pytorch	231.3M	20	1280	256	Decoder-only
TimesFM 3 (Jain & Sen, 2026)	google/timesfm-3.0-pytorch	330.7M	20	1280	256	Decoder-only
Toto-Open-Base 1.0 (Cohen et al., 2025)	Datadog/Toto-Open-Base-1.0	151.3M	12	768	256	Decoder-only
Moirai (Woo et al., 2024)	Salesforce/moirai-1.1-R-small	13.8M	6	384	256	Masked encoder
Sundial (Liu et al., 2025b)	thuml/sundial-base-128m	128.3M	12	768	256	Decoder-only
Time-MoE (Shi et al., 2025)	Maple728/TimeMoE-200M	453.2M	12	768	256	Decoder-only, MoE
TTM-R2 (Ekambaram et al., 2024)	ibm-granite/granite-timeseries-ttm-r2	805K	3	192	512	MLP-mixer
Timer-XL (Liu et al., 2025a)	thuml/timer-base-84m	84.1M	8	1024	288	Decoder-only

Table 2: Forecast heads and sampling settings.

Model	Forecast head	Paths	Statistic	Seeding
Chronos-2	21-quantile head	–	Median quantile	–
TimesFM 2.5	Flip-averaged pinball head	–	Median quantile	–
TimesFM 3	Quantile head	–	Median quantile	–
Toto-Open-Base 1.0	Autoregressive mean	–	Distribution mean	–
Moirai	Mixture head	64	Path mean	Shared across arms
Sundial	Flow matching, 50 steps	64	Path mean	Shared across arms
Time-MoE	Greedy regression	–	Point output	–
TTM-R2	MLP-mixer	–	Point output	–
Timer-XL	Greedy regression	–	Point output	–

A IMPLEMENTATION DETAILS

This appendix specifies the model checkpoints, controlled generators, reference computations, data splits, selection procedures, and intervention settings. The organization follows representation accessibility, forecast response, and causal geometry.

A.1 MODELS AND FORECAST OUTPUTS

Checkpoints and Architectures. Table 1 reports the checkpoint, parameter count, depth L , hidden width d , context length T , and architecture of each TSFM. Parameter counts follow the published Hugging Face safetensors metadata, while L , d , and T follow our evaluation configurations. The models span patched and masked encoders, decoder-only Transformers, a mixture-of-experts model, and an MLP-mixer, ranging from approximately 0.8M to 453M parameters and three to twenty layers. Time-MoE’s count includes all experts, with two active per token. Context length is $T = 256$ for seven models, $T = 288$ for Timer-XL, and $T = 512$ for TTM-R2.

Forecast Heads and Monte Carlo Sampling. Table 2 summarizes the forecast statistic used for each model. For stochastic backends, we average 64 sampled paths and use the same seed across matched arms to reduce Monte Carlo variance.

A.2 CONTROLLED DYNAMICAL LAWS

The thirteen controlled laws instantiate the state-space framework of §3.1. Each sequence contains a history of length T and a future of length H . The first ten laws each have one designated *Dynamical Parameter*, while the remaining three contain multiple parameters that are varied individually. All other quantities are treated as nuisance variables and held fixed within each matched pair.

A.2.1 GENERAL STATE-SPACE FORM

Each law is represented by a latent state s_t , input u_t , and scalar clean trajectory f_t :

$$s_t = As_{t-1} + Bu_t + G\eta_t, \quad f_t = h(s_t),$$

where η_t denotes process noise. Eleven laws use the linear readout $h(s_t) = Cs_t$. Saturating growth and transient pulse instead apply fixed nonlinear readouts to the clock state $s_t = s_{t-1} + 1$. The designated parameters enter the transition, input, or readout, as summarized in Table 3.

State-space coordinates are not unique because the transformation $s_t \mapsto Ms_t$ preserves the observations when the transition and readout are transformed accordingly. We therefore define the

Table 3: Controlled laws in the common state-space form.

Law	Code	dim s_t	Process noise	Input u_t	Readout h	Parameter location
Linear trend	TR	1	–	Constant	Linear	Input drift S
Oscillation	OS	2	–	–	Linear	Transition angle ω
Mean reversion	MR	1	✓	Constant	Linear	Transition eigenvalue ρ
Random walk with drift	RW	1	✓	Constant	Linear	Input drift d
Exponential relaxation	RX	1	–	Constant	Linear	Transition eigenvalue $e^{-\gamma}$
Saturating growth	SG	1	–	Constant	Logistic	Readout growth rate r
Level shift	LS	1	–	Impulse at τ	Linear	Input gain Δ
Damped oscillation	DO	2	–	–	Linear	Transition modulus $e^{-\gamma}$
Seasonal autoregression	SA	P	✓	Constant	Linear	Lag- P coefficient ρ_s
Transient pulse	TP	1	–	Constant	Gaussian	Readout width w
Dual tone	DT	4	–	–	Linear	Angles ω_L, ω_H , ratio κ
Linear state space	SS	2	✓	Constant	Linear	Modulus r , angle θ
Forced diffusion	FD	Distributed	–	Periodic	Linear	Frequency ω , diffusion operator Λ

Table 4: Dimensionless *Dynamical Parameters* and law-identification ranges.

#	Law	Code	Physical	Coordinate $z_{m,j}$	Range
1	Linear Trend	TR	S	$\tilde{S} = \text{sign}(S) \sigma_{\text{clean}} / \sigma_\varepsilon$	$\pm [1, 6]$
2	Oscillation	OS	ω	$C = \omega T / 2\pi$	$[2, 8]$
3	Mean Reversion	MR	ρ	ρ	$[0.3, 0.9]$
4	Random Walk with Drift	RW	d	$D = d\sqrt{T} / \sigma_\eta$	$\pm [0.5, 3]$
5	Exponential Relaxation	RX	γ	$g = \gamma T$	$[0.8, 2.5]$
6	Saturating Growth	SG	r	$R = rT$	$[5, 12]$
7	Level Shift	LS	Δ	$\tilde{\Delta} = \text{sign}(\Delta) \sigma_{\text{clean}} / \sigma_\varepsilon$	$\pm [1, 6]$
8	Damped Oscillation	DO	γ	$g = \gamma T$	$[0.6, 2.8]$
9	Seasonal Autoregression	SA	ρ_s	ρ_s	$[0.3, 0.9]$
10	Transient Pulse	TP	w	w/T	$[0.04, 0.2]$
11	Dual Tone	DT (TL, TH, TW)	$\omega_L, \omega_H, A_H/A_L$	C_L, C_H, κ	$[2, 8], [10, 26], [0.4, 2.5]$
12	Linear State Space	SS (SR, SC)	$ \lambda(A) , \arg \lambda(A)$	r, C_m	$[0.85, 0.98], [12, 32]$
13	Forced Diffusion	FD (WC, WD)	$\omega, z^2/\kappa_d$	C_d, Λ	$[2, 8], [0.002, 0.1]$

designated parameters through the generator dynamics or observed trajectory rather than arbitrary state coordinates. Examples include eigenvalue moduli and angles, forcing frequencies and gains, and readout-shape parameters.

A.2.2 INDIVIDUAL DYNAMICAL LAWS

Table 4 summarizes the parameter coordinates and law-identification ranges, followed by the generator definitions. Parameters of each multi-parameter law are sampled independently. For accessibility, we use $\rho, \rho_s \in [0.02, 0.98]$, $R \in [3, 13]$, $g \in [0.1, 3]$ for damped oscillation, and $w/T \in [0.025, 0.2]$ for transient pulse. Exponential relaxation uses the same range for law identification and accessibility.

1. Linear Trend (TR).

$$x_t = b + St + \varepsilon_t, \quad z_{m,1} = \tilde{S} = \text{sign}(S) \frac{\sigma_{\text{clean}}}{\sigma_\varepsilon}.$$

Here, b is the baseline, S is the slope, and ε_t is observation noise with standard deviation σ_ε . The quantity σ_{clean} is the standard deviation of the noiseless history, so $\tilde{S}$ measures signed trend strength relative to observation noise.

2. Oscillation (OS).

$$x_t = b + A \sin(\omega t + \phi) + \varepsilon_t, \quad z_{m,1} = C = \frac{\omega T}{2\pi}.$$

Here, b is the baseline, A the amplitude, ω the angular frequency, ϕ the phase, and ε_t observation noise. The designated parameter C is the number of cycles within the observed context. Amplitude, phase, and baseline are nuisance variables.

3. Mean Reversion (MR).

$$s_t = \mu + \rho(s_{t-1} - \mu) + \eta_t, \quad x_t = s_t + \varepsilon_t, \quad z_{m,1} = \rho.$$

Here, μ is the long-run mean, ρ controls persistence, η_t is process noise, and ε_t is observation noise. The latent state is initialized from its stationary distribution. Larger ρ produces slower reversion toward μ and longer-lasting effects on future values.

4. Random Walk with Drift (RW).

$$s_t = s_{t-1} + d + \eta_t, \quad x_t = s_t + \varepsilon_t, \quad z_{m,1} = D = \frac{d\sqrt{T}}{\sigma_\eta}.$$

Here, d is the drift, η_t is process noise with standard deviation σ_η , and ε_t is observation noise. The designated parameter D measures cumulative drift relative to diffusion over the observed context. Unlike linear trend, process innovations accumulate through time. This law is included in dynamical-law identification but excluded from parameter-level analyses because its conditional reference is unstable under the evaluated finite-history sampling regime.

5. Exponential Relaxation (RX).

$$x_t = b + A \exp(-\gamma t) + \varepsilon_t, \quad z_{m,1} = g = \gamma T.$$

Here, b is the equilibrium level, A the initial displacement, γ the decay rate, and ε_t observation noise. The designated parameter g expresses the decay rate relative to the context length, with larger values producing faster relaxation. The sampled range excludes trajectories that appear nearly linear or reach equilibrium before the forecast origin.

6. Saturating Growth (SG).

$$x_t = b + \frac{A}{1 + \exp[-r(t - \tau)]} + \varepsilon_t, \quad z_{m,1} = R = rT.$$

Here, b is the initial level, A the total increase, r the growth rate, τ the transition midpoint, and ε_t observation noise. The designated parameter R expresses growth rate relative to context length, with larger values producing a sharper transition. The independently sampled midpoint τ remains fixed within each matched intervention pair.

7. Level Shift (LS).

$$x_t = b + \Delta \mathbf{1}[t \geq \tau] + \varepsilon_t, \quad z_{m,1} = \tilde{\Delta} = \text{sign}(\Delta) \frac{\sigma_{\text{clean}}}{\sigma_\varepsilon}.$$

Here, b is the initial level, Δ the signed shift magnitude, τ the change point, and ε_t observation noise. The change point lies within the observed context so that both levels are visible and remain distinguishable despite observation noise. After history standardization, $\tilde{\Delta}$ represents signed shift strength relative to observation noise rather than raw jump magnitude. This law contributes only to dynamical-law identification, with no parameter-level scores reported.

8. Damped Oscillation (DO).

$$x_t = b + A \exp(-\gamma t) \sin(\omega t + \phi) + \varepsilon_t, \quad z_{m,1} = g = \gamma T.$$

Here, b is the baseline, A the amplitude, γ the decay rate, ω the angular frequency, ϕ the phase, and ε_t observation noise. The designated parameter g expresses envelope decay relative to context length, with larger values producing faster damping. Frequency, phase, amplitude, and baseline remain fixed within each matched intervention pair.

9. Seasonal Autoregression (SA).

$$s_t = \mu + \rho_s(s_{t-P} - \mu) + \eta_t, \quad x_t = s_t + \varepsilon_t, \quad z_{m,1} = \rho_s.$$

Here, μ is the long-run mean, P the seasonal period, η_t process noise, and ε_t observation noise. The designated parameter ρ_s controls dependence on the state one period earlier, with larger values producing stronger seasonal persistence. The period P is selected so that several repetitions occur within the context and remains fixed within each matched pair.

10. Transient Pulse (TP).

$$x_t = b + A \exp\left[-\frac{(t - \tau)^2}{2w^2}\right] + \varepsilon_t, \quad z_{m,1} = \frac{w}{T}.$$

Here, b is the baseline, A the pulse amplitude, τ the pulse center, w the width, and ε_t observation noise. The designated parameter w/T measures pulse duration relative to context length, with larger values producing broader transients. The center τ lies near the history-to-forecast boundary and remains fixed within each matched pair.

Multi-Parameter Laws. The final three laws contain multiple parameters that vary jointly across histories. Each matched intervention changes one designated parameter while holding the remaining parameters, phases, and stochastic realization fixed.

11. Dual Tone (DT) — Designated Parameter Codes: TL, TH, and TW.

$$x_t = b + A[\cos(\omega_L t + \phi_L) + \kappa \cos(\omega_H t + \phi_H)] + \varepsilon_t, \quad z_m = (C_L, C_H, \kappa),$$

where $C_L = \omega_L T/2\pi$ and $C_H = \omega_H T/2\pi$ are the numbers of low- and high-frequency cycles per context. The parameter codes are low-tone frequency C_L (TL), high-tone frequency C_H (TH), and amplitude ratio κ (TW). Here, b is the baseline, A the low-tone amplitude, ϕ_L and ϕ_H the phases, and ε_t observation noise. Following dual-tone multi-frequency signaling (ITU-T, 1988), the frequencies are drawn from disjoint, resolvable bands. The non-designated frequencies, phases, amplitude, and baseline remain fixed within each matched pair.

12. Linear State Space (SS) — Designated Parameter Codes: SR and SC.

$$s_t = A s_{t-1} + B \eta_t, \quad x_t = C s_t + \varepsilon_t, \quad z_m = (r, C_m),$$

where A has eigenvalues $r e^{\pm i\theta}$ and $C_m = \theta T/2\pi$. The parameter codes are modal persistence r (SR) and modal frequency C_m (SC). Here, A governs state evolution, B loads the process noise η_t , C maps the state to the observation, and ε_t is observation noise. The observed component satisfies

$$v_t = 2r \cos(\theta) v_{t-1} - r^2 v_{t-2} + \eta_t$$

and is initialized from its exact stationary distribution. The designated parameters are invariant to changes in state coordinates. The sampled range avoids regimes where persistence and frequency are difficult to distinguish. The non-designated parameter remains fixed within each matched pair.

13. Forced Diffusion (FD) — Designated Parameter Codes: WC and WD.

The forced-diffusion law is the periodic solution of the one-dimensional heat equation in a semi-infinite medium (Carslaw & Jaeger, 1959):

$$x_t = b + A \sum_{k \in \{1,3,5\}} \frac{e^{-\zeta \sqrt{k}}}{k} \times \sin(k(\omega t + \phi) - \zeta \sqrt{k}) + \varepsilon_t, \\ z_m = (C_d, \Lambda).$$

The parameter codes are drive frequency $C_d = \omega T/2\pi$ (WC) and normalized diffusion time $\Lambda = z^2/(\kappa_d T)$ (WD), with $\zeta = \sqrt{\pi C_d \Lambda}$. Here, b is the baseline, A the drive amplitude, ϕ the phase, and ε_t observation noise. The variable z denotes physical depth and κ_d thermal diffusivity, distinct from the parameter vector z_m . The first three odd harmonics approximate the square-wave surface drive across the observed and forecast horizons. Diffusion attenuates and delays harmonic k according to $\zeta \sqrt{k}$, allowing Λ to be identified from waveform dispersion. This construction follows Ångström’s periodic method for measuring thermal diffusivity (Ångström, 1863). Drive phase, amplitude, and baseline remain fixed within each matched pair throughout drive-frequency and diffusion-time interventions.

A.2.3 SEQUENCE STANDARDIZATION AND CONTEXT-LENGTH MATCHING

Dataset Scope. The following construction applies to law identification and the parameter-accessibility datasets. Trend and oscillation use separate raw-scale priors and fixed noise settings, while the response experiments use separate parameter ranges. Matched shape-law interventions share the baseline arm’s normalization statistics across both arms.

History Standardization. Let $\mu_{1:T}$ and $\sigma_{1:T}$ denote the mean and standard deviation of the clean trajectory f_m over the observed history. For parameter z and nuisance variables ν , the standardized trajectory used to construct the observed sequence is

$$\tilde{f}_m(t, z, \nu) = \frac{f_m(t, z, \nu) - \mu_{1:T}}{\sigma_{1:T}}.$$

Table 5: Generator constraints and scoring coverage. LID denotes law identification, ACC representation accessibility, RSP forecast response, and GEO causal geometry.

Law	Construction criterion	LID	ACC	RSP	GEO
Linear trend (TR)	Both slope signs remain visible above observation noise.	✓	✓	✓	✓
Oscillation (OS)	At least two unaliased cycles occur.	✓	✓	✓	✓
Mean reversion (MR)	Persistence remains separated from white noise and a unit root.	✓	✓	✓	✓
Random walk with drift (RW)	Drift remains visible without dominating process innovations.	✓	–	–	–
Exponential relaxation (RX)	Relaxation remains incomplete near the forecast origin.	✓	✓	–	–
Saturating growth (SG)	The inflection lies within the context before saturation.	✓	✓	✓	✓
Level shift (LS)	Observations occur on both sides of the change point.	✓	–	–	–
Damped oscillation (DO)	Several cycles remain visible under the decaying envelope.	✓	✓	✓	✓
Seasonal autoregression (SA)	Several repetitions occur for every sampled period.	✓	✓	✓	✓
Transient pulse (TP)	The pulse reaches the forecast window without becoming impulsive.	✓	✓	✓	✓
Dual tone (DT)	Both tones remain visible, resolvable, and unaliased.	✓	✓	✓	✓
Linear state space (SS)	The process remains stationary with an identifiable spectral mode.	✓	✓	✓	✓
Forced diffusion (FD)	Surviving harmonics reveal frequency-dependent attenuation and delay.	✓	✓	✓	✓

Each clean history has zero mean and unit variance before scale, offset, and observation noise are applied. Matched response pairs use the same baseline normalization statistics.

Common Scale and Noise. For law identification, the scale variable V and nominal signal-to-noise ratio SNR follow common distributions across laws. The observed sequence is

$$x_t = \text{scale} \tilde{f}_m(t) + c + \sigma_\varepsilon \varepsilon_t, \quad \sigma_\varepsilon = \frac{V}{\sqrt{1 + \text{SNR}^2}}, \quad \text{scale} = \text{SNR} \sigma_\varepsilon.$$

The offset c is proportional to V . This construction matches the offset, scale, and nominal SNR distributions across laws.

Context-Length Matching. Dimensionless parameters are defined relative to T when needed, giving them comparable meanings across context lengths. Because longer noisy histories provide more evidence, the sampled signal-to-noise ratio is multiplied by

$$\sqrt{\frac{T_{\text{ref}}}{T}}, \quad T_{\text{ref}} = 256,$$

which approximately preserves $\text{SNR}^2 T$ across models.

Trend and Level-Shift Parameters. History standardization removes the raw scale of linear trend and level shift, so law identification expresses their strengths relative to observation noise. Trend parameter analyses use separate slope and noise settings. Level shift is included only in law identification.

A.2.4 SCORING COVERAGE AND GENERATOR CONSTRAINTS

Parameter ranges keep the laws distinguishable, avoid degenerate regimes, and ensure that each designated parameter affects the forecast horizon. All thirteen laws enter dynamical-law identification. Random walk with drift lacks a stable parameter reference, while level shift is included only in law identification. Exponential relaxation has an accessibility score but no response or causal-geometry score because its response reference is unstable. Table 5 summarizes the construction criteria and parameter-level coverage for all thirteen controlled laws.

A.3 REPRESENTATION ACCESSIBILITY IMPLEMENTATION

A.3.1 PROBE TRAINING AND MODEL SELECTION

All TSFM weights remain frozen. Table 6 summarizes the probe targets, penalty grids, and conditional reference correlations. Law-identification probes use the final exposed sequence position. Parameter probes concatenate this position with available positions 1, 2, 4, 8, and 16 steps earlier. This yields three positions for Toto-Open-Base 1.0 and Timer-XL, four for TimesFM 2.5 and TTM-R2, five for TimesFM 3 and Sundial, and six for Chronos-2, Moirai, and Time-MoE. Features are standardized using training statistics before fitting any layer-specific probe models.

At each layer, the penalty maximizing validation accuracy or correlation is selected. The best validation layer is then evaluated once on the independent test dataset. Parameter datasets contain

Table 6: Probe targets, penalty grids, and conditional reference correlations.

Task	Probe target	Penalty grid	Reference r^*		
			$T = 256$	$T = 288$	$T = 512$
Law ID	13 balanced classes	$\lambda \in \{10^{-4}, \dots, 10^3\}$	—	—	—
TR	$S \sim \mathcal{N}(0, 0.02^2)$	$\alpha \in \{10^0, \dots, 10^6\}$	0.922	0.930	0.959
OS	$\omega = 2\pi/P$, $\log P \sim \mathcal{U}[\log 2.5, \log 4096]$	$\alpha \in \{10^0, \dots, 10^6\}$	0.988	0.986	0.971
MR	$\rho \sim \mathcal{U}[0.02, 0.98]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.976	0.978	0.984
RX	$g \sim \mathcal{U}[0.8, 2.5]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.899	0.901	0.901
SG	$R \sim \mathcal{U}[3, 13]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.983	0.983	0.983
DO	$g \sim \mathcal{U}[0.1, 3]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.986	0.986	0.986
SA	$\rho_s \sim \mathcal{U}[0.02, 0.98]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.976	0.978	0.984
TP	$\log(w/T) \sim \mathcal{U}[\log 0.025, \log 0.2]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.997	0.997	0.997
TL	$C_L \sim \mathcal{U}[2, 8]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	1.000	1.000	1.000
TH	$C_H \sim \mathcal{U}[10, 26]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	1.000	1.000	1.000
TW	$\log \kappa \sim \mathcal{U}[\log 0.4, \log 2.5]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.991	0.991	0.991
SR	$r \sim \mathcal{U}[0.85, 0.98]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.894	0.904	0.944
SC	$C_m \sim \mathcal{U}[12, 32]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.988	0.987	0.975
WC	$C_d \sim \mathcal{U}[2, 8]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	1.000	1.000	1.000
WD	$\log \Lambda \sim \mathcal{U}[\log 0.002, \log 0.1]$	$\alpha \in \{10^0, \dots, 10^{10}\}$	0.991	0.991	0.991

131,070 training, 16,380 validation, and 32,770 test sequences. Law identification uses 131,040 training, 16,380 validation, and 32,760 test sequences, with 2,520 test sequences per law. Independent random streams generate all three splits.

Seven models use context length $T = 256$, Timer-XL uses $T = 288$, and TTM-R2 uses $T = 512$. Linear trend draws baseline $b \sim \mathcal{N}(0, 10^2)$ and uses observation-noise standard deviations 10, 11.2, and 20, respectively. Oscillation uses amplitude $A = 5$, uniform phase, and noise standard deviations 9.43, 10.14, and 14.71. At $T = 256$, the remaining laws draw $\log V \sim \mathcal{U}[\log 0.5, \log 5]$, $\log \text{SNR} \sim \mathcal{U}[\log 1, \log 6]$, and $c \sim \mathcal{N}(0, V^2)$. For other context lengths, the signal-to-noise ratio is scaled by $\sqrt{256/T}$ to approximately equalize finite-history statistical evidence.

The thirteen law classes are balanced, giving chance accuracy 1/13. For TTM-R2 law identification, extending the penalty grid to 10^{-8} leaves the selected accuracy unchanged.

A.3.2 LAW- AND NUISANCE-CONDITIONED PARAMETER REFERENCE

For law m , parameter $z_{m,j}$, and nuisance variables ν , the conditional posterior estimate and reference correlation used throughout the representation-accessibility analysis below are

$$\hat{z}_{m,j}^* = \mathbb{E}[z_{m,j} \mid x_{1:T}, m, \nu], \quad r_{m,j}^* = \text{corr}(\hat{z}_{m,j}^*, z_{m,j}).$$

The posterior mean $\hat{z}_{m,j}^*$ estimates the parameter from the observed history given the law and documented nuisance variables. The reference $r_{m,j}^*$ is its correlation with the true parameter across sequences and measures finite-history recoverability. Higher values indicate that the observed history identifies the parameter more reliably. Each accessibility score $D_{m,j}(\ell)$ normalizes the probe correlation by the reference computed at the corresponding model context length.

The conditional parameter reference conditions on the sequence scale, offset, noise level, and documented nuisance variables. For multi-parameter laws, it also receives the non-designated parameters and phases. The TSFM receives only the observed history, so $r_{m,j}^*$ provides an informed reference for conditional parameter recoverability under the corresponding controlled generator distribution.

The trend posterior mean is computed in closed form. Other single-parameter references use generator likelihoods evaluated over 257 parameter values. The grids increase to 513 values for low-tone frequency (TL), high-tone frequency (TH), and drive frequency (WC). Oscillation uses a grid of 2,048 frequencies and 256 phases. For linear state space, the stationary Gaussian second-order autoregression supplies the likelihood. The non-designated parameter uses a 33-point lattice, while the designated parameter is integrated over its evaluation grid.

A.4 FORECAST RESPONSE IMPLEMENTATION

A.4.1 MATCHED PARAMETER INTERVENTIONS

For each law m and designated parameter $z_{m,j}$, a baseline vector z_m^A is paired with an intervention vector z_m^B . The two vectors differ only in the designated coordinate, while all nuisance quantities, initial conditions, and generator randomness remain fixed. The resulting histories x^A and x^B therefore differ only through the designated parameter.

The frozen TSFM produces the forecast change

$$\Delta \hat{y}_{m,j} = F_\theta(x^B) - F_\theta(x^A),$$

while the corresponding future trajectories define the reference response

$$\Delta y_{m,j}^{\text{ref}} = y^B - y^A.$$

Comparing these quantities measures whether the forecast expresses the expected consequence of the parameter change in direction, magnitude, and temporal structure.

Mean reversion, seasonal autoregression, and linear state space contain future process innovations that no forecaster can anticipate. For these laws, $\Delta y_{m,j}^{\text{ref}}$ is computed from conditional-mean futures by setting post-origin innovations to zero in both arms. The two histories are processed independently, while stochastic forecasting backends use the same sampling seed across arms to reduce Monte Carlo variation without changing the underlying model predictions.

A.4.2 INTERVENTION MAGNITUDES

[Table 7](#) summarizes the parameter ranges, signed intervention magnitudes, and conditional reference responses. Baseline values are sampled so that both arms remain within the specified range. Linear trend uses changes of ± 0.02 and ± 0.05 , while oscillation uses ± 1 , ± 2 , and ± 4 cycles per context. Other parameters use $\pm 2\%$ and $\pm 5\%$ of their sampled ranges. Transient pulse (TP), amplitude ratio (TW), and diffusion time (WD) use the corresponding logarithmic ranges. For oscillation, every intervention magnitude uses the same baseline-frequency distribution.

