SPIMPUTE: STRUCTURE-PRESERVING HISTORY IMPUTATION FOR FROZEN TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series foundation models forecast from history, but missing intervals can change their predictions. Existing imputers often focus on reconstruction error without considering how well downstream models can use the data. We propose SPImpute to fill missing values from observed history and a given period. Its sparse quadratic reconstruction combines local differences with reference increments across periods. Decomposing the correction penalty identifies when missing intervals interact and when the penalty reduces to degree-weighted reference-value fidelity. For eligible long gaps, a same-phase candidate corrects donor levels when all available donors lie on one side of the missing position. Masked validation tasks re-estimate drift. If correction leaves the actual candidate unchanged, SPImpute retains the original procedure. Extensive experiments show its effectiveness. On nineteen series at five missing rates, SPImpute ranks first among 21 imputers under two foundation models, with mean MSE ranks of 5.39 and 6.38. MSE reductions relative to Linear, averaged equally across data sources, are 10.81%, 10.11%, and 2.69% under Sundial, TimesFM-2.5, and Chronos-2, respectively. Independent experiments confirm conditional coupling gains for repeated pulse gaps. We compare same-phase correction with the uncorrected procedure using 1,152 synthetic histories. Across 24 drift settings near the forecast origin, mean relative MSE reductions are 92.9% and 94.9% under TimesFM-2.5 and Chronos-2, respectively; all 576 interior-gap inputs remain unchanged. Median CPU imputation time is 78.66 ms.

Keywords: time series imputation; reference increments; observation constraints; frozen forecasters

1 INTRODUCTION

Pretrained time series models forecast across domains from historical observations (Das et al., 2024; Liu et al., 2025). However, real-world time series can have quality problems. Equipment outages and communication failures, for example, can leave gaps in historical data that require imputation (Du et al., 2023). The imputed history determines the values and temporal relationships supplied to a predictor at inference. Different imputers can affect downstream forecasting quality, so they must preserve reliable observations and reconstruct changes within missing intervals.

Existing methods include interpolation and state-space approaches based on neighboring observations and dynamics (Moritz & Bartz-Beielstein, 2017), difference penalties that regulate local variation (Eilers, 2003; Tibshirani, 2014), and deep models that learn temporal dependencies (Du et al., 2023; Zhou et al., 2023; Talukder et al., 2024). When reconstruction error drives optimization or candidate selection, the criterion measures numerical agreement with ground truth. It does not directly assess whether the imputed history benefits a downstream task. Reconstruction error may therefore not accurately reflect downstream forecasting performance.

To address this, we propose SPImpute, an imputation algorithm for forecasting with foundation models, using observed history and a given period to fill missing intervals. It fixes observed entries and solves a sparse quadratic problem over missing entries, combining local differences with reference increments across periods. The reference supplies recurring changes, while observations constrain gap boundaries. For eligible long gaps, a same-phase candidate corrects donor levels using the median observed lagged slope when donors lie entirely on one side of a missing position. Masked historical

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.

validation selects weights and compares candidates, rebuilding references and drift after masking. SPImpute thus uses temporal structure to reconstruct the model's historical inputs, aiming to improve downstream forecasting.

Consider daily electricity consumption with recurring peaks and valleys whose overall level changes across days (Figure 1). An outage removes a peak and valley but leaves observations on both sides. The preceding day provides a recurring shape at a different level. Linear interpolation attenuates the variation; copying the previous cycle carries over its historical level and creates boundary mismatches. SPImpute fixes the observed boundaries and uses reference increments to guide changes within the gap. The example shows how current observations and periodic context jointly constrain the reconstructed history.

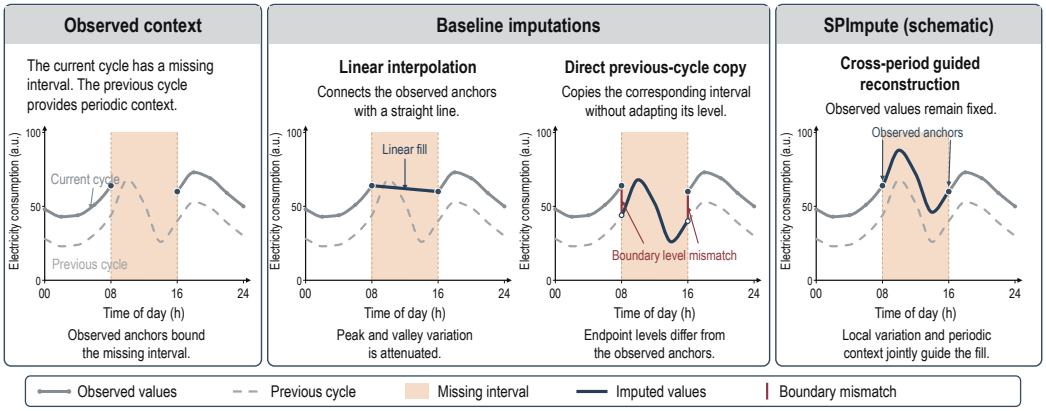


Figure 1: Observation-guided repair of a missing consumption interval. From left to right: observed context, linear interpolation, direct previous-cycle copying, and a schematic SPImpute reconstruction. Solid gray curves show current observations, dashed gray curves show the previous cycle, and shading marks the gap. Linear interpolation attenuates peak and valley variation; direct copying introduces boundary level mismatches, marked in red. The SPImpute sketch combines local variation and periodic context while preserving observed values. Curves illustrate reconstruction behavior, with consumption in arbitrary units.

The contributions are as follows.

- We define a mathematical problem for filling gaps: keep observed values unchanged and estimate missing values using changes between nearby time points and reference changes across periods.
- We develop SPImpute to fill gaps using observed history and periodic patterns. It tests repairs on temporarily hidden observations to choose weights and a repair, and adjusts reference values available on only one side for eligible long gaps. It does not query or retrain the downstream model.
- In comparisons across 19 time series, SPImpute ranks first among 21 imputation methods under Sundial and TimesFM-2.5. Compared with Linear, it reduces forecasting MSE by 10.81%, 10.11%, and 2.69% under Sundial, TimesFM-2.5, and Chronos-2, respectively. These results show that SPImpute is effective across the tested data and forecasting models.

2 Problem Definition

Let $x \in \mathbb{R}^n$ denote a regularly sampled univariate history. An observation mask $o \in \{0,1\}^n$ defines $O = \{i : o_i = 1\}$ and $M = \{i : o_i = 0\}$; only x_O is available to the imputer. A given period p specifies the temporal lag, and a frozen predictor F maps a complete history to the next H values. The task is to construct a history $\hat{x}$ of the same length satisfying

$$\hat{x}_i = x_i \quad (i \in O), \qquad \hat{y} = F(\hat{x}), \qquad O \cup M = \{1, \ldots, n\}, \quad O \cap M = \varnothing. \tag{1}$$

Repair uses the current observed values, missing positions, given period, and a fixed random seed. Observed neighbors may lie on either side of a gap, provided they precede the forecasting origin. Candidate selection has no access to future values or predictor outputs. Known historical targets and

complete future labels are retained only for evaluating reconstruction and downstream forecasting. The reconstruction objective encodes temporal relations within the history. We evaluate its forecasting effect using the same frozen predictor and future labels for every method, with predictor parameters unchanged.

3 SPIMPUTE OVERVIEW

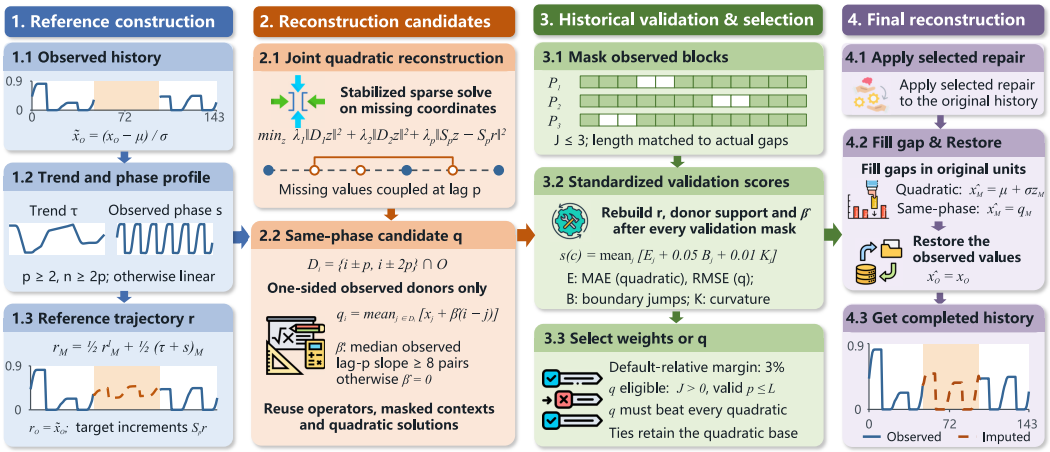


Figure 2: SPImpute overview: reference construction, reconstruction candidates, historical validation and selection, and final reconstruction. The selected quadratic solution is returned to original units, whereas the same-phase candidate q is used directly; observations remain exact. L is the target validation length. Example curves show hourly Australia Solar values divided by 10^6.

SPImpute combines quadratic reconstruction with a separate same-phase candidate (Figure 2). Masked historical validation selects the repair supplied to the frozen predictor.

3.1 REFERENCE TRAJECTORY AND CHANGES ACROSS PERIODS

The observed mean μ and standard deviation σ standardize observations as $\tilde{x}_i = (x_i - \mu)/\sigma$, with $\sigma = 1$ when nonfinite or below 10^{-8}. Linear interpolation constructs r^{l}. For $p \geq 2$ and $n \geq 2p$, a rolling trend τ captures slow changes, and a phase profile s averages observed residuals. At missing positions, r averages r^{l} and $\tau + s$; at observations, $r_O = \tilde{x}_O$. Lagged differences $S_p r$ retain changes between successive periods. Appendix B.1 gives the construction and fallbacks. Quadratic variables are standardized; same-phase formulas use original units, with $\tilde{q} = (q - \mu)/\sigma$ used for scoring. All candidates and validation tasks share the observed-input normalization.

