BEYOND MISSING VALUES: HOW IMPUTATION CHANGES OBSERVED REPRESENTATIONS IN TIME SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Imputation is commonly evaluated by reconstruction error at missing coordinates. In time series foundation models with input-dependent normalization, filling a gap also changes the statistics used to transform the history. It can therefore alter the representations of observed values that remain unchanged in the raw history. Reconstruction scores alone do not describe the changes actually presented to the forecaster. Assessing downstream utility requires understanding how these changes arise and under which observation conditions they are amplified. We analyze the propagation of imputation error through normalization in frozen forecasters. Preserved observations constrain the variance of every feasible completion within each statistical support used by normalization. For positive observed dispersion, we derive a uniform bound on normalized error under ordinary standardization, with a coefficient determined before imputation and sharp when at least two coordinates are missing. Experiments with three frozen forecasters separate reconstruction error, normalized input change, forecast displacement, and forecast loss. In 10% of the 190 baseline conditions, selection by reconstruction score gives forecast RMSE more than 10.14% above the best candidate. For TimesFM, perturbations of 10^{-5} observed-history standard deviations yield mean forecast displacement above 0.01 in the same scale in 454 of 518 constant-patch configurations; fixing historical statistics reduces the mean displacement in every configuration below 4.31×10^{-5}. Imputation evaluation should report observed dispersion and completed scale within the model's statistical supports, and measure forecast loss directly.

Keywords: time series imputation; time series forecasting; error propagation

1 INTRODUCTION

Time series forecasting estimates future values from historical observations, supporting electricity scheduling, traffic management, and computing resource planning. Time series foundation models such as TimesFM and Sundial are pretrained on large collections of sequences and can forecast new series across datasets without updating their parameters (Das et al., 2024; Liu et al., 2025).

Real-world time series often contain missing values. Sensor outages, communication delays, and missing logs can leave contiguous gaps in historical inputs. A common procedure fills these gaps before passing the completed sequence to a forecasting model. Imputation methods are usually compared by MAE or RMSE at masked coordinates. These metrics measure reconstruction quality, while the forecaster uses values normalized by input-dependent statistics. Equal reconstruction errors can therefore produce different changes in the model input.

Figure 1 gives a concrete example. In a fixed Exchange Rate history, retaining only the last two observations with distinct values in the first 32-step patch yields a normalized input change of approximately 0.002 for a small change at the imputed coordinates. When only the last observation is retained, the normalized change after constant imputation is approximately 1. Both cases use the same relative perturbation budget and fixed model weights. A photovoltaic history shows similar behavior: several observations remain in the first patch, but their values are identical. Counts of

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

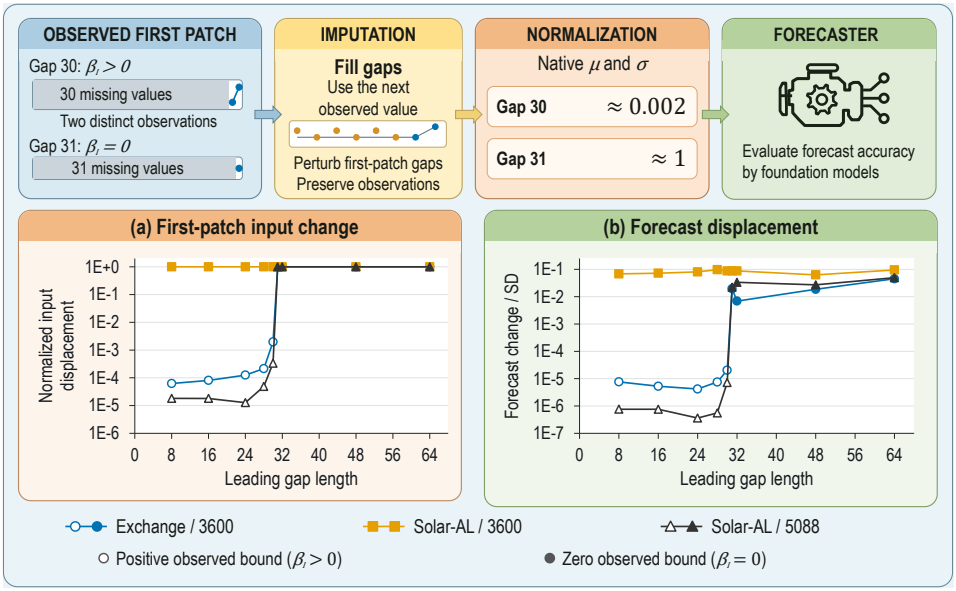


Figure 1: Imputation sensitivity in TimesFM. Top: the Exchange Rate example. Bottom: normalized first-patch changes and forecast displacement versus gap length. Perturbations affect first-patch gaps only. Displacement uses baseline completed-history SD units; forecast loss is evaluated separately. Filled and hollow markers denote zero and positive observed bounds.

missing and observed coordinates alone therefore cannot explain how imputation error changes the normalized input.

Task-oriented imputation studies show that reconstruction quality should be evaluated alongside downstream utility (Wang et al., 2024b; Du et al., 2024; Sartipi et al., 2025; Le Naour et al., 2026). Research on normalization and forecasting robustness examines the effects of input statistics, low variance, and perturbations (Kim et al., 2022; Fu & Yang, 2025; Li et al., 2026; Cohen et al., 2025; Khwaja et al., 2026; Zhang et al., 2025). We study the constraints imposed by preserving observed values during imputation. Before an imputation method is chosen, these values already constrain the variance of every feasible completion within the relevant statistical support. This constraint relates reconstruction error to normalized error through a bound whose coefficient depends on the support used by the model.

We connect these constraints to forecasting by studying imputed histories, normalization, and model responses together. Imputation can produce a low-variance support even when the complete history varies. Statistics computed on that support can change the representations of both imputed and unchanged observed coordinates. Forecasts also depend on output-scale restoration and the model's response to the normalized input. We vary imputed values, historical statistics, and context while holding model weights and future targets fixed, measuring normalized input change, forecast displacement, and forecast loss separately.

Our contributions are as follows.

1. We derive an upper limit on how much ordinary standardization can amplify imputation error. This limit is computed before imputation from unchanged observed values, provided they are not all equal within the normalization range.
2. We compare the same forecasting case while changing filled values, normalization statistics, or context and keeping model parameters fixed. We measure changes in normalized inputs, predictions, and prediction errors separately.
3. Experiments show that small imputation changes can be amplified when observed values vary little. Unchanged observations can affect forecasts through altered normalized values. Larger forecast changes need not increase forecast error.

Related work is discussed in Appendix A.

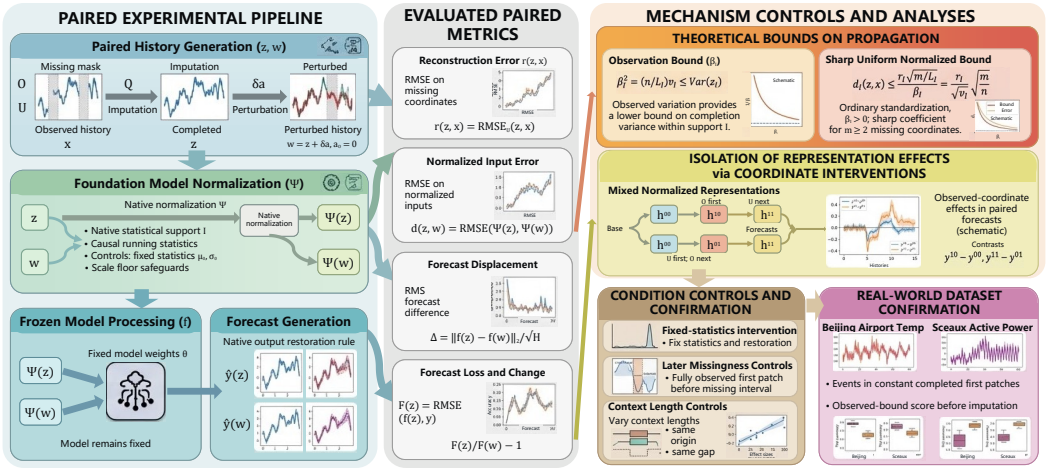


Figure 2: Paired evaluation and mechanism analysis of imputation. Observation-preserving histories pass through native normalization and a frozen forecaster. Reconstruction error, normalized input change, forecast displacement, and forecast loss are measured separately. Mixed normalized states h^{ab} use observed-coordinate representations from state a and missing-coordinate representations from state b; 0 and 1 denote baseline and perturbed states. Coordinate interventions fix centering, scaling, and output restoration to baseline values. The small plots illustrate the analysis design schematically.

2 Evaluation Framework

Given a forecast origin t, let $x = x_{t-L:t}$ denote the history and $y = x_{t:t+H}$ the future, where each interval includes its start and excludes its end. Let O and U denote the observed and missing historical coordinates. A completion z can use only values visible before the forecast origin and must satisfy $z_i = x_i$ for $i \in O$. Complete history is used only to evaluate reconstruction at artificially masked coordinates and to construct reference interventions; future labels are used only to evaluate forecast loss. Let f denote the frozen forecaster, $\Psi(z)$ the actual normalized history, and $\hat{y} = f(z)$ the forecast. The parameters remain fixed, while the original implementation still computes input-dependent statistics and applies its inference rules.

Figure 2 compares completions z and w that agree at observed coordinates. We record each model's statistics, normalized input, and restored forecast. Controls fix historical statistics, change context around the same physical gap, or replace selected normalized coordinates. Each control uses its own baseline, with model weights and future targets fixed.

We measure reconstruction, normalized inputs, and forecasts separately. Raw reconstruction error r is measured at missing coordinates. Normalized error d compares the historical values actually received by the model. For inputs z and w, forecast displacement is $\|f(z) - f(w)\|_2/\sqrt{H}$, reported in a common reference scale. Forecast loss is $F(z) = \mathrm{RMSE}(f(z), y)$, and its relative change is $F(z)/F(w) - 1$. A negative loss change indicates lower forecast error. Displacement measures how much predictions change; loss change measures whether they become more accurate.

Each normalization operation uses a set of historical positions, which we call its statistical support. The models use different supports. Sundial inference computes the population standard deviation over the entire history and adds 10^{-5}. In addition to full-history preprocessing, TimesFM-2.5 uses causal running statistics over 32-step patches. Toto 2.0 uses double-precision causal statistics, a sample-variance correction, a continuous scale floor, and an asinh transformation. We record the actual statistical support I for each experiment. In particular, the observed variance in Sundial's first 32 steps is not treated as a condition on its full-history normalization.

We evaluate forecast accuracy using point-forecast RMSE. Sundial's point forecast is the mean of 100 samples. The TimesFM-2.5 adapter used in this study returns the 0.5 quantile as its point forecast, and Toto 2.0 also uses the 0.5 quantile. Candidate rankings, the best candidate under each condition, and the additional error from method selection all use this definition of forecast loss.

3 Observation Conditions and Error Propagation

3.1 Observed Variance and Feasible Completions

Consider a statistical support I of length L_I, with n observed coordinates and m missing coordinates, where $n + m = L_I$. Let v_I be the population variance of the visible observations. All compared completion vectors agree at these observed coordinates. Define β_I as the standard-deviation lower bound provided by the observations, with $\beta_I = 0$ when there are no observations.

$$\beta_I^2 = \frac{n}{L_I} v_I \leq \mathrm{Var}(z_I). \tag{1}$$

Proposition 1. The population variance of any observation-preserving completion on I is at least β_I^2. Decomposing observed squared deviations around the completed mean gives within-observation variation and a nonnegative mean-shift term. Missing coordinates contribute a nonnegative term, yielding Eq. (1). For $n > 0$, filling all missing coordinates with the observed mean attains the bound. With no observations, any constant completion attains zero. Thus, two distinct observations exclude zero-variance completions, whereas a large observation count alone does not ensure a large lower bound.

The bound β_I can be computed before imputation from the observations within the current statistical support. When $\beta_I = 0$, the observations provide no positive uniform lower bound, although a particular completion can still have nonzero variance. The SAITS control in Appendix E.4 has the same observation conditions but does not produce a constant first patch.

3.2 A Sharp Bound on Normalized Error

Consider ordinary standardization without numerical-protection branches:

$$N_I(z) = \frac{z_I - \mathrm{mean}(z_I)\mathbf{1}}{\sigma(z_I)},$$

where σ is the population standard deviation. If $\beta_I > 0$, two completions z and w that preserve the same observations satisfy

$$\|N_I(z) - N_I(w)\|_2 \leq \frac{\|z_I - w_I\|_2}{\beta_I}. \tag{2}$$

Proposition 2. If $\beta_I > 0$, Eq. (2) holds uniformly over all completions that preserve the observations on I. If at least two coordinates are missing, the coefficient $1/\beta_I$ is sharp.