The reported response score averages the model and conditional-reference coefficients separately across the magnitude set $\mathcal{M}_{m,j}$ before computing their final normalized ratio:

$$\bar{U}_{m,j} = \frac{\sum_{\delta \in \mathcal{M}_{m,j}} U_{m,j}(\delta)}{\sum_{\delta \in \mathcal{M}_{m,j}} U_{m,j}^*(\delta)}.$$

Each magnitude generally uses 32,768 matched pairs. Moirai uses 131,072 pairs for each single-parameter response law, while Timer-XL uses 131,072 pairs for oscillation. These larger datasets reduce sensitivity to heavy-tailed per-pair coefficients.

A.4.3 CONDITIONAL POSTERIOR-PREDICTIVE RESPONSE REFERENCE

A finite noisy history may not fully determine the future response to a parameter change. We therefore define the conditional posterior-predictive forecast

$$F_m^*(x, \nu) = \mathbb{E}[y \mid x, m, \nu],$$

where m is the known law and ν contains the documented nuisance variables. For the two matched histories, the conditional forecast change is

$$\Delta y_{m,j}^* = F_m^*(x^B, \nu) - F_m^*(x^A, \nu).$$

This change is scored against the same generator response and over the same matched pairs as the TSFM forecast change, yielding $U_{m,j}^*(\delta)$. The normalized response across intervention magnitudes is

$$\bar{U}_{m,j} = \frac{\sum_{\delta \in \mathcal{M}_{m,j}} U_{m,j}(\delta)}{\sum_{\delta \in \mathcal{M}_{m,j}} U_{m,j}^*(\delta)}.$$

The conditional response reference therefore corresponds to $\bar{U}_{m,j} = 1$ by construction.

Table 7: Intervention settings and mean conditional response references.

Law	Parameter	Range	Signed change	Mean reference $\bar{U}_{m,j}^*$		
				$T = 256$	$T = 288$	$T = 512$
Linear trend (TR)	Slope	$S \in [-0.45, 0.45]$	$\pm 0.02, \pm 0.05$	0.995	0.995	0.997
Oscillation (OS)	Frequency	$C \in [0.0625, 102.4]$	$\pm 1, \pm 2, \pm 4$	0.870	0.869	0.833
Mean reversion (MR)	Persistence	$\rho \in [0.02, 0.98]$	$\pm 2\%, \pm 5\%$	0.760	0.766	0.755
Saturating growth (SG)	Growth rate	$R \in [3, 13]$	$\pm 2\%, \pm 5\%$	0.945	0.950	0.952
Damped oscillation (DO)	Decay rate	$g \in [0.1, 3]$	$\pm 2\%, \pm 5\%$	0.954	0.954	0.954
Seasonal autoregression (SA)	Persistence	$\rho_s \in [0.02, 0.98]$	$\pm 2\%, \pm 5\%$	0.845	0.841	0.807
Transient pulse (TP)	Pulse width	$w/T \in [0.025, 0.2]$	$\pm 2\%, \pm 5\%$	0.982	0.981	0.979
Dual tone (DT)	Low-tone frequency (TL)	$C_L \in [2, 8]$	$\pm 2\%, \pm 5\%$	1.000	0.999	0.999
	High-tone frequency (TH)	$C_H \in [10, 26]$	$\pm 2\%, \pm 5\%$	1.000	1.000	1.000
	Amplitude ratio (TW)	$\kappa \in [0.4, 2.5]$	$\pm 2\%, \pm 5\%$	0.972	0.972	0.972
Linear state space (SS)	Modulus (SR)	$r \in [0.85, 0.98]$	$\pm 2\%, \pm 5\%$	0.819	0.820	0.824
	Modal frequency (SC)	$C_m \in [12, 32]$	$\pm 2\%, \pm 5\%$	0.866	0.858	0.831
Forced diffusion (FD)	Drive frequency (WC)	$C_d \in [2, 8]$	$\pm 2\%, \pm 5\%$	1.000	1.000	1.000
	Diffusion time (WD)	$\Lambda \in [0.002, 0.1]$	$\pm 2\%, \pm 5\%$	0.957	0.957	0.956

The conditional response reference conditions on each sequence’s scale, offset, noise level, and documented nuisance variables. For multi-parameter laws, it also receives the non-designated parameters and phases. For mean reversion (MR), seasonal autoregression (SA), and the linear state-space parameters eigenvalue modulus (SR) and modal frequency (SC), it receives the standard deviation used to normalize the latent shape. For saturating growth (SG), transient pulse (TP), and damped oscillation (DO), it receives the inflection time, pulse center, and oscillation frequency and phase, respectively.

Each candidate shape is normalized using the standard deviation implied by its candidate parameter q before comparison with either observed history. Because the intervened history retains the baseline history’s normalization scale, a candidate for the intervention arm uses the corresponding baseline parameter $I^{-1}(q, \delta)$, where I is the intervention map. This inverse is $q - \delta$ for additive changes and $q \exp(-\delta_{\log})$ for logarithmic changes. The resulting coefficient is a nuisance-informed normalization rather than a strict upper bound on signed aligned gain. A model can exceed this coefficient while producing greater full-vector error. For damped oscillation, additionally inferring frequency and phase changes $U_{m,j}^*$ by less than 0.005 at every magnitude.

Trend-Slope Reference. For the centered trend system

$$x_t = M + (t - c)S + \varepsilon_t, \quad c = \frac{T + 1}{2},$$

M is the level at the context midpoint, S is the slope, and c centers the time indices. Define the observation vector $\mathbf{x} = (x_1, \dots, x_T)^\top$ and centered time vector $\mathbf{g}$, with $g_t = t - c$. Because $\sum_t g_t = 0$, the slope can be estimated independently of the unknown level M . Its least-squares sufficient statistic under this centered finite-history regression model is

$$\hat{S} = \frac{\mathbf{g}^\top \mathbf{x}}{\mathbf{g}^\top \mathbf{g}}, \quad \mathbf{g}^\top \mathbf{g} = \sum_{t=1}^T (t - c)^2.$$

For independent observation noise $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, the estimator satisfies

$$\hat{S} \mid S \sim \mathcal{N}\left(S, \sigma_{\hat{S}}^2\right), \quad \sigma_{\hat{S}} = \frac{\sigma_\varepsilon}{\sqrt{\mathbf{g}^\top \mathbf{g}}}.$$

Thus, $\sigma_{\hat{S}}$ quantifies the uncertainty in the slope estimate induced by finite context length and observation noise for the assumed independent Gaussian observations.

Under the uniform prior $S \in [-s_{\max}, s_{\max}]$, the Gaussian likelihood is truncated at the prior boundaries. Define the standardized lower and upper boundaries

$$a_- = \frac{-s_{\max} - \hat{S}}{\sigma_{\hat{S}}}, \quad a_+ = \frac{s_{\max} - \hat{S}}{\sigma_{\hat{S}}}.$$

The posterior mean is

$$\mathbb{E}[S | \mathbf{x}] = \hat{S} + \sigma_{\hat{S}} \frac{\phi(a_-) - \phi(a_+)}{\Phi(a_+) - \Phi(a_-)},$$

where ϕ and Φ denote the standard normal density and cumulative distribution functions. The correction term pulls estimates near either prior boundary toward the interior of the allowed slope range while leaving interior estimates nearly unchanged.

A matched slope intervention changes S by δ while preserving the history mean. Because the centered statistic changes by the same amount, $\hat{S}^B = \hat{S}^A + \delta$. The conditional-reference coefficient for the corresponding matched slope intervention is therefore

$$U_{\text{TR}}^*(\delta) = \frac{\mathbb{E}[S | \mathbf{x}^B] - \mathbb{E}[S | \mathbf{x}^A]}{\delta}.$$

A coefficient near one indicates that the posterior mean preserves the imposed slope change. At $T = 256$ and $s_{\max} = 0.45$, the average coefficient is 1.000 when $|S^A| < 0.20$ and ranges from 0.986 to 0.994 at larger baseline magnitudes because of truncation near the prior boundaries.

References Across Context Lengths. At a fixed context length T , the matched-pair distribution and conditional response reference do not depend on the TSFM. The seven models using $T = 256$ therefore share the same conditional reference at each intervention magnitude. Timer-XL and TTM-R2 use separately computed references at $T = 288$ and $T = 512$, respectively. [Table 7](#) reports the mean reference for each context length under the corresponding matched intervention distributions.

A.5 CAUSAL GEOMETRY IMPLEMENTATION

A.5.1 FITTING, VALIDATION, AND TEST DATASETS

The causal-geometry analysis uses separate datasets for basis estimation, model selection, and final evaluation. The fitting dataset $\mathcal{D}_{\text{disc}}$ estimates the shared causal basis and fits the coordinate readout. The validation dataset selects admissible layers, subspace ranks, and readout penalties. The selected configuration is evaluated once on an independent test dataset. No test response contributes to basis estimation, readout fitting, or configuration selection.

Each of the 126 model–parameter cells uses 16,384 fitting, 4,096 validation, and 4,096 test matched pairs. The datasets are generated at the model’s context length with horizon $H = 16$ and the intervention magnitude reported in [Table 8](#). Independent random streams are used across the three splits. Generator streams are shared across models for a given parameter, except for Moirai on seasonal autoregression, whose streams are offset by one while preserving the generator and sample sizes. Among admissible layers, selection maximizes the validation projected response over the evaluated ranks. All reported responses are then computed on the test dataset.

A.5.2 LOCAL CAUSAL DIRECTIONS AND STEP CALIBRATION

[Table 8](#) summarizes the parameter changes, calibrated hidden-state steps, and layer-level calibration outcomes for every evaluated model and target.

Regularized Local Directions. The regularized Jacobian inversion in [Equation 10](#) uses

$$\lambda = 10^{-4} \frac{\text{tr}(J_i^\ell J_i^{\ell\top})}{H}.$$

This scaling adapts the regularization strength to the average Jacobian row energy at each input and layer. The direction is computed through the output-space system

$$v_i^\ell = J_i^{\ell\top} (J_i^\ell J_i^{\ell\top} + \lambda I_H)^{-1} \Delta y_i^{\text{ref}}.$$

The inversion targets the complete uncentered reference response. Centering is applied only when the realized nonlinear response is scored using [Equation 11](#).

Intervention-Step Calibration. At each layer, the intervention step α_ℓ is selected on validation data through geometric search followed by log-midpoint bisection. The search begins from the step accepted at the previous layer, or from $\alpha_\ell = 1$ at the first measured layer, and evaluates at most nine

Table 8: Hidden-intervention calibration across parameter targets for every evaluated model and layer. The final two columns report layer counts out of 90.

Law	Parameter	δ	Calibrated step α_ℓ		Layer count	
			Median	Range	$\alpha_\ell < 1$	Converged
Linear trend (TR)	Slope	0.02	0.58	0.021–1.00	86/90	85/90
Oscillation (OS)	Frequency	2	0.19	0.037–0.76	90/90	90/90
Mean reversion (MR)	Persistence	5%	1.00	0.19–1.00	26/90	90/90
Saturating growth (SG)	Growth rate	5%	0.58	0.084–1.00	57/90	90/90
Damped oscillation (DO)	Decay rate	5%	1.00	0.33–1.00	42/90	90/90
Seasonal autoregression (SA)	Persistence	5%	0.11	0.012–1.00	86/90	83/90
Transient pulse (TP)	Pulse width	5%	0.25	0.11–0.76	90/90	90/90
Dual tone (DT)	Low-tone frequency (TL)	5%	1.00	0.76–1.00	19/90	90/90
	High-tone frequency (TH)	5%	1.00	0.22–1.00	34/90	90/90
	Amplitude ratio (TW)	5%	1.00	0.33–1.00	21/90	90/90
Linear state space (SS)	Modulus (SR)	5%	1.00	0.19–1.00	32/90	90/90
	Modal frequency (SC)	5%	0.19	0.064–0.44	90/90	90/90
Forced diffusion (FD)	Drive frequency (WC)	5%	1.00	0.15–1.00	28/90	90/90
	Diffusion time (WD)	5%	0.76	0.25–1.00	49/90	90/90

steps. Seasonal-autoregression runs for Chronos-2, Moirai, Timer-XL, Toto-Open-Base 1.0, and TTM-R2 instead restart from $\alpha_\ell = 1$ at every layer. The selected step minimizes the larger deviation of the pooled and median per-pair local responses from one, with preference given to steps placing both responses within 0.03 of one.

Direction Capping and Layer Selection. Directions whose root-mean-square norm exceeds three times the fitting-dataset median at that layer are rescaled to this threshold. The same factor is applied to the corresponding reference response, preserving the direction–target relationship. A layer is admissible when its pooled and median per-pair local responses lie within $[0.85, 1.05]$ and its projected response does not exceed

$$\max\{1.05, U_{\text{local}}\}$$

on validation data. The median is computed over pairs whose capped reference-response norm is at least 5% of the median norm for that model, target, and layer.

Evaluated Layers. Blocks 1 through L are evaluated for each model, except TimesFM 2.5, for which layers 1, 5, 9, 12, 16, and 20 are evaluated. Time-MoE’s final block is excluded because its exposed hidden state follows the final normalization and cannot be modified through the intervention interface. These choices govern all causal-geometry measurements.

Connecting Hidden Edits to Input-Induced Changes. For

$$\Delta h_i^\ell = h^\ell(x_i^B) - h^\ell(x_i^A),$$

the local approximation

$$\Delta F_i \approx J_i^\ell \Delta h_i^\ell$$

holds with downstream bypass quantities fixed. The actual input intervention can also change the bypass quantities b_i and move the model beyond the local linear regime. The local causal direction v_i^ℓ is constructed to produce the reference response, whereas Δh_i^ℓ is the state change induced by the parameter intervention. Their alignment, row- and null-space energy, and induced gains indicate whether the input change follows a forecast-sensitive hidden-state direction.

A.5.3 SHARED CAUSAL SUBSPACE AND RANK SELECTION

The shared basis and projected directions follow Equation 12. For each layer, the numerical rank r_ℓ counts squared singular values greater than 10^{-10} times the largest. The evaluated ranks are

$$\mathcal{K}_\ell = \{2^0, 2^1, \dots, 2^{10}\} \cap [1, r_\ell].$$

Thus, the largest evaluated rank is 1,024, or the largest admissible power of two when $r_\ell < 1,024$. The compactness statistic $k_{90}(\ell)$ is the smallest evaluated rank that preserves at least 90% of $U_{\text{local}}(\ell)$

on validation data. If no rank reaches the threshold, the largest evaluated rank is reported as a lower bound for that layer and parameter target.

State-Dimension Normalization. Panel-wide comparisons normalize k_{90} by the intervened-state dimension d_h , defined as the hidden width multiplied by the number of modified token positions. The intervention modifies 3 positions for Timer-XL (TX), 4 for Toto-Open-Base 1.0 (TO), 8 for TimesFM 2.5 (F2) and TTM-R2 (TT), 10 for TimesFM 3 (F3), 16 for Sundial (SU), 17 for Moirai (MO), 18 for Chronos-2 (C2), and 256 for Time-MoE (ME). The resulting dimension d_h ranges from 1,536 for TTM-R2 (TT) to 196,608 for Time-MoE (ME). Consequently, a large absolute k_{90} can still occupy a small fraction of a model’s intervened hidden state.

A.5.4 CAUSAL-COORDINATE READOUT AND ORACLE FIXED-DIRECTION CONTROL

Causal-Coordinate Readout. For layer ℓ and rank k , the coordinate readout in Equation 13 is fitted by ridge regression on standardized fitting-dataset representations:

$$\left(\tilde{P}_{\ell,k}, \tilde{b}_{\ell,k}^c\right) = \arg \min_{P,b} \left\{ \sum_{i \in \mathcal{D}_{\text{disc}}} \left\| c_i^{\ell,k} - P D_{\ell}^{-1} (h_i^{\ell} - \mu_{\ell}) - b \right\|_2^2 + \lambda_c \|P\|_F^2 \right\}.$$

Here, $c_i^{\ell,k}$ contains the coordinates of the local causal direction in the rank- k shared basis. The vector μ_{ℓ} and diagonal matrix D_{ℓ} contain the coordinate-wise mean and standard deviation of the fitting representations. The intercept is not penalized.

Transforming the fitted coefficients back to the original representation scale gives

$$P_{\ell,k} = \tilde{P}_{\ell,k} D_{\ell}^{-1}, \quad b_{\ell,k}^c = \tilde{b}_{\ell,k}^c - P_{\ell,k} \mu_{\ell}.$$

Regularization therefore applies to the standardized-feature coefficients $\tilde{P}_{\ell,k}$.

Penalty and Rank Selection. The penalty λ_c is selected independently at each rank from 33 values logarithmically spaced between 10^{-4} and 10^{12} . Selection maximizes validation R^2 pooled across the predicted coordinates. For mean reversion (MR), Timer-XL (TX) and TTM-R2 (TT) use 25 values between 10^{-4} and 10^8 . For seasonal autoregression (SA), Chronos-2 (C2), Timer-XL (TX), Toto-Open-Base 1.0 (TO), and TTM-R2 (TT) use the same reduced grid. Moirai (MO) on seasonal autoregression (SA) uses 11 values between 10^1 and 10^6 .

The readout rank $\hat{k}(\ell)$ is the smallest evaluated rank whose validation deployed response retains at least 90% of the local response. If no evaluated rank reaches this threshold, the largest rank is selected. The test deployed response and projected-response validity condition are then evaluated at the selected rank for each layer and parameter target.

At validation and test time, the deployed direction requires only the input’s hidden representation and quantities estimated from the fitting dataset. It receives neither the input-specific Jacobian nor the reference response. The shared basis, standardization statistics, and readout coefficients use only fitting data, while the penalty and rank use only validation data.

Oracle Fixed-Direction Control. The fixed-direction control is deliberately oracle-assisted. For an evaluation dataset $\mathcal{D}_{\text{eval}}$, its common direction is the mean local causal direction

$$\bar{v}^{\ell} = \frac{1}{|\mathcal{D}_{\text{eval}}|} \sum_{i \in \mathcal{D}_{\text{eval}}} v_i^{\ell}.$$

For each example i , this direction is rescaled to the norm of its oracle local direction:

$$v_i^{\ell, \text{fixed}} = \|v_i^{\ell}\|_2 \frac{\bar{v}^{\ell}}{\|\bar{v}^{\ell}\|_2}.$$

The control therefore receives both the evaluation-dataset mean oracle direction and each example’s oracle perturbation norm. The comparison asks whether an input-dependent readout improves upon a strong input-independent direction despite this oracle assistance. Because the deployed readout predicts direction and magnitude jointly, the comparison does not isolate directional inference from norm prediction. Both components determine its reported performance.

Table 9: Normalized law-emergence depth and first-layer decodability across models.

Model	TR	OS	MR	RW	RX	SG	LS	DO	SA	TP	DT	SS	FD	Mean	$D_m(1)$
Chronos-2 (C2)	0.08	0.08	0.17	0.17	0.08	0.08	0.08	0.08	0.08	0.08	0.17	0.17	0.17	0.12	0.91
TimesFM 2.5 (F2)	0.10	0.10	0.10	0.15	0.05	0.05	0.10	0.05	0.05	0.05	0.10	0.10	0.10	0.08	0.90
TimesFM 3 (F3)	0.20	0.15	0.15	0.20	0.05	0.10	0.15	0.05	0.15	0.15	0.15	0.15	0.20	0.14	0.81
Toto-Open-Base 1.0 (TO)	0.25	0.25	0.33	0.42	0.17	0.08	0.25	0.17	0.17	0.17	0.33	0.33	0.33	0.25	0.75
Moirai (MO)	0.33	0.33	0.33	0.67	0.33	0.33	0.33	0.50	0.33	0.50	0.33	0.33	0.33	0.38	0.55
Sundial (SU)	0.25	0.33	0.25	0.33	0.17	0.08	0.25	0.17	0.08	0.17	0.25	0.25	0.25	0.22	0.72
Time-MoE (ME)	0.25	0.33	0.33	0.50	0.25	0.25	0.33	0.25	0.25	0.25	0.33	0.50	0.33	0.32	0.27
TTM-R2 (TT)	0.67	0.67	0.67	0.67	0.33	0.33	0.67	0.67	0.67	0.67	1.00	0.67	0.67	0.64	0.67
Timer-XL (TX)	0.25	0.25	0.25	0.38	0.25	0.12	0.38	0.25	0.12	0.12	0.25	0.25	0.25	0.24	0.81
Mean	0.26	0.28	0.29	0.39	0.19	0.16	0.28	0.24	0.21	0.24	0.32	0.31	0.29	0.27	

B DETAILED DYNAMICAL-PARAMETER RESULTS

This appendix presents layer-wise and model-level results for law identification, representation accessibility, forecast response, and causal geometry. Single-parameter results are followed by multi-parameter and cross-measure analyses.

Common Protocol. Matched-input analyses use paired datasets spanning signed intervention magnitudes. Causal geometry uses independent fitting, validation, and test datasets. Each forecast horizon is evaluated separately because some forecasting backends are not prefix-consistent.

B.1 DYNAMICAL-LAW IDENTIFICATION

Complementing the validation-selected test results in §4.1, we examine layer-wise validation profiles and test-set confusion patterns. The multinomial probe uses the final exposed sequence position, and the balanced test dataset contains 2,520 sequences per law. Each multi-parameter law is treated as one class, and law decodability scores report normalized above-chance accuracy.

B.1.1 LAYER-WISE DYNAMICAL-LAW DECODABILITY

For each TSFM, we evaluate $D_m(\ell)$ across all layers and thirteen dynamical laws. The layer-wise curves use validation data because layer selection is also performed on the validation dataset. All panels share the same vertical range, allowing direct comparison across models. Table 9 summarizes emergence depth, while Figure 8 presents the complete layer-wise profiles.

For each model and law, emergence depth is the first layer at which $D_m(\ell)$ reaches 95% of its final-layer value, expressed as a fraction of model depth. Table 9 also reports mean first-layer decodability to characterize how early law information appears.

The layer-wise profiles reveal three properties not captured by the validation-selected summary in §4.1 because it reports only one layer.

Laws Become Decodable at Different Depths. Random walk with drift (RW) emerges latest, at 0.39 of model depth on average, compared with 0.16–0.32 for the remaining laws. It is the last or joint-last law to emerge in eight of nine models. For TTM-R2 (TT), whose three layers make the depth measure coarse, dual tone (DT) emerges last. The three multi-parameter laws emerge between 0.29 and 0.32 of model depth on average. These jointly varying laws therefore become linearly distinguishable relatively late, although law classification does not require recovery of every parameter. Saturating growth (SG), exponential relaxation (RX), and seasonal autoregression (SA) emerge earliest on average. The models broadly agree on which laws require greater depth before becoming linearly accessible across their otherwise distinct hidden representations.

Depth Requirements Vary More Than Final Decodability. Mean first-layer decodability ranges from 0.27 for Time-MoE (ME) to 0.91 for Chronos-2 (C2). Chronos-2 (C2) and TimesFM 2.5 (F2) gain only 0.05 and 0.06 across their layers, whereas Time-MoE (ME) gains 0.67. Models with similar final decodability can therefore differ substantially in where law information becomes accessible.

Selected Layers Typically Lie Beyond Saturation. For every model, the validation-selected layer lies at or beyond the point where mean law decodability first reaches 95% of its final-layer value. Mean decodability at the selected layer is within 0.026 of its final-layer value. The selected layer

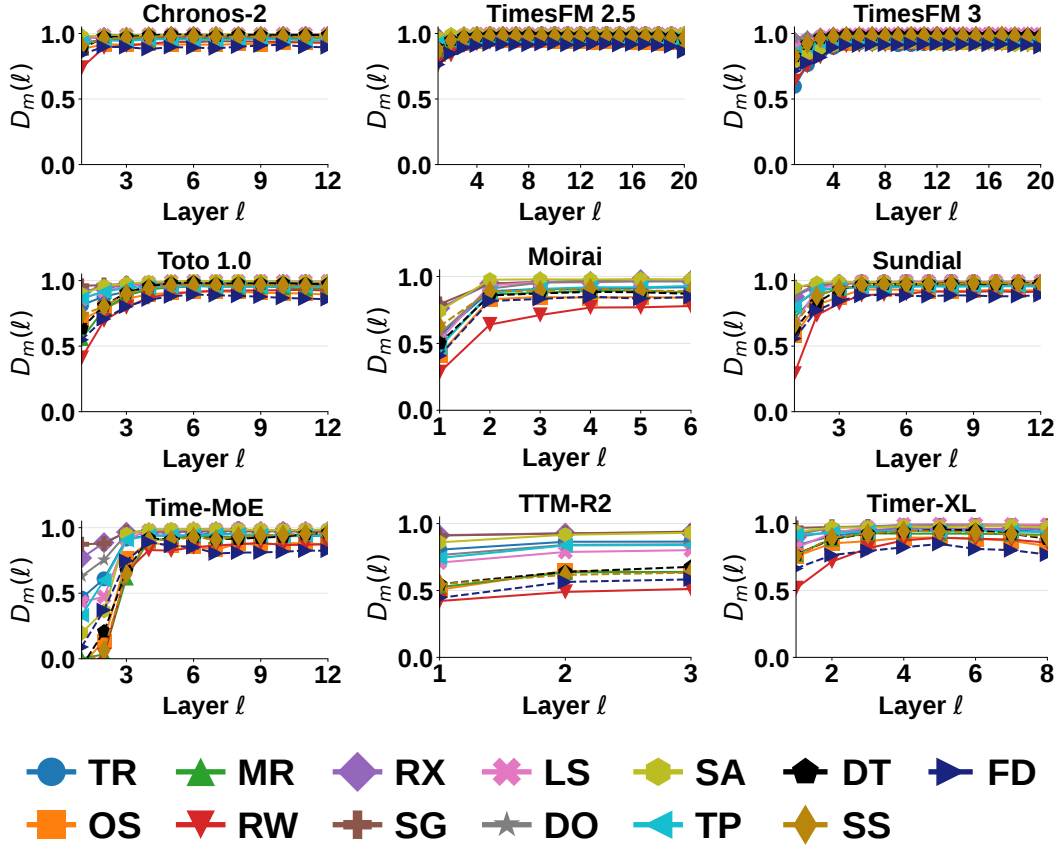


Figure 8: Validation law decodability across model depth. Panels show all nine frozen TSMFs. Dashed curves denote multi-parameter laws.

therefore represents a state supporting strong law identification rather than the precise point at which law identity first becomes accessible.