3.2 JOINT RECONSTRUCTION UNDER OBSERVATION CONSTRAINTS

Let D_1, D_2 denote first- and second-difference operators, and define $(S_p z)_t = z_{t+p} - z_t$. For nonnegative weights $\lambda = (\lambda_1, \lambda_2, \lambda_p)$, the base reconstruction solves

$$\min_z \ \lambda_1 \|D_1 z\|_2^2 + \lambda_2 \|D_2 z\|_2^2 + \lambda_p \|S_p z - S_p r\|_2^2 \quad \text{subject to } z_O = \tilde{x}_O. \tag{2}$$

Local terms regulate adjacent changes and curvature, including gap boundaries. The periodic term follows increments in r. For $r_t = at + s_t$ with $s_{t+p} = s_t$, its target is ap, whereas periodic consistency $\|S_p z\|^2$ targets zero.

To distinguish increment fidelity from reference-value fidelity, write $\delta = z - r$, so $\delta_O = 0$. Let $d_i = (S_p^\top S_p)_{ii}$ and let E_{MM} contain periodic edges whose two endpoints are missing. The graph-energy expansion (Shuman et al., 2013) gives

$$\|S_p \delta\|^2 = \sum_{i \in M} d_i \delta_i^2 - 2 \sum_{(i,j) \in E_{MM}} \delta_i \delta_j. \tag{3}$$

Without missing–missing edges, the cost equals degree-weighted value fidelity. Otherwise, the cross-products couple corrections across gaps. In the running example, two missing peaks can move

together relative to their reference while local terms connect them to observed boundaries. The matched value control retains the same degrees, local penalties, observations, and stabilization; it removes the cross-products. This identity concerns a fixed candidate. Temporary validation masks may still cause the two objectives to select different weights. Appendix C.1 provides the derivation; forecasting effects are evaluated separately.

3.3 Correcting Same-Phase Donor Levels

For missing position i, let $\mathcal{D}_i = \{i-p, i+p, i-2p, i+2p\} \cap O$ be its available same-phase donors. The uncorrected estimate $q_i^{(0)}$ averages x_j over $\mathcal{D}_i$, using linear interpolation when no donor exists. Let $\chi_i = 1$ when this set is nonempty and lies entirely before or entirely after i; otherwise $\chi_i = 0$. We estimate a common slope from observed pairs:

$$\mathcal{P} = \{t : t, t+p \in O\}, \qquad \hat{\beta} = \begin{cases} \operatorname{median}_{t\in\mathcal{P}} \dfrac{x_{t+p} - x_t}{p}, & |\mathcal{P}| \geq 8, \\ 0, & |\mathcal{P}| < 8. \end{cases} \tag{4}$$

The candidate translates donors only at positions with $\chi_i = 1$:

$$q_i = \begin{cases} \dfrac{1}{|\mathcal{D}_i|} \displaystyle\sum_{j\in\mathcal{D}_i} \big[x_j + \hat{\beta}(i-j)\big], & i \in M,\ \chi_i = 1, \\ q_i^{(0)}, & i \in M,\ \chi_i = 0, \\ x_i, & i \in O. \end{cases} \tag{5}$$

At an enabled coordinate, $q_i - q_i^{(0)} = \hat{\beta}\,|\mathcal{D}_i|^{-1} \sum_{j\in\mathcal{D}_i}(i-j)$. For example, donors at $i-p$ and $i-2p$ receive an average correction of $3p\hat{\beta}/2$. Positions with donors on both sides retain $q^{(0)}$, even if the donor counts differ. No donors or insufficient slope pairs also retain the original estimate. Only originally observed values serve as donors; filled values are never reused. Selecting q replaces the quadratic repair on missing coordinates. The uncorrected $q^{(0)}$ is a component ablation.

Property 1 (Trend equivariance on enabled coordinates). Fix the mask, with at least eight observed lag-p pairs and $\chi_i = 1$. For $x'_t = x_t + a + bt$, the candidate satisfies $q_i(x') = q_i(x) + a + bi$. Each observed pair slope increases by b, so its median does too; every translated donor then increases by $a + bi$. The property applies only to supported one-sided coordinates and gives no guarantee for selection, the fallback procedure, or forecasting.

3.4 Weight Selection Using Masked Blocks

If the actual candidates q and $q^{(0)}$ agree elementwise, SPImpute preserves the complete selection and output of the procedure using $q^{(0)}$. This procedure uses the weights, block sampling, RMSE fitting term, eligibility, and switching rule below, substituting $q^{(0)}$ for q. Temporary validation masks therefore cannot alter a deployed result whose actual candidate is unchanged. Matched objective controls use the same rule.

Otherwise, the target validation length is the median actual missing-run length, clipped to $[2, 128]$. SPImpute masks observed blocks to form $J \leq 3$ tasks. Each task rebuilds r, $q^{(0)}$, and q after masking, recomputes χ_i from its surviving donor set, and re-estimates $\hat{\beta}$ from its surviving observed pairs. Neither the actual donor classification nor a slope estimated before masking is reused. Appendix B.2 specifies sampling, shortening, and insufficient-observation cases.

For candidate c, fitting error E_j is measured on the hidden validation positions after common standardization. Quadratic candidates use MAE and q uses RMSE. The fixed boundary term B_j averages absolute jumps between missing and observed neighbors, while K_j measures absolute second differences around the validation positions:

$$s(c) = \frac{1}{J} \sum_{j=1}^{J} \big[E_j(c) + 0.05 B_j(c) + 0.01 K_j(c)\big]. \tag{6}$$

Valid periods activate eight quadratic weight triples (Table B1). Let c_0 be the default and $c_{\min}$ the lowest-scoring quadratic candidate. The selected base switches from c_0 only when

$$s(c_0) - s(c_{\min}) > 0.03 \max\big(|s(c_{\min})|, 10^{-12}\big). \tag{7}$$

The additional q candidate is eligible when validation tasks exist and p does not exceed the target validation length. It replaces the selected base only if it scores strictly below every quadratic candidate and satisfies Eq. (7) with $c_{\min} = q$. Ties retain the base. Unavailable scores retain the default. The matched controls use these same candidates, tasks, score conventions, and switching rule. Selection uses historical reconstruction scores and does not query the predictor.

3.5 Sparse Solution and Computation Reuse

Eliminating the observed coordinates leaves

$$(Q_{MM} + \eta d I) z_M = b_M - Q_{MO} \tilde{x}_O, \tag{8}$$

where $Q = \lambda_1 D_1^\top D_1 + \lambda_2 D_2^\top D_2 + \lambda_p S_p^\top S_p$ and $b = \lambda_p S_p^\top S_p r$. The diagonal scale d and $\eta = 10^{-8}$ stabilize the system. The scale is the mean absolute diagonal of Q_{MM}, replaced by one when nonpositive or nonfinite. Observed coordinates contribute through the right-hand side and are excluded from the unknown vector. For each fixed candidate, the stabilized matrix is positive definite. Operators require linear storage; factorization cost depends on period and missingness geometry. Candidate evaluation reuses operators, masked contexts, and base solutions. The selected base is returned to the original scale, or replaced by q, and observations are restored exactly. Appendix B.3 describes numerical fallbacks; Algorithm B1 gives the procedure.

4 Experimental Setup

4.1 Data and Evaluation Protocol

The main benchmark comprises nineteen univariate series from ten sources, five block-missing rates from 10% to 50%, and three mask seeds. At each rolling origin, a history of at most 2,880 values predicts the next 720 steps. Sundial and TimesFM-2.5 share masks; Chronos-2 uses independent masks. The latter comparison therefore changes both predictor and missing positions. Imputation receives only the current observed history and the period supplied with the dataset. Appendix D specifies origins, masks, and checkpoints.

Confirmation uses 1,152 synthetic histories from 72 conditions and sixteen new seeds per condition. Histories span ten or fourteen periods, with drift strengths 0, 0.75, and 1.5; pulse phase, width, and noise vary across seeds. The additional-source study uses 225 history/seed tasks from five sources, with three origins, three task seeds, and five patterns per source. Development and confirmation inputs are separate; confirmation future intervals do not overlap those used in development. The rule and analysis were fixed before forecasting, and confirmation results are reported separately from the main benchmark.

4.2 Baselines and Attribution Controls

The main comparison includes twenty external baselines under Sundial and TimesFM-2.5. Chronos-2 compares SPImpute, Linear, BiSmooth-Kalman, SAITS, and FIM (Seifner et al., 2025). Appendix D.2 records the evaluated implementations and training budgets. FIM's complete procedure and subset without linear fallback are distinguished.

Matched controls replace the increment objective with degree-weighted value fidelity or periodic consistency while retaining the same reference, weights, tasks, one-sided correction, and unchanged-candidate rule. The $q^{(0)}$ ablation removes donor-level correction and re-evaluates its score. Base-solution comparisons isolate same-phase replacement. CPU repair costs use identical inputs, warmup, and interleaved repetitions. Related work appears in Appendix A.

4.3 Metrics and Statistical Units

The primary metric is 720-step median-forecast MSE; MAE is secondary. Main-benchmark errors aggregate windows, seeds, and missing rates within each series. Relative reductions are computed per series, averaged within sources, then across equally weighted sources. Paired uncertainty is estimated across sources for real data and across generation seeds for synthetic data. Individual bootstrap

intervals and correction families are stated with each study; windows are not independent samples. Appendix D details aggregation and statistical tests.

5 Experimental Results

We evaluate forecasting performance and imputation cost, then use matched controls to examine the reconstruction objective and phase estimate. Complete results and source-level comparisons appear in the appendices.

5.1 Overall Forecasting Performance

Figure 3 summarizes mean MSE ranks in the 21-method comparison under the two main predictors. SPImpute ranks first, with source-balanced means of 5.388 under Sundial and 6.375 under TimesFM-2.5. The latter is close to BiSmooth-Kalman's 6.378. Across the two predictors, SPImpute and BiSmooth-Kalman have mean ranks of 5.88 and 7.09, respectively. Table E1 includes every method and both MSE and MAE ranks.