The derivative of standardization has operator norm at most $1/\sigma$. Every point on the segment joining z and w preserves the observations, so Proposition 1 bounds σ below by β_I along the entire segment. Integration gives Eq. (2). When missing values are filled with the observed mean, a zero-sum perturbation at two missing coordinates asymptotically attains the coefficient. With at least two missing coordinates, the observations therefore determine the best uniform coefficient for the full normalized vector, independently of the imputation method. The same bound holds when the normalized difference is restricted to observed coordinates.

When $\beta_I > 0$ and $m > 0$, let $w = x$ be the complete historical truth, r_I the reconstruction RMSE at missing coordinates in the support, and d_I the normalized RMSE over the entire support. Then

$$d_I(z, x) \leq \frac{r_I \sqrt{m/L_I}}{\beta_I} = \frac{r_I}{\sqrt{v_I}} \sqrt{\frac{m}{n}}. \tag{3}$$

Equation (3) bounds normalized error using reconstruction error, observed dispersion, and the missingness proportion. The coefficient is available before imputation, whereas the standard deviation of a completed history depends on the chosen method. Applying the bound to naturally missing values requires a valid bound on r_I. Forecast accuracy additionally depends on the model response and

future target. For a restricted set of coordinates, such as the original observations, the corresponding RMS bound uses the square root of that set's size as its denominator.

Appendix D gives the proof details and the corresponding bounds for continuous scale protection.

3.3 Constant Imputation and Numerical Limits

If imputation produces a constant support $c\mathbf{1}$, ideal standardization satisfies the following identity for any nonconstant direction a and any nonzero δ:

$$N(c\mathbf{1} + \delta a) = \operatorname{sgn}(\delta)\frac{a - \bar{a}\mathbf{1}}{\sigma(a)}. \tag{4}$$

Standardization cancels the magnitude of δ, leaving a normalized pattern determined by a and the sign of δ. At an unchanged coordinate with $a_j = 0$, the standardized value of the perturbed input is $-\operatorname{sgn}(\delta)\operatorname{mean}(a)/\operatorname{sd}(a)$. If exactly q coordinates receive an equal positive increment, this value is $-\sqrt{q/(L_I - q)}$. The expression describes values after perturbation. Ideal standardization is undefined at a constant baseline, which actual models handle through a protective branch or a positive scale floor.

Actual models use finite precision, scale floors, or conditional branches. For sufficiently small δ, numerical protection becomes active. Equation (4) therefore does not imply a fixed response to arbitrarily small changes. The main mechanism controls use relative amplitudes 10^{-3} and 10^{-5}. An additional study (Appendix C.5) examines eight amplitudes from 10^{-2} to 10^{-9} on the existing six physical scenarios. We check the actual float32 inputs and values passed to the model in every comparison, including inputs unchanged by rounding and forecasts with zero response.

Causal statistics also place inputs and outputs on different scales. An early imputed patch can use a small local scale, while forecasts are restored using a longer history. The relationship between numerical errors before and after normalization therefore does not directly extend to forecast loss. Applying the same projection onto nonnegative outputs to both forecasts cannot increase their Euclidean distance. If an input change switches whether that projection is enabled, however, its effect requires separate analysis. The initial-gap-length, random-perturbation, truth-correction, and context controls in this study check that this rule does not switch.

4 Experimental Design

4.1 Datasets

The benchmark contains 19 univariate target series from ten source collections. Five block-missingness rates (10%–50%) and three seeds cover 301 forecast origins and 4,515 window–mask configurations. Block length is 50, maximum history length is 2,880, and forecast horizon is 720. Two additional datasets were not used to develop the mechanism experiments: hourly Beijing airport temperature and minute-level Sceaux household active power (Chen, 2015; Hebrail & Berard, 2006). Whether they appeared in model pretraining is unknown. Twenty origins per source, chosen by timestamp and data availability before forecasting, combine with three missingness rates and Linear and Mean imputation to give 240 configurations per model. All 142 configurations with fully observed first patches and 98 with missing first-patch coordinates are retained.

4.2 Evaluation Metrics

Forecast RMSE is computed within each seed, then averaged across seeds. Reconstruction scores average the method rankings from MAE and RMSE averaged across windows and seeds. We compare selected and best-candidate RMSE by their ratio minus one, averaging tied selections. Section 2 defines normalized input change and forecast displacement. On the additional datasets, an event has mean displacement above 0.01 observed-history standard deviations over three fixed directions at budget 10^{-5}. Diagnostics use the 98 configurations with missing first-patch coordinates, with nonconstant patches examined separately.

4.3 Baseline Methods

We compare 19 imputation implementations: eight constant, propagation, or interpolation rules; six statistical methods; and five learned methods (SAITS, GPT4TS, TimeMixer, TOTEM, and HELIX). All preserve observations. Learned methods are fitted in PyPOTS using only visible history at each origin. TimesFM and Sundial receive identical imputed inputs. Toto is used in mechanism controls with external numerical imputation and its dedicated missingness branch disabled. Appendix B lists all methods, settings, and implementation limits.

4.4 Experimental Protocols and Controls

Temporal selection uses the first half of origins to select methods and the remainder for evaluation; selection labels end no later than the first evaluation origin. We resample all configurations from each source together. All 518 configurations with constant Linear first patches are retained, including development cases. Sources, origins, and selection rules are fixed before forecasting.

Formula controls vary observed dispersion and count, holding perturbation energy fixed within each of three budgets (480 pairs; 1,440 comparisons). A fixed subset of 64 pairs per model tests TimesFM and Toto over 128 steps. Coordinate interventions use all 42 constant Linear first-patch configurations with at least one observed coordinate in that patch (33 histories; ten sources). They separately replace observed and missing representations, fixing baseline centering and scaling. The primary horizon is 128 steps; the 720-step extension also fixes TimesFM's future RevIN statistics. Appendix C gives all parameter ranges and protocols.

5 Experimental Results

5.1 Reconstruction-Based Selection and Forecast Loss

Across the 190 baseline conditions, reconstruction and forecast rankings have a mean Spearman correlation of 0.534, with a 95% bootstrap interval of $[0.431, 0.654]$ from resampling source collections. For methods selected by reconstruction score, forecast RMSE is at least 2.13% above the best candidate in half the conditions and more than 10.14% above it in 10% of conditions; 71.0% of selections are within 5% of the best candidate. With selection on earlier windows and evaluation on later windows, these gaps are 2.19% and 16.86%, respectively. Even with identical absolute errors at every imputed coordinate, four error patterns yield different forecast RMSEs. Across windows, the median difference between the highest and lowest RMSE is 6.14% for Sundial and 4.37% for TimesFM, relative to the lowest RMSE in each window.

Finding 1. Reconstruction accuracy only partly predicts forecast accuracy. Equal imputation-error magnitudes can produce different forecast losses; the following controls examine the role of normalization.

5.2 Sensitivity to Observed Dispersion

Linear imputation produces constant first patches in 518 of 4,515 configurations (11.5%). The corresponding complete first patches are nonconstant in 508 of these configurations. At a perturbation budget of 10^{-5} observed-history standard deviations, 454 configurations have mean forecast displacement above 0.01, or 89.5% after equal weighting across sources. Fixing historical statistics, including output restoration, reduces all mean displacements to at most 4.31×10^{-5} (Appendix E.3). Input normalization and output restoration thus contribute to amplification in these configurations.

Figure 3 extends the analysis to small positive observation bounds. In controlled histories with $\beta_I = 2 \times 10^{-6}$ and perturbation RMS 10^{-5}, neither model activates its scale safeguard before or after perturbation. The median normalized RMS changes at unchanged observations are nevertheless 1.865 for TimesFM and 1.285 for Toto, over the two observation counts. Their median 128-step forecast displacements are 0.00736 and 0.02216 in the fixed synthetic scale. At $\beta_I = 0.1$, these displacements fall to 3.56×10^{-7} and 1.54×10^{-6}.

All defined formula comparisons satisfy their bounds. Bound utilization is the normalized input change divided by its upper bound. Along the locally tight direction in Proposition 2, utilization

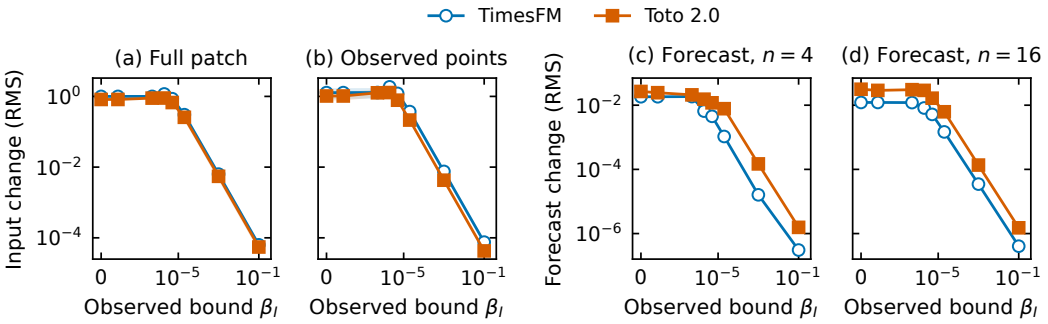


Figure 3: Native responses as the observed scale lower bound varies. (a,b) Normalized RMS changes in the full first patch and its unchanged observed coordinates; lines and shading give the median and range over $n = 4$ and $n = 16$ observations. (c,d) The corresponding 128-step forecast displacements for each observation count. Within each observation-count setting, a fixed positive direction perturbs only missing coordinates, with RMS 10^{-5} over the first 32 steps; subsequent history is unchanged. Both models use a synthetic reference scale of one. All eight bound levels, including zero, are retained.

exceeds 0.99996 when perturbation RMS is at most $0.01\beta_I$. Random directions have median utilization 0.5054; individual directions need not attain the upper bound. On inputs from the additional datasets with positive observation bounds and nonzero perturbations, median utilization is 0.231–0.392 across the four model–source groups (Appendix E.11). Appendix E.8 reports all formula comparisons and the models' numerical branches.

Finding 2. Small observed dispersion permits amplification without triggering a scale safeguard. The response depends on the observed scale lower bound relative to the perturbation budget, as the $\beta_I = 2 \times 10^{-6}$ controls show.

5.3 Forecast Effects of Observed-Coordinate Changes

To isolate the effect of unchanged observations, we separate the normalized history into originally observed and missing coordinates. Forecasts y_{00}, y_{10}, y_{01}, and y_{11} use baseline representations, perturbed representations at originally observed coordinates only, perturbed representations at originally missing coordinates only, and perturbed representations at both sets of coordinates, respectively. Centering, scaling, and output restoration use the same baseline statistics. The contrasts $y_{10} - y_{00}$ and $y_{11} - y_{01}$ measure the observed-coordinate effect with missing-coordinate representations held at their baseline and perturbed values, respectively. Restoring observed representations reverses the sign of the second contrast. The primary comparison covers the first 128 forecast steps.

Across the 33 histories, the median observed-only displacement is 0.0571 for TimesFM and 0.0177 for Toto. This displacement exceeds 0.01 in 32 and 26 histories, respectively. The conditional observed effect has medians of 0.0172 and 0.00376 and exceeds 0.01 in 21 and five histories, respectively. Every source has a positive median for both contrasts (Figure 4). The interaction term has median displacements of 0.0559 and 0.0192; the norms of these effects cannot be added to obtain percentage contributions. Appendix E.9 reports both amplitudes and forecast lengths.

Finding 3. Observed values can remain unchanged while their representations alter forecasts. These effects occur across all ten sources and depend on the representations at missing coordinates.

5.4 Forecast Sensitivity across Datasets

On Beijing airport temperature and Sceaux household active power, large forecast responses are concentrated in constant supports (Figure 5). At budget 10^{-5}, TimesFM has 11 events among 120 Beijing configurations and 13 among 120 Sceaux configurations; Toto has 13 and 12, respectively. The subsets with missing first-patch coordinates contain 54 and 44 configurations. All events occur in the 27 configurations with constant completed first patches. Neither model has an event among the 71 perturbable nonconstant configurations.