B.1.2 RANDOM-WALK CONFUSIONS ACROSS DRIFT MAGNITUDES

Random walk with drift (RW) and forced diffusion (FD) have the lowest average identification scores in Figure 1. We examine which laws receive the random-walk errors and whether those errors vary with the amount of drift. This isolates dominant finite-history ambiguity patterns.

Random Walk Is Primarily Confused with Other Drifting Laws. Table 10 reports the classifications assigned to test random-walk sequences. Errors concentrate on laws whose finite histories can depart persistently from their initial levels. Linear trend (TR) receives between 2.1% and 12.6% of these sequences, while exponential relaxation (RX) receives between 1.0% and 6.4%. TTM-R2 (TT) additionally assigns 14.0% to level shift (LS). Linear trend is the largest incorrect classification for seven models. TimesFM 3 (F3) instead assigns marginally more sequences to mean reversion (MR), while TTM-R2 (TT) assigns more to level shift (LS). The dominant confusions therefore remain concentrated among laws with similar finite-history trajectories.

Confusions Increase with Drift Magnitude. Table 11 divides random-walk sequences into three drift-magnitude bins for two models with strong identification and two with weaker identification. Recall decreases as drift increases for every model. It falls from 95.4% to 89.6% for Chronos-2 (C2), from 94.7% to 86.1% for Toto-Open-Base 1.0 (TO), from 83.1% to 59.4% for Moirai (MO), and from 60.3% to 44.8% for TTM-R2 (TT). Most of the lost recall shifts to linear trend (TR) and then exponential relaxation (RX), consistent with a strongly drifting random walk resembling a noisy line over a finite context. TTM-R2’s confusion with level shift (LS) remains between 12% and 15% across drift bins. This persistent confusion is consistent with its level-shift law decodability score of 0.83 in Figure 1, compared with 0.97 or higher for every other model.

Table 10: Test-set classifications of random-walk sequences at each model’s validation-selected layer. Correct classifications are bold.

Model	True random walk classified as (%)													Layer
	TR	OS	MR	RW	RX	SG	LS	DO	SA	TP	DT	SS	FD	
Chronos-2 (C2)	2.3	0.0	1.2	94.0	1.0	0.3	0.3	0.1	0.0	0.6	0.0	0.0	0.0	10/12
TimesFM 2.5 (F2)	2.5	0.0	1.0	94.4	1.1	0.2	0.2	0.0	0.0	0.5	0.0	0.0	0.0	10/20
TimesFM 3 (F3)	2.1	0.0	2.5	92.7	1.3	0.2	0.2	0.2	0.1	0.5	0.0	0.1	0.1	16/20
Toto-Open-Base 1.0 (TO)	2.9	0.0	1.2	92.7	1.5	0.3	0.6	0.2	0.0	0.5	0.0	0.0	0.1	6/12
Moirai (MO)	10.8	0.6	1.1	76.2	4.2	1.5	2.5	1.1	0.0	1.4	0.2	0.0	0.4	5/6
Sundial (SU)	3.4	0.0	1.0	92.6	1.2	0.4	0.5	0.1	0.0	0.6	0.0	0.0	0.1	12/12
Time-MoE (ME)	4.9	0.0	1.5	88.3	2.2	0.6	1.5	0.2	0.0	0.6	0.0	0.0	0.1	12/12
TTM-R2 (TT)	12.6	1.0	0.8	55.0	6.4	2.4	14.0	1.5	0.1	3.7	1.3	0.1	1.2	3/3
Timer-XL (TX)	4.0	0.0	0.9	90.7	2.0	0.6	1.0	0.2	0.0	0.6	0.1	0.0	0.1	5/8

Table 11: Test-set classifications of random-walk sequences by drift magnitude for two stronger and two weaker models. All reported class entries are percentages.

Model	$ D $	n	Random walk	Trend	Relaxation	Shift	Other
Chronos-2 (C2)	[0.5, 1)	950	95.4	1.2	0.6	0.4	2.4
	[1, 2)	981	95.4	1.3	0.7	0.3	2.2
	[2, 3]	589	89.6	5.9	2.2	0.2	2.0
Toto-Open-Base 1.0 (TO)	[0.5, 1)	950	94.7	0.9	0.6	0.8	2.8
	[1, 2)	981	94.6	1.7	1.1	0.5	2.0
	[2, 3]	589	86.1	8.1	3.4	0.3	2.0
Moirai (MO)	[0.5, 1)	950	83.1	3.8	2.1	3.2	7.9
	[1, 2)	981	79.6	9.0	3.9	2.3	5.2
	[2, 3]	589	59.4	25.3	8.0	1.7	5.6
TTM-R2 (TT)	[0.5, 1)	925	60.3	5.6	3.8	14.6	15.7
	[1, 2)	1,010	56.0	11.7	6.5	14.7	11.1
	[2, 3]	585	44.8	25.3	10.3	12.0	7.7

B.2 SINGLE-PARAMETER REPRESENTATION ACCESSIBILITY

This subsection examines how parameter accessibility develops across depth and compares representations with raw and matched-input controls. It compares ridge probes with shallow nonlinear probes for oscillation frequency, seasonal dependence, and the TimesFM 3 trend-slope result.

B.2.1 LAYER-WISE PARAMETER ACCESSIBILITY

Parameter accessibility is evaluated for eight of the ten single-parameter laws. Random walk with drift (RW) is excluded because its finite-history conditional parameter reference is unstable. Level shift (LS) is retained only for law identification because its standardized construction does not provide a comparable continuous target for parameter-level scoring. The remaining eight laws have stable conditional references and continuous parameter targets.

Table 12 summarizes conditional recoverability, emergence depth, and input controls, while Figure 9 shows parameter accessibility across model depth.

The main comparison in §4.2.1 uses histories from the forecast-response regime. Here, we retain the original accessibility-prior depth analysis for the eight laws with parameter-accessibility scores, including exponential relaxation (RX). Random walk with drift (RW) lacks a stable conditional parameter reference, while level shift (LS) is retained only for law identification. Relative to the main comparison, this analysis uses a narrower trend prior and the original law-specific sampling settings.

For each law, the table reports the range of $D_{m,j}$ across nine models, first-layer accessibility, emergence depth, largest subsequent decline, and input controls. The controls are properties of the generator rather than individual models. Their ranges reflect the three context lengths used in the panel. Seven models use $T = 256$, Timer-XL (TX) uses $T = 288$, and TTM-R2 (TT) uses $T = 512$.

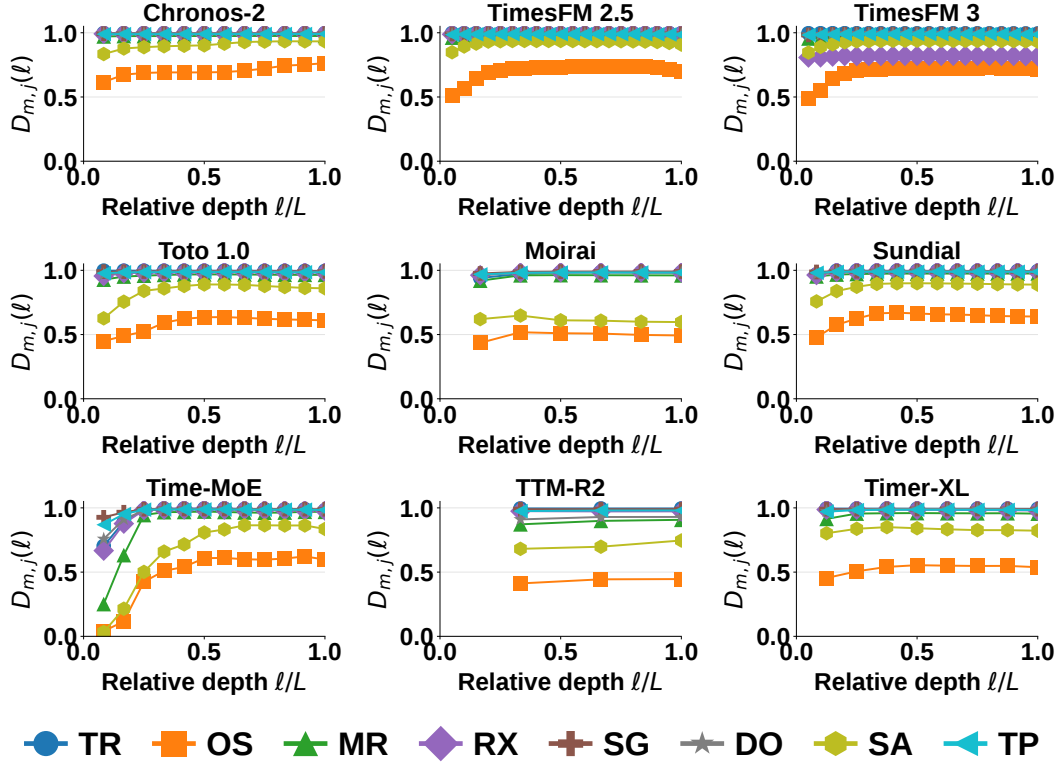


Figure 9: Test-set parameter representation accessibility across relative depth for eight single-parameter laws and nine evaluated TSFM architectures.

Table 12: Single-parameter references, accessibility profiles, and input controls across nine TSFMs. Brackets show minima and maxima across models or context lengths.

Law	Representation accessibility			Depth profile		Input control	
	Reference r^*	Selected D	First layer $D(\ell_1)$	Emergence	Decline	Raw	Matched
Linear trend (TR)	[0.92, 0.96]	[0.985, 0.997]	[0.70, 1.00]	0.08	0.013	1.00	N/A
Oscillation (OS)	[0.97, 0.99]	[0.444, 0.763]	[0.04, 0.62]	0.38	0.041	[0.00, 0.01]	[0.65, 0.72]
Mean reversion (MR)	0.98	[0.907, 0.976]	[0.25, 0.97]	0.08	0.009	0.00	0.00
Exponential relaxation (RX)	0.90	[0.821, 0.993]	[0.67, 0.99]	0.08	0.009	[0.83, 0.84]	0.99
Saturating growth (SG)	0.98	[0.990, 0.993]	[0.93, 0.99]	0.08	0.003	0.83	[0.98, 0.99]
Damped oscillation (DO)	0.99	[0.929, 0.996]	[0.76, 1.00]	0.08	0.008	[-0.01, 0.00]	[0.00, 0.01]
Seasonal autoregression (SA)	0.98	[0.648, 0.938]	[0.03, 0.85]	0.25	0.051	-0.01	[0.00, 0.01]
Transient pulse (TP)	1.00	[0.980, 0.990]	[0.87, 0.99]	0.08	0.004	[-0.01, 0.01]	[-0.01, 0.00]

Three properties of the layer-wise profiles are not visible in the validation-selected summary of Figure 2, which contains one layer per TSFM.

Six Parameters Are Accessible Early, While Two Require Greater Depth. Linear trend (TR), mean reversion (MR), exponential relaxation (RX), saturating growth (SG), damped oscillation (DO), and transient pulse (TP) reach 95% of their maximum by the second layer in eight of nine models. For most corresponding model-law cells, first-layer accessibility is already between 0.91 and 1.00. Oscillation frequency (OS) and seasonal autoregression (SA) are the exceptions, reaching 95% of their maxima at median relative depths of 0.38 and 0.25. Oscillation reaches this threshold only at 0.83 of Chronos-2’s (C2) depth, while seasonal autoregression reaches it at the final layer of TTM-R2 (TT). These two parameters therefore exhibit larger depth-dependent gains in linear accessibility.

Time-MoE Begins with Lower Linear Accessibility. Time-MoE’s (ME) first-layer accessibility averages 0.53 across the eight laws, compared with 0.85–0.92 for every other model. Its first-layer scores are near chance for oscillation frequency at 0.04 and seasonal autoregression at 0.03. Accessibility then increases by 0.40 on average, including gains of 0.83 for seasonal autoregression and 0.72 for mean reversion. Six laws reach the range of the remaining models by one quarter of the network depth. Seasonal autoregression and oscillation frequency reach 95% of their maxima

Table 13: Linear and nonlinear probe accessibility across 19 model-law cells. Accessibility differences are computed separately for each cell.

Parameter	Models	Cells	Median Accessibility		Accessibility Difference	
			Linear D	MLP D	Median ΔD	Maximum ΔD
Oscillation frequency (OS)	All nine	9	0.636	0.708	+0.046	+0.078
Seasonal dependence (SA)	All nine	9	0.890	0.919	+0.029	+0.115
Trend slope (TR)	TimesFM 3 (F3)	1	0.372	0.363	−0.010	−0.010

only at relative depths of 0.58 and 0.50. The law-identification profiles above show a similar delayed emergence, indicating a broader property of Time-MoE’s early layers.

Accessibility Profiles Generally Saturate Rather Than Peak. No curve declines by more than 0.051 after reaching its maximum. The largest decline occurs for Moirai (MO) on seasonal autoregression, followed by 0.041 for TimesFM 2.5 (F2) on oscillation frequency. Linear accessibility is therefore largely retained across later measured layers. TimesFM 3 (F3) provides the principal flat outlier: exponential-relaxation accessibility remains between 0.81 and 0.82 across all 20 layers. The low end of the exponential-relaxation range in Table 12 therefore does not result from layer selection.

Together, the profiles show early and persistent linear accessibility for most parameters, with later emergence for oscillation frequency and seasonal dependence.

B.2.2 LINEAR AND NONLINEAR PROBE COMPARISON

Probe Setting. We compare the ridge probe with a one-hidden-layer GELU MLP using the same features, targets, datasets, and selected layers. The MLP uses hidden width 64 or 256, with dropout and weight decay selected on validation data. The comparison contains 19 distinct model-law cells: nine for oscillation frequency (OS), nine for seasonal dependence (SA), and one for TimesFM 3 (F3) trend slope (TR), using one independent result per cell.

For cell i , the nonlinear accessibility difference is

$$\Delta D_i = D_{i,\text{MLP}} - D_{i,\text{linear}}.$$

A positive value indicates higher accessibility under the MLP, while a negative value favors the linear probe. The differences are computed for individual cells before aggregation. Table 13 reports the median linear and nonlinear accessibility scores together with the median and maximum cellwise differences across the evaluated models and parameters.

Control Probes. Random-label and permutation controls select their probe configurations using their corresponding validation labels. The largest absolute correlation between control predictions and control labels is 0.1422. The maximum correlation between control predictions and the true labels is 0.2905 and serves as a separate leakage diagnostic.

Nonlinear Accessibility Differences. The MLP increases median accessibility by 0.046 for oscillation frequency (OS) and by 0.029 for seasonal dependence (SA). The maximum cellwise improvements are 0.078 and 0.115, respectively. For TimesFM 3 (F3) trend slope (TR), accessibility instead changes from 0.372 to 0.363. This MLP therefore recovers modest additional accessibility for oscillation and seasonal dependence but does not improve the wide-slope trend result. The main accessibility-response comparison retains the matched-regime ridge scores.

Sampling Uncertainty. Across 71 original-regime model-law cells, accessibility standard errors have median 0.0005 and maximum 0.0034. These values summarize within-cell sampling variation for the original accessibility datasets. The main matched-regime analysis contains a separate set of 63 model-law cells with its own response-regime sampling design.

B.3 SINGLE-PARAMETER FORECAST RESPONSE

Forecast response is evaluated for seven of the ten single-parameter laws. Random walk with drift (RW) and level shift (LS) have no parameter-level score, while exponential relaxation (RX) is excluded because its conditional response reference is unstable. This subsection relates representation accessibility to matched-input forecast response under a common generative regime, examines

Table 14: Parameter accessibility on the original and response-regime datasets. Each model entry reports original→response-regime accessibility, and the final row reports the corresponding conditional reference r^* computed from the same history distribution.

Model	TR	OS	MR	SG	DO	SA	TP
Chronos-2 (C2)	0.997 → 0.999	0.763 → 0.796	0.975 → 0.972	0.993 → 0.992	0.996 → 0.996	0.934 → 0.927	0.990 → 0.989
TimesFM 2.5 (F2)	0.997 → 0.999	0.738 → 0.770	0.976 → 0.973	0.993 → 0.992	0.996 → 0.995	0.936 → 0.931	0.989 → 0.987
TimesFM 3 (F3)	0.997 → 0.372	0.726 → 0.764	0.976 → 0.973	0.990 → 0.989	0.996 → 0.995	0.938 → 0.934	0.989 → 0.988
Toto-Open-Base 1.0 (TO)	0.997 → 0.999	0.636 → 0.686	0.968 → 0.965	0.992 → 0.991	0.992 → 0.991	0.890 → 0.884	0.986 → 0.985
Moirai (MO)	0.985 → 0.998	0.518 → 0.592	0.962 → 0.959	0.990 → 0.989	0.984 → 0.981	0.648 → 0.629	0.983 → 0.982
Sundial (SU)	0.997 → 0.999	0.670 → 0.719	0.976 → 0.973	0.993 → 0.992	0.995 → 0.994	0.899 → 0.894	0.989 → 0.988
Time-MoE (ME)	0.995 → 0.999	0.622 → 0.680	0.971 → 0.968	0.992 → 0.991	0.992 → 0.991	0.864 → 0.856	0.988 → 0.986
TTM-R2 (TT)	0.996 → 0.998	0.444 → 0.489	0.907 → 0.902	0.990 → 0.990	0.929 → 0.927	0.746 → 0.733	0.980 → 0.980
Timer-XL (TX)	0.995 → 0.999	0.550 → 0.621	0.959 → 0.956	0.991 → 0.990	0.993 → 0.992	0.850 → 0.841	0.985 → 0.985
Conditional reference r^*	0.922 → 0.999	0.988 → 0.990	0.976 → 0.975	0.983 → 0.982	0.986 → 0.986	0.976 → 0.975	0.997 → 0.997

stability across intervention magnitudes, evaluates alternative response estimators, and audits the exclusion of exponential relaxation through complementary robustness and sensitivity analyses.

B.3.1 ACCESSIBILITY AND RESPONSE UNDER A COMMON REGIME

Accessibility predicts relative response primarily where it varies across models. For oscillation frequency (OS) and seasonal autoregression (SA), the Spearman correlations are 0.92 and 0.82, although their largest normalized responses reach only 0.64 and 0.71. Among the 42 cells with accessibility above 0.95, 20 have response below 0.5, and five respond oppositely. Linear trend (TR) provides the clearest example: eight models have accessibility between 0.998 and 0.999, including all three with negative trend responses.

For a distribution-matched comparison, we repeat the accessibility probes on histories from the response experiments using the original splits and validation-based model selection. For the six paired laws, both arms enter the accessibility dataset with their respective parameter values. The accessibility and response experiments otherwise share the generator, parameter and nuisance distributions, noise, context length, and horizon. Linear trend instead uses independent histories from the wider response-regime slope distribution, whose standard deviation is 0.260 rather than 0.020. Each score is normalized by the conditional reference r^* computed on the corresponding history distribution. Table 14 reports the resulting scores for all models and laws.

Five Paired Laws Retain Their Accessibility. Across mean reversion (MR), saturating growth (SG), damped oscillation (DO), seasonal autoregression (SA), and transient pulse (TP), the largest changes are 0.019 for seasonal autoregression and 0.005 for mean reversion. All remaining changes are at most 0.003, and model ordering is effectively unchanged. The accessibility–response correlations remain 0.92 for mean reversion and 0.82 for seasonal autoregression.

Oscillation Becomes Easier without Changing Model Order. Every model gains between 0.032 and 0.074 on oscillation frequency (OS), while r^* changes by at most 0.002. The response histories contain fewer examples with less than one observed cycle, increasing accessibility without changing model order. The accessibility–response correlation therefore remains 0.92.

Trend Remains Nearly Complete Except for TimesFM 3. Under the response-regime slope distribution, eight models reach accessibility between 0.998 and 0.999. These include Time-MoE (ME), TTM-R2 (TT), and Timer-XL (TX), whose trend responses are negative. TimesFM 3 (F3) instead declines from 0.997 on the original dataset to 0.372 under the response regime, remaining low across all layers despite having the strongest trend response at 1.01. Its separate detrending pathway distinguishes selected-state accessibility from information used elsewhere in the forecasting pipeline.

Therefore, the common-regime results confirm that high parameter accessibility can coexist with weak or reversed forecast response across models and controlled dynamical settings.

B.3.2 RESPONSE STABILITY ACROSS INTERVENTION MAGNITUDES

Figure 10 reports the conditional-reference-normalized response at every signed intervention magnitude for the seven scored laws. The main score $\bar{U}_{m,j}$ averages the model and conditional-reference coefficients separately across these magnitudes.

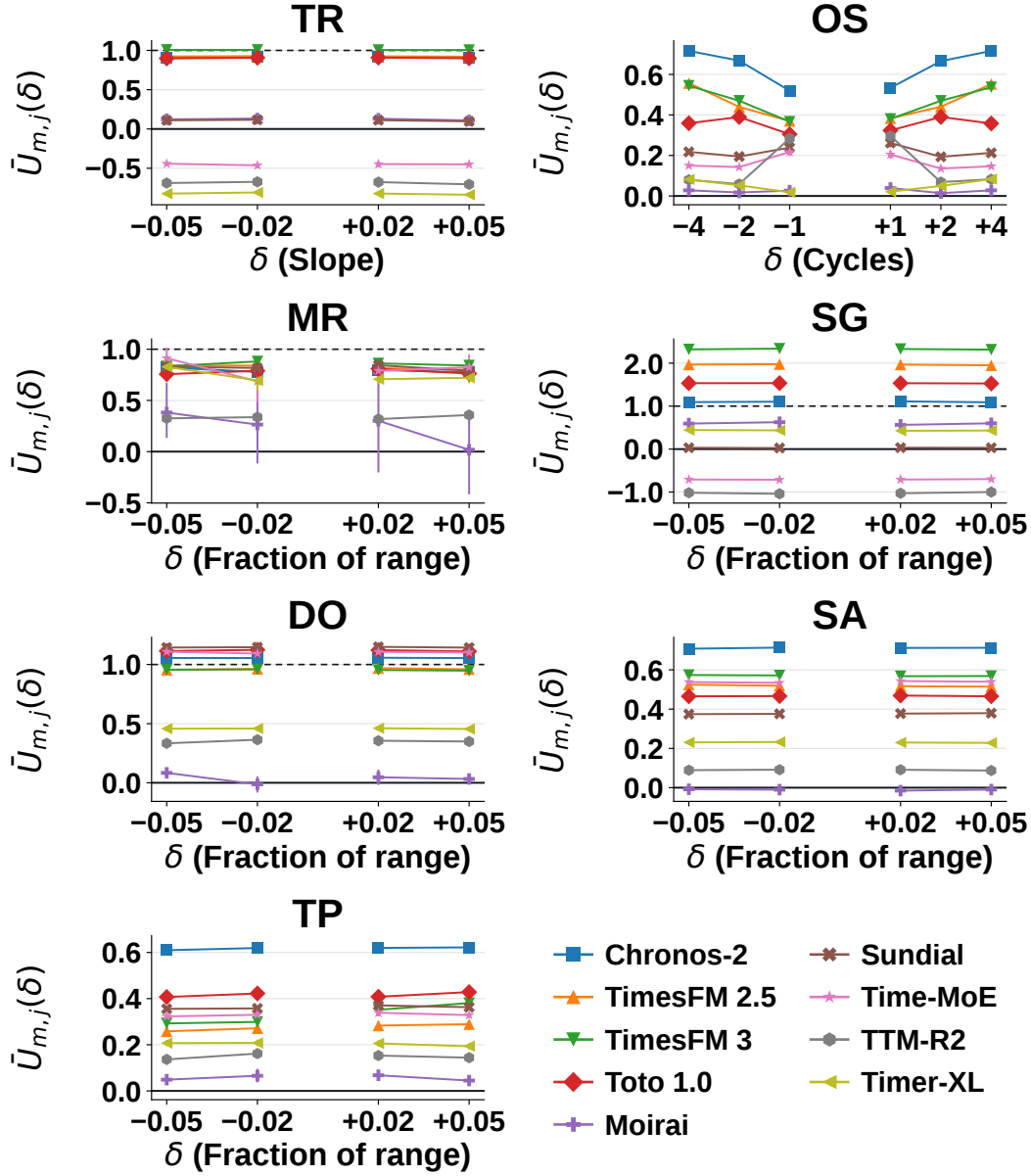


Figure 10: Conditional-reference-normalized response across signed intervention magnitudes. The dashed line and bands mark the conditional reference and 95% bootstrap intervals.

Six Laws Are Stable Across Magnitudes. For linear trend (TR), mean reversion (MR), saturating growth (SG), damped oscillation (DO), seasonal autoregression (SA), and transient pulse (TP), the median response spread ranges from 0.006 to 0.037. After conditional normalization, the spread ranges from 0.006 to 0.052, while the median fitted change from the smallest to largest magnitude remains within 0.01 of zero. The negative trend responses of Time-MoE (ME), TTM-R2 (TT), and Timer-XL (TX) persist at every magnitude, as do the negative saturating-growth responses of Time-MoE (ME) and TTM-R2 (TT), confirming these effects are not dose-specific.

Oscillation Frequency Is the Main Exception. Oscillation frequency (OS) has a median spread of 0.073 before and 0.085 after conditional normalization. Chronos-2 (C2), TimesFM 2.5 (F2), and TimesFM 3 (F3) respond more strongly to larger frequency shifts, whereas TTM-R2 (TT) falls from 0.29 to 0.06 between shifts of one and two cycles. These model-specific trends largely cancel, leaving the mean response near 0.27 at one cycle and 0.30 at four cycles despite changes in model ordering across the evaluated frequency intervention magnitudes.

Table 15: Scale-aware and unpaired response diagnostics. Panel A summarizes seven laws across nine models, while Panel B reports cells with $\bar{U}_{m,j} < 0$ or $\bar{U}_{m,j} > 1.5$. Bracketed entries report model minima and maxima; unbracketed γ , $\bar{\rho}$, and R^2 entries in Panel A report medians.