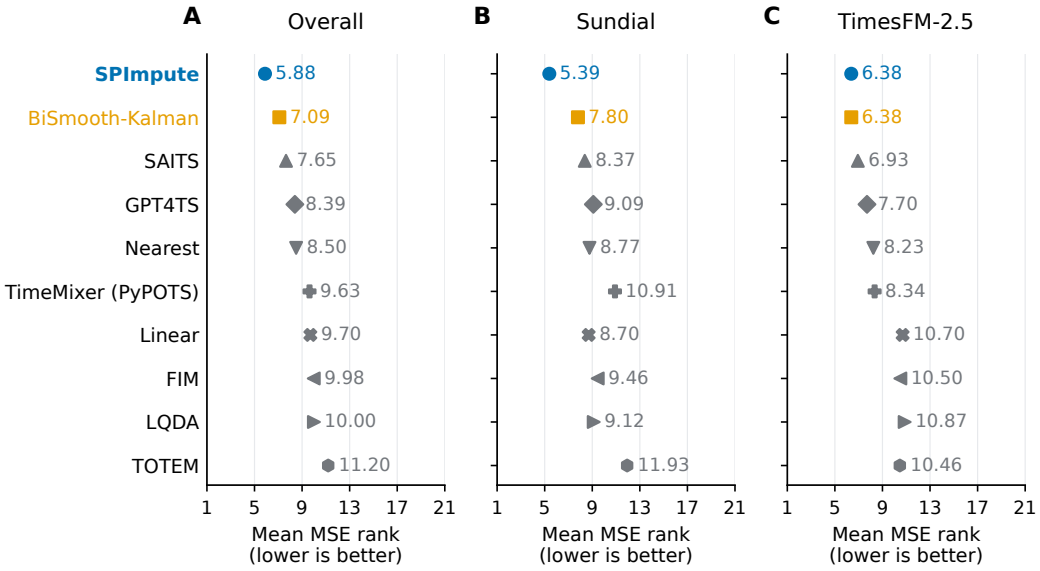


Figure 3: Mean forecasting MSE ranks in the 21-method comparison. A averages the two main predictors; B and C show each predictor. Ten representative methods are displayed, with ranks computed in the complete pool. Sources receive equal weight. Lower values are better.

Source-balanced mean relative MSE reductions against Linear are 10.81% under Sundial and 10.11% under TimesFM-2.5. The five-method Chronos-2 evaluation gives a mean rank of 2.31 and a 2.69% reduction relative to Linear. Its masks differ from the main evaluation, so the results are reported separately in Appendix E.3. Advantages vary across series and predictors; Appendix E.3 gives per-series errors and source-level intervals.

Relative to the complete FIM adapter, including fallback, MSE reductions are 13.74%, 7.82%, and 3.25% under Sundial, TimesFM-2.5, and Chronos-2. On histories where FIM needs no fallback, they are 12.82%, 10.96%, and 6.90%; each individual 95% source-bootstrap interval is above zero. This subset has a different gap distribution and is reported separately (Table E8).

5.2 Matched Objectives and Components

With the phase branch disabled, increment fidelity reduces source-mean MSE by 7.21%, 13.08%, and 9.55% relative to local smoothing. Its differences from value fidelity are much smaller (Table F1). The BM30 comparison gives every objective the same corrected phase candidate, validation tasks, eligibility, and switching rule. All 903 histories in each mask set are retained, along with the phase-only and fixed-weight controls.

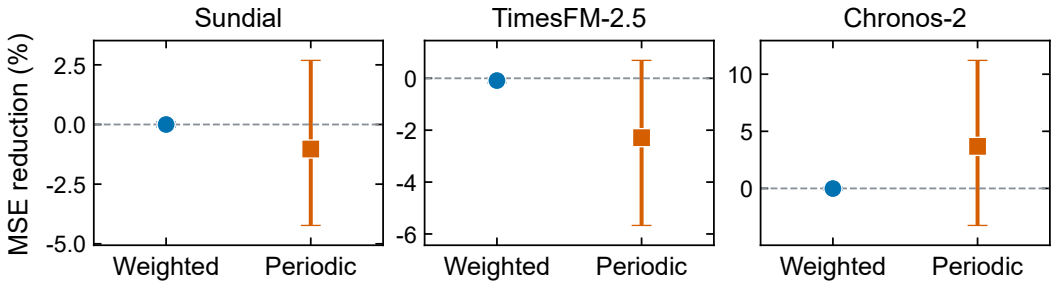


Figure 4: Matched BM30 objective comparisons. Positive reductions favor SPImpute over the indicated objective with the same corrected phase option. Bars are individual 95% bootstrap intervals over ten data sources.

Absolute source-mean differences from degree-weighted value guidance remain below 0.1% under each predictor (Figure 4). Periodic consistency is also competitive. The matched table and expanded rankings are in Appendix F; the leading ranks in Figure 3 concern its stated external comparison pool. The main BM30 mask set selects the phase branch in 141 of 903 histories. We therefore examine increment coupling on repeated short gaps, where the phase branch is unavailable, and phase correction on eligible long gaps.

5.3 Repeated Short Gaps Under Drift

We compare recurring pulses at three drift strengths under one, three, or six aligned gaps and a staggered control. The experiment uses 64 generation seeds excluded from development. All compared objectives retain their historical weight selection. Gaps are shorter than the period, making the phase branch ineligible and isolating quadratic reconstruction.

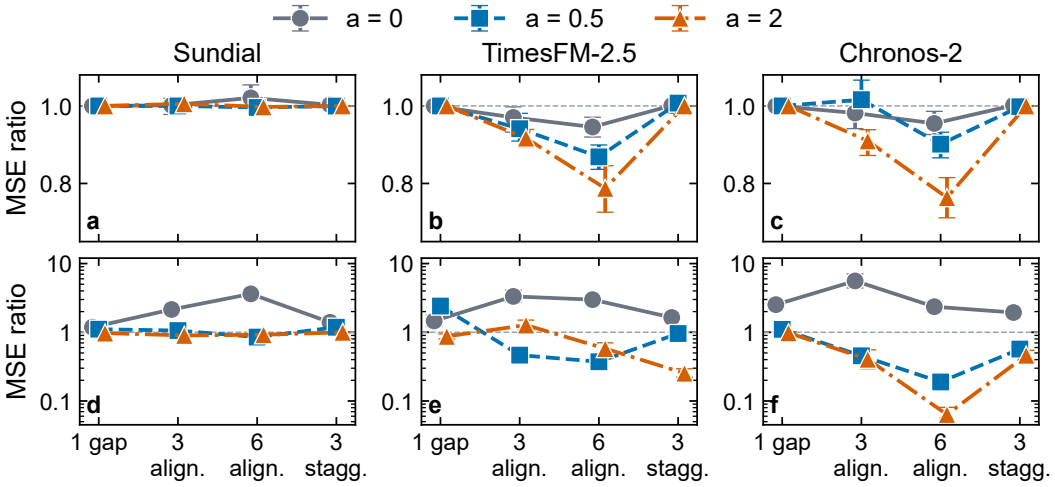


Figure 5: Independent pulse experiment with short gaps. Ratios below one favor SPImpute over degree-weighted value fidelity (top) or periodic consistency (bottom, log scale). Error bars are individual 95% paired-bootstrap intervals over 64 generation seeds; the metric is 720-step median-forecast MSE.

For six aligned gaps, TimesFM-2.5 and Chronos-2 reduce MSE by 13.13% and 9.86% relative to weighted-value fidelity at drift $a = 0.5$, and by 21.27% and 23.65% at $a = 2$. These four comparisons retain $p < 0.001$ after Holm correction over 210 primary tests. Sundial's mean differences are smaller. Figure 5 includes isolated and staggered gaps, where the coupling benefit diminishes. The real-history geometry comparison in Table F3 further shows that periodic connections alone do not determine a forecasting gain.

5.4 One-Sided Phase Correction

The separate confirmation crosses three periods, two waveforms, three drift strengths, two gap lengths, and near-origin or interior placement. Sixteen new seeds per condition give 1,152 histories; context length, pulse shape, and noise also vary as specified in Appendix F. The complete methods with corrected and uncorrected phase estimates use the same information and base candidates.

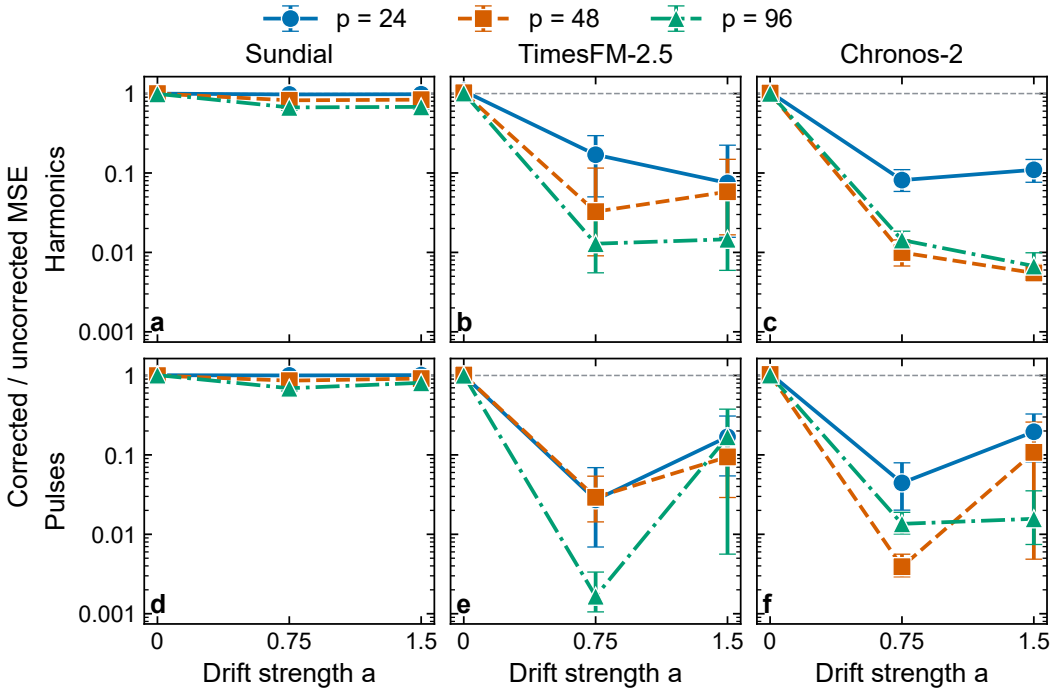


Figure 6: Independent confirmation of one-sided phase correction for near-origin gaps. Each point averages the MSE ratios of two gap lengths at the indicated period and drift. Error bars resample the 16 paired generation seeds jointly across lengths and give individual 95% intervals. Ratios below one favor correction. All 576 separately tested interior-gap inputs and forecasts are unchanged.