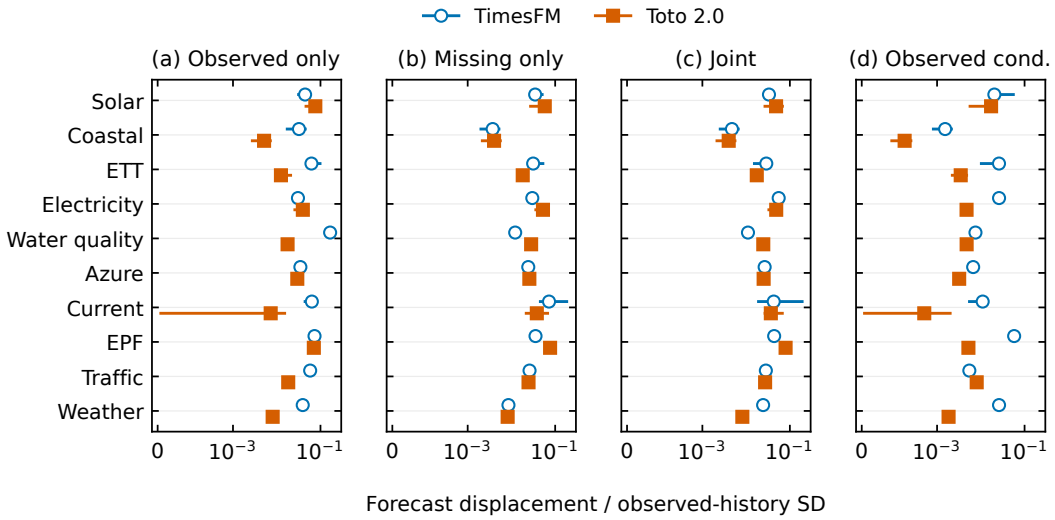


Figure 4: Coordinate interventions at budget 10^{-5} over the first 128 forecast steps. Panels compare observed-only, missing-only, joint, and conditional observed effects for both models and all ten sources. Points are source medians after averaging directions and configurations sharing one history; bars span the middle 50% of histories within each source. All 42 configurations and 33 histories are retained. Observed conditional denotes the displacement between y_{11} and y_{01}. Centering, scaling, and output restoration are fixed to baseline statistics. The common scale is linear below 10^{-4} and logarithmic above it. Solar, Electricity, and Current abbreviate Australia_Solar, OpenElectricity, and CurrentVelocity.

On perturbable inputs, the observed-bound score has average precision 0.750–0.929 across the four model–source groups. Constant-patch detection and completed scale both achieve 0.846–1.000, with higher average precision in each matched group. The observed bound and completed scale answer different questions: β_I is available before imputation and applies across feasible completions, while completed scale characterizes the tested input. Appendix E.10 reports loss changes and fixed-budget comparisons.

5.5 Statistical Support and Forecast Accuracy

Context controls hold the physical missing interval and imputed values fixed. With a 512-step history and budget 10^{-5}, median forecast displacement is 0.31804 for TimesFM and 0.15530 for Toto. With 1,024 steps, both medians are approximately 4×10^{-6}. At 512 steps, the affected patch lies at the beginning of the causal statistical support; at 1,024 steps, its support includes earlier observations with variation. Retaining the 512-step network input while using longer-history statistics reduces TimesFM displacement to approximately 10^{-5}. This control changes the statistics while keeping the network input length fixed (Appendix E.7).

Finding 4. Statistical support can change sensitivity even when the physical gap is fixed. In the six context scenarios, using longer-history statistics reduces TimesFM displacement even with the same 512-step network input.

The effect also occurs after a fully observed first patch: later missingness changes preceding observations within the affected support and changes forecasts while first-patch forward values remain fixed (Appendix E.6). Fixing statistics reduces these TimesFM responses but increases Toto responses because its baseline scale remains at the floor. A scale intervention therefore requires evaluation under the particular model and support.

Among the 518 constant-patch configurations, RMSE rises by more than 1% in 175 configurations and falls by more than 1% in 116. Under native 512-step context processing, RMSE changes range from -12.27% to $+69.91\%$ for TimesFM and from -35.87% to $+4.83\%$ for Toto.

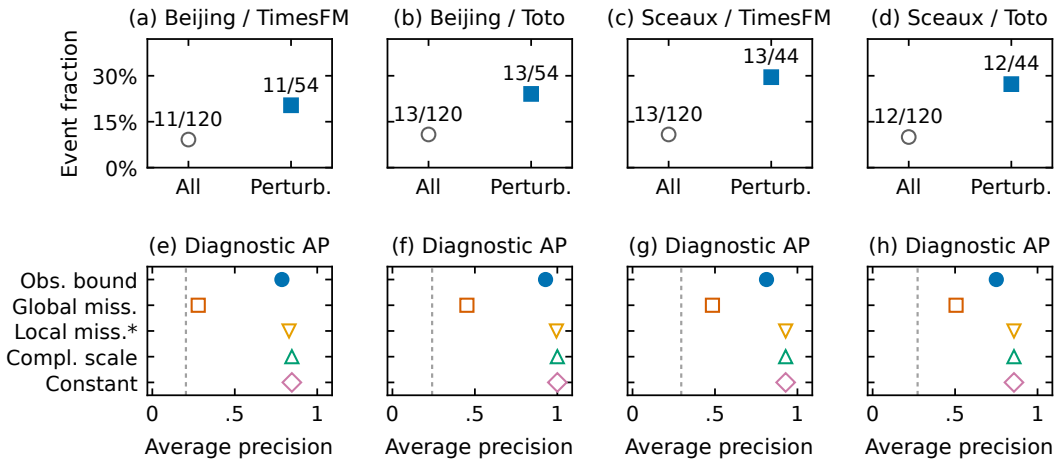


Figure 5: Forecast sensitivity on Beijing airport temperature and Sceaux household active power. Columns distinguish the sources and models. (a–d) Event fractions among all 120 configurations per source and their perturbable subsets; labels give events/inputs. An event has direction-averaged 720-step displacement $D_f > 0.01$ in observed-history scale at budget 10^{-5}. (e–h) Average precision on perturbable patches only. Global missing fraction uses the whole history; local missing fraction (*) uses the first patch and is a secondary comparison. Gray dashed lines give event fractions. Scores use fixed definitions, with tied values grouped.

Finding 5. Forecast displacement does not determine whether forecast accuracy improves. Forecast loss depends on the model and future target and must be measured directly.

6 Discussion and Limitations

The observation-based bound concerns all feasible completions within a support; completed scale describes the chosen input. The bound applies to the stated normalization formula, so each model's scale floors and conditional rules need separate checks. The SAITS control shows that a zero bound does not make every completion sensitive. Output restoration also matters: its absolute effect exceeds the input-normalization effect in 22 of 30 Sundial combinations (Appendix E.1). Replacing native normalization requires direct forecast-loss evaluation.

The experiments use frozen univariate forecasters with artificial block missingness and forecast horizons of 128 and 720 steps. Natural and value-dependent missingness, multivariate forecasting, and native missing-token handling require separate evaluation. The results for constant-patch configurations and coordinate interventions apply to the stated input conditions. The two additional datasets include repeated masks and some overlapping histories; event proportions apply to the evaluated configurations. Modified internal representations may not correspond to any completion of the raw history. Whether the additional datasets were used in pretraining is unknown.

7 Conclusion

We analyzed when normalization amplifies imputation error in frozen time series foundation models. Under ordinary standardization with positive observed dispersion, the observations determine a uniform bound on normalized error over feasible completions. Its coefficient is sharp for the full normalized vector when at least two coordinates are missing. Experiments show amplification at zero and small positive bounds. Coordinate interventions show that altered representations of unchanged observations affect forecasts; context controls identify the role of statistical support. Forecast error can increase or decrease. Evaluation should report observed dispersion and completed scale within the model's statistical supports and measure forecast loss directly.

Reproducibility Statement

https://anonymous.4open.science/r/RECA
The repository provides source code, fixed experimental protocols, dependency records, and numerical summaries. Raw data, model weights, and large saved-run arrays are separate inputs, as described in the README. Dataset and imputer settings are given in Appendix B, experimental protocols and controls in Appendix C, proofs and numerical details in Appendix D, and supplementary results in Appendix E.

AI Usage Statement

Generative AI tools were used to assist with translation and language editing.

References

Ihab Ahmed, Denis Krompaß, Cheng Feng, and Volker Tresp. A comparative study on how data normalization affects zero-shot generalization in time series foundation models. arXiv preprint arXiv:2512.02833v2, 2026. URL https://arxiv.org/abs/2512.02833v2. Version 2 revised January 29, 2026.

Abdul Fatir Ansari, Lorenzo Stella, Ali 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 Bernie Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/734e6bfcd358e25ac1db0a4241b95651-Abstract.html.

Song Chen. Beijing PM2.5. UCI Machine Learning Repository, 2015. URL https://archive.ics.uci.edu/dataset/381/beijing+pm2+5+data. Dataset; temperature column used in this study.

Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990. URL https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/stl-a-seasonal-trend-decomposition-procedure-based-on-loess.pdf.

Ben Cohen, Emaad Khwaja, Youssef Doubli, Salahidine Lemaachi, Chris Lettieri, Charles Masson, Hugo Miccinilli, Elise Ramé, Qiqi Ren, Afshin Rostamizadeh, Jean 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 *Advances in Neural Information Processing Systems*, volume 38, pp. 50907–50951. Curran Associates, Inc., 2025. doi: 10.52202/085713-1698. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/48ea942b60e6c52b50ca90e4a40d8ac2-Abstract-Conference.html.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Wenjie Du, David Côté, and Yan Liu. SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023. doi: 10.1016/j.eswa.2023.119619. URL https://doi.org/10.1016/j.eswa.2023.119619.

Wenjie Du, Jun Wang, Linglong Qian, Yiyuan Yang, Zina Ibrahim, Fanxing Liu, Zepu Wang, Haoxin Liu, Zhiyuan Zhao, Yingjie Zhou, Wenjia Wang, Kaize Ding, Yuxuan Liang, B. Aditya Prakash, and Qingsong Wen. TSI-Bench: Benchmarking time series imputation. arXiv preprint arXiv:2406.12747v2, 2024. URL https://arxiv.org/abs/2406.12747v2.

Fanzhe Fu and Yang Yang. Noise or signal? deconstructing contradictions and an adaptive remedy for reversible normalization in time series forecasting. arXiv preprint arXiv:2510.04667v1, 2025. URL https://arxiv.org/abs/2510.04667v1.

Google Research. TimesFM-2.5-200M PyTorch model card. Hugging Face model documentation, 2025. URL https://huggingface.co/google/timesfm-2.5-200m-pytorch/blob/1d952420fba87f3c6dee4f240de0f1a0fbc790e3/README.md. Updated October 2, 2025. Accessed September 16, 2026.

Georges Hebrail and Alice Berard. Individual Household Electric Power Consumption. UCI Machine Learning Repository, 2006. URL https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption. Dataset.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.

Jason Kemp, Stefan C. Kremer, Genevieve Ali, and Daniel Gillis. Learning from missing values: Encoding missingness in representation-space for LSTM time series forecasting. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=DMmCMIrrez.

Fareed Khan. Rebuilding Google TimesFM from scratch—the complete end-to-end story. GitHub project documentation, 2026. URL https://github.com/FareedKhan-dev/timesfm-from-scratch/blob/185c94d04fca4cf7120110697dea396c7054d087/docs/PROJECT_JOURNEY.md. Updated June 25, 2026. Accessed September 16, 2026.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era. arXiv preprint arXiv:2605.20119v2, 2026. URL https://arxiv.org/abs/2605.20119v2.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Etienne Le Naour, Tahar Nabil, Adrien Petralia, and Ghislain Agoua. Are time-indexed foundation models the future of time series imputation? *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=cTk56KpsP5.

Yujie Li, Zezhi Shao, Chengqing Yu, Yisong Fu, Tao Sun, Yongjun Xu, and Fei Wang. APT: Affine prototype-timestamp for time series forecasting under distribution shift. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(18):15180–15188, 2026. doi: 10.1609/aaai.v40i18.38542. URL https://ojs.aaai.org/index.php/AAAI/article/view/38542.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25be.html.

Tong Nie, Guoyang Qin, Wei Ma, Yuewen Mei, and Jian Sun. ImputeFormer: Low rankness-induced transformers for generalizable spatiotemporal imputation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2260–2271. ACM, 2024. doi: 10.1145/3637528.3671751. URL https://doi.org/10.1145/3637528.3671751.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. doi: 10.7551/mitpress/3206.001.0001. URL https://gaussianprocess.org/gpml/.

Amir Sartipi, Joaquín Delgado Fernández, Sergio Potenciano Menci, and Alessio Magitteri. Bridging smart meter gaps: A benchmark of statistical, machine learning and time series foundation models for data imputation. In *2025 IEEE Kiel PowerTech*, pp. 1–7. IEEE, 2025. doi: 10.1109/PowerTech59965.2025.11180477. URL https://doi.org/10.1109/PowerTech59965.2025.11180477.

Sabera J. Talukder, Yisong Yue, and Georgia Gkioxari. TOTEM: TOkenized time series EMbeddings for general time series analysis. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QlTLkH6xRC.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, volume 2024, pp. 38626–38652, 2024a. URL https://openreview.net/forum?id=7oLshfEIC2.

Zhixian Wang, Linxiao Yang, Liang Sun, Qingsong Wen, and Yi Wang. Task-oriented time series imputation evaluation via generalized representers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 137403–137431. Curran Associates, Inc., 2024b. doi: 10.52202/079017-4365. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f88264fcc54775ee1706116e90fe351a-Abstract-Conference.html.

Fengming Zhang, Wenjie Du, Huan Zhang, Ke Yu, and Shen Qu. HELIX: Hybrid encoding with learnable identity and cross-dimensional synthesis for time series imputation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FN20iuPnEU.