PANEL A: LAW-LEVEL SUMMARIES

Law	Paired Intervention					Unpaired Check	
	$\bar{U}_{m,j}$	$U^{(3)}$	γ	$\bar{\rho}$	$U^{*(3)}/U^*$	Coefficient	R^2
Linear trend (TR)	[−0.82, 1.01]	[−0.87, 1.22]	−0.021	–	0.93	[−1.54, 0.89]	–
Oscillation (OS)	[0.03, 0.64]	[0.02, 0.49]	−0.100	0.42	0.82	[−0.15, 0.48]	0.52
Mean reversion (MR)	[0.24, 0.85]	[0.24, 0.79]	−0.007	0.68	0.81	[0.59, 1.26]	0.78
Saturating growth (SG)	[−1.02, 2.33]	[−0.99, 0.63]	−0.002	0.85	1.01	[−1.03, 1.14]	0.87
Damped oscillation (DO)	[0.04, 1.15]	[0.02, 0.91]	−0.005	0.97	1.01	[0.04, 1.22]	0.95
Seasonal autoregression (SA)	[−0.01, 0.71]	[−0.01, 0.43]	−0.003	0.33	0.73	[−0.04, 0.36]	0.73
Transient pulse (TP)	[0.06, 0.62]	[0.04, 0.45]	−0.006	0.91	1.08	[0.02, 0.50]	0.67

PANEL B: WEAK, REVERSED, OR AMPLIFIED RESPONSES

Law	Model	$\bar{U}_{m,j}$	$U^{(3)}$	γ	$\bar{\rho}$	Coefficient	R^2
Linear trend (TR)	Time-MoE (ME)	−0.45	−0.19	−0.103	–	−1.41	–
Linear trend (TR)	TTM-R2 (TT)	−0.69	−0.75	−0.003	–	−1.27	–
Linear trend (TR)	Timer-XL (TX)	−0.82	−0.87	−0.021	–	−1.54	–
Saturating growth (SG)	TimesFM 2.5 (F2)	1.97	0.36	−0.003	0.99	1.02	0.99
Saturating growth (SG)	TimesFM 3 (F3)	2.33	0.39	−0.005	0.99	1.14	0.99
Saturating growth (SG)	Toto-Open-Base 1.0 (TO)	1.53	0.40	−0.002	0.99	0.96	0.98
Saturating growth (SG)	Time-MoE (ME)	−0.71	−0.30	−0.009	0.85	−0.21	0.87
Saturating growth (SG)	TTM-R2 (TT)	−1.02	−0.99	0.000	0.80	−1.03	0.70
Seasonal autoregression (SA)	Moirai (MO)	−0.01	−0.01	−0.165	0.19	−0.04	0.07

For linear trend (TR), the reference forecast and reference change are collinear across the horizon, so the refit does not uniquely separate shape from scale. Dashes denote unavailable or non-identifiable diagnostics.

Signs Are Stable under Alternative Aggregations. Among the 43 cells whose intervals exclude zero at every magnitude, 34 differ by less than 0.02 between positive and negative interventions. Most larger asymmetries shrink under a 10%-trimmed average, indicating sensitivity to heavy-tailed per-pair coefficients. Across the five laws with stored per-pair coefficients, trimmed means preserve the sign and remain within 0.1 of $\bar{U}_{m,j}$ in 36 of 45 cells, compared with 28 cells under median aggregation. Energy-weighted pooled ratios differ more often, especially for transient pulse, because they emphasize pairs with the largest reference responses.

Overall, weak and reversed responses persist across intervention magnitudes and reasonable pair-weighted aggregations across evaluated models and dynamical laws.

B.3.3 RESPONSE-METRIC DIAGNOSTICS

We evaluate whether the matched-input score reflects the designated response rather than rescaling of the existing forecast or an artifact of the paired construction. For pair i , let q_i denote the model forecast change, $r_i = \Delta y_i^{\text{ref}}$ the reference change, and $b_i = \mathcal{C}F_m(x_i^A)$ the centered arm-A forecast. To separate these two contributions explicitly, the scale-aware fit is

$$(a_i, U_i^{(3)}, \gamma_i) \in \arg \min_{a, u, \gamma} \|q_i - a\mathbf{1} - ur_i - \gamma b_i\|_2^2. \quad (14)$$

Here, $U_i^{(3)}$ measures the reference-aligned response after controlling for a component proportional to the existing forecast. The unpaired diagnostic separately compares each model forecast with the conditional forecast for the same history, without matched pairs or response normalization. Table 15 summarizes both diagnostics across evaluated model and law combinations.

Opposing Trend Responses Persist. Under the scale-aware refit, the negative trend responses remain −0.19, −0.75, and −0.87 for Time-MoE (ME), TTM-R2 (TT), and Timer-XL (TX), respectively. Their fitted forecast-slope ratios in the unpaired noisy-ramp check are −1.41, −1.27, and −1.54. The opposing responses therefore persist without matched pairs or conditional normalization.

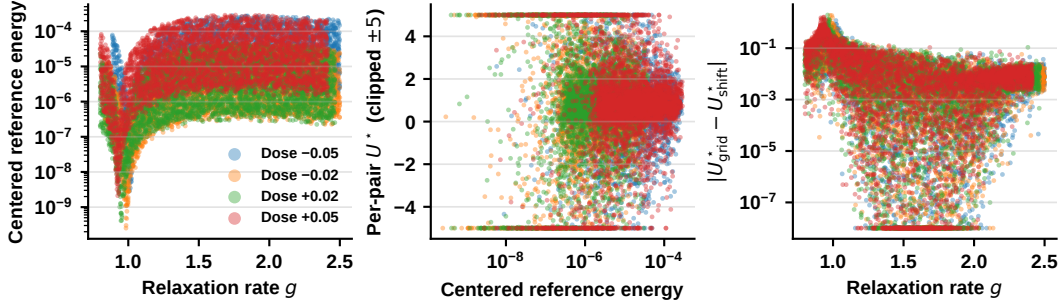


Figure 11: Exponential-relaxation reference stability across the original and restricted parameter ranges under multiple numerical robustness checks.

Saturating Growth Separates Alignment from Scale. Chronos-2 (C2), Toto-Open-Base 1.0 (TO), TimesFM 2.5 (F2), and TimesFM 3 (F3) have saturating-growth matched-input responses between 1.10 and 2.33. The scale-aware refit places the same models between 0.36 and 0.40. The two regressors have median absolute correlation $\bar{\rho} = 0.99$, whereas the unpaired coefficients remain between 0.96 and 1.14, with shape R^2 between 0.98 and 0.99. These models therefore track the reference response shape but overstate its matched-input magnitude, while the scale-aware refit removes part of the aligned component.

Across the remaining laws, the paired and unpaired diagnostics generally preserve the direction of the primary response. Alternative estimators therefore do not remove the separation between parameter accessibility and forecast expression.

B.3.4 EXCLUDED-LAW REFERENCE AUDIT

Exponential relaxation (RX) remains in law identification and representation accessibility but is excluded from normalized forecast-response scoring because many matched futures contain little centered reference change over the forecast horizon. As shown in Figure 11, 33.5% of pairs on the original parameter range fail the relative nondegeneracy criterion. Across quadrature budgets, grid shifts, independent datasets, and aggregation rules, the reference summary changes by as much as 0.228. Restricting the range to its central 80% and 60% reduces this variation to 0.054 and 0.034, but still leaves 29.4% and 22.2% degenerate pairs. No evaluated range reaches the required 90% nondegenerate-pair threshold with a converged reference. The exclusion therefore reflects an unstable response reference rather than unfavorable model performance across tested numerical and range restrictions in every evaluated robustness configuration considered.

B.4 SINGLE-PARAMETER CAUSAL GEOMETRY

Causal geometry uses the same seven laws as the forecast-response analysis to explain why accessible parameters can remain weakly expressed in matched-input forecasts. We first validate local edits that produce the reference response, characterize their shared subspaces and target specificity, compare the required edits with input-induced hidden-state changes, and evaluate input-dependent coordinate prediction. The final analysis measures full-vector fidelity for matched-input, projected, and deployed responses across every evaluated model and parameter.

B.4.1 LOCAL CAUSAL DIRECTIONS AND NUMERICAL ROBUSTNESS

The Jacobian-derived local direction is constructed to produce the reference forecast change under a first-order approximation. The nonlinear response $U_{\text{local}}(\ell)$ tests whether the resulting edit remains effective in the full TSFM. Figure 12 reports selected-layer validity and the fraction of validation-admissible layers, while Figure 13 shows the complete depth profiles.

Validation and Test Agree at Most Layers. Validation and test admissibility differ at 11 of 630 measured layers. Seven discrepancies occur at Time-MoE (ME) trend-slope layers 1–7. The remaining four occur at early transient-pulse (TP) layers whose test local responses lie between 1.05 and 1.08. Using test rather than validation admissibility would not change any selected layer.

C2	0.99	1.01	1.01	1.03	0.97	1.02	1.02
F2	1.02	1.00	1.00	1.01	1.00	1.00	0.99
F3	0.99	1.00	1.00	1.01	1.00	1.03	1.00
TO	1.02	1.00	1.02	1.02	1.00	1.03	1.00
MO	1.02	0.99	1.03	1.02	1.00	1.03	1.02
SU	0.98	1.00	1.01	1.02	0.99	1.01	1.02
ME	0.92	1.00	0.99	0.93	1.00	1.00	0.93
TT	0.99	0.98	1.00	1.02	1.00	1.00	1.00
TX	0.99	1.02	1.01	1.03	1.00	1.02	1.00
	TR	OS	MR	SG	DO	SA	TP

(a) Selected-layer test response.

C2	12/12	12/12	12/12	12/12	12/12	12/12	12/12
F2	6/6	6/6	6/6	6/6	6/6	6/6	6/6
F3	19/20	20/20	20/20	20/20	20/20	20/20	20/20
TO	12/12	12/12	12/12	12/12	12/12	12/12	12/12
MO	5/6	6/6	6/6	6/6	6/6	6/6	6/6
SU	11/12	12/12	12/12	12/12	12/12	11/12	12/12
ME	11/11	11/11	11/11	11/11	11/11	11/11	11/11
TT	3/3	3/3	3/3	3/3	3/3	3/3	3/3
TX	7/8	8/8	8/8	8/8	8/8	8/8	8/8
	TR	OS	MR	SG	DO	SA	TP

(b) Validation-admissible layers.

Figure 12: Selected-layer test response and validation-admissible layer fraction for Jacobian-derived hidden-state edits across seven single-parameter laws.

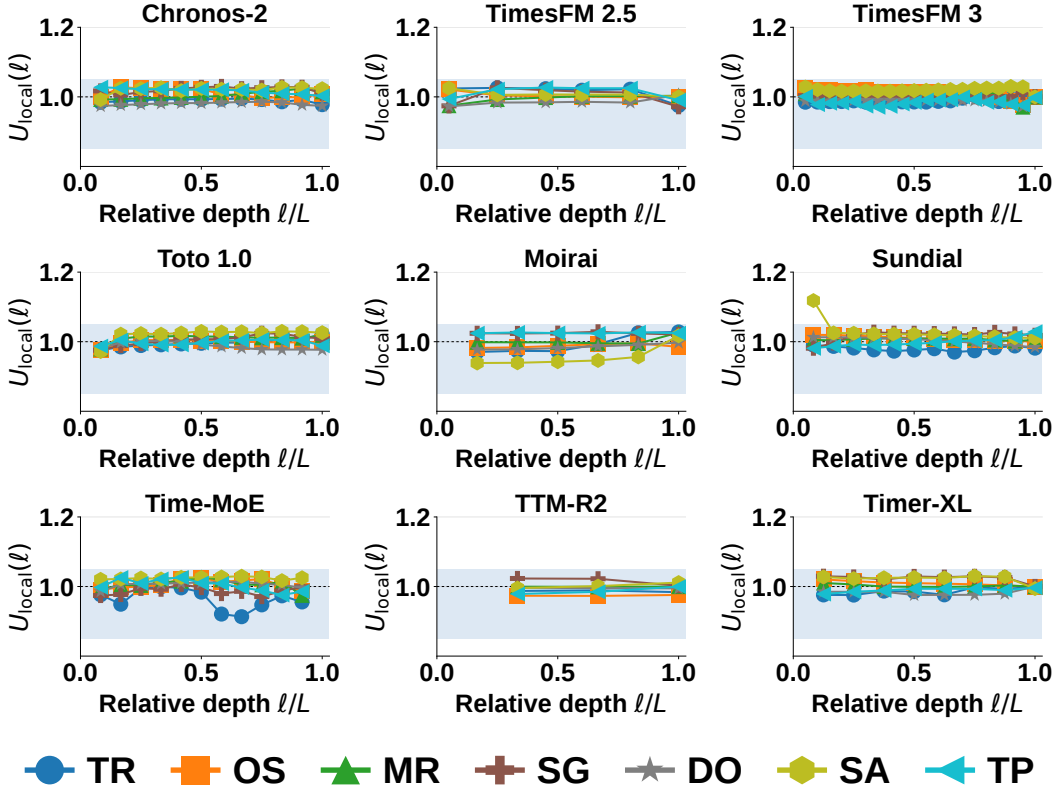


Figure 13: Validation local response across model depth for nine TSFMs and seven single-parameter laws. The shaded band denotes the admissible interval $[0.85, 1.05]$.

Expert Routing Explains Much of the Time-MoE Trend Exception. Table 16 reports the layer-wise comparison between live and frozen expert routing. At the calibrated step, freezing Time-MoE (ME) expert selections raises its linear-trend (TR) local response from between 0.67 and 0.92 to between 0.89 and 0.98. At one-tenth of the step, frozen-routing responses remain between 0.95 and 0.98. The calibrated edit reroutes between 49% and 97% of pairs across the measured layers.

Numerical Sensitivity Is Cell-Dependent. For 16 cells formed by Chronos-2 (C2), TimesFM 3 (F3), Time-MoE (ME), and TTM-R2 (TT) crossed with linear trend (TR), oscillation (OS), seasonal autoregression (SA), and transient pulse (TP), we vary the inversion ridge over six values and the hidden-state step over five multiples of its calibrated value. As shown in Figure 14, projected and deployed responses remain stable in some cells but vary substantially in others. Projected response ranges from 0.89 to 0.93 for Chronos-2 (C2) on linear trend (TR) and from 0.48 to 0.96 for Time-

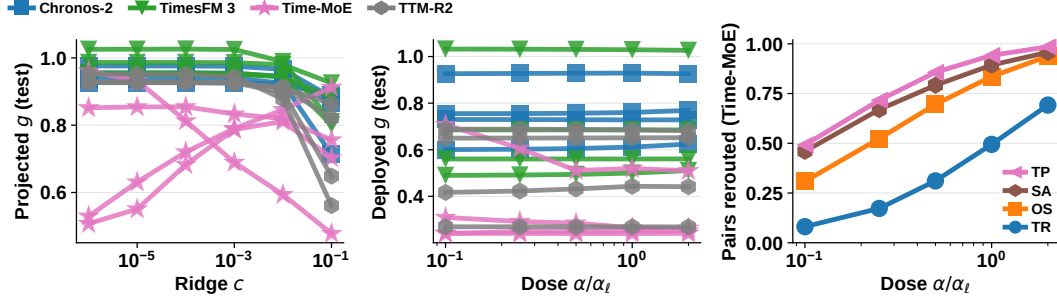


Figure 14: Numerical sensitivity of causal-geometry interventions. The left and center panels vary the inversion ridge and relative hidden-state dose. The right panel reports Time-MoE (ME) expert rerouting for linear trend (TR), oscillation (OS), seasonal autoregression (SA), and transient pulse (TP) across the evaluated intervention strengths.

Table 16: Time-MoE (ME) linear-trend (TR) local response and expert rerouting under live and frozen routing at the calibrated step α_ℓ and one-tenth of that step.

Layer	α_ℓ	Response at α_ℓ		Response at $0.1\alpha_\ell$		Rerouted Pairs	
		Live	Frozen	Live	Frozen	α_ℓ	$0.1\alpha_\ell$
1	0.0370	0.79	0.96	0.39	0.97	94%	37%
2	0.0642	0.84	0.98	0.70	0.97	97%	47%
3	0.0214	0.67	0.92	0.55	0.98	75%	19%
4	0.0281	0.75	0.94	0.44	0.98	77%	20%
5	0.0281	0.79	0.93	0.65	0.98	74%	18%
6	0.0642	0.81	0.91	1.06	0.97	89%	29%
7	0.0642	0.77	0.90	0.75	0.97	86%	27%
8	0.0642	0.85	0.90	1.01	0.96	79%	20%
9	0.0642	0.87	0.90	0.99	0.96	73%	18%
10	0.0642	0.92	0.89	0.98	0.96	65%	13%
11	0.0559	0.92	0.92	1.01	0.95	49%	8%

MoE (ME) on oscillation (OS). Larger Time-MoE (ME) interventions also reroute more pairs, limiting extrapolation beyond the calibrated local step. These sweeps are diagnostic, and the primary results retain the validation-selected settings used throughout all reported causal-geometry comparisons.

In summary, the selected layers therefore support calibrated edits that produce the reference response, while routing boundaries account for the principal local-validity exception.

B.4.2 SHARED CAUSAL SUBSPACE AND TARGET CONTROLS

This subsection examines how compact shared causal subspaces develop across model depth, whether their compactness is specific to the correct parameter target, and how both properties vary with forecast horizon. Figure 15 presents the depth profiles, Figure 16 compares parameter targets with controls, Figure 17 shows representative response-retention curves, and Figure 18 reports horizon sensitivity across all evaluated models and laws.

For each admissible layer, the shared basis is estimated from the fitting dataset. The required subspace dimension $k_{90}(\ell)$ is the smallest evaluated rank satisfying

$$\frac{U_{\text{projected}}(\ell, k)}{U_{\text{local}}(\ell)} \geq 0.9.$$

Because intervened-state dimensions differ across models, panel-wide comparisons also report $k_{90}(\ell)/d_h$. Figure 15 traces the required dimension or corresponding lower bound across model depth for every evaluated law and model.

Required Subspace Dimensions Usually Contract with Depth. Across the 63 model–parameter cells, endpoint comparisons give 56 decreases and seven unchanged outcomes. Seven models contract on all seven laws, while TTM-R2 (TT) contracts on six. Time-MoE (ME) is the main exception: only damped oscillation (DO) contracts, and no evaluated dimension reaches the threshold for linear

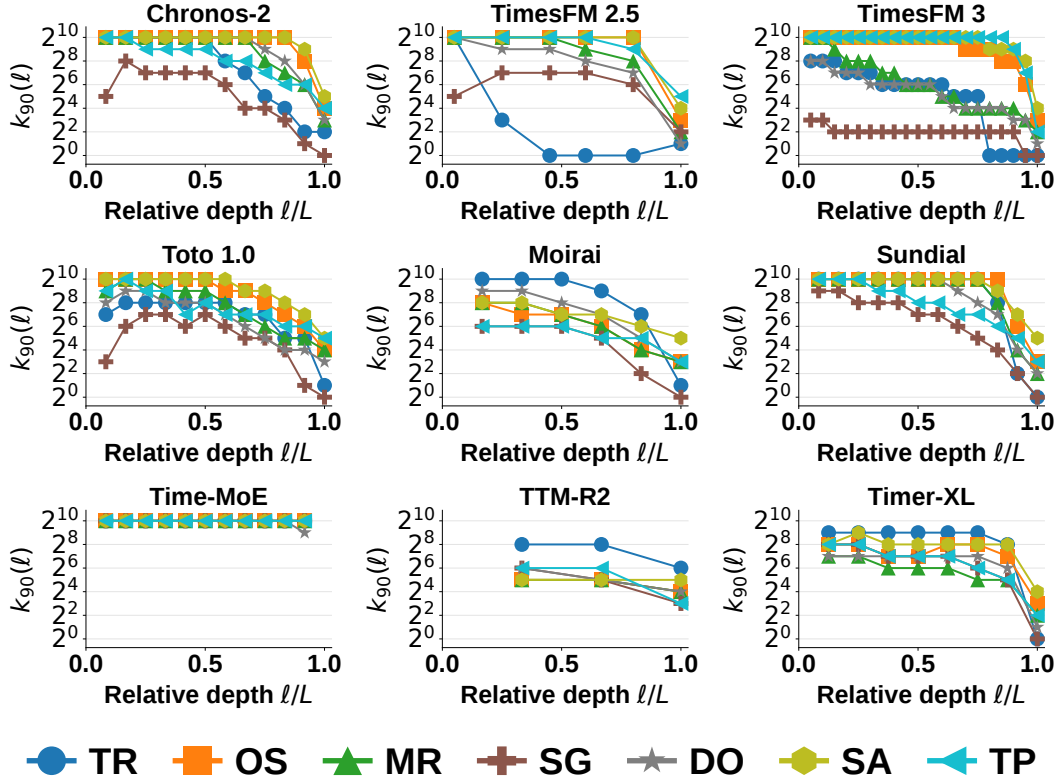


Figure 15: Shared causal subspace rank k_{90} across relative model depth for nine TSFMs and seven single-parameter laws. Unreached 90% response-retention thresholds are shown as lower bounds.

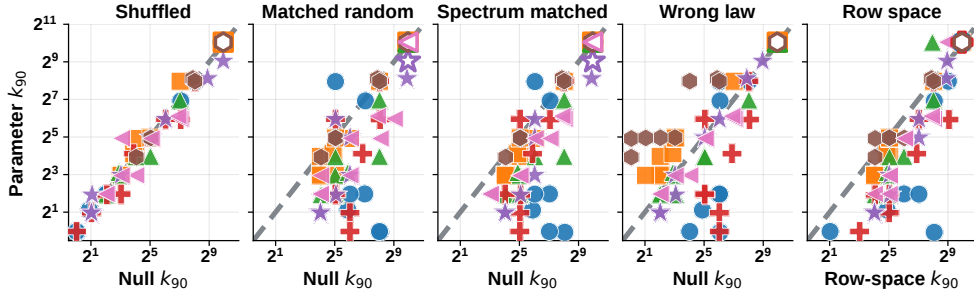


Figure 16: Required subspace dimension k_{90} for parameter targets relative to shuffled, matched-random, spectrum-matched, wrong-law, and generic row-space controls. Hollow markers indicate unreached 90% response-retention thresholds and corresponding lower bounds.

trend (TR), oscillation (OS), or seasonal autoregression (SA). At the selected layers, the median is $k_{90} = 16$, corresponding to 0.26% of the intervened-state dimension.

Compactness Is Not Consistently Parameter Specific. The target-control experiment holds the selected layer, Jacobian, datasets, regularization, edit cap, and evaluated dimensions fixed while changing the forecast-space target. Controls include shuffled input–target pairings, norm-matched random targets, spectrum-matched targets, wrong-law targets, and a generic basis constructed from pooled Jacobian row spaces. Figure 16 compares the required dimensions across all target-control cells, while Figure 17 shows the complete response-retention curves for four representative cells. The correct parameter target improves both required dimension and rank-curve area over matched-random, spectrum-matched, and generic row-space controls in 40, 41, and 44 of 63 cells. It improves both measures over the shuffled target in only six cells and ties its k_{90} in 51 cells. The compact dimensions therefore reflect substantial generic shared causal-subspace structure beyond the correct input–response pairing across most evaluated models and laws.

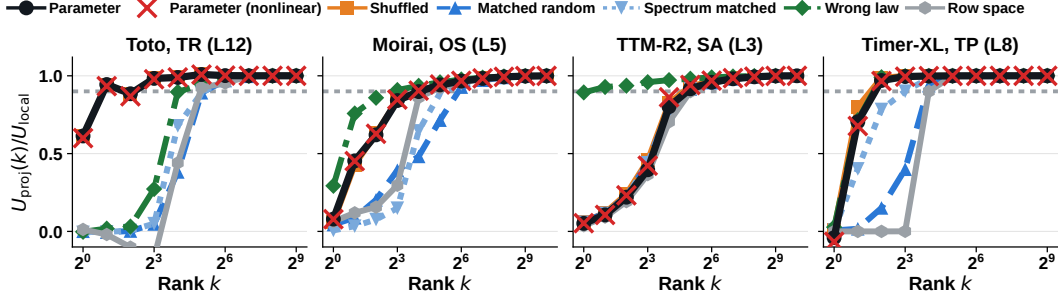


Figure 17: Linearized response retention across subspace dimension k for parameter targets, control targets, and the generic Jacobian row-space basis in four representative model-parameter cells.

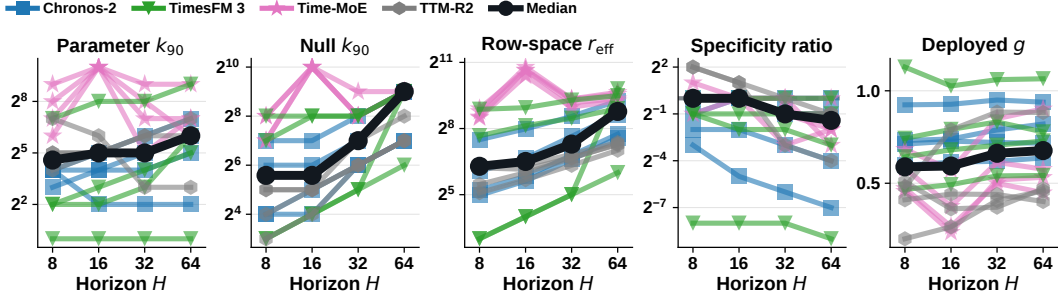


Figure 18: Parameter-target and matched-control subspace dimensions, specificity, and deployed gain across four forecast horizons.

Longer Horizons Increase Dimension but Can Sharpen the Control Comparison. Figure 18 summarizes the parameter-target and matched-control dimensions, specificity, and deployed gain at horizons 8, 16, 32, and 64. The median parameter and matched-random dimensions are 24/48, 32/48, 32/128, and 64/512, respectively. The parameter target has lower required dimension than its matched-random control in 44%, 31%, 69%, and 81% of runs. Median deployed gain rises from 0.59 to 0.68, while centered error falls from 0.70 to 0.55. Both the absolute subspace dimension and the parameter-control comparison therefore depend on forecast horizon.

B.4.3 INPUT-INDUCED HIDDEN-STATE CHANGES

A successful reference-derived edit does not establish that changing the input follows the same hidden-state route. At each selected layer, we form

$$\Delta h_i = h^\ell(x_i^B) - h^\ell(x_i^A)$$

on 4,096 test matched pairs. We measure its cosine with the required local edit, its energy in the shared k_{90} -dimensional basis, and its row- and null-space energy under the baseline Jacobian J_i . The gain of $J_i \Delta h_i$ measures the forecast change predicted by the local linearization. Figure 19 relates matched-input gain to the shared-subspace fraction, linearized gain, null-space fraction, nonlinear remainder, and intervention size for every evaluated model and parameter.