Across the 24 drifting near-origin settings, condition-mean MSE decreases under both TimesFM-2.5 and Chronos-2. Their mean relative reductions are 92.9% and 94.9%, respectively; 14 and 19 comparisons remain significant after Holm correction over 216 tests. Sundial's corresponding mean reduction is 14.6%. The stationary and interior conditions are included in the complete results. Interior gaps have two-sided phase observations and retain the original procedure exactly. In stationary near-origin settings, the phase branch retains substantial benefits over quadratic reconstruction alone.

An additional 225 history/seed tasks from five real sources use future intervals disjoint from development. Source-mean reductions relative to Linear are 16.61%, 19.86%, and 12.93% under the three predictors. Figure 7 reports source-level differences. This five-source evaluation is descriptive; Appendix F provides paired intervals and statistical tests.

5.5 Imputation Time

Figure 8 compares all ten methods displayed in Figure 3 on the same 57 histories, with 171 timed calls per method. Linear, Nearest, LQDA, SPImpute, and BiSmooth-Kalman run on the two-thread CPU; FIM, SAITS, GPT4TS, TimeMixer, and TOTEM run on the RTX 3060. SPImpute takes 78.66 ms at the median, compared with 12.17 ms for LQDA and 208.11 ms for BiSmooth-Kalman. GPU medians range from 75.96 ms for TOTEM to 512.30 ms for SAITS. These costs include per-history training where required. Appendix G.2 specifies the two measurement batches and weight-loading costs.

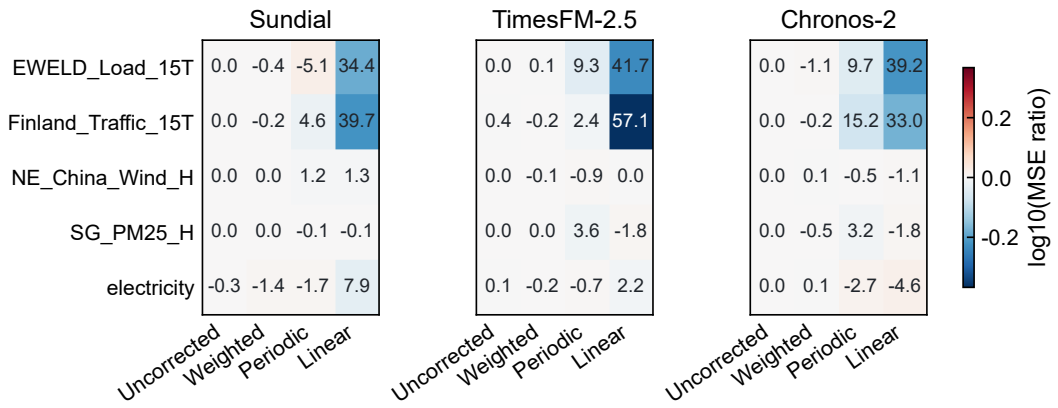


Figure 7: Confirmation effects by source and predictor. Cell annotations are MSE reductions (%), positive for SPImpute; color uses $\log_{10}$ of the final/comparator MSE ratio. Matched weighted and periodic objectives use the same corrected candidate and equality rule.

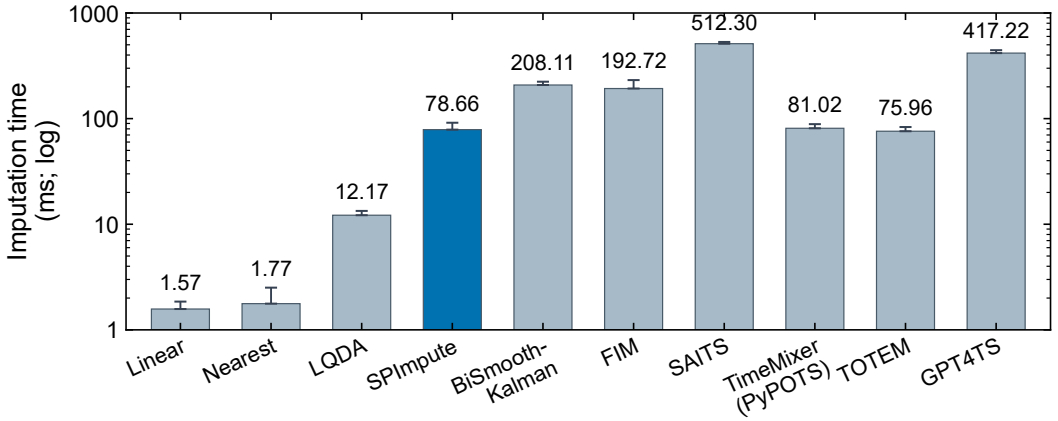


Figure 8: Imputation time for the ten methods displayed in Figure 3. Bars and labels show medians; upper whiskers reach P95 over 171 calls per method on 57 matched histories. The shared axis is logarithmic. Devices and training costs are specified in the text.

6 Discussion

Abrupt or nonlinear changes can make the common lagged slope inaccurate, and asymmetric donor times can bias two-sided averages. Trend equivariance applies to enabled candidate coordinates; it gives no guarantee for selection or forecasting. The experiments cover regularly sampled univariate histories with supplied periods, primarily under controlled artificial gaps. The additional real inputs retain pre-existing missing values, but do not isolate a natural missingness mechanism. Unknown or changing periods, multivariate dependence, and irregular sampling require separate evaluation. The confirmation studies characterize the tested conditions; their prevalence in deployment remains unmeasured.

7 Conclusion

SPImpute repairs incomplete histories for frozen forecasters using constrained reference-increment reconstruction and a same-phase estimate with one-sided drift correction. The correction decomposition identifies when missing intervals interact. The phase rule preserves the original procedure whenever correction leaves its actual candidate unchanged. Experiments establish coupling gains for repeated pulse gaps and phase-correction benefits in new drifting conditions near the forecast origin. SPImpute remains competitive in the complete real-data comparisons, with a median imputation time of 78.66 ms on the measured CPU inputs. Candidate selection uses historical observations and requires no forecasting-model queries.

Reproducibility Statement

https://anonymous.4open.science/r/SPImpute/
Data and evaluation settings are in Appendix D, implementation details and candidate weights in Appendix B, analyses and proofs in Appendix C, and runtime protocols and training budgets in Appendix G.

AI use statement

AI tools assisted with translation and language polishing.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A Benchmark For General Time Series Forecasting Model Evaluation. *arxiv preprint arxiv:2410.10393*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From Univariate to Universal Forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Christoph Bergmeir, Rob J. Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, Apr 2018. ISSN 0167-9473. doi: 10.1016/j.csda.2017.11.003. URL http://dx.doi.org/10.1016/j.csda.2017.11.003.

George E. P. Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, fifth edition, 2015. ISBN 9781118675021.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Mar 2004. ISBN 9780511804441. doi: 10.1017/cbo9780511804441. URL http://dx.doi.org/10.1017/CBO9780511804441.

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.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 21–27 Jul 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, June 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.119619. URL http://dx.doi.org/10.1016/j.eswa.2023.119619.

Wenjie Du, Yiyuan Yang, Linglong Qian, Jun Wang, and Qingsong Wen. PyPOTS: A Python Toolkit for Machine Learning on Partially-Observed Time Series. *arXiv preprint arXiv:2305.18811v2*, 2025. URL https://arxiv.org/abs/2305.18811v2.

Paul H. C. Eilers. A Perfect Smoother. *Analytical Chemistry*, 75(14):3631–3636, May 2003. ISSN 1520-6882. doi: 10.1021/ac034173t. URL http://dx.doi.org/10.1021/ac034173t.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash Time Series Forecasting Archive. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*,

volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/eddea82ad2755b24c4e168c5fc2ebd40-Paper-round2.pdf.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, fourth edition, 2013. doi: 10.56021/9781421407944.

Google Research. TimesFM-2.5-200M-PyTorch model card, 2025. URL https://huggingface.co/google/timesfm-2.5-200m-pytorch. Accessed: 2026-09-19.

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002. ISBN 0-89871-521-0.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, Mar 1960. ISSN 0021-9223. doi: 10.1115/1.3662552. URL http://dx.doi.org/10.1115/1.3662552.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liu25be.html.

Steffen Moritz and Thomas Bartz-Beielstein. imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, 9(1):207–218, 2017. doi: 10.32614/RJ-2017-009.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, July 2003. ISSN 1557-7368. doi: 10.1145/882262.882269. URL http://dx.doi.org/10.1145/882262.882269.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. MIT Press, Cambridge, MA, 2006. ISBN 026218253X. doi: 10.7551/mitpress/3206.001.0001.

Yousef Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, second edition, Jan 2003. ISBN 9780898718003. doi: 10.1137/1.9780898718003. URL http://dx.doi.org/10.1137/1.9780898718003.

Patrick Seifner, Kostadin Cvejoski, Antonia Körner, and Ramsés J. Sánchez. Zero-shot Imputation with Foundation Inference Models for Dynamical Systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NPSZ7V1CCY.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, May 2014. ISBN 9781107298019. doi: 10.1017/cbo9781107298019. URL http://dx.doi.org/10.1017/CBO9781107298019.

D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, May 2013. ISSN 1053-5888. doi: 10.1109/msp.2012.2235192. URL http://dx.doi.org/10.1109/MSP.2012.2235192.

Sabera J Talukder, Yisong Yue, and Georgia Gkioxari. TOTEM: TOkenized Time Series EMbeddings for General Time Series Analysis. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QlTLkH6xRC.

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, Feb 2014. ISSN 0090-5364. doi: 10.1214/13-aos1189. URL http://dx.doi.org/10.1214/13-AOS1189.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 3839–3848. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d54e99a6c03704e95e6965532dec148b-Paper.pdf.

Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29, pp. 847–855. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/85422afb467e9456013a2a51d4dff702-Paper.pdf.

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 *Forty-third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.02278.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One Fits All: Power General Time Series Analysis by Pretrained LM. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *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/file/86c17de05579cde52025f9984e6e2ebb-Paper-Conference.pdf.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919, 2003. URL https://mlg.eng.cam.ac.uk/pub/pdf/ZhuGhaLaf03a.pdf.