Jiawen Zhang, Zhenwei Zhang, Shun Zheng, Xumeng Wen, Jia Li, and Jiang Bian. Are time-series foundation models deployment-ready? a systematic study of adversarial robustness across domains. arXiv preprint arXiv:2505.19397v2, 2025. URL https://arxiv.org/abs/2505.19397v2.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained LM. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43322–43355. Curran Associates, Inc., 2023. doi: 10.52202/075280-1877. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/86c17de05579cde52025f9984e6e2ebb-Abstract-Conference.html.

APPENDICES

Appendix A reviews related work. Appendix B describes the datasets, imputation methods, and implementation settings. Appendix C specifies the experimental protocols and controls, and Appendix D provides proofs and numerical implementation details. Supplementary experimental results are reported in Appendix E.

A RELATED WORK

Pretrained models, including TimesFM and Sundial, forecast directly from numerical histories (Ansari et al., 2024; Das et al., 2024; Liu et al., 2025; Google Research, 2025). RevIN removes and restores instance statistics to address distribution changes (Kim et al., 2022). Subsequent work

examines outliers, robust statistics, and affine adjustments (Fu & Yang, 2025; Li et al., 2026), while Ahmed et al. (2026) compare normalization scope and the scale effects of training losses during foundation-model pretraining. Toto and Toto 2.0 use causal statistics with explicit numerical safeguards (Cohen et al., 2025; Khwaja et al., 2026). First-patch scaling instability has also been documented in an independently trained, unofficial TimesFM reproduction (Khan, 2026). We derive the variance constraint shared by all completions that preserve the available observations and examine how imputation affects normalization within the statistical supports of frozen models.

Adversarial robustness studies examine position-dependent forecast responses to small perturbations that preserve trend or seasonal structure (Zhang et al., 2025). Our perturbations are restricted to prespecified missing coordinates and directions, and are independent of future labels. Corrections toward the complete historical values do not increase reconstruction error. The Min–Std study investigates numerical encodings of missingness together with external scaling in LSTM forecasting (Kemp et al., 2026). We derive error bounds from the observed values within each statistical support and use interventions in frozen forecasters to distinguish the effects of imputed content, statistical support, and output restoration.

Missing-value imputation has developed from propagation and interpolation to state-space models, seasonal decomposition, and learning-based modeling (Kalman, 1960; Rasmussen & Williams, 2006; Cleveland et al., 1990; Cao et al., 2018; Tashiro et al., 2021; Du et al., 2023; Nie et al., 2024; Zhou et al., 2023; Talukder et al., 2024; Wang et al., 2024a; Zhang et al., 2026). TSI-Bench and benchmarks of foundation models for imputation compare recovery quality under different missingness conditions (Du et al., 2024; Sartipi et al., 2025; Le Naour et al., 2026). Task-oriented evaluation further examines how imputed data affect downstream learning (Wang et al., 2024b). We evaluate reconstruction quality and forecast accuracy separately, focusing on the historical inputs and normalization operations of frozen forecasters.

B Datasets and Imputation Methods

Table 1 lists the 19 univariate target series in the baseline benchmark, grouped by the source collections used for joint resampling. H denotes hours and T denotes minutes; structural periods are measured in sampling steps. The additional Exchange Rate and Solar-AL sources are used only in prespecified mechanism experiments and are not included in the original 190 conditions.

Table 1: The 19 baseline target series and their domains. Shading separates source collections.

Configuration	Freq.	Target	Length	Period	Domain
Australia_Solar_H	H	solar_63726	35,064	24	Energy
Coastal_T_S_15T	15T	PSAL	23,134	96	Natural environment
Coastal_T_S_20T	20T	TEMP	8,198	72	Natural environment
Coastal_T_S_5T	5T	PRES_REL	105,408	288	Natural environment
Coastal_T_S_H	H	PRES_REL	8,784	24	Natural environment
ETTh1	H	OT	17,420	24	Energy
ETTh2	H	OT	17,420	24	Energy
ETTm1	15T	OT	69,680	96	Energy
ETTm2	15T	OT	69,680	96	Energy
OpenElectricity_ NEM_5T	5T	demand__QLD1	43,488	288	Energy
Water_Quality_ Darwin_15T	15T	DOX2	18,985	96	Natural environment
azure2019_D_5T	5T	avg_cpu	8,639	288	Computing
azure2019_I_5T	5T	avg_cpu	8,634	288	Computing
current_velocity_15T	15T	TEMP	17,953	96	Natural environment
current_velocity_20T	20T	ABSIC1	16,759	72	Natural environment

Continued on next page

Table 1: The 19 baseline target series and their domains. Shading separates source collections. (Continued)

Configuration	Freq.	Target	Length	Period	Domain
current_velocity_5T	5T	PRES_REL	26,486	288	Natural environment
epf_electricity_price_H	H	target	52,416	24	Energy
traffic	H	OT	17,544	24	Traffic
weather	10T	OT	52,696	144	Natural environment

H denotes hours and T minutes; periods are measured in sampling steps. The four ETT series and the two azure2019 variants each share one source collection.

Propagation and interpolation in the baseline benchmark may use observations after a missing interval but before the forecast origin; they do not use future forecast labels. Learned imputers are fitted to the visible history at each origin and then restore the original observed values. BiSmooth-Kalman and LQDA are local adapters; the algorithms associated with their historical result keys are not reproduced in full. The Seasonal implementation adds the decomposed components back after linear initialization. Its float32 inputs are elementwise identical to Linear in 4,408 of 4,515 windows, so only Linear is retained as a formal candidate.

The original weather timestamps contain one duplicate and one 100-minute interval. Historical and future arrays follow the saved observation order, while the model adapter uses a nominal 10-minute grid. Regular-grid times in the original exported records therefore cannot be interchanged directly with row positions in the original CSV. We use the positions of cached inputs and labels as the reference and provide the corresponding original calendar times. The exported time alignment differs by eight steps in this configuration, but the future values used to compute elementwise forecast errors agree with the cached labels. The irregular timestamps mean that this configuration cannot be treated as a regularly sampled sequence without gaps.

The recorded environment for the original benchmark is Python 3.10.20, PyTorch 2.5.1, NumPy 2.1.3, and pandas 2.2.3; the original forecasts ran on CUDA. The TimesFM and Sundial controls use CPU inference with the same Python and numerical libraries. Toto uses a separate environment with Python 3.13.5, PyTorch 2.7.1 CPU, and toto-2 2.0.0. All three forecasters retain fixed model parameters.

Table 2: Imputation rules of the 19 candidate implementations.

Method	Imputation rule	Family
Zero	Zero at missing coordinates	Constant
Mean	Mean of the current history's observations	Constant
Forward	Nearest preceding observation	Propagation
Backward	Nearest subsequent observation	Propagation
Nearest	Available observation nearest in time	Interpolation
Linear	Linear interpolation between adjacent observed endpoints	Interpolation
Polynomial	Local polynomial curve through observed segments	Interpolation
Spline	Piecewise smooth spline across a missing block	Interpolation
Kalman-Struct	Local-linear-trend state-space smoothing	Statistical
Kalman-ARIMA	Stable AR(p) state-space approximation with Kalman smoothing; $p \leq 3$	Statistical
STL-Kalman	Robust STL or seasonal profile with Kalman residual smoothing	Statistical

Continued on next page

Table 2: Imputation rules of the 19 candidate implementations. (Continued)

Method	Imputation rule	Family
GP-RBF	One-dimensional RBF Gaussian process on time indices; at most 512 observations	Statistical
BiSmooth-Kalman	Local deterministic bidirectional smoothing combined with AR-Kalman	Statistical
LQDA	Local deterministic quantile-distribution adapter	Statistical
SAITS	Overlapping univariate windows	Learned
GPT4TS	Overlapping univariate windows	Learned
TimeMixer (PyPOTS)	Overlapping univariate windows	Learned
TOTEM	Overlapping univariate windows	Learned
HELIX	Overlapping single-feature windows	Learned

All five learned implementations use PyPOTS. HELIX is evaluated on a single feature; cross-feature modeling is not evaluated. Fixed settings and numerical handling are given in Table 3.

Table 3: Fixed settings, fitting budgets, and numerical handling.

Method	Configuration	Epochs	Handling and scope
Zero, Mean Forward, Backward	pandas constant/propagation rules. Forward: `ffill`→`bfill`. Backward: `bfill`→`ffill`.	–	No training.
Linear, Nearest Polynomial, Spline	pandas/SciPy interpolation. Polynomial order 2; Spline order 3.	–	Nearest, Polynomial and Spline: bidirectional linear fallback on exceptions.
Kalman-Struct Kalman-ARIMA STL-Kalman	Local-linear-trend, stable AR(p) state-space ($p \leq 3$), or robust STL/seasonal profile with Kalman residual smoothing, respectively.	–	Univariate, data-adaptive estimation.[a]
GP-RBF	RBF Gaussian process on normalized time indices; ≤ 512 observations. Length scale $\max(10 \times \text{median spacing}, 0.03)$; noise 10^{-4}.	–	Sampling uses the imputer seed. Cholesky failure returns linear initialization.
SAITS	PyPOTS; 2 layers; $d_{\text{model}} = 64$; $d_{\text{ffn}} = 128$; 4 heads; $d_k = d_v = 16$; zero dropout. Window 96; stride 48.	10	Patience 3.
HELIX	PyPOTS single-feature adapter; $d_{\text{pe}} = 8$; $d_{\text{feature_embed}} = 1$; $d_{\text{model}} = 64$; 4 heads; 1 layer; zero dropout. ORT_weight = MIT_weight = 1.	3	Cross-feature synthesis is not evaluated.
GPT4TS	PyPOTS with GPT-2 weights; patch 16; stride 8; 1 retained layer; $d_{\text{ffn}} = 64$; fixed embeddings; `freq=h`; `train_gpt_mlp=False`; zero dropout.	3	Numerical failures may use Linear.[b]
TimeMixer (PyPOTS)	PyPOTS; 1 layer; $d_{\text{model}} = 32$; $d_{\text{ffn}} = 64$; top_k = 3; moving_avg = 3; 1 downsampling layer with window 2; univariate channel independence; zero dropout.	3	Historical key: `timemixerpp`.[c]

Continued on next page

Table 3: Fixed settings, fitting budgets, and numerical handling. (Continued)

Method	Configuration	Epochs	Handling and scope
TOTEM	PyPOTS; $d_{\text{block_hidden}} = 32$; 1 residual layer; $d_{\text{residual_hidden}} = 32$; $d_{\text{embedding}} = 16$; codebook 32; commitment cost 0.25; compression factor 4.	3	No external pretrained checkpoint.
BiSmooth-Kalman LQDA	Local deterministic adapters. BiSmooth-Kalman: 4 bidirectional denoising iterations; average with AR-Kalman at weights 0.5 and 0.5. LQDA: 6 local-quantile updates; distribution weight 0.35; rolling window 7.	–	No training.[d]

All methods restore observed values after imputation. Neural adapters use only windows from the currently visible history, with batch size at most 32. Epochs gives the training limit; – denotes a method without an epoch budget.

[a]Kalman-ARIMA is the historical name of an AR state-space approximation, not a general-purpose ARIMA implementation. [b]GPT4TS outputs equal Linear in 22 of 4,515 cached inputs; this does not determine the actual fallback count. [c]The key `timemixerpp` does not denote TimeMixer++. [d]BiSmooth-Kalman and LQDA use the historical keys `ssd_ts` and `temporal_wasserstein`, respectively; they do not implement the original SSD-TS or TWI papers.

C Experimental Setup and Controls

C.1 Observed Dispersion and Sample Size

The formula study uses 32-coordinate supports with $n \in \{2, 4, 8, 16, 30\}$ observed values and $\beta_I \in \{0, 10^{-8}, 5\times10^{-7}, 2\times10^{-6}, 5\times10^{-6}, 2\times10^{-5}, 10^{-3}, 0.1\}$. The observed values are equally spaced, centered, and rescaled to attain the specified lower bound. Missing coordinates are filled with the mean of the observed values. Perturbation RMS $\epsilon = \|\Delta z\|_2/\sqrt{32}$ is 10^{-7}, 10^{-5}, or 10^{-3} in a reference scale of one. Within each budget, this fixes total perturbation energy when the observation count changes. We use one zero-sum direction on two missing coordinates, two fixed Gaussian directions, and one fixed positive uniform direction. Each direction is reused across bound levels and budgets. The zero-sum direction tests sharpness and is summarized separately from the random directions.

The 480 pairs are evaluated with ordinary standardization, population standardization with 10^{-5} added to the denominator, and sample standardization with the denominator floored at 10^{-6} followed by asinh. Cases with a constant baseline remain in the results, with ordinary standardization marked as undefined. The native-model subset uses $n \in \{4, 16\}$, budgets 10^{-5} and 10^{-3}, and the zero-sum and positive directions, giving 64 pairs per model. Each first patch is followed by the same 480-step, zero-mean, unit-variance sinusoidal tail. History length is 512 and forecast horizon is 128. No synthetic future labels are supplied, so this study measures response without assessing forecast accuracy. Both the original-sign and sign-reversed TimesFM histories are recorded; numerical protection and branch changes are recorded separately from ideal-formula results.