Input Changes Mostly Follow Forecast-Insensitive Directions. Input-induced changes place median energy of 99.6% in the Jacobian null space and only 1.1% in the learned shared subspace. Their median linearized aligned gain is 0.28. Among the 32 cells with matched-input gain below 0.5, 31 have either linearized gain below 0.5 or nonlinear remainder above 0.5. The accessibility-response gap is therefore accompanied by a mismatch between input-induced motion and forecast-sensitive motion across all seven scored single-parameter laws.

B.4.4 CAUSAL-COORDINATE READOUT AND TRANSFER

The deployed intervention predicts input-specific coordinates in the shared causal basis from the unperturbed hidden state. The oracle-assisted fixed-direction control instead applies the mean local direction to every input while receiving each example’s local-direction norm. This subsection examines behavior across model depth, improvement over the fixed-direction control, and transfer

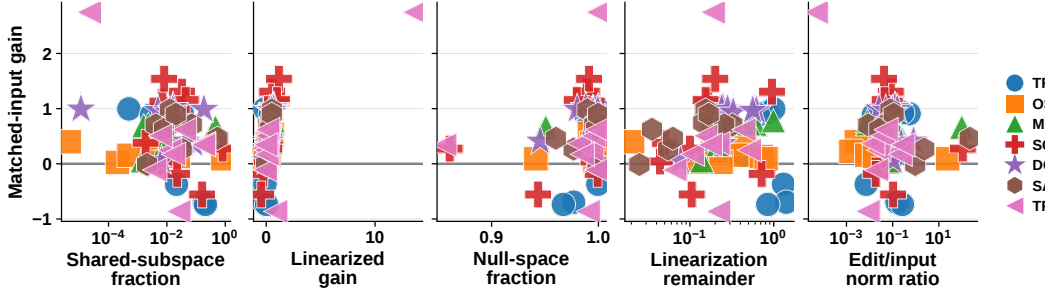


Figure 19: Matched-input aligned gain relative to geometric diagnostics of input-induced hidden-state changes across 63 model-parameter cells. The panels report shared-subspace fraction, linearized gain, Jacobian-null-space fraction, nonlinear remainder, and intervention size.

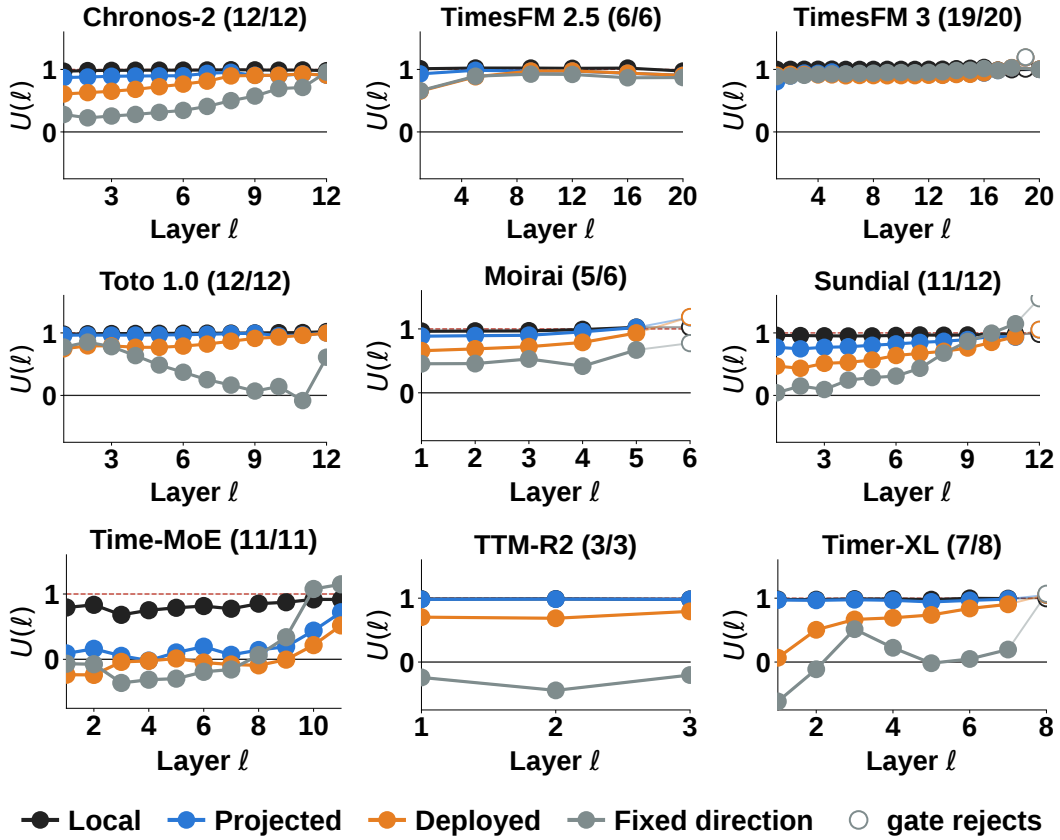


Figure 20: Local, projected, deployed, and oracle-assisted fixed-direction responses for linear trend (TR) across model depth in nine TSFMs. Hollow markers denote validation-rejected layers, and parentheses report admissible and evaluated layer counts.

across distribution shifts. The depth, fixed-control, and transfer results appear in Figure 20, Figure 21, and Figure 22, respectively, for all evaluated model-parameter settings.

Deployed Trend Response Generally Improves with Depth. Figure 20 compares local, projected, deployed, and fixed-direction responses for linear trend (TR) across model depth. For Chronos-2 (C2), Toto-Open-Base 1.0 (TO), Moirai (MO), Sundial (SU), and Timer-XL (TX), deployed response rises substantially across admissible layers and reaches between 0.90 and 1.00. TimesFM 2.5 (F2) and TimesFM 3 (F3) remain near 0.9 or above at later layers, while TTM-R2 (TT) remains between 0.69 and 0.79. Time-MoE (ME) remains near zero through layer 9 and reaches 0.52 at layer 11.

Input-Specific Coordinates Usually Improve upon a Fixed Direction. As shown in Figure 21, the deployed response exceeds the fixed-direction response in 56 of 63 cells at the selected layer and

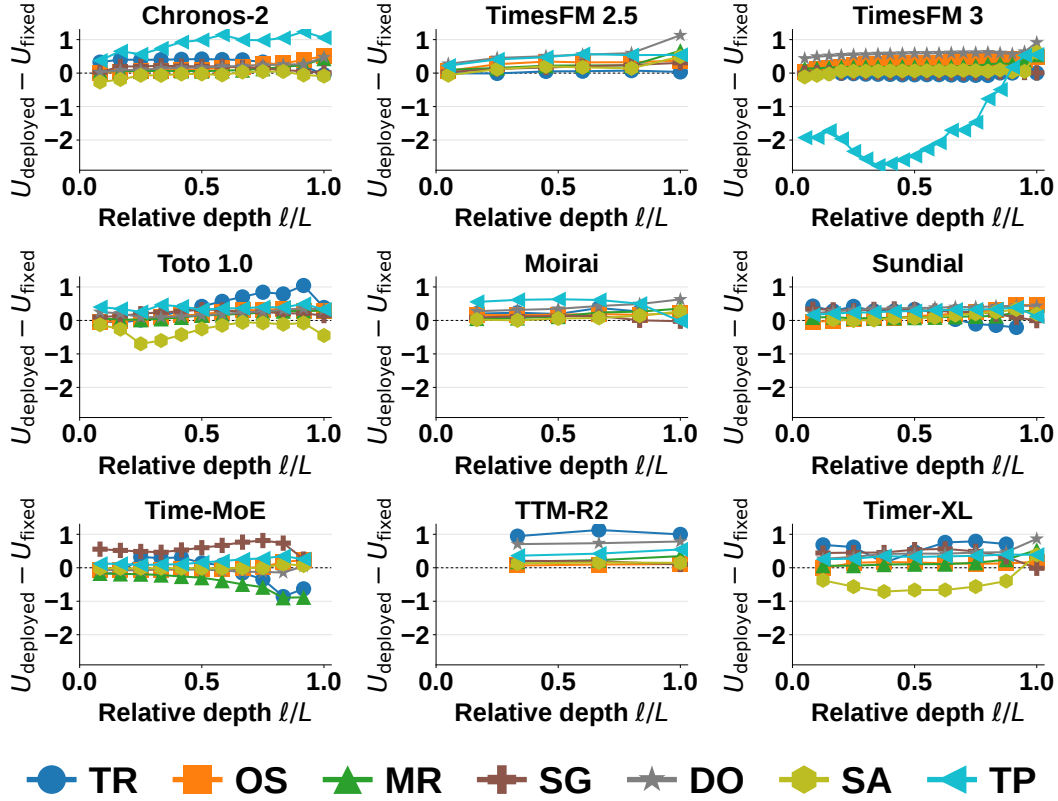


Figure 21: Difference between deployed and oracle-assisted fixed-direction responses across validation-admissible layers for nine TSFMs and seven single-parameter laws. Positive values indicate an advantage for the input-specific coordinate readout.

subspace dimension, with a median difference of 0.26. Across all 625 validation-admissible layers, the median difference is 0.18, and the deployed response remains higher at the final admissible layer in 53 cells. The improvement is most consistent for mean reversion (MR) and damped oscillation (DO) and least consistent for seasonal autoregression (SA).

The Coordinate Readout Transfers Partially. Figure 22 evaluates transfer across intervention magnitude, noise, and parameter-range shifts without target-regime refitting. Panels (a)–(d) show Chronos-2 (C2) on seasonal autoregression (SA) across source and target intervention magnitudes. Panel (a) shows that aligned gain varies from 0.38 to 1.01, with stronger transfer when the target remains near the source intervention regime. Panel (b) shows comparatively stable cosine similarity between 0.75 and 0.83, indicating that response direction transfers more consistently than response magnitude. Panel (c) reports centered error between 0.56 and 0.70, demonstrating that substantial full-vector error remains even when directional alignment is preserved. Panel (d) compares the deployed readout with the oracle-assisted fixed-direction control. The difference ranges from -0.31 to 0.38 , so the input-specific coordinates help most for nearby or same-sign target interventions but do not dominate uniformly across all evaluated intervention combinations.

Panel (e) aggregates aligned gain across all 16 evaluated model–parameter cells. Median deployed gain is 0.59 in regime, 0.51 across intervention magnitude, 0.42 across noise strata, and 0.16 across parameter-range strata. The corresponding fixed-direction gains are 0.20, 0.09, 0.11, and 0.07. A direct hidden-to-output map reaches 0.45 in regime and 0.52 across magnitude, while a probe-derived direction reaches 0.03 in regime. The coordinate readout therefore retains an advantage over the fixed-direction control under the tested shifts, although its magnitude fidelity and absolute response decrease outside the source parameter range for the most challenging transfers.

Overall, the unperturbed hidden state predicts useful input-specific edit coordinates that generally outperform the oracle-assisted fixed-direction control. However, the predicted edits remain incomplete and degrade under distribution shifts, particularly outside the source parameter range.

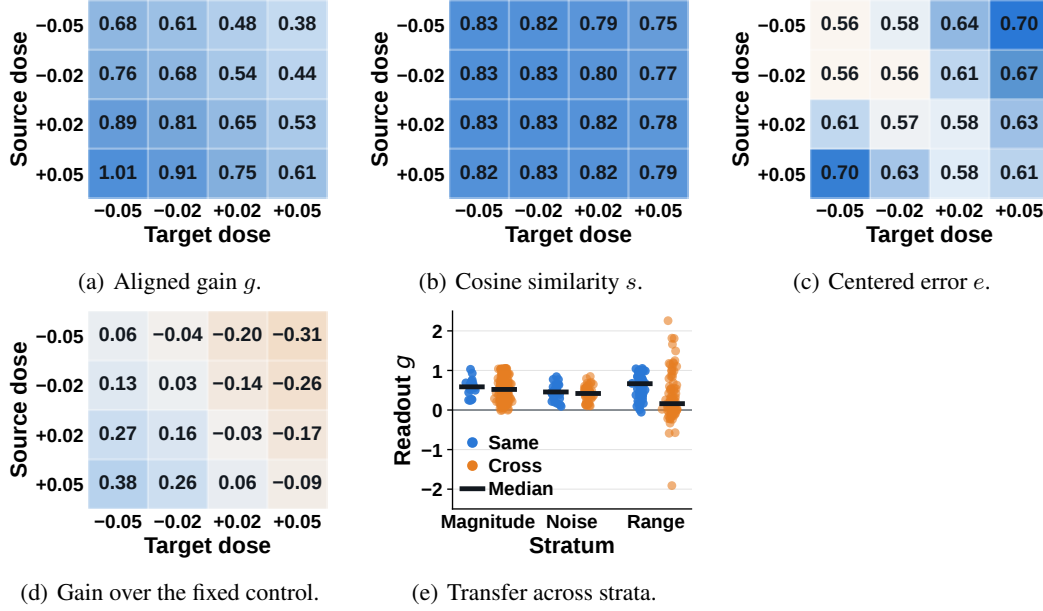


Figure 22: Causal-coordinate transfer across intervention magnitude, noise, and parameter-range shifts without target-regime refitting. The top row reports aligned gain, cosine similarity, and centered error for Chronos-2 (C2) on seasonal autoregression (SA). The second row reports improvement over the oracle-assisted fixed-direction control and readout gain across all evaluated transfer cells.

Table 17: Median pooled response-fidelity metrics across 63 model–parameter cells. Hidden-state interventions use calibrated steps and capped targets.

Condition	Centered fidelity					Raw fidelity	
	g	s	e	$e_{\perp}$	e_{level}	g^{raw}	e^{raw}
Matched input	0.42	0.54	1.06	0.55	1.70	0.84	0.54
Local	1.00	1.00	0.06	0.04	0.09	1.00	0.03
Projected	0.93	0.99	0.18	0.12	0.22	0.96	0.10
Deployed	0.58	0.75	0.68	0.29	0.77	0.77	0.48
Fixed direction	0.23	0.34	1.06	0.54	1.54	0.52	0.93
Conditional reference	0.97	0.99	0.14	0.10	0.84	0.97	0.14

B.4.5 FULL-VECTOR RESPONSE FIDELITY AND CENTERING

Aligned gain in §3.3 measures only the component of a forecast change along the reference response. We therefore compare the complete response vectors for matched-input and hidden-state intervention responses to characterize their full-trajectory fidelity comprehensively.

We evaluate six conditions: matched input, reference-derived local edit, rank-restricted projected edit, deployed coordinate-readout edit, fixed-direction control, and conditional posterior-predictive reference. Matched-input and conditional-reference responses use the response test dataset. Hidden-edit responses use the independent geometry test dataset at the validation-selected layer, calibrated step, and capped target for consistent comparison across intervention conditions.

For response q and reference r , let $\tilde{q}$ and $\tilde{r}$ subtract their respective horizon means. We report centered aligned gain g , centered cosine similarity s , total relative error e , orthogonal error $e_{\perp}$, and level error e_{level} to capture complementary aspects of fidelity:

$$\begin{aligned}
 g &= \frac{\langle \tilde{q}, \tilde{r} \rangle}{\|\tilde{r}\|_2^2}, & s &= \frac{\langle \tilde{q}, \tilde{r} \rangle}{\|\tilde{q}\|_2 \|\tilde{r}\|_2}, \\
 e &= \frac{\|\tilde{q} - \tilde{r}\|_2}{\|\tilde{r}\|_2}, & e_{\perp} &= \frac{\|\tilde{q} - g\tilde{r}\|_2}{\|\tilde{r}\|_2}, & e_{\text{level}} &= \frac{\sqrt{H} |\text{mean}(q) - \text{mean}(r)|}{\|\tilde{r}\|_2}.
 \end{aligned}$$

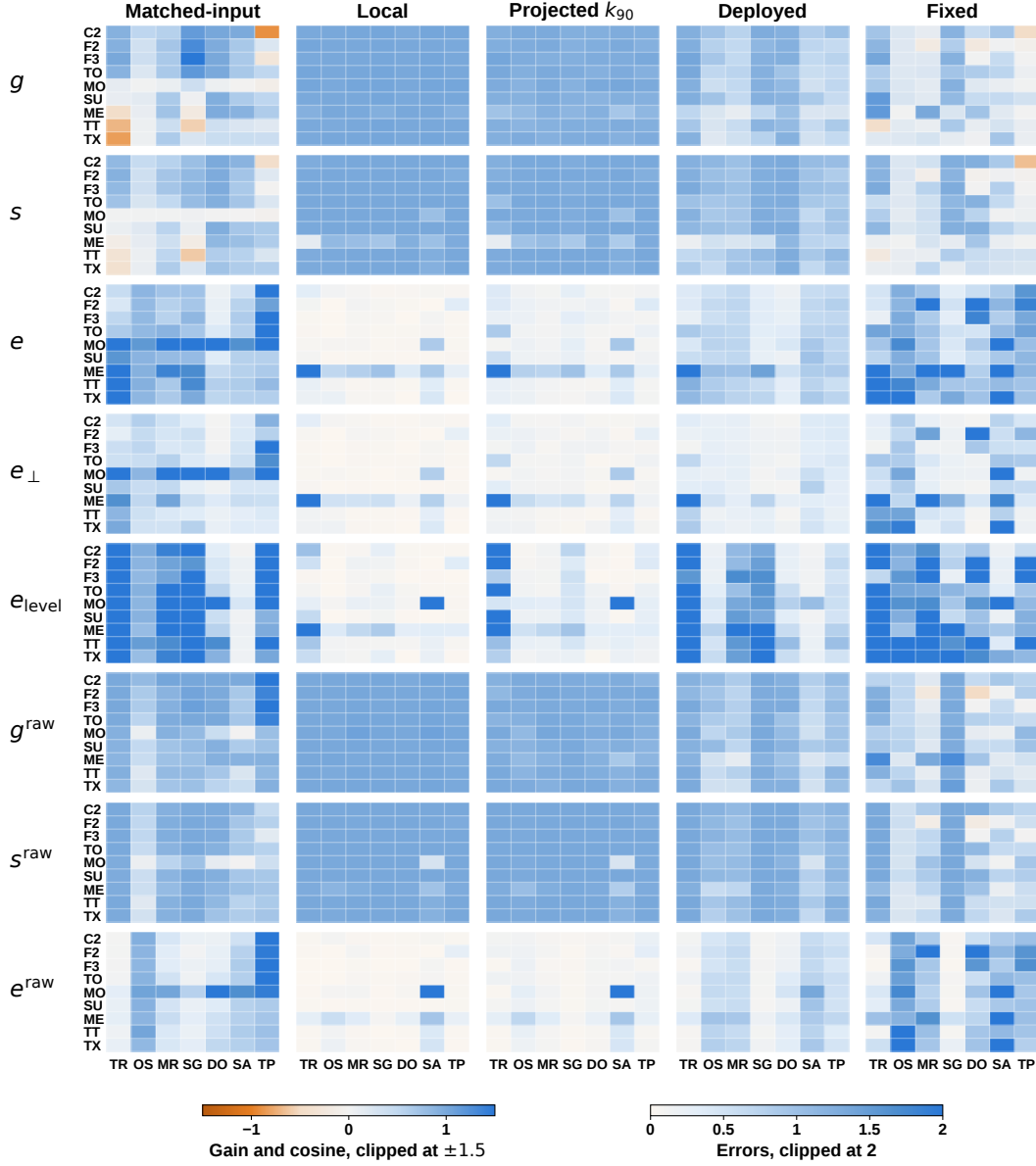


Figure 23: Centered and raw response-fidelity metrics across the single-parameter panel. Columns compare matched-input, local, projected, deployed, and fixed-direction responses. Centered metrics capture response alignment and trajectory error, while raw metrics use uncentered forecast changes.

The raw metrics g^{raw} and e^{raw} use the uncentered vectors. The identity $e^2 = (g - 1)^2 + e_{\perp}^2$ separates aligned-magnitude error from off-direction error.

Matched-Input Gain Does Not Imply Full-Trajectory Agreement. As shown in Table 17 and Figure 23, matched-input responses have median centered gain 0.42 and centered error 1.06. The heatmaps further show substantial variation across model-parameter cells, while local and projected responses remain consistently closer to the reference across the panel. Deployed responses provide an intermediate level of fidelity, whereas the fixed-direction control generally retains larger errors. For matched input, the median raw gain rises to 0.84, showing that level agreement is substantially stronger than centered dynamical-shape agreement. Only 8% of matched-input pairs satisfy all three conditions: $0.8 \leq g \leq 1.2$, $s \geq 0.9$, and $e \leq 0.5$.

Local and Projected Edits Recover the Full Target Most Closely. As shown in Table 17 and Figure 23, local edits achieve median centered gain 1.00 and error 0.06, while projected edits retain gain 0.93 with error 0.18. The heatmaps show that this high fidelity is broadly consistent across

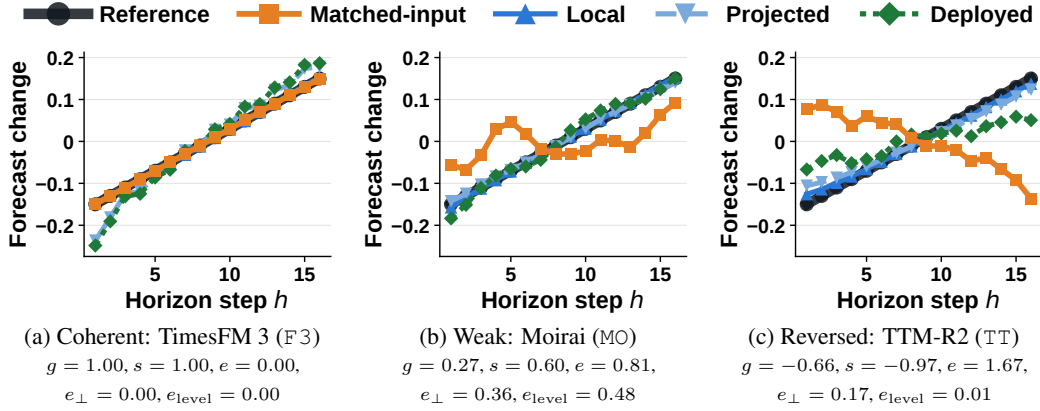


Figure 24: Matched linear-trend (TR) response examples illustrating coherent, weak, and reversed matched-input behavior across TimesFM 3 (F3), Moirai (MO), and TTM-R2 (TT). Each panel compares the reference forecast response with the matched-input, local, projected, and deployed responses across representative model-specific hidden-state response patterns.

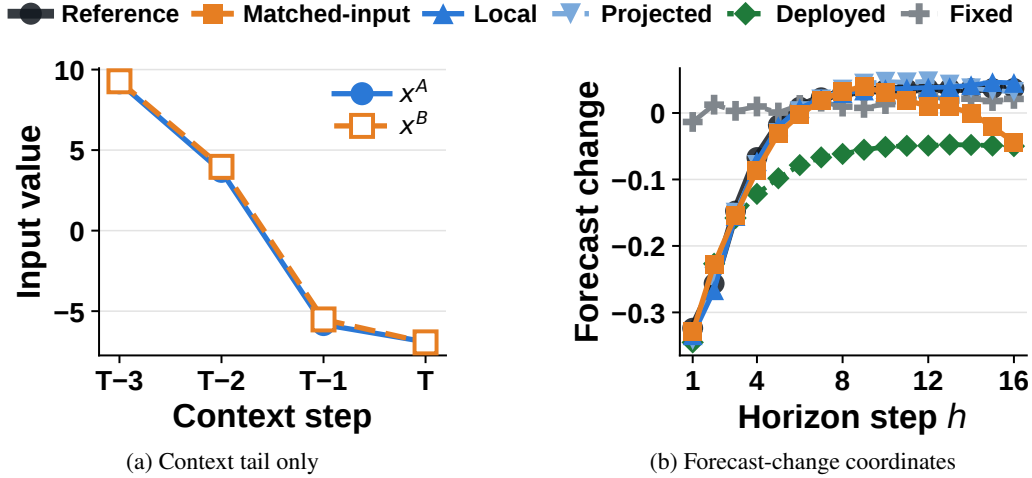


Figure 25: One Chronos-2 (C2) mean-reversion (MR) pair at layer 12, with $\rho^A = 0.4006$, $\rho^B = 0.4486$, and $\delta q = +0.0480$, traced through the matched-input and hidden-state intervention responses.

model-parameter cells rather than being driven by a small subset of cases. The deployed edit partially recovers the target at gain 0.58 and error 0.68, outperforming the fixed direction at gain 0.23 and error 1.06. The simultaneous fidelity criterion is met by 100% of local pairs, 88% of projected pairs, 21% of deployed pairs, and 3% of fixed-direction pairs. Together, these results separate the high fidelity of reference-derived edits from the partial recovery achieved by predicted coordinates.

Representative Trend Responses Show Distinct Matched-Input Behavior. Figure 24 provides three representative examples for the linear trend (TR) law. Panel (a) shows a coherent matched-input response for TimesFM 3 (F3), panel (b) shows a weak response for Moirai (MO), and panel (c) shows a reversed response for TTM-R2 (TT). The centered fidelity metrics below each panel quantify these differences in gain, direction, trajectory error, and level error. Despite the distinct matched-input behavior, the local and projected responses remain close to the reference in all three panels. The deployed responses recover the reference to varying degrees, indicating that predicted hidden-state coordinates recover only part of the reference response. These examples illustrate how similar hidden-state response capacity can coexist with substantially different responses to the corresponding input change across models with distinct forecasting behaviors.

A Single-Pair Trace Separates the Intervention Stages. Figure 25 follows one Chronos-2 (C2) mean-reversion (MR) pair through the matched-input, local, projected, deployed, and fixed responses. Panel (a) shows the matched context tails x^A and x^B generated by the change from $\rho^A = 0.4006$ to $\rho^B = 0.4486$. Panel (b) compares the corresponding forecast-change trajectories against the reference response. The matched-input response captures part of the target, while the local and

Table 18: Multi-parameter summary across nine TSFMs. Entries for representation accessibility, matched-input response, and deployed aligned gain report ranges across models. The six-model group contains all models except Moirai (MO), TTM-R2 (TT), and Timer-XL (TX), which form the three-model group. Compact cells report the descriptive compactness count from the run summary.