APPENDICES

Appendix A reviews related work. Appendix B gives implementation details and the complete algorithm, and Appendix C establishes the stated properties. Appendix D records data, baselines, and statistical protocols. Complete forecasting and attribution results appear in Appendices E and F. Appendix G covers computation and training budgets.

A RELATED WORK

Seasonal decomposition, state-space models, and Gaussian processes describe trend, recurrence, dynamic states, and covariance relationships (Cleveland et al., 1990; Kalman, 1960; Box et al., 2015; Rasmussen & Williams, 2006). Univariate imputation combines these models with interpolation (Moritz & Bartz-Beielstein, 2017). Squared-difference smoothing penalizes local variation quadratically (Eilers, 2003); trend filtering uses absolute differences for local adaptation (Tibshirani, 2014). Temporal regularization connects incomplete series across time (Yu et al., 2016). SPImpute fixes observations and combines local quadratic terms with reference increments. Its matched degree-weighted control isolates interactions between missing coordinates.

SAITS, HELIX, GPT4TS, and TOTEM learn temporal representations for imputation (Du et al., 2023; Zhang et al., 2026; Zhou et al., 2023; Talukder et al., 2024). We also evaluate PyPOTS TimeMixer (Du et al., 2025), with per-history initialization and training budgets in Appendix D.2. FIM integrates estimated temporal derivatives over missing intervals using weights pretrained on synthetic dynamical systems (Seifner et al., 2025). We separately evaluate its complete adapter and inputs without fallback. SPImpute uses current observations to construct references and select candidates, requiring no pretrained imputation weights or predictor queries.

Pretrained forecasters transfer temporal representations across tasks (Das et al., 2024). We freeze Sundial, TimesFM-2.5, and Chronos-2 (Liu et al., 2025; Google Research, 2025; Ansari et al., 2025); GIFT-Eval and Monash provide cross-domain evaluation resources (Aksu et al., 2024; Godahewa et al., 2021). Paired histories and future labels measure imputation's effect on each predictor. Lipschitz analysis can characterize neural-network sensitivity (Virmaux & Scaman, 2018); we do not estimate these constants.

B Implementation Details

B.1 Reference construction and normalization

Figure 2 distinguishes the reference for quadratic reconstruction from the same-phase candidate. Quadratic reconstruction uses $\tilde{x}_i = (x_i - \mu)/\sigma$, where the mean and population standard deviation are computed from observed values. A nonfinite standard deviation or one below 10^{-8} is replaced by one. Linear interpolation extends the nearest finite observations at the ends and uses the observed median when needed. An empty input is returned unchanged; an entirely missing input receives the finite linear reference.

For $p \geq 2$ and $n \geq 2p$, a centered rolling mean of the linear reference estimates τ, using a window of length p and at least $\max(2, \lfloor p/4 \rfloor)$ values. The phase profile averages $\tilde{x}_i - \tau_i$ only at observed positions, interpolates unobserved phases, and is centered. Thus $r_O = \tilde{x}_O$ and $r_M = 0.5 r^1_M + 0.5(\tau + s)_M$. Invalid periods, fewer than p observations, or an unavailable profile use the linear reference. Prescribed differences and fixed observations also underlie guided reconstruction (Pérez et al., 2003).

The quantities $q^{(0)}$, q, and $\hat{\beta}$ in Eqs. (4)–(5) use original data units. Donors are observed values at offsets $-p, +p, -2p, +2p$ from the target. Correction applies only when all available donors lie on the same side. Two-sided or absent donor sets retain $q^{(0)}$; with fewer than eight observed lag-p pairs, the slope is zero. When the actual q and $q^{(0)}$ arrays are elementwise equal, the method uses the complete procedure with $q^{(0)}$. Otherwise, each validation task hides its labels before rebuilding donor support, references, and the slope. Its same-phase output is converted to $\tilde{q} = (q - \mu)/\sigma$ for comparison with standardized quadratic repairs. Generated values are never reused as donors.

B.2 Candidate set and selection

Table B1: Fixed weight triples; an input activates five nonperiodic or eight periodic candidates.

ID	λ_1	λ_2	λ_p	Activation	Role
C1	0	0.5	0	Both	Local smoothing
C2	0.01	1	0	Both	Local variation
C3	0.05	1	0	Both	Local variation
C4	0.10	0.5	0	Both	Local variation
C5	0.02	1	0	Nonperiodic	Default
C6	0.01	1	0.25	Periodic	Reference guidance
C7	0.02	1	0.75	Periodic	Default
C8	0.05	1	0.75	Periodic	Reference guidance
C9	0.02	0.5	1.5	Periodic	Reference guidance

The target validation length is the rounded median length of the actual missing runs, clipped to $[2, 128]$ and bounded by $n-2$. A task-local seed samples up to three fully observed blocks, preferring interior starts. Sampling may repeat a start when fewer than three are available. If necessary, the length is shortened to two, followed by a fallback set of observed interior coordinates. With fewer than eight observations, the method uses the default without validation. The interpretation of time-series validation depends on serial dependence and the task being assessed (Bergmeir et al., 2018).

Each quadratic candidate uses MAE at the hidden positions. The same-phase candidate uses RMSE. Both add 0.05 times the missing-to-observed boundary jump and 0.01 times local absolute second differences; scores are averaged over tasks. A valid period and p no larger than the target validation length make the same-phase candidate eligible. Its score must beat every quadratic score and pass Eq. (7) against the default. The retained quadratic weights follow the same default-relative threshold. Ties and unavailable scores preserve the corresponding base/default decision. The weighted-value and periodic controls use the same one-sided correction and actual-candidate equality rule.

B.3 Algorithm

Algorithm B1 specifies the selection and reconstruction steps.

Algorithm B1 SPImpute with one-sided reference correction

Input: Incomplete history x, observation mask o, period p, seed s
Output: Repaired history $\hat{x}$ in original units, with $\hat{x}_O = x_O$

1: $n \leftarrow |x|$; $O \leftarrow \{i : o_i = 1\}$; $M \leftarrow \{i : o_i = 0\}$
2: **if** $n = 0$ or $O = \varnothing$ or $M = \varnothing$ or not VALIDPERIOD(p, n) **then**
3: **return** ORIGINAL(x, o, p, s)
4: $(\mu, \sigma) \leftarrow$ OBSERVEDSTATS(x_O)
5: $\tilde{x}_O \leftarrow (x_O - \mu)/\sigma$; keep $\tilde{x}_M$ missing
6: $q^{(0)} \leftarrow \mu + \sigma$ PHASEMEAN$(\tilde{x}, O, p)$; $q^{(0)}_O \leftarrow x_O$
7: $q \leftarrow$ ONESIDED$(x, O, p, q^{(0)})$
8: **if** $q = q^{(0)}$ elementwise **then**
9: **return** ORIGINAL(x, o, p, s)
10: $(\Lambda, \lambda_0) \leftarrow$ active weight triples and default in Table B1
11: $L \leftarrow$ TARGETLENGTH(o); $\mathcal{V} \leftarrow$ SAMPLETASKS(O, L, s)
12: $S_\lambda \leftarrow [\,]$ for each $\lambda \in \Lambda$; $S_q \leftarrow [\,]$
13: **for** each validation index set $V_j \in \mathcal{V}$ **do**
14: $(O_j, \tilde{x}^{(j)}) \leftarrow$ HIDE$(O, \tilde{x}, V_j)$
15: $r_j \leftarrow$ REFERENCE$(\tilde{x}^{(j)}, O_j, p)$; $q^{(0)}_j \leftarrow$ PHASEMEAN$(\tilde{x}^{(j)}, O_j, p)$
16: $q_j \leftarrow$ ONESIDED$(\tilde{x}^{(j)}, O_j, p, q^{(0)}_j)$
17: **for** each $\lambda \in \Lambda$ **do**
18: $z_{j,\lambda} \leftarrow$ SOLVEFIXED$(\tilde{x}^{(j)}, O_j, p, r_j, \lambda, \eta)$
19: append $\mathcal{L}_{\mathrm{MAE}}(z_{j,\lambda}; j)$ to S_λ
20: append $\mathcal{L}_{\mathrm{RMSE}}(q_j; j)$ to S_q
21: $s_\lambda \leftarrow \mathrm{Avg}(S_\lambda)$ for $\lambda \in \Lambda$; $s_q \leftarrow \mathrm{Avg}(S_q)$
22: $\lambda^* \leftarrow \lambda_0$; $\lambda_{\min} \leftarrow \arg\min_{\lambda \in \Lambda} s_\lambda$
23: **if** $\mathrm{Improve}(s_{\lambda_0}, s_{\lambda_{\min}})$ **then**
24: $\lambda^* \leftarrow \lambda_{\min}$
25: **if** $\mathcal{V} \neq \varnothing$ and $p \leq L$ and $s_q < \min_{\lambda \in \Lambda} s_\lambda$ and $\mathrm{Improve}(s_{\lambda_0}, s_q)$ **then**
26: $\hat{x} \leftarrow q$
27: **else**
28: $r \leftarrow$ REFERENCE$(\tilde{x}, O, p)$; $z \leftarrow$ SOLVEFIXED$(\tilde{x}, O, p, r, \lambda^*, \eta)$
29: $\hat{x} \leftarrow \mu + \sigma z$
30: $\hat{x}_O \leftarrow x_O$
31: Replace nonfinite entries of $\hat{x}$ with the corresponding linear-reference values
32: **return** $\hat{x}$

SPImpute first handles degenerate inputs and standardizes the observed history. It then constructs the uncorrected and one-sided phase estimates. If their actual outputs are identical, the algorithm returns ORIGINAL, which uses $q^{(0)}$ as its phase candidate for both repair and validation, so masking cannot introduce a new correction (lines 6–9). The reference and phase constructions follow Appendix B.1.

Otherwise, observed blocks are masked to select candidate weights. Each task rebuilds references, donor support, and drift from surviving observations while retaining the original normalization (lines 13–20). The loss $\mathcal{L}_m$ combines MAE or RMSE with the boundary and curvature terms in Eq. (6); Avg averages task scores and returns $+\infty$ for an empty list or a nonfinite mean.

Selection starts from the default quadratic weights. $\mathrm{Improve}(a, b)$ requires finite b and the 3% margin in Eq. (7). The phase candidate replaces the base only when validation exists, $p \leq L$, its score beats every quadratic candidate, and it passes the default-relative test (lines 22–26). Here L is the target validation length, even when sampling uses shorter blocks.