C.2 Coordinate Interventions

All 42 configurations with constant Linear first patches and at least one original observed first-patch coordinate are retained from the set of 518 configurations. They cover ten sources and 33 distinct source–origin histories, each of length 2,880. Perturbations affect only missing coordinates in the first patch, at budgets 10^{-3} and 10^{-5} in observed-history scale. Three positive random directions use the original perturbation seeds, $294912 + 10000 \times$ original case index + repeat index. The first six configurations, drawn from distinct sources, verify the implementation before the remaining configurations are processed; they remain in the final results.

Table 4: Inputs, interventions, and fixed conditions in the mechanism experiments.

Experiment	Evaluation inputs	Intervention	Fixed conditions
Observation conditions	480 formula pairs; 64 native pairs per model	Observed dispersion and count; budget and direction	Support length; energy within each budget; unchanged remaining history
Coordinate effects	42 configurations; 33 histories; 10 sources; 2 models	Observed and missing normalized coordinates	Baseline centering, scaling, output restoration and masks
Additional datasets	2 datasets; 20 origins each; 240 configurations per model	Three fixed perturbation directions	Origin rule, masks, imputers, event threshold and primary scores
Equal absolute error	10 sources $\times$ 1 window; 2 models	Four error directions	Equal absolute error at every coordinate
Factorial statistics	10 sources $\times$ 3 imputers; 2 models	Input content and historical statistics	Model weights and forecast labels
Initial gap length	2 sources $\times$ 3 origins; 3 models	Missing length and perturbation amplitude	Origin, imputation rule and perturbed coordinates
Real windows	TimesFM; 10 sources $\times$ 2 window groups $\times$ 3 imputers	Positive random perturbation; fixed statistics	Observed scale and direction within a window
Truth-directed correction	TimesFM; the same 20 Linear windows	Local correction proportion	No increase in error at any corrected coordinate
Context boundary	6 scenarios $\times$ 3 repeats; 3 models	History length (steps): 512, 1,024 or 2,880	Physical gap, imputed values and perturbation
Numerical range	Repeat 0 of the same 6 scenarios; 2 models	Eight amplitudes from 10^{-2} to 10^{-9}	Direction; 512-step history; common scale
All constant patches	TimesFM; 518 windows with constant Linear first patches	Three random directions; fixed statistics	All windows retained; observations preserved
Later missingness	2 histories with fully observed first patches; 2 models	Small changes at missing coordinates in the second patch	First-patch forward values; common inputs and directions

For a baseline and perturbed input, let h^0 and h^1 be their native normalized histories. The four states h^{ab} use h^a on original observed coordinates and h^b on original missing coordinates, for $a, b \in \{0, 1\}$. They retain native mask channels and are evaluated on the baseline raw input with fixed centering and scaling statistics, giving forecasts y_{ab}. We report observed-only effect $y_{10} - y_{00}$, missing-only effect $y_{01} - y_{00}$, joint effect $y_{11} - y_{00}$, and interaction $y_{11} - y_{10} - y_{01} + y_{00}$. The conditional observed effect $y_{11} - y_{01}$ is the negative of the effect of restoring the baseline representations at observed coordinates while leaving missing-coordinate representations perturbed. The Euclidean norms of these effects cannot be added to obtain percentage contributions. These mixed histories intervene on internal representations; they are not imputations of raw values.

TimesFM replays the baseline means and scales in all 28 RevIN calls for its 720-step forecast, including full-history preprocessing, both signed historical passes, and future forward and inverse transformations. Only historical value channels are replaced; mask channels and the transformer's internal RMS normalization remain native. The primary analysis of the first 128 forecast steps directly reflects the intervention on historical coordinates. Its 720-step extension also holds the future RevIN schedule fixed. Consequently, the difference between the joint mixed-state forecast and the native perturbed forecast includes changes in future statistics over the longer forecast horizon. Toto replaces the post-asinh historical values in a single forecast block and retains baseline output location and scale. Both models preserve their native quantile processing and output constraints.

We check that baseline replay reproduces the original forecast, that the intervention code agrees with the model's forecasting routine, and that unchanged inputs preserve outputs across all ten TimesFM

channels and all nine Toto quantiles. Effects and signed losses are recomputed from saved arrays. We first average directions within a configuration and then configurations sharing a physical history. Figure 4 reports within-source medians and interquartile ranges over these histories. Across-history summaries contain 33 histories; repeated masks are not counted as independent histories.

C.3 Evaluation on Additional Datasets

The two additional datasets and their input rules were fixed before forecasting. The Beijing target is hourly airport temperature from the UCI Beijing PM2.5 dataset during 2010–2014. Sceaux uses minute-level global active power in kilowatts from the UCI household dataset during 2006–2010. Both datasets are distributed under CC BY 4.0. Candidate origins require 2,880 preceding and 720 following finite measurements at consecutive timestamps. Twenty origins are selected evenly from each source's chronologically ordered eligible list, using measurement availability without filtering values, variance, masks, or forecast outcomes. Naturally missing measurements are never imputed as ground truth. Beijing histories overlap at 19 origin pairs, whereas future intervals do not overlap; Sceaux histories do not overlap.

Block masks with length 50 and nominal rates 0.1, 0.3, and 0.5 are generated on each full source timeline independently of its values, then sliced at the selected histories. Blocks may therefore cross a history boundary. Linear and Mean use only the visible history and preserve observations exactly. This yields 240 configurations, all with positive observed-history scale. The first patch contains missing coordinates in 98 configurations and is fully observed in the remaining 142; each of the latter contributes three unchanged-input records. There are 27 constant completed first patches and 30 zero-bound configurations. These counts refer to configurations; multiple configurations can share the same forecast origin.

The primary perturbation budget is 10^{-5} observed-history standard deviations over the first 32 coordinates, with three fixed positive uniform directions shared between imputers. The event threshold is a direction-averaged 720-step forecast displacement greater than 0.01 in that scale. Prespecified scores are whole-history missing fraction, exact first-patch constancy, negative completed first-patch scale, and negative observed bound, with scales divided by the same observed-history standard deviation. Tied scores are grouped when computing average precision. For inspection budgets of 10%, 20%, and 30%, we report the expected number of detected events when configurations tied at the cutoff are selected uniformly at random. No score or event threshold is fitted to forecasts on the additional datasets.

We also report the first-patch missing fraction to compare quantities on the same support. This additional comparison uses the missing-coordinate locations recorded before forecasting and was added after Toto results and the TimesFM pilot were available. It leaves every primary score, sample, and threshold unchanged. All comparisons separate the full set of configurations, perturbable patches, nonconstant patches, and their intersection. For subsets containing only events or only nonevents, we report counts without interpreting how well the scores separate the two groups. Forecast RMSE, signed and absolute changes, and distribution summaries retain all directions and zero responses. Independence here refers to mechanism development; it does not establish exclusion from model pretraining.

C.4 Perturbation and Normalization Controls

Random perturbations in the first-patch experiments affect only missing coordinates in that patch. Their weights are sampled from $\mathrm{Uniform}(0.1, 1)$ and normalized by within-patch RMS. The real-window study defines the budget using observed-history standard deviation and uses the same direction across imputers within a window. The initial-gap-length study uses a fixed positive alternating pattern. The two relative amplitudes are 10^{-3} and 10^{-5}. Truth-directed correction replaces missing coordinates in the first patch with $(1-\alpha)z + \alpha x$, for $\alpha \in \{10^{-5}, 10^{-3}, 1\}$, leaving all other coordinates unchanged. This reference intervention uses historical ground truth and cannot be deployed directly at naturally missing coordinates.

The constant-patch evaluation includes all 518 windows with constant Linear first patches. Three fixed random directions and a budget of 10^{-5} compare native processing with fixed historical statistics. Responses are first averaged over the three directions within each window and then aggregated by source. We average the within-source proportions equally across ten sources. This set includes

development and previously probed windows and covers all inputs meeting the condition within this benchmark. All perturbed inputs equal their baselines in float32 in 28 windows, which remain in every aggregate. Across the full 4,515-configuration input audit, the first nonpadding patch has valid width 26 or 31 in 30 short-history configurations and width 32 otherwise. All 3,108 response records for this set were recomputed from saved arrays, and ten previously studied native baselines were reproduced exactly.

The later-missingness control inspects subsequent statistical supports in 20 windows with fully observed, constant first patches. Eighteen meet the condition but correspond to only two distinct histories. One window is chosen for each history by ordering the earliest affected patch, missingness rate, and seed. The two histories, denoted A and B, are both from Australia_Solar. Their forecast origins are 20,664 and 26,424, their block-missingness rates are 20% and 30%, and both mask seeds are 43. Raw perturbations change only 2 and 21 missing coordinates, respectively, in the second patch. Both models use the same 12 sets of actual inputs. Native TimesFM processing produces small changes in first-patch forward values. An additional control therefore fixes those values to the baseline in both signed forward passes, while every other historical forward value remains elementwise identical to native processing and all unperturbed output channels retain their baseline values. Toto's native first-patch forward values are already unchanged.

The context experiment fixes a physical missing interval of length 64 at 512 steps before the forecast origin. It is imputed only once, after which different suffixes are taken from the same 2,880-step history. All three context lengths retain all 32 perturbed coordinates. The perturbation budget and forecast displacement use the same visible scale from the longest history. Each scenario uses three prespecified random directions. Sundial also varies the forecast-sampling seed, pairing it across baselines, perturbed inputs, and contexts. Direction and sampling-seed effects are not estimated separately. An additional TimesFM control retains a 512-step network input but replaces its forward values with the normalized values corresponding to the longer history. Output restoration retains the native short-history statistics, and displacement is evaluated against that control's own baseline.

The numerical-range study reuses the six scenarios from repeat zero of the context experiment. It fixes the perturbation direction, a 512-step history, and the common scale, and evaluates TimesFM and Toto separately. Each model performs 54 forecasts, producing 48 response records. Earlier baselines and the inputs, first-patch forward values, and point forecasts at the two previously tested amplitudes are reproduced exactly, giving 18 checks of exact agreement per model. Although the two groups of real windows are matched by configuration, missingness rate, and seed, they occur at different origins. Their between-group differences are descriptive; the effects of normalization are identified primarily through interventions using the same method and window.

The fixed-statistic interventions use the original forecasting entry points and first verify that recomputation with an input's own statistics reproduces its native forecast. TimesFM checks all ten output channels, and Toto checks all nine quantiles. Random generation uses paired seeds. Records include original inputs, masks, normalized histories, forecasts, labels, and code and weight versions. All mechanism controls use sequential CPU inference with fixed model parameters.

Column zero is fixed for both additional sources. Forecast origins are 3,600, 5,040, and 6,480 for Exchange, and 3,600, 5,088, and 6,576 for Solar-AL, all indexed by array position. The forecast horizon is 720, and history length is 2,880 in the initial-gap study. Initial missing lengths are 8, 16, 24, 28, 30, 31, 32, 48, and 64. TimesFM and Toto share 54 sets of inputs. Sundial verifies 18 sets at the three prespecified lengths 30, 31, and 32. Error budgets at both amplitudes are computed consistently from the actual first-32-step input.

For each of ten sources, real-window selection matches a constant Linear first patch to a positive observed-variance-bound condition, prioritizing equal configuration, missingness rate, and mask seed, with selection seed 275912. All 20 selected windows satisfy this preferred matching condition. Input selection excludes the ten cases used during early hypothesis development. The three imputers share one observed-history scale and random direction within each window. This experiment uses only one realization of the random direction. The context experiment separately uses three directions per scenario; the two experiments are not combined into a single population estimate.

Direction seeds in the complete set of constant-patch configurations are $294912 + 10000 \times$ window index + repeat index, with all three repeats retained. History length is 2,880 in 517 windows and 2,879 in one window, whose actual first-patch width is 31. Direction seeds in the later-missingness control are $310912 + 10000 \times$ original audit-candidate index + repeat index; candidate indices for

A and B are 7 and 13. The later-missingness inputs and first-patch-fixing protocol were finalized on September 12, 2026 at 04:10:25 and 06:49:32 UTC, respectively. Every selected condition is reported. The statistical control includes historical output restoration, whereas first-patch fixing replaces only the first-patch values in the neural network's historical input.

The context gap is fixed at $[t-512, t-448)$ and filled with the observation at $t-448$, which is still within history. Perturbations affect only $[t-512, t-480)$, and all three history lengths retain every affected coordinate. Direction seeds are $286912 + 100 \times$ scenario index + repeat index. Sundial forecast seeds are $287912 + 100 \times$ scenario index + repeat index; scenario and repeat indices start at zero. The TimesFM long-history-normalization control replaces the historical values supplied to the neural network before forecasting. It adds no long-history tokens to the network and does not replace subsequently generated forecast values.