Parameter	Accessibility	Matched-Input Response $\bar{U}_{m,j}$		Compact	Deployed	Deployed > Fixed
	$D_{m,j}$	Six models	Three models	Cells	Gain	Cells
<i>Dual Tone</i>						
Low-tone frequency (TL)	0.86–0.99	0.82–1.00	0.04–0.16	7/9	0.65–0.98	9/9
High-tone frequency (TH)	0.93–0.99	0.75–0.99	−0.16–0.12	6/9	0.40–0.97	9/9
Amplitude ratio (TW)	0.94–1.00	0.99–1.18	0.06–0.35	6/9	0.44–0.94	8/9
<i>Linear State Space</i>						
Eigenvalue angle (SC)	0.91–0.99	0.70–0.83	0.07–0.26	7/9	0.28–0.63	8/9
Eigenvalue modulus (SR)	0.61–0.91	0.85–1.02	0.07–0.28	6/9	0.17–0.59	5/9
<i>Forced Diffusion</i>						
Drive frequency (WC)	0.98–1.00	0.32–0.56	−0.04–0.05	6/9	0.16–0.77	9/9
Diffusion time (WD)	0.62–0.89	0.76–0.96	0.10–0.51	7/9	0.51–0.84	8/9

Table 19: Multi-parameter recoverability, accessibility, depth profiles, and input controls across nine TSFMs. Brackets indicate minima and maxima across models or context lengths.

Parameter	Accessibility		First Layer	Depth Profile		Input Control	
	r^*	$D_{m,j}$	$D_{m,j}(\ell_1)$	Emergence	Decline	Raw	Matched
<i>Dual Tone</i>							
Low-tone frequency (TL)	1.00	[0.857, 0.995]	[0.10, 0.96]	0.25	0.039	0.00	[0.00, 0.01]
High-tone frequency (TH)	1.00	[0.927, 0.994]	[0.07, 0.97]	0.17	0.028	[0.01, 0.02]	[0.00, 0.01]
Amplitude ratio (TW)	0.99	[0.940, 0.996]	[0.08, 0.99]	0.17	0.037	0.01	0.00
<i>Linear State Space</i>							
Eigenvalue angle (SC)	[0.98, 0.99]	[0.909, 0.995]	[0.09, 0.99]	0.08	0.012	[−0.02, 0.00]	[−0.01, 0.00]
Eigenvalue modulus (SR)	[0.89, 0.94]	[0.608, 0.912]	[0.17, 0.84]	0.25	0.064	[−0.01, 0.00]	[−0.01, 0.01]
<i>Forced Diffusion</i>							
Drive frequency (WC)	1.00	[0.976, 0.999]	[0.18, 1.00]	0.12	0.004	0.00	[0.00, 0.01]
Diffusion time (WD)	0.99	[0.623, 0.895]	[0.36, 0.88]	0.25	0.076	[−0.01, 0.00]	0.00

projected traces closely follow the reference across the forecast horizon. The deployed trace preserves much of the response direction but retains a noticeable level offset, whereas the fixed direction largely misses the target change. For the matched-input, local, projected, deployed, and fixed responses, respectively, the raw aligned gains are 0.941, 1.041, 1.029, 0.883, and 0.029. The corresponding centered gains are 0.897, 1.042, 1.032, 0.714, and 0.061. The differences between raw and centered gains illustrate why response alignment should be interpreted together with trajectory, orthogonal, and level errors across the distinct hidden-state intervention stages.

Together, these analyses show that the selected hidden states can produce reference-aligned forecast changes even when the corresponding input-induced changes fail to produce the same response. Local and projected interventions recover the target with high full-vector fidelity, while the predicted coordinate readout recovers only part of the required response. This separation distinguishes response capacity in the hidden state from the response naturally induced by changing the input.

B.5 MULTI-PARAMETER LAWS

§4.3 summarizes representation accessibility, matched-input response, and deployed response for seven parameters across three multi-parameter laws. This subsection provides the detailed analyses supporting those results. We first examine parameter accessibility across models and depth, together with a matched single-tone control. We then evaluate matched-input response across intervention magnitudes and full-vector response fidelity. Finally, we analyze causal geometry through local intervention validity, shared causal-subspace structure, and the advantage of input-specific deployed directions over the fixed-direction control. The parameter ranges, conditional references, and calibrated intervention settings are summarized in Table 6, Table 7, and Table 8.

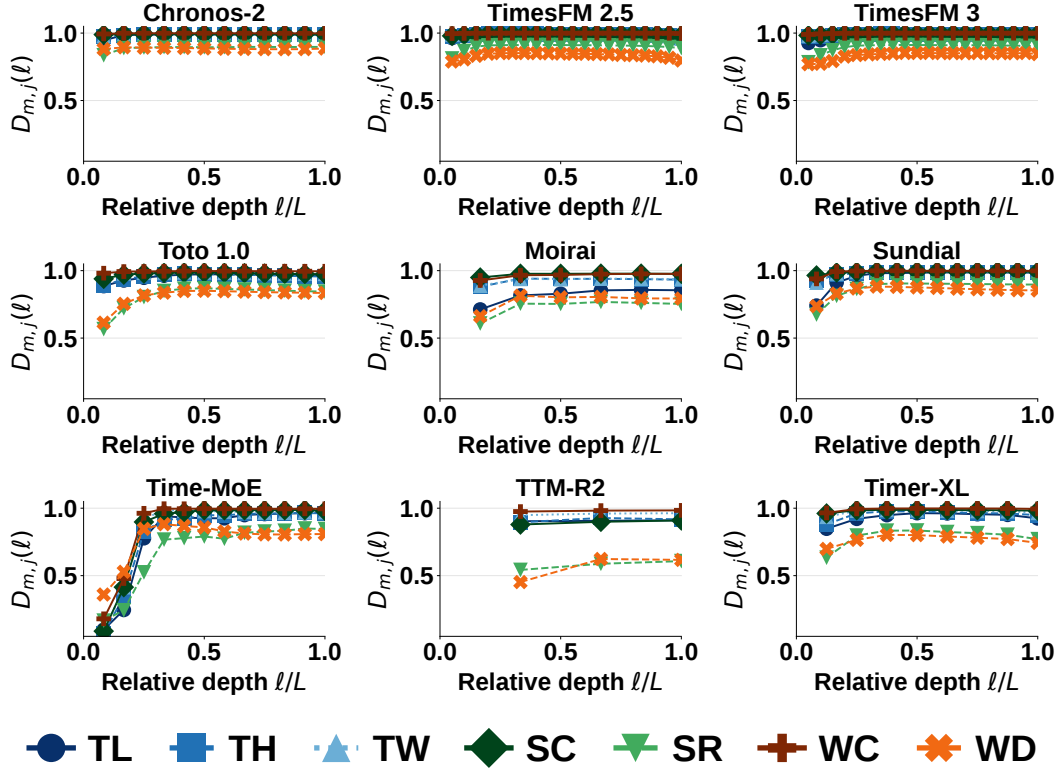


Figure 26: Test-set parameter accessibility across relative depth for seven jointly identified parameters. Panels show all nine frozen TSFMs.

Table 20: Matched single-tone control for the dual-tone low-frequency parameter (TL). Accessibility and matched-input response compare the single-tone setting with TL, with $\Delta = \text{TL} - \text{single tone}$.

Model	Accessibility $D_{m,j}$			Matched-Input Response $\tilde{U}_{m,j}$		
	Single	TL	Δ	Single	TL	Δ
Chronos-2 (C2)	0.999	0.992	-0.007	1.02	0.97	-0.05
TimesFM 2.5 (F2)	1.000	0.994	-0.006	0.97	0.97	0.00
TimesFM 3 (F3)	1.000	0.995	-0.005	1.01	1.00	0.00
Toto-Open-Base 1.0 (TO)	0.998	0.973	-0.025	1.03	0.82	-0.21
Moirai (MO)	0.983	0.857	-0.126	0.08	0.04	-0.04
Sundial (SU)	0.999	0.987	-0.012	1.03	0.96	-0.08
Time-MoE (ME)	0.998	0.965	-0.033	0.99	0.89	-0.10
TTM-R2 (TT)	0.989	0.911	-0.078	0.22	0.16	-0.06
Timer-XL (TX)	0.999	0.963	-0.036	0.27	0.12	-0.15

B.5.1 REPRESENTATION ACCESSIBILITY

This analysis examines how representation accessibility for the seven jointly identified parameters varies across models, depth, and input controls. Table 18 summarizes parameter-level measurements, while Table 19 reports recoverability, depth profiles, and input controls. Table 20 compares the dual-tone low-frequency parameter with a matched single-tone control. Finally, Figure 26 shows accessibility across model depth for all seven parameters. These analyses further examine the differences summarized in §4.3 across the three multi-parameter dynamical laws.

Frequencies Are Accessible Early, While Rates Remain Lower. As shown in Table 18, Table 19, and Figure 26, the four frequency parameters have strong accessibility across models and are already accessible at early layers. Their median first-layer accessibility is 0.886 for TL, 0.895 for TH, 0.961 for SC, and 0.975 for WC. They reach 95% of their maximum accessibility by median relative depths between 0.08 and 0.25. The two rate parameters start lower, at 0.633 for SR and 0.698 for WD, and their median peak accessibility values of 0.869 and 0.852 remain below the median first-layer

accessibility of every frequency parameter. The gap between modal frequency and eigenvalue modulus reported in §4.3 is therefore present throughout model depth rather than arising from layer selection. Time-MoE (ME) has the lowest first-layer accessibility, ranging from 0.07 to 0.36, and exhibits the largest increases across the seven parameters, ranging from 0.52 to 0.91. The depth profiles are otherwise largely stable after reaching their maxima, with the largest subsequent decline equal to 0.076 across all evaluated model–parameter combinations.

The Second Tone Has Limited Effect on Accessibility but Reduces Response. The matched single-tone control in Table 20 isolates the effect of adding the second tone, while Table 18 places the resulting low-tone accessibility and response within the full multi-parameter panel. Relative to the single-tone setting, low-tone frequency (TL) loses between 0.005 and 0.126 in accessibility. For the six-model response group, the accessibility reduction is at most 0.033. Matched-input response decreases for every model, ranging from less than 0.001 for TimesFM 2.5 (F2) to 0.209 for Toto-Open-Base 1.0 (TO). The three largest response reductions occur for Toto-Open-Base 1.0 (TO), Timer-XL (TX), and Time-MoE (ME), at 0.209, 0.154, and 0.104, respectively. These reductions exceed their corresponding accessibility losses of 0.025, 0.036, and 0.033, showing that the added tone affects forecast response more strongly than parameter accessibility in these models.

TSFM Processing Makes the Joint Parameters Linearly Accessible. Table 19 shows that the raw and matched-input controls remain near zero for all seven parameters, with absolute values reaching at most 0.02, while TSFM accessibility ranges from approximately 0.61 to 1.00. For low-tone frequency (TL), high-tone frequency (TH), and drive frequency (WC), the conditional reference is 1.00, yet the corresponding input controls remain near zero. The selected hidden representations therefore expose these parameters substantially more linearly than the tested raw-input summaries across models and jointly varying settings. This comparison does not establish that the parameter information is absent from the input or uniquely computed by the network.

Together, these results show that joint identification preserves strong accessibility for the frequency parameters while maintaining parameter-specific differences within each dynamical law. The matched single-tone control further shows that adding a second frequency component has a modest effect on accessibility but can produce a larger reduction in matched-input forecast response.

B.5.2 FORECAST RESPONSE

This analysis examines matched-input response for the seven jointly identified parameters across full-vector fidelity, intervention magnitudes, and response diagnostics. Figure 27 compares centered and raw full-vector fidelity across response conditions. Figure 28 then shows conditional-reference-normalized response across signed intervention magnitudes, while Table 21 reports scale-aware and unpaired diagnostics. Together, these analyses provide the detailed response results underlying §4.3.

Full-Vector Fidelity Separates the Response Conditions. Figure 27 compares centered and raw fidelity for matched-input, local, projected, deployed, and fixed-direction responses across the seven jointly identified parameters. The local and projected interventions remain closest to the reference across the fidelity metrics, showing that the selected hidden states can support the reference response and that much of this response is retained after projection into the shared subspace. Median pooled centered errors are 0.93, 0.14, and 0.61 for matched-input, projected, and deployed responses, respectively. The deployed responses therefore recover a substantial part of the reference response but retain larger gain, directional, and trajectory errors than the reference-derived projected interventions. Matched-input responses show the largest variation across parameters and models, while the fixed-direction control generally provides weaker fidelity than the input-specific deployed directions. Differences between centered and raw metrics further indicate that agreement in overall forecast level can coexist with disagreement in the dynamical shape of the response. Together, the figure separates response capacity in the hidden state from the fidelity achieved by natural input changes and predicted intervention coordinates across jointly varying multi-parameter dynamical systems.

Weak Response Persists Across Tested Magnitudes. As shown in Figure 28, the median spread across four intervention magnitudes is at most 0.044 for every parameter, while the largest spread among the six responsive models is 0.098 for low-tone frequency (TL). The largest overall spreads occur for TTM-R2 (TT) on drive frequency (WC) at 0.194 and low-tone frequency (TL) at 0.140, where responses remain near zero. The difference between positive and negative interventions of equal magnitude is at most 0.095. For drive frequency (WC), the six responsive models remain

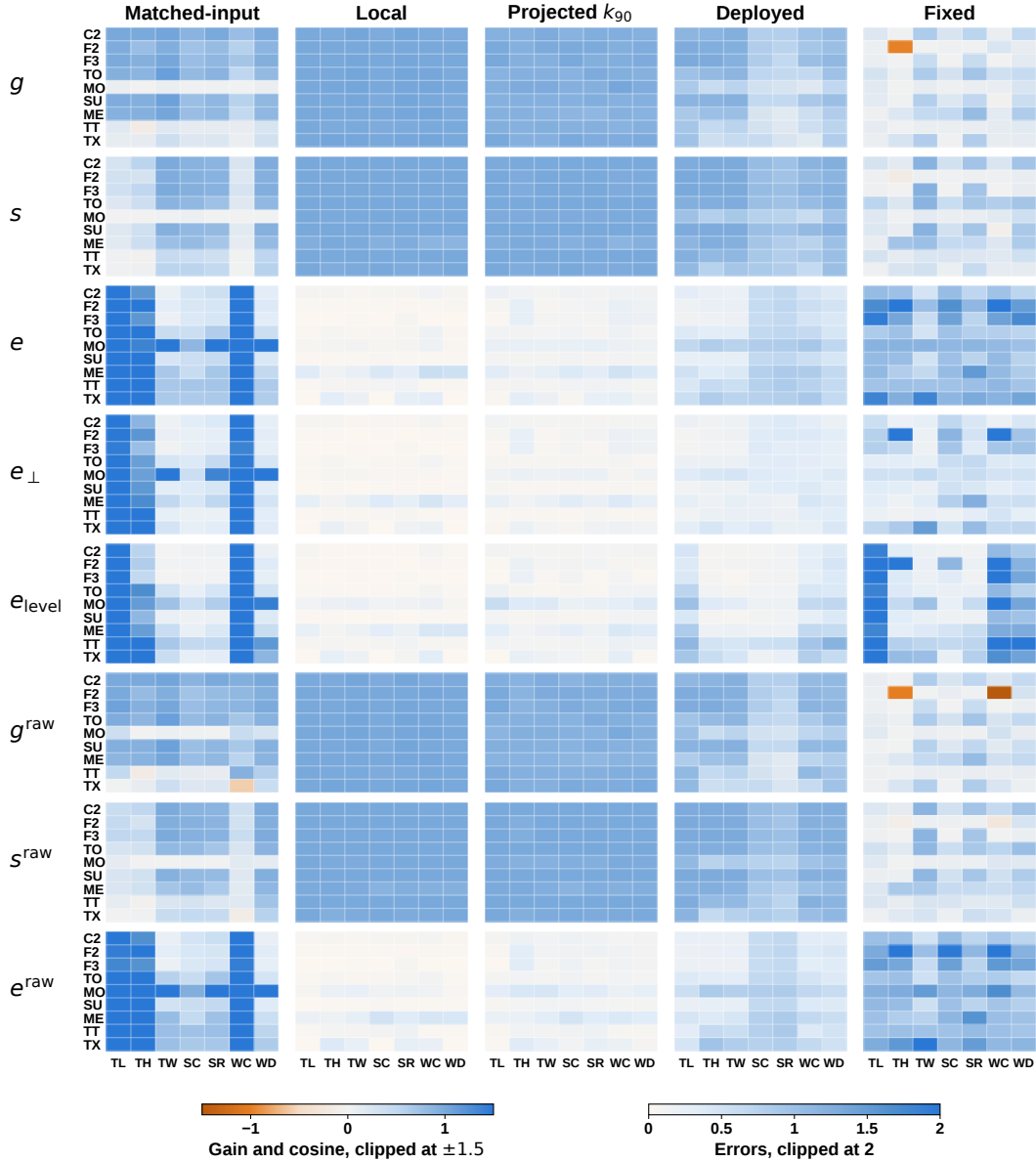


Figure 27: Centered and raw response-fidelity metrics across the multi-parameter panel. Columns compare matched-input, local, projected, deployed, and fixed-direction responses. Centered metrics capture response alignment and trajectory error, while raw metrics use uncentered forecast changes.

between 0.31 and 0.58 at every magnitude, with a within-model spread of at most 0.063. The weak drive-frequency response therefore persists across the evaluated intervention magnitudes rather than arising from a particular magnitude.

Drive-Frequency Response Is Weaker Than Unpaired Waveform Tracking. Table 21 shows that the six responsive models track the unpaired conditional forecast substantially better than they respond to a controlled drive-frequency change. For drive frequency (WC), the unpaired coefficient ranges from 0.59 to 0.88, with shape R^2 between 0.92 and 0.96. The two tone frequencies show comparable unpaired tracking, with coefficients between 0.85 and 0.98 and R^2 between 0.92 and 0.97. By contrast, matched-input drive-frequency response remains between 0.32 and 0.56. These models therefore track much of the unpaired waveform shape while responding more weakly to a controlled change in its frequency across the evaluated multi-parameter settings.

Some Response Components Are Associated With Window Scale. The scale-aware diagnostics in Table 21 separate the reference-aligned component from a component associated with window scale.

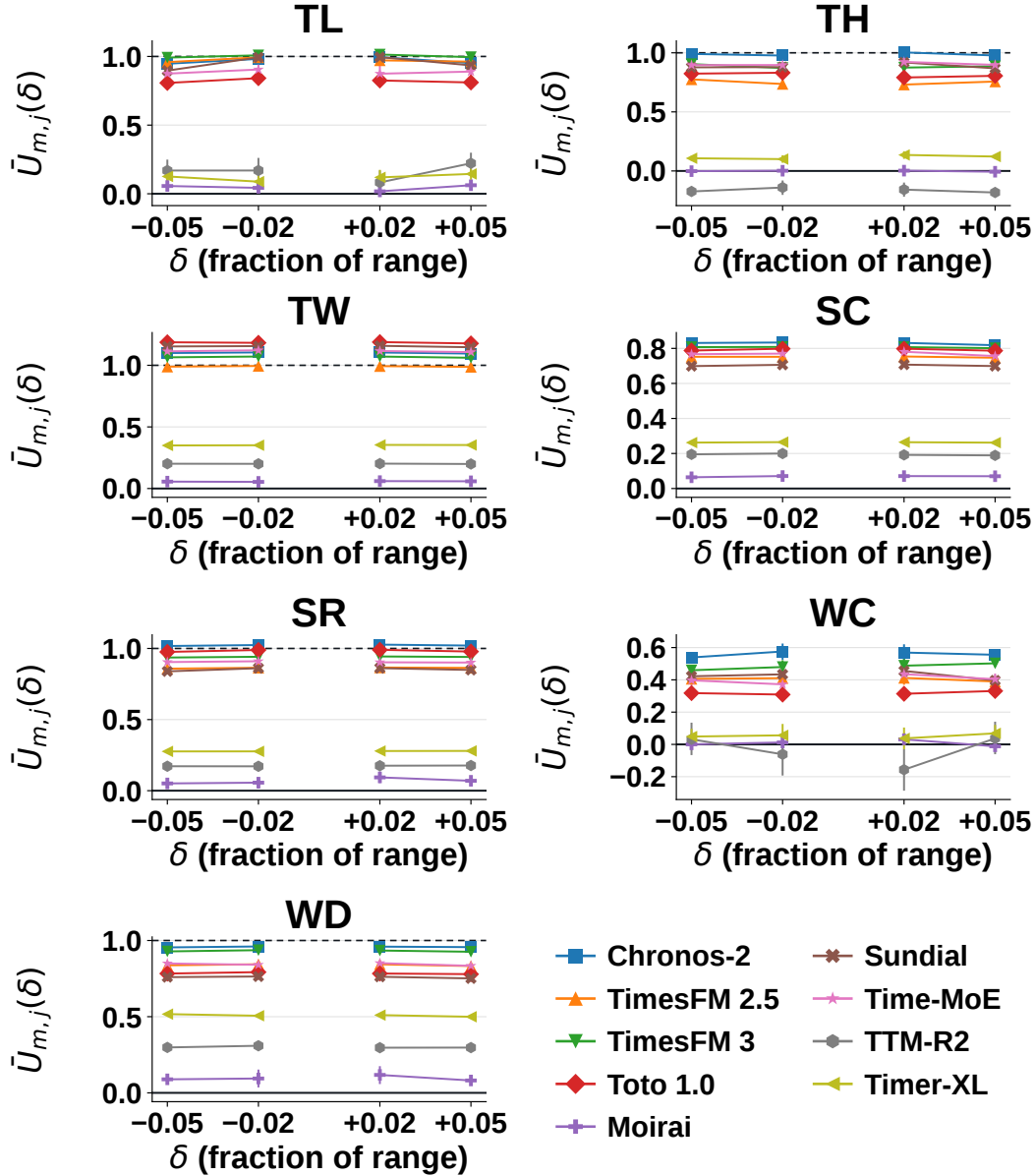


Figure 28: Conditional-reference-normalized multi-parameter response across signed intervention magnitudes. The dashed unit-reference line is shown where it lies within the panel range.

For the six responsive models, adding the scale term reduces eigenvalue-modulus (SR) responses from 0.85–1.02 to 0.61–0.75, and diffusion-time (WD) responses from 0.76–0.96 to 0.52–0.76. For amplitude ratio (TW), the maximum decreases from 1.18 to 1.09, while the frequency parameters change less. The responses remain positive, indicating that the scale term explains part, but not all, of the reference-aligned component. This regression comparison is diagnostic and does not isolate a causal normalization pathway. The four cells whose signs change under the refit all have primary responses near zero, with $|\bar{U}_{m,j}| \leq 0.12$.

Heavy-Tailed Cells Do Not Change the Conclusions. A small number of model–parameter cells have heavy-tailed per-pair response coefficients. In five of the 63 cells, $\bar{U}_{m,j}$ differs by more than 0.2 from the corresponding ratio formed using per-pair medians. Four involve weakly responding models: TTM-R2 (TT) on high-tone frequency (TH), drive frequency (WC), and low-tone frequency (TL), and Timer-XL (TX) on diffusion time (WD). Under median aggregation, the three TTM-R2 responses remain between 0.35 and 0.38, still well below the conditional-reference level. The fifth cell is TimesFM 2.5 (F2) on high-tone frequency (TH), where $\bar{U}_{m,j} = 0.75$, the median-based value is 1.01,

C2	0.99	0.97	1.02	0.97	1.02	0.98	0.97	C2	12/12	12/12	12/12	12/12	12/12	12/12	12/12	12/12
F2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	F2	6/6	6/6	6/6	6/6	6/6	6/6	6/6	6/6
F3	1.00	1.00	1.01	1.00	1.01	1.00	1.00	F3	20/20	20/20	20/20	20/20	20/20	20/20	20/20	20/20
TO	1.00	1.00	1.02	0.99	1.03	1.01	0.98	TO	12/12	12/12	12/12	12/12	12/12	12/12	12/12	12/12
MO	1.00	1.02	1.03	1.00	1.01	1.02	1.00	MO	6/6	6/6	6/6	6/6	6/6	6/6	6/6	6/6
SU	1.00	1.00	1.02	0.98	1.02	1.00	0.98	SU	12/12	12/12	12/12	12/12	12/12	12/12	12/12	12/12
ME	1.00	1.00	1.00	1.00	1.01	1.02	0.98	ME	11/11	11/11	11/11	11/11	11/11	11/11	11/11	11/11
TT	1.00	0.98	0.99	0.97	0.98	1.00	0.99	TT	3/3	3/3	3/3	3/3	3/3	3/3	3/3	3/3
TX	1.00	1.03	1.02	1.00	1.01	1.02	1.00	TX	8/8	8/8	8/8	8/8	8/8	8/8	8/8	8/8
	TL	TH	TW	SC	SR	WC	WD		TL	TH	TW	SC	SR	WC	WD	

(a) Selected-layer local response

(b) Validation-admissible layer fraction

Figure 29: Local validity of Jacobian-derived hidden-state interventions across seven jointly identified parameters. Panel (a) reports test $U_{\text{local}}(\ell)$ at the validation-selected layer, while panel (b) reports the fraction of measured layers satisfying the local-validity criterion.

Table 21: Scale-aware and unpaired response diagnostics for seven jointly identified parameters. Panel A summarizes results across nine TSFMs, with brackets indicating model minima and maxima. Panel B reports weak, reversed, or amplified response cells.

PANEL A: PARAMETER-LEVEL SUMMARY

Parameter	Model	Matched-Input Diagnostics						Unpaired Check	
		$\bar{U}_{m,j}$	$U^{(3)}$	γ	$\bar{\rho}$	$U^{*(3)}/U^*$		Coefficient	R^2
Low-tone frequency (TL)	Nine models	[0.04, 1.00]	[-0.04, 1.01]	-0.011	0.50	1.00		[0.03, 0.98]	0.94
High-tone frequency (TH)	Nine models	[-0.16, 0.99]	[-0.52, 1.01]	-0.008	0.25	1.00		[0.03, 0.98]	0.94
Amplitude ratio (TW)	Nine models	[0.06, 1.18]	[0.04, 1.09]	-0.001	0.91	0.97		[0.03, 0.98]	0.94
Eigenvalue angle (SC)	Nine models	[0.07, 0.83]	[0.05, 0.79]	-0.016	0.43	1.03		[0.09, 0.81]	0.78
Eigenvalue modulus (SR)	Nine models	[0.07, 1.02]	[0.06, 0.75]	-0.001	0.66	0.89		[0.09, 0.78]	0.76
Drive frequency (WC)	Nine models	[-0.04, 0.56]	[-0.14, 0.49]	-0.089	0.43	1.00		[-0.07, 0.88]	0.92
Diffusion time (WD)	Nine models	[0.10, 0.96]	[0.05, 0.76]	-0.001	0.76	0.99		[-0.08, 0.88]	0.92

PANEL B: WEAK, REVERSED, OR AMPLIFIED RESPONSES

High-tone frequency (TH)	Moirai (MO)	-0.00	0.00	-0.140	0.25	-	0.03	0.17
High-tone frequency (TH)	TTM-R2 (TT)	-0.16	-0.52	-0.137	0.64	-	0.16	0.63
Drive frequency (WC)	TTM-R2 (TT)	-0.04	-0.13	-0.099	0.59	-	-0.07	0.69

and the 10%-trimmed value is 0.81. The heavy tail therefore understates rather than overstates the response in this case without altering the overall response pattern.