SOLVEFIXED solves the stabilized system in float64 with $\eta = 10^{-8}$. A sparse-direct exception invokes conjugate gradients; a nonzero CG status or nonfinite solution entries trigger reference fallback. Without SciPy, dense solution is limited to $n \leq 2,048$, with least squares after `LinAlgError`. The chosen repair is returned to original units, observations are restored exactly, and remaining nonfinite values use the linear reference (lines 26–32).

C Analysis

C.1 Quadratic solution

Expanding the objective gives $z^\top Q z - 2b^\top z$ up to a constant. Eliminating fixed observed coordinates yields

$$A z_M = b_M - Q_{MO}\tilde{x}_O, \qquad A = Q_{MM} + \eta d I. \tag{C.1}$$

The matrices $D_1^\top D_1$, $D_2^\top D_2$, and $S_p^\top S_p$ are positive semidefinite. Their nonnegative weighted sum and its principal submatrix are also positive semidefinite. Since $\eta d > 0$, A is positive definite and the fixed base candidate has a unique solution (Boyd & Vandenberghe, 2004, Sections 4.4 and 10.1). The observation equalities hold both in standardized and original units.

Sensitivity to the reference. With observations, weights, period, and stabilization fixed, replacing r by $r + e$ changes only the right-hand side:

$$A(z'_M - z_M) = \lambda_p (S_p^\top S_p e)_M. \tag{C.2}$$

This follows by subtracting the two systems (Higham, 2002, Chapter 7); $S_p e = 0$ therefore leaves the base solution unchanged. Candidate reselection and same-phase replacement lie outside this identity.

Missing-coordinate interactions. For $\delta = z - r$, the observation constraint gives $\delta_O = 0$. Writing each row of S_p as an edge in E_p,

$$\|S_p \delta\|^2 = \sum_{(i,j)\in E_p} (\delta_j - \delta_i)^2. \tag{C.3}$$

Expanding the squares gives Eq. (3): squared terms accumulate according to vertex degree, and every cross-term with an observed endpoint vanishes. This is the graph Dirichlet-energy identity (Shuman et al., 2013); fixing observed vertices also appears in harmonic reconstruction (Zhu et al., 2003). If no edge joins two missing coordinates, the increment penalty equals degree-weighted value fidelity, including its diagonal and stabilization scale. Equal fixed weights therefore give equal base solutions. Different validation masks can still produce different selected weights.

C.2 One-sided correction

Let the mask be fixed, let at least eight lag-p pairs be observed, and consider a missing coordinate i whose nonempty donor set lies entirely on one side. For any affine addition $x'_t = x_t + a + bt$, each observed pair slope increases by b, so $\hat{\beta}' = \hat{\beta} + b$. Every corrected donor obeys

$$x'_j + \hat{\beta}'(i - j) = x_j + \hat{\beta}(i - j) + a + bi. \tag{C.4}$$

Averaging over the unchanged donor set gives $q_i(x') = q_i(x) + a + bi$. This identity holds in original units at supported one-sided coordinates, including when the slope before transformation is zero. It does not cover two-sided or unsupported coordinates, candidate selection, or forecasts. Standardization for scoring leaves these conditions unchanged.

When the actual arrays q and $q^{(0)}$ agree elementwise, the algorithm uses the complete procedure with $q^{(0)}$. Temporary validation masks therefore cannot change its output for that input. For a single internal gap of length at most $2p$, with observed context of at least $2p$ on both sides, every missing coordinate has donors on both sides. Both the candidate and the complete procedure are then unchanged.

C.3 Error bounds

For a complete standardized history z^* with the same observations, define $\rho = A z^*_M - (b_M - Q_{MO}\tilde{x}_O)$. Then $z_M - z^*_M = -A^{-1}\rho$, and the standard spectral relation gives

$$\frac{\|\rho\|_2}{\lambda_{\max}(A)} \le \|z_M - z^*_M\|_2 \le \frac{\|\rho\|_2}{\lambda_{\min}(A)}. \tag{C.5}$$

The bound applies to a fixed linear system (Golub & Van Loan, 2013, Chapter 2). Historical validation tasks are dependent and do not establish the uniform approximation to forecasting risk required by generalization arguments (Shalev-Shwartz & Ben-David, 2014, Chapter 4). We therefore evaluate forecasting effects directly.

D Data and Evaluation Protocols

D.1 Datasets and rolling forecast origins

The nineteen main series are complete before artificial masking. Periods come from fixed metadata. Full-series masks place blocks of length 50 until the target missing count is reached, shortening the final block if necessary. Each rolling origin truncates the data and mask before repair: at most 2,880 preceding values enter the imputer, and the following 720 values are forecasting labels. A window's realized missing fraction can differ from its nominal rate.

The missing rates are 10%, 20%, 30%, 40%, and 50%. Sundial and TimesFM-2.5 share mask seeds 42–44; Chronos-2 uses 1042–1044 at the same origins. Each mask set contains 4,515 histories. Comparisons within a predictor use identical inputs and future labels. We record each history's imputation seed separately because it can differ from the mask seed.

Table D1: Main univariate targets, source groups, and fixed periods.

Target	Freq.	Length	Period	Source group
Australia_Solar_H	H	35,064	24	Australia Solar
Coastal_T_S_15T	15T	23,134	96	Coastal T S
Coastal_T_S_20T	20T	8,198	72	Coastal T S
Coastal_T_S_5T	5T	105,408	288	Coastal T S
Coastal_T_S_H	H	8,784	24	Coastal T S
ETTh1	H	17,420	24	ETT
ETTh2	H	17,420	24	ETT
ETTm1	15T	69,680	96	ETT
ETTm2	15T	69,680	96	ETT
OpenElectricity_NEM_5T	5T	43,488	288	OpenElectricity NEM
Water_Quality_Darwin_15T	15T	18,985	96	Water Quality Darwin
azure2019_D_5T	5T	8,639	288	azure2019
azure2019_I_5T	5T	8,634	288	azure2019
current_velocity_15T	15T	17,953	96	current velocity
current_velocity_20T	20T	16,759	72	current velocity
current_velocity_5T	5T	26,486	288	current velocity
epf_electricity_price_H	H	52,416	24	epf electricity price
traffic	H	17,544	24	traffic
weather	10T	52,696	144	weather

D.2 Baseline implementations and predictors

Each trainable neural adapter is initialized separately for each history and normalizes its observed inputs. Training and imputation use length-96 windows with stride 48 and an additional final window; overlapping outputs are averaged. Batch size is at most 32. The main adapters have no independent validation split, and future labels are excluded from fitting and stopping. GPT4TS loads GPT-2, while TOTEM uses random initialization without an external time-series checkpoint. TimeMixer uses the PyPOTS 1.5 interface. BiSmooth-Kalman is a local bidirectional smoother, and LQDA is local quantile-distribution alignment.

Table D2: Neural-adapter training limits and architectures.

Method	Epoch limit	Initialization	Settings
SAITS	10; patience 3	Random	2 layers; width 64; FFN 128; 4 heads
HELIX	3	Random	1 layer; width 64; 4 heads; ORT/MIT weights 1
TimeMixer (PyPOTS)	3	Random	1 layer; width 32; FFN 64; average window 3; downsample factor 2
GPT4TS	3	GPT-2	1 layer; patch 16, stride 8; frozen GPT MLP; FFN 64
TOTEM	3	Random	Width 32; 1 residual layer; embedding 16; codebook 32; compression 4

Seasonal recombines the trend, seasonal, and residual components of a history filled by linear interpolation, reproducing Linear up to numerical precision. The primary external pools retain Linear and omit Seasonal, giving 21 methods under each main predictor and five under Chronos-2. Adding four base objectives creates separate pools of 25 and nine. The SAITS budget study in Appendix G.3 uses independent historical validation.

Table D3: Frozen checkpoints and forecasting outputs (Liu et al., 2025; Google Research, 2025; Ansari et al., 2025).

Predictor	Checkpoint	Output settings
Sundial	thuml/sundial-base-128m	100 samples; predictive median
TimesFM-2.5	google/timesfm-2.5-200m-pytorch	Median forecast
Chronos-2	amazon/chronos-2	Univariate median forecast

Forecasting uses a batch size of eight. Repaired inputs are converted to float32, and all 720 predictions are saved. Random states and input ordering are paired across methods. FIM (Seifner et al., 2025) uses frozen temporal-base and pointwise-base weights, float32, and one history per batch. Its five-window adapter spans 256 positions and supports gaps of length 12–127 with observed boundaries on both sides. Other lengths or absent bilateral boundaries use a recorded linear fallback. Runtime failures are reported without replacement. Results for the complete adapter and the subset without fallback are reported separately.

D.3 Confirmation datasets

Method development and confirmation use separate inputs. The synthetic study fixes 72 conditions: periods 24, 48, and 96; harmonics and pulses; drift $a \in \{0, 0.75, 1.5\}$; gap lengths $1.25p$ and $2p$; and late or internal placement. Seeds 9200–9215 give sixteen paired realizations per condition. Even seeds use ten periods of history and odd seeds fourteen. Harmonics use $\sin(2\pi t/p + \phi) + 0.35\sin(4\pi t/p + 2\phi + 0.4)$, with ϕ uniform on $[-0.15\pi, 0.15\pi]$. Pulses use $\exp(-((t/p \bmod 1 - c)/w)^2)$, with $c \in [0.25, 0.55]$ and $w \in [0.08, 0.16]$ sampled from the fixed seed. Both add $at/(4p)$ and Gaussian noise with seed-specific standard deviation in $[0.02, 0.05]$. Late gaps end four observations before the origin; internal gaps end $4p + 4$ positions before it. Imputation seeds add 60,000 to generation seeds.

Real-data confirmation uses 225 new masking tasks from the five additional sources in Table D4. Complete 720-step futures must be disjoint from development futures, and preceding histories must contain at least 80% observed values. We greedily construct a set of nonoverlapping futures and select its first, middle, and last origin. Three task seeds (9300–9302) and five patterns give 45 tasks per source. The patterns are BM30, aligned and staggered short gaps, and late and internal long gaps. Long-gap positions are fixed across task seeds. These seeds vary candidate selection randomness, so the resulting tasks are not independent histories. Existing missing values remain in the history; the SG_PM25_H inputs contain up to 53 such values before injection. The imputation seed adds 60,000 to the task seed. These sources were available during development; confirmation uses new inputs and disjoint future intervals.