The fixed weight revisions are `1d952420fba87f3c6dee4f240de0f1a0fbc790e3` for TimesFM, `3212e42564493f520593e5414af4367fc4b49226` for Sundial, and `685e4ae3e2be8d8998025e53dd98e7fdcb296a89` for Toto 2.0 22M. Toto's decode-block upper limit is 768 steps. A 720-step forecast is internally padded to 23 patches of 32 steps, or 736 steps, computed in one pass and truncated to the first 720 steps. Its dedicated missingness branch is disabled. The input mask treats every imputed numerical value as available, while the original missingness mask is used only for this study's analysis. The default backfilling of statistics for short observed segments is not enabled. Conclusions are restricted to this released configuration.

Factorial statistical interventions independently use complete or imputed history for historical content and statistics, holding future labels fixed. TimesFM retains native running-statistic updates while generating future values. If content and statistical effects offset each other, both signs are retained and their ratio is not reported as a percentage of explained effect. The visible-value-statistic control is reported separately from the complete-history reference.

The primary benchmark uses source collections as the resampling unit, with 10,000 draws. Dependent configurations, methods, and seeds within a condition are resampled jointly with their source. Windows in the mechanism controls are condition-selected and limited in number. We use paired descriptions and do not construct independent-sample confidence intervals by treating repeated methods or nearby origins as independent observations. All quantiles use linear interpolation. Truth-correction records with unchanged float32 inputs retain zero response and are checked against their own baselines.

C.5 Numerical Precision and Perturbation Size

The numerical-range study tests how small perturbations can become before the implemented models cease to show the response in Eq. (4). Across the six fixed scenarios, TimesFM forecast displacement is zero at both 10^{-8} and 10^{-9}, while Toto retains a maximum displacement of 0.0396 at 10^{-9} on the same inputs. At each of these two amplitudes, four raw inputs are unchanged in float32. The other two raw inputs change, but TimesFM first-patch forward values and point forecasts still reproduce the baseline. The perturbation amplitude below which the response disappears differs between the two models. These finite-range results also do not establish a nonzero limiting response as perturbations approach zero.

D Proofs and Numerical Implementation

For $n > 0$, the minimum in Proposition 1 follows from a mean decomposition. Let a be the observed mean on the support and b the mean of the completed vector. The sum of squared deviations at observed coordinates is $\sum_{i \in O \cap I}(x_i - a)^2 + n(a-b)^2$. Adding the nonnegative squared deviations at missing coordinates and dividing by L_I gives the lower bound. When every missing value equals a, we have $b = a$ and the missing-coordinate term is zero, so the bound is attainable. With only one observation, or identical observations, $v_I = 0$. This conclusion does not depend on whether missing coordinates are contiguous.

For Proposition 2, let C be the centering matrix and $u = Cz/\|Cz\|_2$. The derivative of ordinary standardization is $\sqrt{L_I}(\mathsf{I} - uu^\top)C/\|Cz\|_2$. Centering and orthogonal projections have operator norm at most one, giving the derivative bound $1/\sigma(z)$. The segment connecting two feasible completions is also feasible and thus shares the same positive lower bound. With at least two missing coordinates,

a nonzero zero-sum direction exists on them. At mean completion, this direction is orthogonal to the centered vector, and the local derivative attains $1/\beta_I$.

Continuous scale protection can be included in the corresponding exact-arithmetic analysis. When the denominator is the population standard deviation plus τ, the coefficient becomes $1/(\beta_I + \tau)$. When sample standard deviation is clipped below at τ and followed by asinh, let $\beta_s = \sqrt{nv_I/(L_I - 1)}$; the coefficient is at most $1/\max(\beta_s, \tau)$. The latter result follows because projection onto a Euclidean ball cannot increase Euclidean distance and asinh is 1-Lipschitz. Floating-point rounding requires separate treatment. TimesFM replaces an excessively small divisor directly with 1, so the ordinary-standardization bound does not apply directly across its protective branch.

The ordinary-standardization bound uses the population standard deviation. Sample-standard-deviation implementations require the sample-size correction. Continuous scale floors and asinh also require adjustments for the specific transformation. TimesFM's low-variance branch must be checked separately because its normalization rule is not smooth across that branch. Conclusions about the implemented models use their recorded network inputs and forecasts, with the ideal bounds restricted to the numerical branches for which they apply.

E Supplementary Results

E.1 Reconstruction Error and Forecast Loss

Section 5.1 reports reconstruction-based selection on the full set of configurations and on temporally separated windows. In the temporal comparison, the first $\lfloor W/2 \rfloor$ origins select the imputer and the remaining origins evaluate it. The mean rank correlation is 0.485, with a source-bootstrap 95% interval of $[0.353, 0.612]$. Forecast RMSE is more than 2.19% above the best candidate in half the conditions and more than 16.86% above it in 10% of conditions. Source-bootstrap 95% intervals for these two thresholds are $[0.54\%, 4.10\%]$ and $[7.28\%, 28.85\%]$, respectively. Earlier forecast labels end no later than the first evaluation origin. These temporal results reuse the benchmark sources and are distinct from validation on previously unseen sources.

Separately replacing content and historical statistics in the 30 source–imputer combinations reveals statistical effects in both directions. Statistics increase forecast loss in 21 Sundial combinations and decrease it in nine; the corresponding counts for TimesFM are 16 and 14. Replacing the statistics with visible-value statistics yields equally weighted source RMSE changes of $+0.41\%$ and $+1.84\%$, respectively, with both reported intervals including zero. After remapping output scales for the saved Sundial forecasts, the absolute output-restoration effect exceeds the input-normalization effect in 22 of 30 combinations. The joint statistical effect therefore cannot be attributed entirely to changes in observed-coordinate representations, and direct statistic replacement does not provide a general accuracy advantage.

Using historical forecast errors for calibration does not consistently improve method selection. When calibration uses complete historical reference values from three windows, the median forecast-RMSE gap relative to the best candidate is 2.303% for reconstruction-based selection and 2.273% for forecast-based selection. The difference in the proportion of selections within 5% of the best candidate is five percentage points, with a source-bootstrap interval of $[-4.29, 16.52]$ percentage points. When calibration is constructed only from observed coordinates and restricted to eight basic methods, the corresponding median relative RMSE gaps are 3.410% and 5.202%. These comparisons do not establish a method-selection rule that generalizes across sources.

Seven masks at Solar-AL origin 3,600 yield exact recovery of the complete history before perturbation. They provide a controlled comparison around the same historical truth, with six distinct perturbed inputs per amplitude. At amplitude 10^{-5}, reconstruction RMSE is only $(1\text{–}2) \times 10^{-5}$ historical standard deviations, yet TimesFM forecast RMSE is 3.91%–5.66% above the exact-recovery baseline. Toto changes by -0.89% to $+0.73\%$. Reducing the amplitude from 10^{-3} to 10^{-5} decreases reconstruction error 100-fold, while median forecast-displacement ratios remain 1.003 for TimesFM and 0.9999 for Toto. The corresponding ratio is 0.0132 for the three Sundial controls. Thus, within this numerical range, improving reconstruction need not proportionally reduce the forecast response.

The exact-recovery control uses seven masks from the same Solar-AL history at origin 3,600, with six distinct perturbed numerical inputs per amplitude. Directly comparing the 10^{-5} and 10^{-3} inputs gives TimesFM forecast-RMSE changes of -0.0112% to $+0.00484\%$ relative to the higher-error

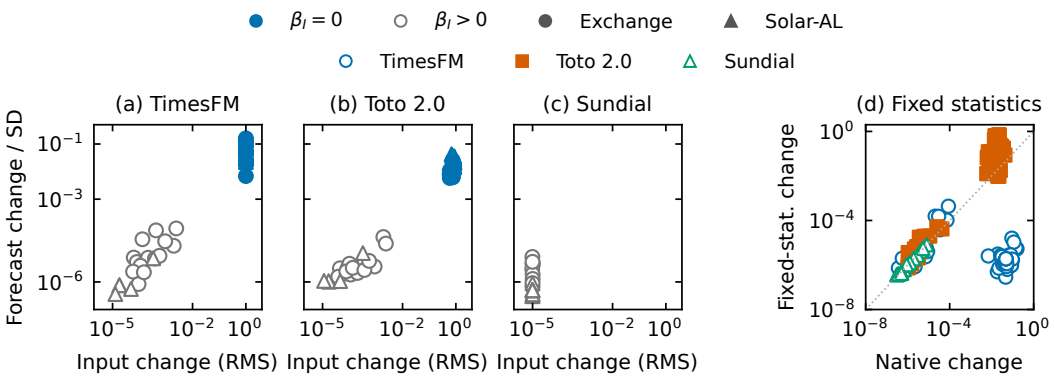


Figure 6: Model responses at perturbation budget 10^{-5}. (a–c) Native normalized first-patch changes and forecast displacements for TimesFM, Toto, and Sundial. Filled and open symbols mark zero and positive observation bounds; circles and triangles distinguish Exchange and Solar-AL. Bounds use Sundial's full-history support and the causal support of the affected first patch in TimesFM and Toto. Each point is a prespecified source, origin, and missing-length condition. (d) Native and fixed-statistics forecast displacements for the same conditions; model colors and markers are identified in the legend. The dotted diagonal marks equal responses.

input. Its 3.91%–5.66% gap from exact recovery persists at the lower amplitude. All 17 model pairs reduce or preserve absolute reconstruction error at every missing coordinate. The shared history makes this a controlled case study, with all matched results retained.

E.2 Constant Supports across Models

Linear imputation produces constant first patches in 518 of the 4,515 window–mask configurations, or 11.5%. The complete first patch is nonconstant in 508 of these cases, showing that imputation creates the constant support. Of the 518 configurations, 476 have no first-patch observations and 42 retain observations. More broadly, 575 configurations with missing first-patch values have a zero observed-variance bound. Even with a zero observed bound, the selected imputer can produce a nonconstant patch.

Of the 54 TimesFM conditions in the initial-gap-length study, 20 have positive observed-variance lower bounds, while 34 have zero bounds and actually constant imputed first patches. At amplitude 10^{-5}, normalized first-patch change is at most 0.00244 in the first group and approximately 1 in every condition of the second group. The corresponding normalized forecast displacements are at most 8.70×10^{-5} and range from 0.00689 to 0.15968, respectively. Some photovoltaic conditions retain 24 true observations with identical values, which still do not yield a positive lower bound.

Toto 2.0 responses also depend on the observed bound in the same 54 inputs. Under zero bounds, first-patch displacement after asinh ranges from 0.591 to 0.870, and median forecast displacements are 0.00958 for Exchange and 0.02198 for Solar-AL. Under positive bounds, their maximum forecast displacements are 4.25×10^{-5} and 1.11×10^{-5}. These results show amplification under both TimesFM's conditional divisor and Toto's continuous scale protection. Sundial's full-history support has a positive bound in all 18 prespecified conditions, with small responses in this comparison.

E.3 All Constant-Patch Configurations

Figure 7 summarizes all 518 windows with constant first patches. Direction-averaged forecast displacement exceeds 0.01 observed-history standard deviations in 454 windows, giving 89.5% after equal weighting across sources. Source median displacements range from 0.00995 to 0.05394. Fixing historical statistics, including output restoration, reduces every window's direction-averaged displacement to at most 4.31×10^{-5}. This comparison identifies the combined effect of input normalization and output restoration as the main source of amplification. Forecast RMSE increases by more than 1% in 175 windows and decreases by more than 1% in 116, corresponding to source-weighted proportions of 34.2% and 20.0%. The source-weighted mean signed change is $+0.76\%$. Forecast accuracy varies across inputs within this amplification condition.

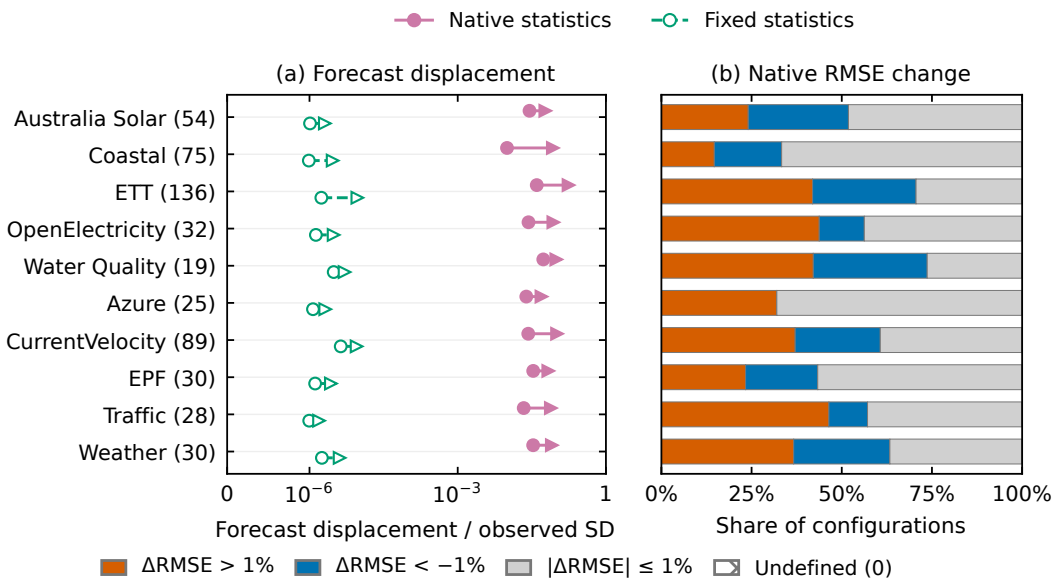


Figure 7: TimesFM responses in all 518 constant-patch configurations. The two panels share source rows; labels give configuration counts. (a) Source medians (circles) and 90th percentiles (triangles) of direction-averaged forecast displacement under native (filled) and fixed (open) historical statistics. The 90th percentile marks the 90% position in the sorted values. Lines connect these summaries within each source; they are not confidence intervals. The scale is linear below 10^{-7} and logarithmic above it, retaining zero displacement. (b) Source shares classified by direction-averaged relative forecast-RMSE change under native processing: above 1%, below -1%, within $\pm 1\%$, or undefined. All relative losses are defined in this cohort. Membership in the $\pm 1\%$ interval does not imply stability in every direction. All 28 coastal windows without actual input changes and all earlier development windows remain included. Fixed statistics include output restoration.