Together, these results show that weak multi-parameter responses persist across intervention magnitudes and alternative response diagnostics. Full-vector fidelity further separates matched-input behavior from the stronger target recovery obtained through hidden-state interventions.

B.5.3 CAUSAL GEOMETRY

This analysis examines whether the jointly identified parameters admit effective and structured hidden-state interventions across model depth. Figure 29 summarizes selected-layer local validity and the fraction of validation-admissible layers, while Figure 30 shows the complete layer-wise local-response profiles. Figure 31 then characterizes how the required shared causal-subspace rank changes with depth. Finally, Figure 32 evaluates whether input-specific deployed directions improve over the fixed-direction control across admissible layers. The calibrated intervention settings are summarized in Table 8, and these analyses provide the detailed causal-geometry results underlying §4.3.

Local Causal Directions Remain Effective Across Depth. Figure 29 shows that the validation local response lies between 0.970 and 1.030 at all 630 measured layers, so every measured layer satisfies the local-validity criterion. At the validation-selected layers, the corresponding test responses also remain between 0.970 and 1.030. Across all measured test layers, the broader range is 0.942 to 1.035, indicating only limited deviation from the reference response outside the selected configurations. The calibrated intervention step converges at every layer, as reported in Table 8. These results show that the Jacobian-derived directions remain effective after nonlinear evaluation rather than succeeding only under the first-order approximation.

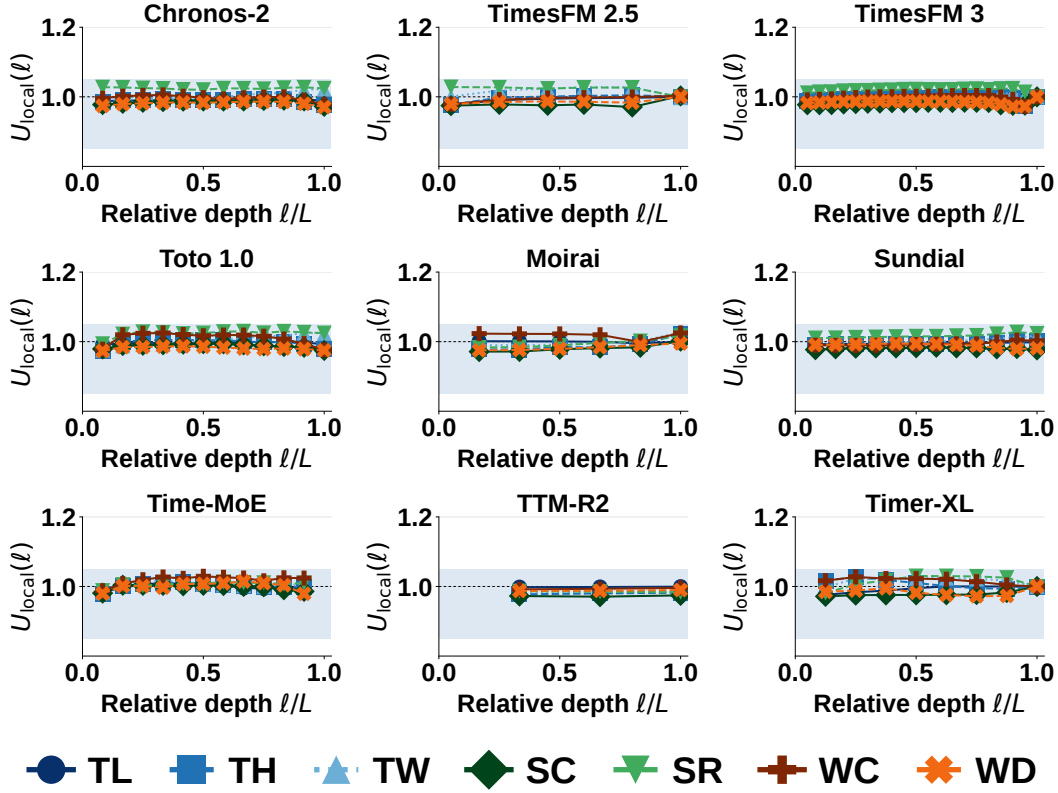


Figure 30: Validation local response across relative depth for seven jointly identified parameters and nine TSFMs. The profiles show the nonlinear effectiveness of Jacobian-derived hidden-state interventions across measured layers.

Layer-Wise Profiles Show Broad Local Validity Rather Than Isolated Valid Layers. Figure 30 provides the full depth profiles behind the aggregate validity results. The local response stays near the unit target across nearly the entire depth of each model and parameter, rather than peaking only at a small number of layers. This stability indicates that hidden-state directions that produce the reference response are available across a broad portion of the representation hierarchy. The selected layer therefore identifies a configuration for the subsequent shared-subspace and deployed analyses rather than creating the local response through a narrow layer-selection effect.

The Shared Causal Subspace Contracts with Depth. Figure 31 shows lower recorded rank-grid outcomes at the last measured layer in 58 of 63 model-parameter cells. The median displayed rank or lower bound decreases from 1,024 at the first measured layer to 8 at the last measured layer. This contraction indicates that later representations often require fewer shared directions to retain the local reference-aligned response. Values at the top grid point remain censored when the 90% response-retention threshold is not reached, so unchanged values of 1,024 do not imply unchanged true subspace dimension. Time-MoE (ME) is the clearest exception, with final-layer displayed ranks or lower bounds between 256 and 1,024, while TTM-R2 (TT) provides only three measured layers. The dominant pattern nevertheless indicates increasing concentration of shared causal-subspace structure with depth across the multi-parameter panel.

Input-Specific Coordinates Usually Improve on a Fixed Direction. Figure 32 compares the deployed coordinate-readout intervention with the fixed-direction control across validation-admissible layers. The median deployed-minus-fixed response difference is +0.19, and the deployed direction remains better at the final admissible layer in 58 of 63 model-parameter cells. For TTM-R2 (TT), the deployed direction exceeds the fixed direction at every admissible layer for all seven parameters, with a minimum advantage of +0.11, despite its weak matched-input responses. This separation indicates that the unperturbed hidden state contains information useful for selecting input-specific coordinates even when the original input intervention does not produce the desired forecast response. The advantage is not uniform across all models and depths. Time-MoE (ME) and Timer-XL (TX) have

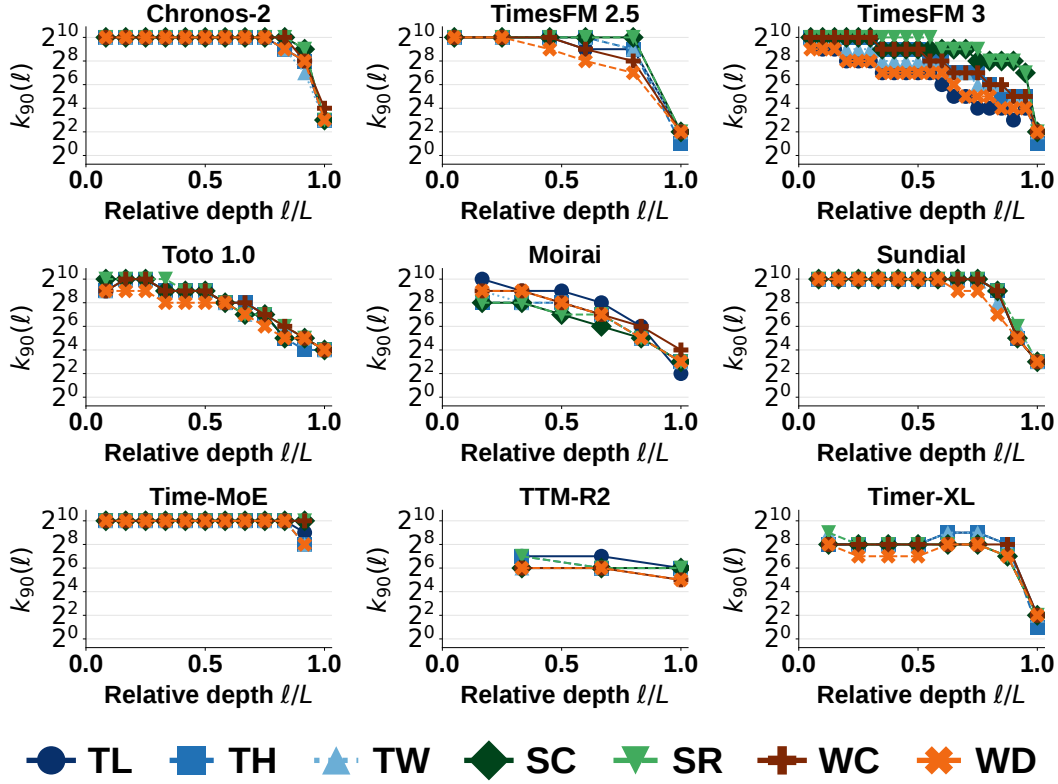


Figure 31: Shared forecast-control rank across relative depth for seven jointly identified parameters and nine TSFMs. Values at the top grid point 2^{10} include unreached response-retention thresholds and therefore represent lower bounds where censored.

only two and one parameters, respectively, with positive deployed-over-fixed gain at every admissible layer, with minimum differences of -0.62 and -0.47 .

Together, the four analyses separate three properties of the multi-parameter causal geometry. Local directions establish that the selected hidden states can support the reference response, shared-subspace results show that these directions often become more compact with depth, and deployed-over-fixed gains show that input-specific coordinates can often be inferred from the unperturbed representation. The jointly identified parameters therefore retain broadly valid hidden-state response pathways whose shared structure becomes more concentrated with depth, while coordinate prediction remains model- and parameter-dependent across the evaluated multi-parameter settings.

B.6 ADDITIONAL RESULTS ON REAL-SERIES BACKGROUNDS

This subsection extends the controlled real-series evaluation in §4.4 from ETTh1 to Electricity and Weather. All datasets use the same linear trend and sinusoid injections. Injection strength is the component root-mean-square (RMS) amplitude relative to the context standard deviation, evaluated at 0.25, 0.5, and 1.0. Solid lines denote linear trend and dashed lines denote sinusoid. Figure 33 shows the complete cross-dataset results across all settings.

ETTh1. Panel (a) shows substantial model variation under the same injected dynamics. Several models retain negative linear trend responses across all three injection strengths, while sinusoid responses are generally closer to the expected positive direction. At strength 1.0, the pooled linear trend gain remains -0.62 , whereas the sinusoid gain reaches 0.67 . Increasing the injected amplitude therefore improves the response for many models but does not eliminate the trend-response mismatch.

Figure 34 shows one representative ETTh1 matched pair at injection strength 0.5. The left panel shows that the two inputs share the same ETTh1 background and differ only in the controlled linear trend component. The right panel compares the resulting Chronos-2 (C2) forecast change with the analytic reference response across the forecast horizon. The model follows the same response

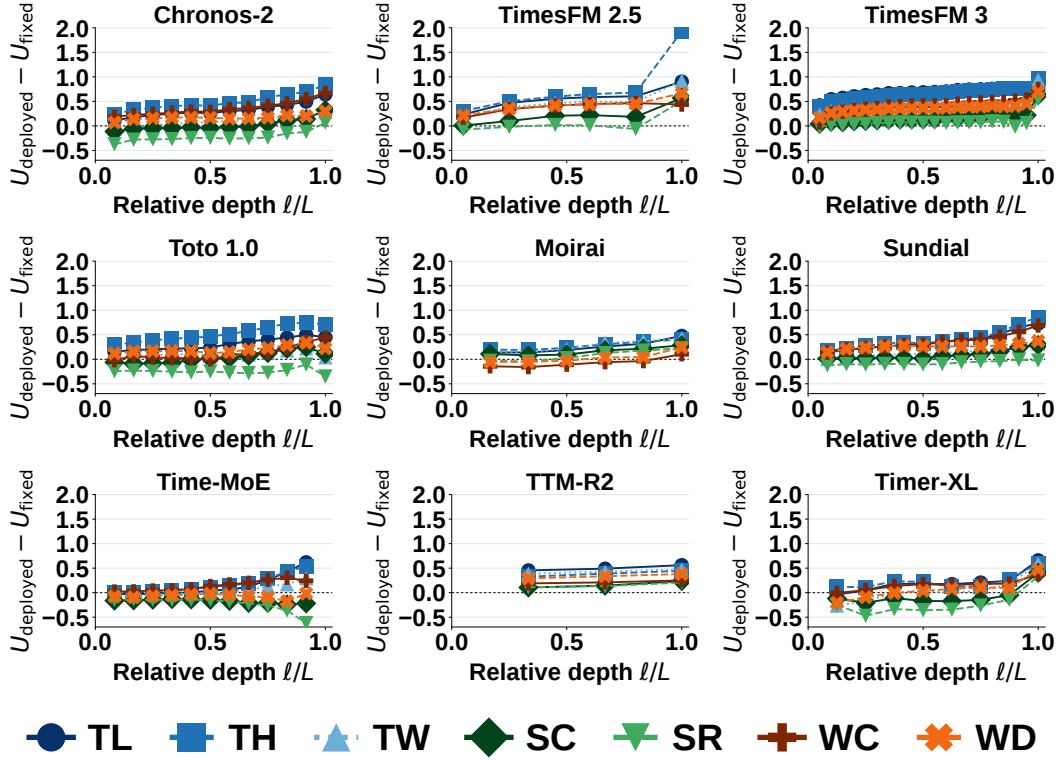


Figure 32: Deployed-over-fixed response gain across validation-admissible layers for seven jointly identified parameters and nine TSFMs. Positive values indicate an advantage of the input-specific deployed direction over the fixed-direction control.

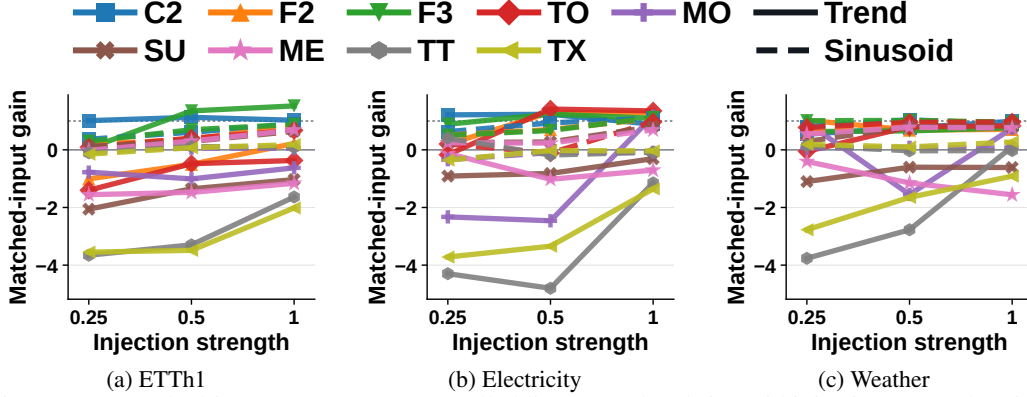


Figure 33: Matched-input responses to controlled linear trend and sinusoid injections on real-series backgrounds. Injection strength is component root-mean-square (RMS) amplitude relative to context standard deviation for all three evaluated real datasets.

direction but remains more negative than the analytic reference across much of the horizon. This example illustrates that a precisely controlled input change need not produce the corresponding analytic continuation uniformly across forecast steps.

Electricity. Panel (b) shows a stronger dependence on injection strength. Several models have strongly negative linear trend responses at strengths 0.25 and 0.5, but move substantially toward the injected continuation at strength 1.0. The pooled linear trend gain reaches 0.89 at strength 1.0, although some models remain below the reference response. Sinusoid responses are generally positive and vary less strongly with injection strength.

Weather. Panel (c) shows a similar but model-dependent transition. The pooled linear trend gain reaches 0.71 at strength 1.0, while sinusoid gain reaches 0.77. Several models nevertheless retain

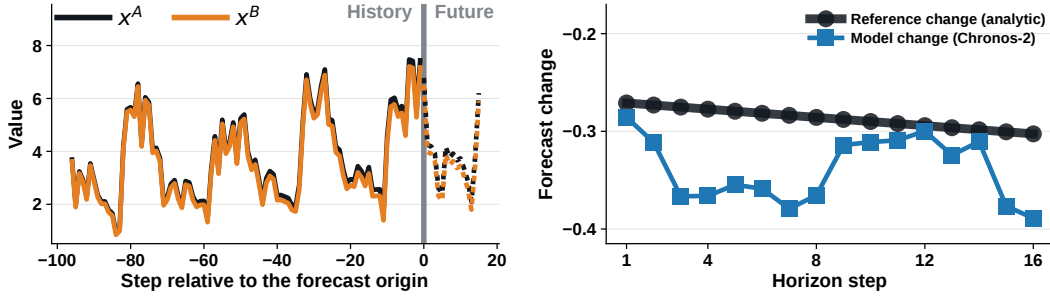


Figure 34: Illustrative ETTh1 linear trend injection and Chronos-2 (C2) forecast-change traces at injection strength 0.5 for one representative matched input pair.

Table 22: Thresholded agreement and cross-model correlations among representation accessibility, matched-input response, and deployed response. Counts use original-bank cells with $D_{m,j} \geq 0.95$, so the trend comparison is exploratory rather than common-regime.

Law	Cells With $D_{m,j} \geq 0.95$				Spearman ρ	
	All	$\bar{U}_{m,j} < 0$	$0 \leq \bar{U}_{m,j} < 0.5$	$\bar{U}_{m,j} > 1.5$	$D_{m,j}, \bar{U}_{m,j}$	$U_{\text{deployed}}, \bar{U}_{m,j}$
Linear trend (TR)	9	3	2	0	0.62	0.78
Oscillation (OS)	0	—	—	—	0.92	0.78
Mean reversion (MR)	8	0	1	0	0.92	−0.03
Saturating growth (SG)	9	2	2	3	0.18	0.45
Damped oscillation (DO)	8	0	2	0	0.38	−0.20
Seasonal autoregression (SA)	0	—	—	—	0.82	0.32
Transient pulse (TP)	9	0	8	0	0.78	−0.03
All	43	5	15	3	0.33	0.37

weak or negative linear trend responses, particularly at smaller injection strengths. Increasing component amplitude can therefore strengthen the expected response without producing uniform behavior across models.

Cross-Dataset Comparison. Across the three backgrounds, the same controlled dynamical change produces substantially different forecast responses across models and datasets. Stronger injections often move the response toward the injected continuation, but both its magnitude and direction remain dependent on the underlying background. The ETTh1 example in Figure 34 further shows that this mismatch can vary across the forecast horizon within a single matched pair. Together, these results extend the main-text finding by showing that high representation accessibility can coexist with weak, reversed, or horizon-dependent forecast responses when controlled dynamics are embedded in real-series backgrounds across the evaluated real-series settings.

B.7 CROSS-MEASURE AND MECHANISTIC DIAGNOSTICS

The main evaluation shows that representation accessibility, forecast response, and deployed response can diverge for the same Dynamical Parameter. This subsection tests the robustness of this separation and examines mechanisms underlying model-specific trend responses. We first compare the three measurements across the model panel and test whether weak response can be explained by ordinary forecasting competence. We then examine reversed linear trend (TR) responses using law-identification controls, TimesFM 3 (F3) detrending, continuation behavior, a shape-and-scale decomposition, and normalization replay across several targeted mechanistic diagnostic settings.

B.7.1 CROSS-MEASURE ROBUSTNESS

This analysis tests whether the separation among representation accessibility, matched-input forecast response, and deployed aligned gain persists across the single-parameter panel and among models with stronger ordinary forecasting performance. Table 22 summarizes thresholded disagreements and cross-model correlations among the three measurements. Table 23 and Figure 35 then examine whether weak matched-input response can be explained by ordinary forecast competence. Because

the cross-measure comparison uses the original accessibility datasets, the common-regime results in §4.2.2 remain the primary comparison between accessibility and forecast response.

The Three Measurements Remain Distinct Across the Panel. As shown in Table 22, among the 43 model–parameter cells with accessibility $D_{m,j} \geq 0.95$, five have negative matched-input response, 15 have response between 0 and 0.5, and three exceed 1.5. These disagreements vary substantially across dynamical laws. Accessibility and matched-input response are rank-correlated across models for oscillation (OS), mean reversion (MR), and seasonal autoregression (SA), with correlations between 0.82 and 0.92, while the correlation falls to 0.38 for damped oscillation (DO). Deployed and matched-input responses also need not agree, indicating that natural forecast expression and response capacity under predicted hidden-state intervention remain distinct measurements across the evaluated model–parameter cells and across different dynamical response regimes.

Forecast Competence Does Not Eliminate Response Mismatch. We next test whether weak matched-input response simply reflects poor ordinary forecasting. For each law, normalized forecast regret is defined relative to the conditional reference and fixed baseline as

$$\frac{L(F_\theta, y) - L(F_m^*, y)}{L(F_{\text{base}}, y) - L(F_m^*, y) + \epsilon},$$

where L is horizon mean-squared error, F_m^* is the conditional reference, and F_{base} is a fixed law-specific baseline. A separate validation dataset of 4,096 histories per law determines the forecast-competence subsets before accessibility and response are evaluated on their respective test datasets.

As shown in Figure 35, each point represents one model–parameter cell, with color indicating the dynamical law and marker shape indicating the TSFM. The horizontal axis reports test forecast regret, where smaller values indicate forecasts closer to the conditional reference relative to the fixed baseline. Filled markers denote cells selected as forecast competent using validation regret below one, while hollow markers denote the remaining cells. Because competence is selected on the validation dataset, the filled and hollow markers need not fall on opposite sides of the vertical test-regret line. The highlighted TimesFM 2.5 (F2) markers compare its two forecast heads under the same response evaluation for a controlled within-model head comparison.

The left panel plots matched-input aligned gain g against test forecast regret. A gain near 1 indicates a response with the expected reference-aligned magnitude, a value near 0 indicates little aligned response, and a negative value indicates an opposing response. Although lower-regret cells tend to have larger gain, cells with similar forecast regret span a wide range of response gains. Several validation-selected competent cells still have weak or negative gain, showing that accurate ordinary forecasts do not guarantee a reference-aligned response to the parameter intervention.

The right panel plots centered response error e , where smaller values indicate closer agreement with the full reference trajectory. Lower-regret cells generally have smaller response error, but substantial variation remains even among the competent cells. Several filled markers have centered error near or above 1, indicating that their forecast-change trajectories can remain far from the reference despite competent ordinary forecasting. The overlap of competent and non-competent cells in both panels therefore shows that forecast quality alone does not determine matched-input response fidelity.

Consistent with the scatter plots, test regret and matched-input gain have Spearman correlation $\rho = -0.41$, with a 95% interval of $[-0.60, -0.20]$. Among the 49 cells with validation regret below one, five still have negative matched-input gain, while median accessibility remains 0.976. As summarized in Table 23, their median matched-input gain and centered error are 0.64 and 0.92, respectively. The stricter regret-below-0.5 and within-law best-tercile subsets increase median gain only to 0.65 and 0.69, while centered errors remain 0.92 and 0.83. Ordinary forecasting competence therefore reduces, but does not remove, the separation between parameter accessibility and response fidelity across the evaluated single-parameter model panel.

B.7.2 TREND-RESPONSE MECHANISMS

The reversed linear trend (TR) responses provide a concrete setting for examining why accessible parameter information need not produce the expected forecast change. We first test whether the reversing models confuse trend with other dynamical laws using Table 24. We then examine the TimesFM 3 (F3) detrending pathway in Table 25 and compare model continuation behavior in Table 26. Next, Table 27 separates the standardized-shape and window-scale contributions to

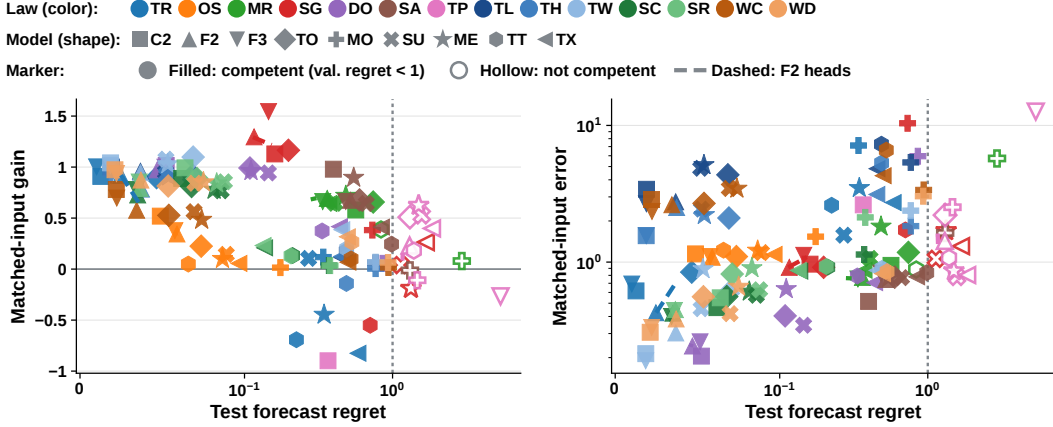


Figure 35: Forecast regret versus matched-input response fidelity across the single-parameter panel. The figure highlights validation-selected competent forecasters and compares the two TimesFM 2.5 (F2) forecast heads under the same response evaluation setting.

Table 23: Single-parameter cells stratified by forecast competence. $D_{m,j}$ denotes original-bank representation accessibility, while g and e are pooled matched-input response metrics. Subset thresholds use validation regret, with metrics evaluated on the corresponding test banks.

Subset	Cells	Median $D_{m,j}$	Median g	Median e
All	63	0.985	0.42	1.06
Regret < 1	49	0.976	0.64	0.92
Regret < 0.5	36	0.990	0.65	0.92
Best tercile	21	0.989	0.69	0.83

the trend response. Finally, Table 28 tests whether the standard normalization pipeline conceals or alters the observed response. Together, these diagnostics examine the mechanisms underlying the trend-response patterns reported in §4.2.2.