Table D4: Additional-source confirmation inputs; origins are zero-based indices.

Source	Frequency	Period	Forecast origins
EWELD_Load_15T	15T	96	3600, 10080, 18552
Finland_Traffic_15T	15T	96	3600, 17280, 33048
NE_China_Wind_H	H	24	3600, 4320, 6902
SG_PM25_H	H	24	4122, 30462, 36609
electricity	H	24	3600, 12960, 24312

MetroPT-3_5T had no eligible future disjoint from the development intervals and was excluded before confirmation forecasting. No source, condition, or seed was removed according to forecasting performance.

D.4 Metrics, uncertainty, and multiple comparisons

The primary metric is the MSE of the 720-step predictive median; its MAE is secondary. Main errors pool evaluation points within conditions, then average seeds and missing rates within each series. For methods a, b, series effects are $100(E_b - E_a)/E_b$. We average related series within sources and weight the ten sources equally. Ranks use seed-averaged errors, midranks for ties, and equal weights across rates and sources. Ranks from differently sized pools are reported separately.

Real-data uncertainty resamples sources 10,000 times. Exact two-sided sign-flip tests enumerate 2^{10} assignments for the main/BM30 data and 2^5 for confirmation, assuming symmetric source effects under the null. With five sources the minimum attainable two-sided value is 0.0625. Confirmation effects first average mask conditions within each source. Synthetic effects compare paired mean errors, with 10,000 seed-bootstrap resamples and two-sided paired Wilcoxon tests. Mean-effect intervals and Wilcoxon tests concern different statistics. All intervals are individual 95% intervals, without simultaneous-coverage guarantees.

Table D5: Correction families for the reported primary MSE comparisons.

Study	Paired unit	Holm family
Base objectives	10 sources	9 comparisons
BM30 matched controls	10 sources	15 primary; 12 secondary
Controlled real masks	10 sources	30 primary
Independent short-gap pulses	64 generation seeds	210 primary
Synthetic confirmation	16 generation seeds	216 vs uncorrected; 216 vs base
Real confirmation	5 sources	12 comparisons

The pulse family comprises fourteen conditions, five comparators, and three predictors. In BM30, the primary comparators are value+q, weighted value+q, periodic+q, q alone, and Linear; base increment, local+q, fixed increment, and the uncorrected procedure form the secondary family. FIM and representative-baseline tables report descriptive effects and source-bootstrap intervals. Supplementary 96-step and saved point-output results remain separate from the primary metric. Source seeds are 260911 for original objective/representative summaries, 260912 for FIM, and 260921 for the new BM30 and confirmation summaries.

E Detailed Forecasting Results

E.1 Ranks across external baselines

Tables E1 and E2 report every method in the corresponding comparison pools. Ranks are computed within each series, missing rate, and predictor before aggregation by source. The supplementary Chronos-2 comparison uses independent masks and a smaller pool, so its raw ranks are reported separately.

Table E1: Complete external-method ranks, weighting ten sources equally. Shaded bold and underlined entries mark the two lowest distinct unrounded values per column.

	Sundial		TimesFM-2.5	
Method	MSE	MAE	MSE	MAE
SPImpute	**5.388**	**5.392**	**6.375**	**5.595**
BiSmooth-Kalman	<u>7.797</u>	8.892	<u>6.378</u>	<u>6.987</u>
SAITS	8.370	<u>7.918</u>	6.932	7.030
Linear	8.697	8.458	10.697	9.552
Nearest	8.767	8.212	8.232	7.458
GPT4TS	9.090	9.802	7.700	7.877
LQDA	9.123	8.633	10.868	9.923
Backward	9.200	8.737	10.748	10.183
FIM	9.457	9.145	10.503	9.763
STL-Kalman	9.740	8.875	11.630	11.163
Forward	9.923	9.287	9.417	8.793
Mean	10.893	12.005	9.515	11.090
TimeMixer (PyPOTS)	10.910	11.738	8.342	9.883

Table E1 (continued)

Method	Sundial MSE	Sundial MAE	TimesFM-2.5 MSE	TimesFM-2.5 MAE
Kalman-ARIMA	11.208	12.123	9.712	10.600
HELIX	11.455	11.525	9.907	9.918
TOTEM	11.933	13.280	10.460	12.290
Polynomial	13.387	13.572	14.498	14.522
Kalman-Struct	13.868	13.192	15.738	15.373
GP-RBF	14.785	13.870	16.382	16.268
Spline	16.997	16.627	17.385	17.278
Zero	20.012	19.718	19.582	19.452

Table E2: Complete five-method Chronos-2 ranks under independent masks. Shaded bold and underlined entries mark the two lowest distinct unrounded values per column.

Method	MSE	MAE
SPImpute	**2.313**	**2.117**
BiSmooth-Kalman	3.063	3.328
Linear	3.087	2.778
FIM	3.105	3.085
SAITS	3.432	3.692

E.2 Sensitivity to aggregation

Table E3 compares weighting by source and by series, including pools that retain the duplicate Seasonal output. These comparisons use the same errors for each condition.

Table E3: Mean MSE ranks under different pool and aggregation definitions.

Predictor	Method	21 methods Sources	21 methods Series	22 methods Sources	22 methods Series
Sundial	SPImpute	5.388	6.295	5.607	6.516
Sundial	FIM	9.457	9.032	9.883	9.463
TimesFM-2.5	SPImpute	6.375	7.084	6.637	7.389
TimesFM-2.5	FIM	10.503	9.305	10.928	9.716

E.3 Results by series

The series have different units and value ranges, so their raw MSE values can differ greatly and are difficult to compare directly. We therefore report each method's MSE relative to Linear on the same series, after averaging over seeds and missing rates. Linear has a ratio of one, and ratios below one indicate lower error. Using the same denominator preserves the ranking of methods within each row.

Table E4: Sundial: forecasting MSE divided by Linear MSE. Shaded bold and underlined entries mark the two lowest distinct unrounded values per row.

Series	SPImpute	FIM	Linear	SAITS	BiSmooth-Kalman
Australia_Solar_H	**0.726**	0.979	1.000	0.912	0.969
Coastal_T_S_15T	0.983	0.997	1.000	**0.965**	0.980
Coastal_T_S_20T	0.866	0.855	1.000	**0.814**	0.922
Coastal_T_S_5T	1.025	1.002	1.000	1.047	**0.988**
Coastal_T_S_H	0.997	0.990	1.000	**0.960**	1.043
ETTh1	**0.960**	1.000	1.000	1.098	1.053
ETTh2	1.004	1.010	**1.000**	1.127	1.158
ETTm1	0.970	0.962	1.000	**0.938**	0.955
ETTm2	**0.947**	0.949	1.000	0.989	0.947
OpenElectricity_NEM_5T	**0.906**	1.436	1.000	1.128	1.156
Water_Quality_Darwin_15T	1.046	1.055	**1.000**	1.088	1.026
azure2019_D_5T	0.986	1.026	1.000	0.952	**0.935**

Table E4 (continued)

Series	SPImpute	FIM	Linear	SAITS	BiSmooth-Kalman
azure2019_I_5T	**0.909**	0.977	1.000	1.032	0.997
current_velocity_15T	0.953	**0.951**	1.000	0.974	1.040
current_velocity_20T	1.018	1.042	1.000	1.027	**0.989**
current_velocity_5T	**0.961**	1.225	1.000	1.935	1.285
epf_electricity_price_H	**0.970**	0.999	1.000	1.007	1.037
traffic	**0.416**	0.971	1.000	0.610	0.914
weather	0.992	0.994	1.000	1.002	**0.960**

Table E5: TimesFM-2.5: forecasting MSE divided by Linear MSE. Shaded bold and underlined entries mark the two lowest distinct unrounded values per row.

Series	SPImpute	FIM	Linear	SAITS	BiSmooth-Kalman
Australia_Solar_H	**0.665**	1.015	1.000	0.781	0.891
Coastal_T_S_15T	0.983	1.004	1.000	0.952	**0.941**
Coastal_T_S_20T	0.947	**0.844**	1.000	1.245	1.486
Coastal_T_S_5T	1.069	0.999	1.000	1.078	**0.942**
Coastal_T_S_H	0.999	1.002	1.000	**0.991**	1.014
ETTh1	1.025	**0.960**	1.000	1.073	0.979
ETTh2	0.999	**0.998**	1.000	1.054	1.109
ETTm1	1.029	1.010	1.000	**0.880**	0.899
ETTm2	1.051	1.008	1.000	1.012	**0.969**
OpenElectricity_NEM_5T	0.853	0.900	1.000	0.799	**0.743**
Water_Quality_Darwin_15T	0.978	0.977	1.000	**0.907**	0.923
azure2019_D_5T	0.958	0.997	1.000	0.961	**0.924**
azure2019_I_5T	**0.919**	0.988	1.000	1.075	0.996
current_velocity_15T	0.988	**0.985**	1.000	1.182	1.223
current_velocity_20T	1.116	1.017	1.000	1.049	**0.979**
current_velocity_5T	1.619	0.825	1.000	1.061	**0.729**
epf_electricity_price_H	0.968	1.012	1.000	**0.959**	0.965
traffic	**0.311**	0.982	1.000	0.490	0.813
weather	1.010	1.045	1.000	1.042	**0.988**

Table E6: Chronos-2: forecasting MSE divided by Linear MSE. Shaded bold and underlined entries mark the two lowest distinct unrounded values per row.