E.4 Imputer Comparisons under Matched Observation Conditions

The 20 real windows were selected by their observation conditions; no window was replaced based on its forecast result. In the constant-Linear-first-patch group, both Linear and Mean produce constant first patches. Their median forecast displacements under 10^{-5} perturbations are 0.03934 and 0.02112, respectively. SAITS produces no constant first patch in the same ten windows and has median displacement 1.17×10^{-6}. In the positive-bound group, the median displacement of every method is at most 2.03×10^{-6}.

Reducing amplitude from 10^{-3} to 10^{-5} yields median forecast-displacement ratios of 0.999 for Linear, 0.999 for Mean, and 0.013 for SAITS in the constant group. For Linear and Mean, normalization by a small standard deviation largely offsets the reduction in perturbation magnitude. Forecast RMSE falls in four windows and rises in six for Linear; for Mean, it falls in five and rises in five. Lower sensitivity in this comparison does not establish an overall accuracy advantage for SAITS, and a large response does not necessarily increase forecast error.

E.5 Truth-Directed Corrections and Statistical Support

Correcting the first patch toward its historical truth never increases absolute error at a corrected coordinate. In the constant group, $\alpha = 10^{-5}$ reduces first-patch reconstruction RMSE by approximately 0.001%, while median normalized first-patch error relative to complete history falls from approximately 1 to 0.0408. It falls further to 0.000400 at $\alpha = 10^{-3}$. Forecast-RMSE changes at $\alpha = 10^{-5}$ range from -2.71% to $+2.96\%$. Even after all missing values in the first patch are restored exactly, the changes range from -10.36% to $+3.10\%$. Reducing reconstruction error can change normalized inputs substantially without producing a proportional improvement in forecast accuracy.

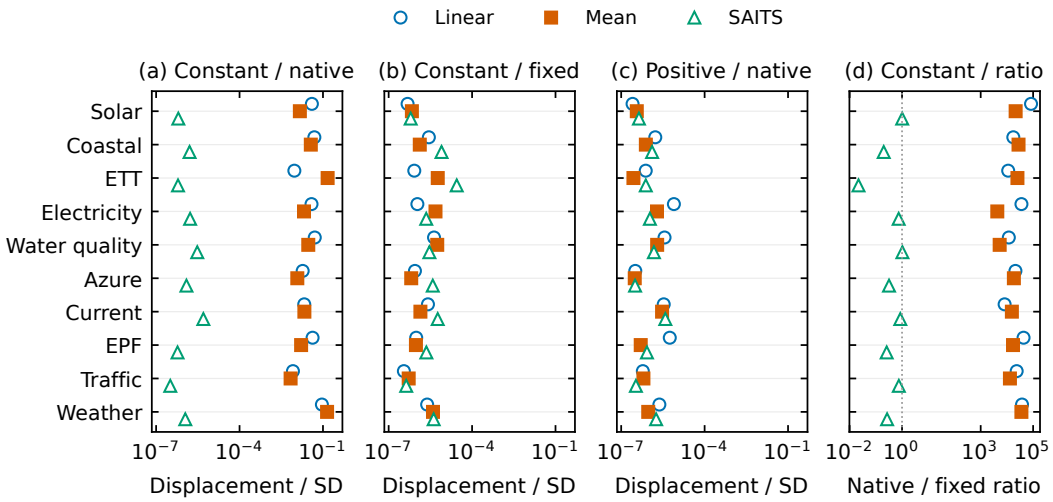


Figure 8: TimesFM responses in preselected real windows at budget 10^{-5}. (a,b) Native and fixed-statistics responses in the constant-Linear group. (c) Native responses in the positive-observation-bound group. (d) Native/fixed-statistics response ratios in the constant-Linear group. Each source contributes one window to each group. The three imputers share the same raw perturbation direction and observed-history scale within a window. Solar denotes Australia_Solar. Statistical controls preserve imputation differences. These selected groups allow comparisons under the stated observation conditions; they do not estimate how often those conditions occur in the population.

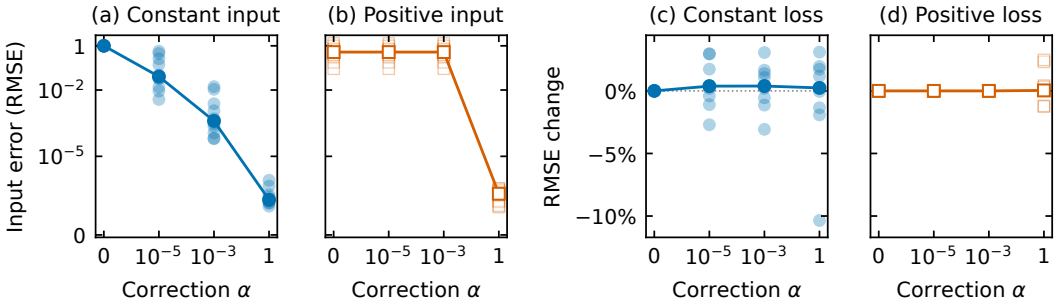


Figure 9: Local corrections of TimesFM inputs toward historical truth. Panels separate constant-Linear and positive-observation-bound groups for (a,b) normalized first-patch error relative to complete history and (c,d) signed forecast-RMSE change. Faint points show individual cases and connected markers their medians. Only originally missing first-patch coordinates are corrected, at discrete proportions. The value $\alpha = 1$ restores those coordinates, without implying that the entire history is complete. Records without actual float32 changes retain their original values. The input-error scale is linear below 10^{-8} and logarithmic above it, retaining exact zero errors.

The affected support can extend beyond the first patch. The Exchange and Solar-AL cases selected by their input conditions have two and four consecutive initial statistical supports, respectively, with no observed variation and constant baseline completions. Applying a common input scale only to the first patch increases Exchange forecast displacement from 0.04522 to 0.05911. Covering both supports reduces it to 2.69×10^{-5}. In Solar-AL, the normalized values of 64 unchanged observations change substantially; covering all four supports reduces forecast displacement from 0.04950 to 7.22×10^{-7}. All output channels of the four scale-intervention baselines match their respective native baselines. Differences are also present in the first 128 steps forecast in a single model pass, so accumulation through longer autoregressive forecasting cannot be their sole explanation.

Sixteen positions in the photovoltaic case receive equal positive perturbations. Equation (4) predicts standardized values of $-1/\sqrt{5}$ and $-1/\sqrt{7}$ at unchanged coordinates in constant statistical supports of lengths 96 and 128, respectively. The third and fourth patches each contain 32 unchanged true

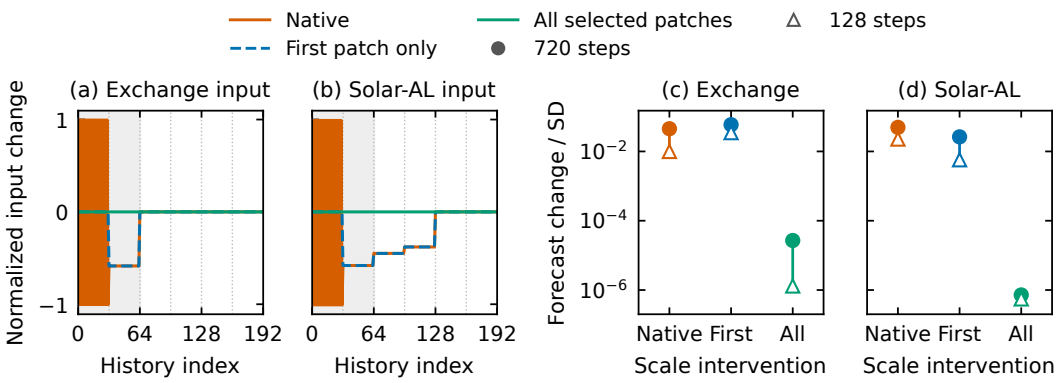


Figure 10: Normalization changes in TimesFM can propagate to unchanged observations. (a,b) Changes in actual historical forward values for Exchange and Solar-AL; gray regions mark missing intervals. (c,d) Forecast responses under native statistics, a first-patch-only scale intervention, and an intervention covering all selected supports. Filled circles and open triangles give 720- and 128-step displacements. Controls preserve raw input, means, output restoration, and model weights, changing only input-scaling divisors on specified supports. Results cover two cases selected by input conditions.

observations. Their actual low-amplitude values are approximately -0.44945 and -0.37976, with maximum absolute deviation from the ideal values at most 0.00224. This calculation uses the existing recorded network inputs, including their finite-precision differences. It is a consistency check on the same experiments.

Constancy across several patches is uncommon in the original set of constant-patch configurations. Of 518 windows, 508 remain constant only in their first valid patch. The other ten all come from Australia_Solar, remain constant for 7–21 patches, and have constant complete first patches as well. This input audit does not establish that all subsequent hidden true values are constant and does not replace forecasting experiments.

E.6 Missing Values after a Fully Observed First Patch

Amplification can also occur after a fully observed first patch. In two Australia_Solar histories selected by the input conditions in Appendix C, we change only positions 62–63 in history A and positions 43–63 in history B. Positions 0–31 remain observed and unchanged. At budget 10^{-5}, TimesFM forecast displacements after fixing the first-patch forward values are 0.07260 and 0.06326; native Toto displacements are 0.05546 and 0.04660. The first-patch-fixed and native TimesFM forecasts differ by at most 2.72×10^{-6} observed-history standard deviations. The main response therefore arises beyond the first-patch forward values.

The two histories contain 30 and 11 observations before the perturbed coordinates whose normalized values change despite unchanged raw values. In history A, observed positions 32–61 and perturbed positions 62–63 share a statistical support. Normalization can therefore affect observations preceding the missing coordinates. Table 5 reports averages over three directions with identical first-patch forward values in both models.

Fixing the statistics affects the two models differently. TimesFM's response falls to the 10^{-6} range, while Toto's displacement increases to 0.1143 and 0.1685. Toto's baseline local scale remains at the 10^{-6} floor in these constant initial supports. Holding it fixed amplifies the perturbation at missing coordinates, giving affected-patch input displacements of 4.224 and 13.195 even after asinh. Unchanged observations retain their representations under this control. The result distinguishes amplification at imputed coordinates from changes in observed-coordinate representations and explains why fixing statistics need not stabilize every model.

Table 5: Later-missingness controls with identical first-patch forward values. The preceding-observation count is the number of original observations before the perturbed coordinates whose representations change. Input displacement measures change in the affected second patch after normalization. Forecast displacement uses the same observed-history scale. TimesFM uses the first-patch-fixed control, and Toto uses native processing.

			Displacement			
Model	History	Preceding observations	Input	Forecast	Fixed statistics	RMSE change
TimesFM	A	30	1.403	7.26×10^{-2}	5.40×10^{-6}	$+2.54\%$
TimesFM	B	11	1.283	6.33×10^{-2}	1.18×10^{-6}	$+5.90\%$
Toto 2.0	A	30	0.623	5.55×10^{-2}	1.14×10^{-1}	$+0.60\%$
Toto 2.0	B	11	0.930	4.66×10^{-2}	1.69×10^{-1}	-2.20%

E.7 Context Length with a Fixed Physical Gap

Figure 11 reports the context comparisons from Section 5, with the physical gap and imputed values fixed.

With the physical gap and imputed values fixed, median low-amplitude forecast displacement at a 512-step history is 0.31804 for TimesFM and 0.15530 for Toto. At a 1,024-step history, both medians are approximately 4×10^{-6}. In the 512-step condition, the affected patch lies at the beginning of the causal statistical support and has a zero observed-variance bound. Longer histories provide preceding observations with variation for that same patch. Sundial median displacements are at most 7×10^{-6} at all three history lengths.