(I) Reversed Trend Response Is Not Explained by Law Confusion. Table 24 uses the trend row of the thirteen-law identification experiment in §4.1 to test whether reversed trend responses coincide with confusion between linear trend (TR) and related dynamics. No model assigns more than 0.7% of true trend sequences to mean reversion (MR), and the three models with negative trend response assign none. Time-MoE (ME), TTM-R2 (TT), and Timer-XL (TX) classify 94.0%, 87.0%, and 96.3% of trend sequences correctly, respectively. For TTM-R2 (TT), the remaining errors fall primarily on random walk with drift (RW) at 4.9% and the remaining laws at 8.1%, rather than on mean reversion. The validation results further show that trend identity becomes substantially more accessible with depth for several models. For example, Time-MoE (ME) increases from 50.4% trend accuracy at the first layer to 94.6% at the selected layer, while TimesFM 3 (F3) increases from 62.8% to 94.8% and Moirai (MO) from 59.1% to 92.2%. Thus, models with weak or reversed trend responses can still develop hidden representations that distinguish the trend law accurately across multiple depths and model architectures. The response reversal therefore does not appear to arise from a general failure to identify linear-trend dynamics in the inspected representations.

(II) TimesFM 3 Uses a Detrending Pathway That Bypasses the Inspected State. TimesFM 3 (F3) provides the opposite pattern from the reversing models. On the wider trend-response histories, its slope accessibility falls to $D_{m,S} = 0.372$, while its conditional-reference-normalized response reaches $\bar{U}_{m,S} = 1.01$. Table 25 shows that the detrending gate fires on 48.0% of the 32,768 response histories, with a sharp dependence on absolute slope. The gate remains inactive for $|S| < 0.2$, fires on only 17.6% of histories in $[0.200, 0.234)$, rises to 92.5% in $[0.234, 0.300)$, and reaches 100% for $|S| \geq 0.300$. This transition explains why the wider slope distribution used for forecast-response evaluation exposes a different representation regime from the narrow-slope accessibility setting. When the gate fires, an ordinary least-squares trend is removed before the transformer, while the fitted trend is extrapolated and added back outside the inspected hidden-state pathway.

The gate-stratified accessibility results further support this mechanism. For histories on which the gate fires, linear slope accessibility remains near zero across layers, ranging from -0.018 to 0.001 .

Table 24: Trend-law classification and validation accuracy across nine TSFMs. Classification columns report the percentage of true linear-trend (TR) sequences assigned to each law group.

Model	True Trend Classified As (%)				Validation Trend Accuracy (%)		
	Trend	Mean Rev.	Rand. Walk	Other	Layer 1	Selected	Selected Layer
Chronos-2 (C2)	97.2	0.0	1.4	1.4	93.7	97.1	10/12
TimesFM 2.5 (F2)	97.6	0.0	0.9	1.5	92.8	97.2	10/20
TimesFM 3 (F3)	94.6	0.7	0.7	4.0	62.8	94.8	16/20
Toto-Open-Base 1.0 (TO)	96.1	0.0	2.0	1.9	82.8	96.2	6/12
Moirai (MO)	92.9	0.0	4.4	2.7	59.1	92.2	5/6
Sundial (SU)	96.2	0.0	1.9	1.9	82.3	97.5	12/12
Time-MoE (ME)	94.0	0.0	3.5	2.6	50.4	94.6	12/12
TTM-R2 (TT)	87.0	0.0	4.9	8.1	82.0	87.5	3/3
Timer-XL (TX)	96.3	0.0	2.3	1.3	84.9	95.0	5/8

Table 25: TimesFM 3 (F3) detrending-gate firing rate across 32,768 trend histories grouped by absolute slope. The gate remains inactive at small slopes and becomes nearly universal once the slope exceeds the analytic crossover region across the evaluated response-history distribution.

Absolute-Slope Bin $ S $	Histories	Gate Fires
All	32,768	48.0%
[0.000, 0.100)	7,324	0.0%
[0.100, 0.200)	7,302	0.0%
[0.200, 0.234)	2,480	17.6%
[0.234, 0.300)	4,798	92.5%
[0.300, 0.450)	10,864	100.0%

For histories on which the gate remains inactive, accessibility instead ranges from 0.990 to 0.993. These two regimes together produce the substantially lower aggregate accessibility of 0.372 on the response-history distribution. Disabling the detrending step restores accessibility to 0.998–0.999 across layers, including 0.999 on histories that would otherwise activate the gate. The detrending pathway therefore accounts for the measured loss of slope accessibility in the inspected transformer states while preserving a separate route through which the fitted trend contributes to the forecast.

(III) Continuation Behavior Tracks the Direction of Trend Response. The trend-response experiment measures how the forecast changes under a finite slope intervention. A complementary diagnostic measures the model’s continuation behavior without intervention. For an unintervened window, we regress the tilt of the standardized forecast on the standardized input slope and denote the resulting continuation gain by k . A value of $k = 1$ indicates a unit average continuation component, $k = 0$ indicates no such component, and $k < 0$ indicates an opposing component.

As shown in Table 26, continuation behavior differs strongly across models. Chronos-2 (C2), TimesFM 2.5 (F2), TimesFM 3 (F3), and Toto-Open-Base 1.0 (TO) have positive continuation gains between 0.79 and 0.94. Time-MoE (ME), TTM-R2 (TT), and Timer-XL (TX) instead have gains of -0.82 , -0.86 , and -1.36 , respectively. Across the nine models, continuation gain and conditional-reference-normalized trend response are rank-correlated at 0.97, showing that the intervention response closely follows the models’ underlying continuation behavior across the evaluated trend-response model panel. The slope-bin results further show that this behavior varies with the baseline slope, particularly for the models with opposing continuation.

TimesFM 2.5 (F2) further isolates the role of the output head. Its pinball-trained median head has continuation gain 0.81, while its MSE-trained mean head has gain 0.54, despite sharing the same trunk and input representation. Their matched-input trend gains are 0.92 and 0.65, respectively. The common trunk therefore does not determine a unique continuation policy, since the two forecast heads produce different continuation and intervention responses.

(IV) Shape and Scale Carry Complementary Parts of the Trend Response. For models using whole-window standardization, the end-to-end trend response contains both a change in the standardized forecast shape and a contribution introduced when the forecast is returned to the original window scale. Writing u_A and u'_B for the two standardized inputs, g for the map from standardized input to standardized forecast, and σ_A and σ_B for the window standard deviations, the raw forecast change

Table 26: Continuation gain across nine TSFMs, reported overall and by absolute-slope bin. The TimesFM 2.5 (F2) rows additionally compare its median and mean forecast heads.

Model	Overall k	Continuation Gain k by $ S $			k^*	$\bar{U}_{\text{TR}}$
		$[0, 0.05)$	$[0.05, 0.2)$	$[0.2, 0.45)$		
Chronos-2 (C2)	+0.87	+0.54	+0.84	+0.89	+1.00	+0.91
TimesFM 2.5 (F2)	+0.81	+0.50	+0.69	+0.86	+1.00	+0.92
Median head (pinball)	+0.81	—	—	—	—	—
Mean head (MSE)	+0.54	—	—	—	—	—
TimesFM 3 (F3)	+0.94	+0.32	+0.80	+0.99	+1.00	+1.01
Toto-Open-Base 1.0 (TO)	+0.79	+0.14	+0.70	+0.84	+1.00	+0.90
Moirai (MO)	−0.20	−0.61	−0.50	−0.09	+1.00	+0.12
Sundial (SU)	−0.18	−1.05	−0.56	−0.02	+1.00	+0.11
Time-MoE (ME)	−0.82	−1.89	−1.53	−0.54	+1.00	−0.45
TTM-R2 (TT)	−0.86	−2.89	−1.10	−0.75	+1.00	−0.69
Timer-XL (TX)	−1.36	−1.00	−1.76	−1.21	+1.00	−0.82

can be written as the sum of two complementary components:

$$\begin{aligned} \sigma_B g(u'_B) - \sigma_A g(u_A) &= \underbrace{\sigma_A [g(u'_B) - g(u_A)]}_{\text{shape}} \\ &\quad + \underbrace{(\sigma_B - \sigma_A) g(u'_B)}_{\text{scale}}. \end{aligned}$$

The first term captures the network response to the standardized-shape change across the forecast horizon. The second captures the contribution of changing the output scale while holding the arm-B standardized forecast fixed. This end-to-end decomposition is distinct from the deployed coordinate-readout intervention defined in §3.4.

As shown in Table 27, the relative contributions of the two channels vary substantially across models and slope regimes. Chronos-2 (C2) remains positive in both channels, with overall shape and scale contributions of +0.39 and +0.53, producing a live response of +0.92 across the evaluated trend-response slope regimes. The three reversing models instead have negative overall scale contributions, ranging from −0.22 for Time-MoE (ME) to −0.77 for Timer-XL (TX).

In the lowest absolute-slope bin, the standardized-shape contribution is itself negative for all three reversing models. It reaches −1.60 for Time-MoE (ME), −2.93 for TTM-R2 (TT), and −0.77 for Timer-XL (TX). The corresponding scale contributions are smaller in magnitude, showing that the strong reversal in this regime arises primarily from the standardized forecast response rather than from output rescaling alone across these low-slope intervention settings.

The balance changes as the baseline slope increases. In the highest slope bin, Time-MoE (ME) and Timer-XL (TX) have positive shape contributions of +0.42 and +0.93, while negative scale contributions of −0.33 and −1.03 reduce the corresponding live responses to +0.09 and −0.10. TTM-R2 (TT) shows the same competition in the middle bin, where a positive shape contribution of +0.57 is offset by a scale contribution of −0.61, leaving a live response near zero at −0.04. The decomposition therefore shows that trend-response reversal can arise through different balances of standardized forecast behavior and output scaling, with those balances changing across models and slope regimes throughout the evaluated trend-response conditions.

(V) Normalization Replay Does Not Reveal a Hidden Positive Trend Response. We finally test whether whole-window normalization conceals the expected linear-trend (TR) response. The replay covers Chronos-2 (C2), Moirai (MO), Sundial (SU), Time-MoE (ME), TTM-R2 (TT), and Timer-XL (TX), whose evaluated interfaces standardize the context using its mean and standard deviation. For the frozen-statistics replay, arm B is forecast using the normalization statistics of arm A. An amplitude-only control separately measures the response induced by presenting the arm-A shape at the arm-B amplitude under the same frozen statistics.

As shown in Table 28, freezing the normalization statistics does not recover a positive trend response for the weak or reversing models. Moirai (MO) changes from +0.12 to −0.46, Sundial (SU) from +0.11 to −1.69, Time-MoE (ME) from −0.45 to −1.25, TTM-R2 (TT) from −0.69 to −2.70, and Timer-XL (TX) from −0.82 to −2.40. Chronos-2 (C2) changes only modestly from +0.91

Table 27: Trend response decomposed into standardized-shape and window-scale contributions across slope bins. Rows report live response together with the corresponding shape and scale components for each evaluated model and slope regime.

Model	Absolute-Slope Bin $ S $	Pairs	Live	Shape	Scale
Chronos-2 (C2)	[0, 0.05)	122	+0.86	+0.76	+0.09
	[0.05, 0.2)	357	+0.96	+0.58	+0.38
	[0.2, 0.45)	545	+0.91	+0.18	+0.73
All		1,024	+0.92	+0.39	+0.53
Time-MoE (ME)	[0, 0.05)	122	-1.33	-1.60	+0.27
	[0.05, 0.2)	357	-0.98	-0.74	-0.23
	[0.2, 0.45)	545	+0.09	+0.42	-0.33
All		1,024	-0.45	-0.23	-0.22
TTM-R2 (TT)	[0, 0.05)	237	-3.22	-2.93	-0.29
	[0.05, 0.2)	709	-0.04	+0.57	-0.61
	[0.2, 0.45)	1,102	-0.56	-0.10	-0.46
All		2,048	-0.69	-0.20	-0.49
Timer-XL (TX)	[0, 0.05)	237	-0.91	-0.77	-0.13
	[0.05, 0.2)	709	-1.92	-1.35	-0.57
	[0.2, 0.45)	1,102	-0.10	+0.93	-1.03
All		2,048	-0.83	-0.06	-0.77

Table 28: Trend response under the standard normalization pipeline, frozen statistics, and an amplitude-only control. The residual reports the difference between the frozen-statistics response and the sum of the live and amplitude-only responses across all evaluated normalization replay models.

Model	Live Pairs	Live	Frozen	Amplitude	Residual
Chronos-2 (C2)	32,768	+0.91	+0.81	-0.10	-0.00
Moirai (MO)	131,072	+0.12	-0.46	-0.56	-0.02
Sundial (SU)	32,768	+0.11	-1.69	-1.76	-0.03
Time-MoE (ME)	32,768	-0.45	-1.25	-0.80	+0.01
TTM-R2 (TT)	32,768	-0.69	-2.70	-2.02	+0.00
Timer-XL (TX)	32,768	-0.82	-2.40	-1.58	+0.00

to +0.81. Thus, holding the normalization statistics fixed generally strengthens rather than removes the weak or reversed behavior.

The amplitude-only control closely accounts for these changes. Its responses are -0.56, -1.76, -0.80, -2.02, and -1.58 for the five weak or reversing models, compared with only -0.10 for Chronos-2 (C2). After subtracting the live and amplitude-only contributions from the frozen-statistics response, the residual remains within 0.03 of zero for every model. The large changes under frozen statistics are therefore closely associated with the accompanying amplitude perturbation rather than with recovery of an expected slope-response pathway. The normalization replay consequently provides no evidence that the standard normalization pipeline is concealing a positive trend response in these models across the evaluated normalization replay settings.

In Summary. Together, these diagnostics rule out several simple explanations for the observed accessibility-response separation. Weak response persists among models with competitive ordinary forecasting performance, and reversed trend response is not explained by confusion between linear trend and mean reversion. The trend-specific analyses instead show that preprocessing pathways, continuation behavior, standardized forecast shape, and output scaling can each alter how a slope change reaches the final forecast across different models and response regimes. These mechanisms help explain why parameter information that is accessible in a hidden representation need not be expressed through the corresponding end-to-end forecast response.

B.8 CONDITIONAL-REFERENCE SENSITIVITY

The main evaluation conditions parameter and response references on the generating law and documented nuisance variables. To test sensitivity to this information, the analysis compares three settings: known law with documented nuisances, known law with nuisances marginalized under

Table 29: Sensitivity of parameter correlation and pair-weighted response-reference gain to nuisance conditioning. Response columns report reference values rather than TSFM scores.

Parameter	r_{cond}^*	r_{hist}^*	U_{cond}^*	U_{hist}^*
Seasonal autoregression (SA)	0.975	0.956	0.848	0.883
Transient pulse (TP)	0.997	0.989	0.983	0.768
Damped oscillation (DO)	0.986	0.986	0.957	0.958
Drive frequency (WC)	1.000	1.000	1.000	0.917
Diffusion time (WD)	0.991	0.942	0.960	0.971

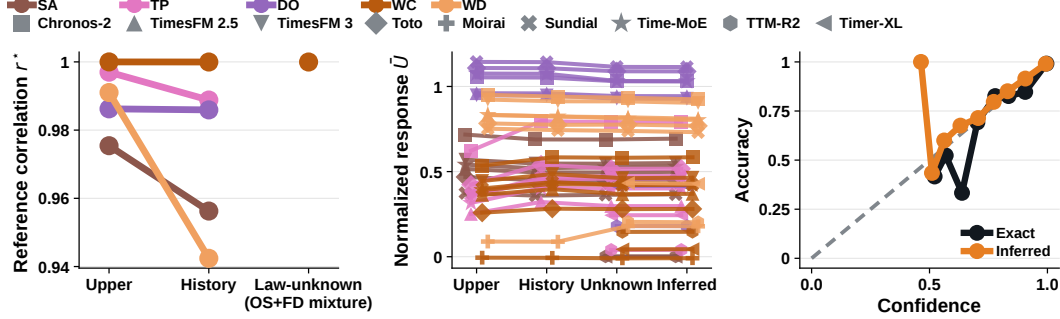


Figure 36: Reference and normalized-score changes when nuisance variables or the generating law are not supplied. Left: conditional and history-only parameter correlations. Middle: model responses normalized by each reference. Right: calibration of exact and classifier-based law posteriors.

the generator priors, and an eleven-law mixture with the law posterior inferred from history. The primary references are $\mathbb{E}[z | x, m, \eta]$ for the parameter and $F_m^*(x; \eta)$ for the response. All settings use identical histories, signed doses, future targets, and scoring rules.

The comparison covers seasonal autoregression (SA), transient pulse (TP), damped oscillation (DO), drive frequency (WC), and diffusion time (WD). Parameter correlations use 8,192 histories, known-law response references use 4,096 pairs per signed dose, and law-unknown references use 2,048 pairs. On a balanced dataset of 2,200 histories, the exact and calibrated discriminative law posteriors achieve accuracies of 0.964 and 0.955, with expected calibration errors of 0.011 and 0.006. Numerical refinement changes every retained metric by at most 0.0064.

Table 29 and Figure 36 show that parameter references are generally more stable than response normalization when conditioning information is reduced.

Parameter References Remain Stable. The left panel of Figure 36 shows that history-only parameter correlations remain high for all five evaluated parameters. Damped oscillation (DO) and drive frequency (WC) are unchanged at 0.986 and 1.000, while seasonal autoregression (SA) and transient pulse (TP) decrease only from 0.975 to 0.956 and from 0.997 to 0.989, respectively. Diffusion time (WD) shows the largest reduction, from 0.991 to 0.942, but remains strongly correlated with the conditional reference. Thus, removing nuisance information changes parameter recovery only modestly across the evaluated laws.

Response References Are More Sensitive. The middle panel of Figure 36 shows larger and more heterogeneous changes when model responses are renormalized using reduced-information references. Consistent with Table 29, seasonal autoregression (SA), damped oscillation (DO), and diffusion time (WD) change only modestly between conditional and history-only response references. Transient pulse (TP) decreases from 0.983 to 0.768, while drive frequency (WC) decreases from 1.000 to 0.917. For transient pulse (TP), pooled gain instead increases from 0.325 to 1.746 because pooling places greater weight on higher-energy pairs. The model-level points in the middle panel further show that these reference changes do not shift all TSFMs uniformly, indicating that the choice of response reference can alter both normalization magnitude and cross-model comparisons.

Law Uncertainty Does Not Dominate the Result. The right panel of Figure 36 compares calibration for the exact and inferred law posteriors. Both remain close to the diagonal over most confidence levels, consistent with accuracies of 0.964 and 0.955 and expected calibration errors of 0.011 and 0.006. Using the eleven-law mixture, the 0.05 agreement criterion is satisfied for

Table 30: Median target and non-target errors across 64 horizon runs. Quantile metrics use linearly extrapolated changes for Chronos-2 (C2) and TimesFM 3 (F3), with coverage change measured for the 50% interval across the two evaluated quantile models.

Condition	Target e	Non-target	Relative Δ CRPS	Δ Coverage
Matched input	1.10	0.325	—	—
Deployed	0.64	0.177	0.000	-0.001
Projected	0.21	0.030	-0.002	-0.001
Fixed	1.12	0.155	0.013	-0.005
Equal-norm random	1.00	0.423	0.025	-0.015

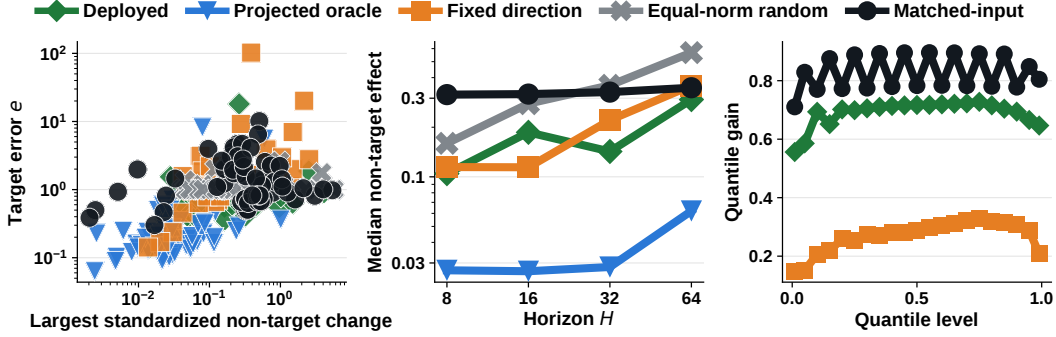


Figure 37: Target fidelity, non-target deviations, and quantile-level response changes for hidden-state edits. Quantile diagnostics apply only to Chronos-2 (C2) and TimesFM 3 (F3).

seasonal autoregression (SA), transient pulse (TP), drive frequency (WC), and diffusion time (WD). Damped oscillation (DO) is the only exception, with a pooled difference of 0.058. These results indicate that the main sensitivity arises more from how the response reference changes under reduced conditioning than from inaccurate inference of the generating law across the evaluated reference settings. Overall, parameter-reference correlation remains comparatively stable, whereas response normalization is more dependent on the information supplied to the reference.

B.9 INTERVENTION SIDE EFFECTS AND OUTPUT SEMANTICS

The causal-geometry evaluation in §4.2.3 measures whether hidden-state edits recover the reference forecast response. To test whether this recovery introduces unintended output changes, we evaluate target fidelity, non-target forecast properties, and distributional effects across 64 combinations of model, parameter, and forecast horizon. Target fidelity is measured by centered error against the desired parameter response. The non-target score is the largest standardized deviation in forecast variance, first-difference energy, or lag-one autocorrelation from the ideal edited forecast. For forced diffusion, edits targeting drive frequency (WC) are also projected onto the sibling diffusion-time (WD) response to measure cross-coordinate effects.

Chronos-2 (C2) and TimesFM 3 (F3) additionally provide quantile forecasts. Their per-unit-step quantile changes are linearly extrapolated before comparing relative continuous ranked probability score (CRPS), 50%-interval coverage, and quantile crossing with the corresponding arm-B forecast. These distributional diagnostics therefore apply only to the two quantile models.

Target Fidelity and Non-Target Effects. Table 30 shows that projected edits achieve the lowest median target error and non-target deviation, at 0.21 and 0.030, respectively. Deployed edits reduce target error from 1.00 for equal-norm random edits to 0.64, while reducing non-target deviation from 0.423 to 0.177. Fixed edits have a slightly smaller non-target deviation of 0.155, but their target error rises to 1.12, indicating that a small side effect alone does not imply successful target control. The left panel of Figure 37 shows the same tradeoff across individual runs. Projected edits concentrate toward low target error and small non-target change, while deployed edits generally lie between the projected condition and the fixed or random controls. Thus, improved target alignment is not explained simply by applying an edit of comparable magnitude.

Side Effects Vary With Forecast Horizon. The middle panel of Figure 37 shows how non-target effects change with forecast horizon. Projected edits remain lowest across horizons, whereas the

Table 31: Computational cost at $H = 16$. Jacobian construction and inversion are measured per pair, deployment per input, and basis construction once. Shared-device contention and backend differences preclude a general speed comparison across the four evaluated TSFM architectures.

Model	Jacobian (s)	Inversion (ms)	Basis (s)	Peak (GB)	Deployment (ms)
Chronos-2 (C2)	0.003	0.04	6.2	11.0	6.3
TimesFM 3 (F3)	0.007	0.02	4.6	14.6	22.2
Time-MoE (ME)	0.036	0.08	11.4	21.1	9.2
TTM-R2 (TT)	0.002	0.03	5.2	10.1	0.2

deviations of deployed, fixed, and equal-norm random edits generally increase as the horizon grows. The deployed condition nevertheless remains below the equal-norm random control across the evaluated horizons. This pattern suggests that selective response control becomes harder to preserve over longer forecasts, even when the target direction is recovered more accurately.

Distributional Outputs Remain Largely Stable. For Chronos-2 (C2) and TimesFM 3 (F3), projected and deployed edits produce median relative CRPS changes near zero and 50%-interval coverage changes of only -0.001 , as shown in Table 30. The right panel of Figure 37 provides a finer view across quantile levels. Deployed edits preserve substantially more of the quantile-level response than the fixed-direction control, although they do not fully reproduce the matched-input response across the quantile range. Thus, small aggregate CRPS and coverage changes can coexist with structured differences in how the intervention response is distributed across quantiles.

Cross-Coordinate Effects. For forced diffusion, the deployed drive-frequency (WC) edit has gain -0.48 along the diffusion-time (WD) response, close to the projected value of -0.46 . Because the two response directions overlap, this quantity does not isolate parameter leakage. Together, these analyses show that hidden-state edits improve target alignment while retaining measurable non-target, horizon-dependent, and distributional effects.

B.10 COMPUTATIONAL COST AND OPERATIONAL BOUNDARY

The hidden-state intervention pipeline incurs different costs during direction discovery and test-time deployment. To quantify this operational boundary, 128 linear trend histories are timed at each selected layer on one NVIDIA H100. Discovery measures per-input Jacobian construction and inversion together with one-time basis construction using 16,384 discovery displacements. Deployment includes readout evaluation, coordinate lifting, and the edited forward pass, without constructing a new Jacobian.

Results. As shown in Table 31, inversion is negligible relative to the other components at $H = 16$, requiring only 0.02–0.08 ms across the four evaluated models. Jacobian construction ranges from 0.002 to 0.036 s, while one-time basis construction ranges from 4.6 to 11.4 s. Deployment is substantially cheaper than basis construction, ranging from 0.2 to 22.2 ms per input.

Across $H = 8$ to 64, Jacobian time increases from 0.001 to 0.008 s for Chronos-2 (C2), 0.004 to 0.021 s for TimesFM 3 (F3), 0.011 to 0.062 s for Time-MoE (ME), and 0.001 to 0.005 s for TTM-R2 (TT). Time-MoE (ME) also has the largest measured memory footprint, ranging from 17.6 to 24.8 GB. Thus, the principal computational burden lies in discovery and validation, while the learned readout can be applied at test time without repeating the per-input Jacobian analysis.

B.11 LIMITATIONS AND FUTURE WORK

The controlled laws isolate designated Dynamical Parameters but do not cover broader multivariate, nonstationary, or exogenous dynamics. Linear probes measure linear representation accessibility and may miss information available only through nonlinear readouts. The conditional reference receives law and nuisance information unavailable to the TSFM, so it serves as an informed response reference rather than a model-equivalent predictor. Hidden-state interventions measure response capacity at calibrated steps and depend on the selected state, family of reference responses, and forecast horizon. Controlled injections on real-series backgrounds preserve realistic temporal structure but do not identify the intrinsic dynamics of the underlying series. Future work could extend the framework to richer dynamical systems, nonlinear accessibility measurements, and training objectives that better align accessible parameter information with forecast response.