To separate normalization from the additional information in a longer context, we retain the 512-step network input and replace its normalized values with those obtained from the longer history. Median forecast displacement falls to approximately 10^{-5}. For these six scenarios, the paired comparison shows that changing the statistical support changes the amplification. Under native 512-step processing, forecast-RMSE changes range from -12.27% to $+69.91\%$ for TimesFM and from -35.87% to $+4.83\%$ for Toto.

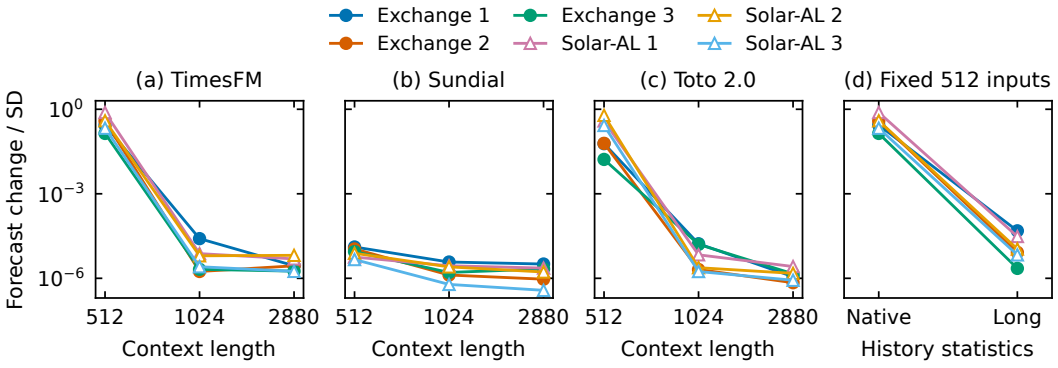


Figure 11: Context controls with a fixed physical missing interval. (a–c) Native responses of three models. Lines give median displacement over three repeats in each prespecified scenario; filled circles denote Exchange and open triangles Solar-AL. Every context retains all perturbed coordinates and uses the observed scale from the longest history. (d) The 512-step TimesFM network input and native output restoration are retained while historical normalized values use longer-history statistics. Each mode uses its own baseline.

E.8 Formula Bounds across Observed Dispersion Levels

All 1,380 defined comparisons satisfy their corresponding bound; 60 ordinary-standardization comparisons at zero bound retain an undefined baseline. In the 30 ordinary-standardization comparisons

along the tight direction with $\epsilon/\beta_I \leq 0.01$, bound utilization is at least 0.9999625. For the 315 random-direction comparisons with positive bounds under ordinary standardization, median utilization is 0.5054. The 90th percentile, marking the 90% position in the sorted ratios, is 0.9943. The sharp coefficient gives the largest local amplification over feasible completions; individual perturbation directions can produce much smaller changes. Recomputing the 76 saved comparisons that have positive observed bounds and use formulas with scale protection gives maximum relative discrepancy 3.91×10^{-13} from their saved values.

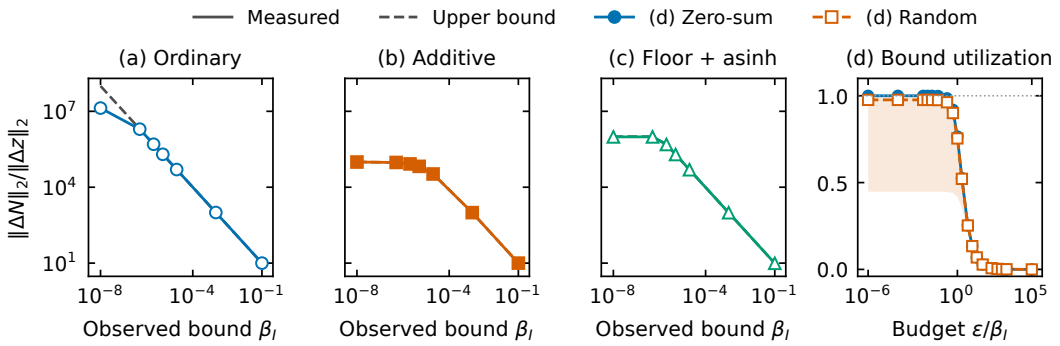


Figure 12: Formula bounds across observed dispersion levels. (a–c) Amplification and its upper bound under three formulas along the tight zero-sum direction at $\epsilon = 10^{-7}$; lines and shading summarize five observation counts. (d) Ordinary-bound utilization against ϵ/β_I for zero-sum and fixed random directions; shading gives the random-direction range. Positive bound levels appear on logarithmic axes. Zero-bound cases remain in the complete results, including 60 undefined ordinary baselines.

The native-model subset contains 64 paired histories per model, including all numerical branches. Each model has 24 pairs that cross a scale-protection boundary, including 16 with a positive observed bound. TimesFM has 40 pairs with ordinary native scaling, 16 with active protection and a defined ordinary comparison, and eight with an undefined ordinary baseline. Figure 3 reports the native responses, including floating-point rounding and the implemented numerical branches. The fixed synthetic reference scale is one; forecast displacements in this experiment are not normalized by the observed scale of a real series.

E.9 Coordinate Effects at Both Forecast Horizons

Table 6 reports the coordinate effects over 128 and 720 steps at both perturbation budgets. The comparisons retain all 42 configurations and all directions, including unchanged inputs. For each model, one of the 33 physical histories has zero coordinate effects at the lower budget. All sources have positive medians for the four contrasts shown in Figure 4. The observed-only effect exceeds 0.01 in 32 TimesFM histories and 26 Toto histories over the first 128 forecast steps; the conditional observed effect exceeds this level in 21 and five histories, respectively. The conditional observed effect depends on the missing-coordinate representations held fixed during the comparison.

At budget 10^{-5} over 128 steps, native RMSE increases by more than 1% in 11 of 33 TimesFM histories and 14 of 33 Toto histories, and decreases by more than 1% in ten histories for each model. A secondary comparison quantifies the relative RMSE change after restoring observed representations as $\mathrm{RMSE}(y_{01}, y)/\,\mathrm{RMSE}(y_{11}, y) - 1$. Ratios are computed per direction, averaged within configurations, and then averaged within histories. For TimesFM, the median change is -0.0863%, with the middle 80% of changes spanning $[-2.9621\%, 3.1483\%]$ (10th to 90th percentiles). For Toto, the median is $+0.0157\%$, with the corresponding range $[-0.4663\%, 0.6048\%]$.

E.10 Forecast Error and Diagnostic Recall on Additional Datasets

The additional datasets provide 240 configurations per model, including all 142 with fully observed first patches. TimesFM has 24 large-response events and Toto has 25. The events occur in 24 and 25 of the 27 constant-completion configurations, respectively. Both models have zero events among the 41 Beijing and 30 Sceaux configurations whose first patches are perturbable and nonconstant. The

Table 6: Coordinate-effect medians over 33 histories. Directions are averaged within each configuration, then configurations sharing a history are averaged. O, U, J, $O \mid U$, and I denote observed-only, missing-only, joint, conditional observed, and interaction effects. All effects use observed-history scale. The 720-step TimesFM extension also fixes future RevIN statistics.

Evaluation setting			Median coordinate effect				
Model	Budget	H	O	U	J	$O \mid U$	I
TimesFM	1×10^{-5}	128	0.05708	0.03046	0.02919	0.01716	0.0559
Toto 2.0	1×10^{-5}	128	0.0177	0.02177	0.02466	0.003759	0.01915
TimesFM	0.001	128	0.05803	0.03066	0.02922	0.01874	0.05949
Toto 2.0	0.001	128	0.02573	0.02177	0.02487	0.004326	0.02606
TimesFM	1×10^{-5}	720	0.05458	0.04117	0.03737	0.02183	0.07785
Toto 2.0	1×10^{-5}	720	0.01955	0.02816	0.02787	0.004257	0.01887
TimesFM	0.001	720	0.05985	0.04679	0.03764	0.03625	0.08734
Toto 2.0	0.001	720	0.03127	0.02857	0.02971	0.004794	0.03161

largest mean displacements in this latter group are 7.82×10^{-4} for TimesFM and 1.15×10^{-3} for Toto. These event counts apply to the fixed budget and sources evaluated here.

Table 7 reports signed and absolute RMSE changes across all configurations, including mean signed changes and percentile summaries. Table 8 reports the prespecified 20% inspection budget on perturbable inputs, corresponding to 11 inspected configurations in Beijing and nine in Sceaux. The expected number of detected events accounts for ties at the cutoff. The source-specific diagnostic comparison in Figure 5 includes all four primary scores and the secondary local-missingness score; the latter uses the same first-patch support as the scale diagnostics. No threshold, combination of scores, or input subset is fitted to the results on the additional datasets.

Table 7: Forecast-RMSE changes on the additional datasets. Signed and absolute relative changes are averaged over directions within a configuration, then summarized across all 120 configurations per source. Absolute values are taken before direction averaging. P05 and P95 mark the 5% and 95% positions in the sorted values, using linear interpolation. Zero responses remain included; positive signed values indicate increased loss.

		Signed change			Absolute change
Model	Source	Mean	P05	P95	P95
TimesFM	Beijing	-0.11%	0.00%	+0.47%	1.33%
TimesFM	Sceaux	-0.11%	-0.44%	0.00%	0.62%
Toto 2.0	Beijing	+0.21%	0.00%	+0.97%	0.97%
Toto 2.0	Sceaux	0.00%	-0.02%	+0.04%	0.29%

Table 8: Event recall with a 20% inspection budget on perturbable configurations. Beijing has 54 configurations and Sceaux has 44; the budget is rounded upward. Recall is the expected fraction of all events found when configurations tied at the score cutoff are selected uniformly at random. Global and local denote whole-history and first-patch missing fraction. Local is an additional comparison; the other four scores were fixed before forecasting.

			Prespecified scores				Secondary
Model	Source	Inspected	Global	Constant	Scale	β_I	Local
TimesFM	Beijing	11	0.182	0.846	0.846	0.786	0.833
TimesFM	Sceaux	9	0.231	0.643	0.643	0.562	0.643
Toto 2.0	Beijing	11	0.308	0.846	0.846	0.786	0.846
Toto 2.0	Sceaux	9	0.333	0.643	0.643	0.562	0.643

E.11 Normalization Bounds on Additional Datasets

We evaluate the corresponding ideal transformations on every saved float32 input pair, using float64 arithmetic for the formulas. Each model retains 720 directions: 294 change the input and 426 are unchanged. Of the directions that change the float32 input, 204 have a positive observed bound, comprising 120 Beijing directions from 40 configurations and 84 Sceaux directions from 28 configurations. Table 9 reports their utilization, a completed-scale reference, and the measured native response.

All 630 defined TimesFM uniform-bound comparisons satisfy the ideal inequality; 81 ordinary baselines are undefined because they are constant, and nine further pairs have a defined formula with zero observed bound and no finite uniform coefficient. Toto's protected formula has 720 finite-bound comparisons, all of which satisfy the inequality. Zero and undefined records remain in these counts.

For TimesFM, we check an exact-arithmetic sufficient condition for ordinary scaling along the entire paired segment at both global and first-patch scales. The global observed sample-scale bound must be at least the protection threshold. The raw first-patch bound must also be at least the local threshold times the larger global sample standard deviation of the two endpoints. Convexity of the standard deviation bounds the global scale along the segment. This joint condition holds for all 630 positive-bound directions, including all 204 with actual perturbations, and none has a native protection flag. Both signed passes have identical branch categories and antisymmetric normalized changes. Over the full set of configurations, 639 directions are ordinary at both endpoints and 81 cross the protection boundary.

Toto has 639 directions above the scale floor at both endpoints, 75 crossing the floor, and six remaining at the floor. For directions that change the input and have a positive observed bound, the median ratio of the implemented model's response to the ideal-formula response ranges from 1.000017 to 1.000309 across the four model–source groups. The baseline completed-scale reference uses the actual imputation and yields larger utilization ratios than the observation-based bound for these configurations. It is a local comparison and does not replace a uniform bound over feasible completions.

Table 9: Normalization bounds on inputs from the additional datasets with a positive observed bound and a nonzero perturbation. D counts directions. Bound utilization is the ideal normalized input change divided by its uniform upper bound. P90 marks the 90% position in the sorted ratios, using linear interpolation. Local denotes the ratio to the baseline completed-scale reference, which is not a uniform bound over completions. Native/ideal reports the median ratio of the actual first-patch response to the ideal-formula response. TimesFM uses ordinary standardization and Toto uses sample-scale protection followed by asinh.

			Bound utilization		Reference ratios	
Model	Source	D	Median	P90	Local median	Native/ideal
TimesFM	Beijing	120	0.2561	0.7012	0.4884	1.000309
TimesFM	Sceaux	84	0.3917	0.8075	0.576	1.000017
Toto 2.0	Beijing	120	0.2313	0.5621	0.4247	1.000047
Toto 2.0	Sceaux	84	0.3408	0.704	0.4685	1.000053