Series	SPImpute	FIM	Linear	SAITS	BiSmooth-Kalman
Australia_Solar_H	**0.760**	1.067	1.000	0.864	0.971
Coastal_T_S_15T	0.994	0.992	1.000	0.980	**0.963**
Coastal_T_S_20T	1.267	**0.940**	1.000	1.286	1.499
Coastal_T_S_5T	0.923	0.974	1.000	1.157	**0.905**
Coastal_T_S_H	0.996	1.009	1.000	**0.976**	1.084
ETTh1	0.988	**0.972**	1.000	1.096	1.069
ETTh2	**0.986**	0.997	1.000	1.085	1.093
ETTm1	1.011	1.039	1.000	**0.927**	0.973
ETTm2	1.072	1.033	1.000	1.018	**0.936**
OpenElectricity_NEM_5T	**0.836**	0.929	1.000	1.114	0.934
Water_Quality_Darwin_15T	1.017	**0.985**	1.000	1.081	1.011
azure2019_D_5T	**0.951**	0.987	1.000	0.990	0.956
azure2019_I_5T	**0.973**	0.984	1.000	1.152	1.031
current_velocity_15T	1.066	1.030	**1.000**	1.128	1.159
current_velocity_20T	1.726	0.966	1.000	**0.921**	0.925
current_velocity_5T	1.348	1.178	**1.000**	2.268	1.204
epf_electricity_price_H	1.020	1.007	**1.000**	1.090	1.021
traffic	**0.688**	1.040	1.000	0.918	1.125
weather	1.008	1.027	1.000	0.995	**0.990**

E.4 Effects by source and missing rate

Figure E1 shows both the mean effects and their variation across ten equally weighted sources. The intervals resample sources. Table E7 separates the missing rates.

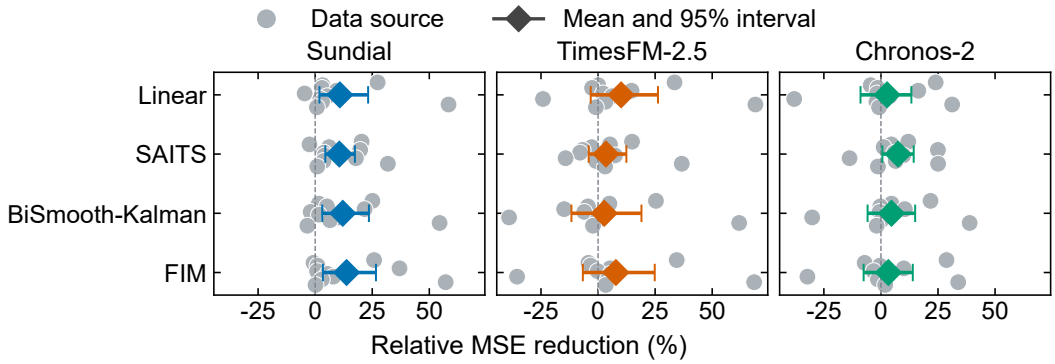


Figure E1: Forecasting effects across ten equally weighted sources. Gray dots are source-level MSE reductions; diamonds and bars give their mean and individual 95% bootstrap intervals. Positive values favor SPImpute; averages include improvements and regressions.

Table E7: Source-mean reduction in SPImpute MSE relative to Linear at every missing rate.

Predictor	Missing rate	Reduction (%)	95% CI
Sundial	10%	6.05	[0.71, 14.66]
Sundial	20%	7.88	[1.39, 17.75]
Sundial	30%	11.37	[2.68, 23.10]
Sundial	40%	12.39	[1.77, 26.63]
Sundial	50%	12.99	[1.54, 27.84]
TimesFM-2.5	10%	1.49	[-4.82, 7.20]
TimesFM-2.5	20%	11.53	[4.29, 20.25]
TimesFM-2.5	30%	14.44	[1.74, 29.76]
TimesFM-2.5	40%	14.88	[2.48, 30.97]
TimesFM-2.5	50%	1.36	[-30.82, 28.23]
Chronos-2	10%	-8.68	[-22.13, 1.57]
Chronos-2	20%	4.84	[0.37, 9.45]
Chronos-2	30%	5.91	[-0.14, 12.45]
Chronos-2	40%	-3.89	[-31.69, 14.85]
Chronos-2	50%	8.17	[-3.50, 21.13]

E.5 FIM results with and without fallback

The complete FIM adapter includes recorded linear fallback. The subset without fallback excludes every history containing a fallback gap and therefore has a different gap distribution from the full evaluation. Table E8 pairs methods separately for these two sets; the subset results apply only to histories supported directly by FIM.

Table E8: SPImpute relative to FIM, with source bootstrap intervals and descriptive series W/T/L counts.

Predictor	Scope	Histories	Reduction (%)	95% CI	W/T/L
Sundial	All	4,515	13.74	[3.38, 26.63]	14/0/5
Sundial	No fallback	2,370	12.82	[2.83, 25.52]	14/0/5
TimesFM-2.5	All	4,515	7.82	[-6.67, 24.83]	9/0/10
TimesFM-2.5	No fallback	2,370	10.96	[1.59, 23.50]	14/0/5
Chronos-2	All	4,515	3.25	[-7.59, 13.95]	10/0/9
Chronos-2	No fallback	2,405	6.90	[2.89, 11.69]	15/0/4

F Ablation and Confirmation Experiments

F.1 Base objectives without same-phase replacement

The base comparison fixes observed values and disables same-phase replacement for every objective. Local smoothing removes the additional guidance term. Value fidelity, periodic consistency, and increment fidelity use the same references, candidate weights, tasks, and default threshold. Table F1 reports the nine primary comparisons. Same-phase correction leaves these base results unchanged.

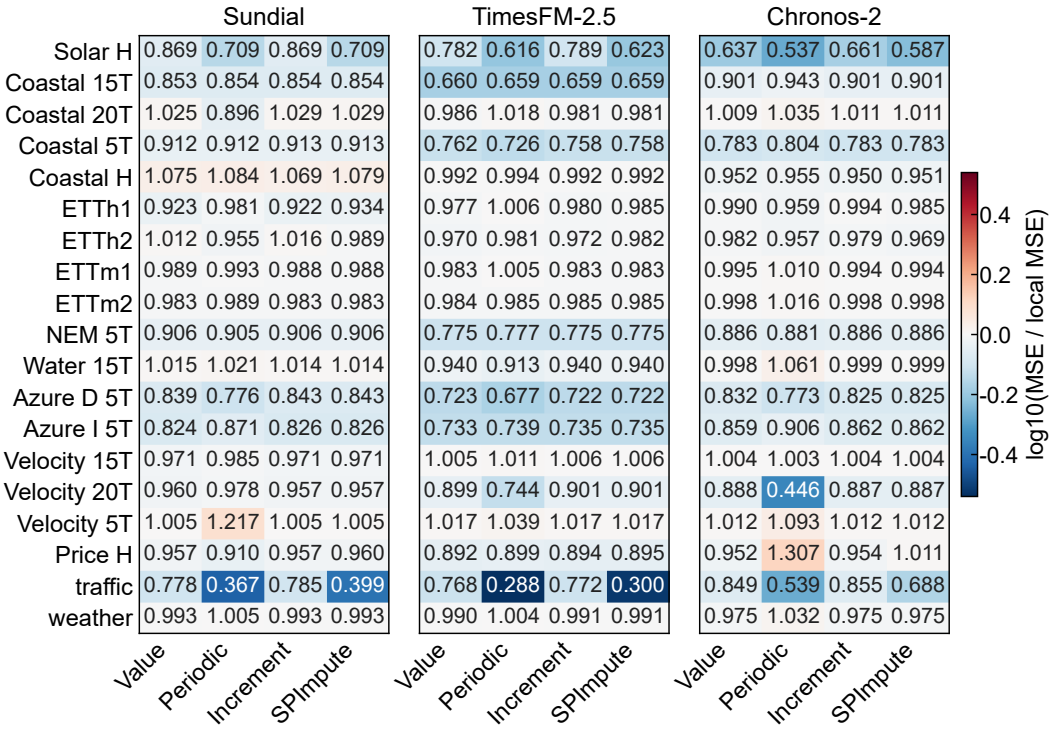


Figure F1: Per-series MSE ratios relative to local smoothing (ratio one). The first three columns disable same-phase replacement; SPImpute includes it. Blue/red indicates lower/higher error on a shared logarithmic scale. Dataset abbreviations follow Table D1.

Table F1: Reduction in base increment MSE relative to other base objectives. All nine tests share Holm correction.

Predictor	Comparator	Reduction (%)	95% CI	Holm p
Sundial	Local smoothing	7.21	[3.03, 11.83]	0.0410
Sundial	Value fidelity	-0.11	[-0.30, 0.02]	0.9062
Sundial	Periodic consistency	-13.79	[-37.39, 0.71]	0.9062
TimesFM-2.5	Local smoothing	13.08	[7.25, 18.85]	0.0176
TimesFM-2.5	Value fidelity	-0.18	[-0.38, -0.02]	0.4453
TimesFM-2.5	Periodic consistency	-20.44	[-54.43, -0.46]	0.4453
Chronos-2	Local smoothing	9.55	[4.26, 16.26]	0.0176
Chronos-2	Value fidelity	-0.47	[-1.28, 0.01]	0.9062
Chronos-2	Periodic consistency	-7.36	[-22.51, 5.60]	0.9062

F.2 Matched controls under BM30

All 903 BM30 histories are evaluated under each predictor; the two main predictors share masks and Chronos-2 uses independent masks. Value, degree-weighted value, periodic, and local objectives receive the same corrected q, equality rule, validation tasks, and selection threshold. The weighted-value control keeps the increment penalty's diagonal and removes missing–missing cross-products.

Standalone q, fixed increment weights, the increment base, and the procedure using $q^{(0)}$ separate the candidate and selection effects.

Table F2: Source-mean reduction in final SPImpute MSE at BM30. The first five comparators per predictor form the 15-test primary family; the last four form the 12-test secondary family. Intervals are individual.

Predictor	Comparator	Reduction (%)	95% CI	Holm p
Sundial	Value + q	0.082	[-0.041, 0.217]	1.0000
Sundial	Weighted value + q	0.000	[-0.104, 0.107]	1.0000
Sundial	Periodic + q	-1.028	[-4.230, 2.686]	1.0000
Sundial	q alone	-1.915	[-8.555, 5.003]	1.0000
Sundial	Linear	11.278	[2.678, 22.166]	0.3516
Sundial	Increment base	7.267	[0.190, 17.758]	1.0000
Sundial	Local + q	4.319	[1.156, 8.087]	0.2148
Sundial	Fixed increment	8.191	[-0.173, 19.159]	1.0000
Sundial	Uncorrected ($q^{(0)}$)	-0.006	[-0.046, 0.022]	1.0000
TimesFM-2.5	Value + q	-0.098	[-0.168, -0.017]	0.5469
TimesFM-2.5	Weighted value + q	-0.088	[-0.168, -0.011]	0.8633
TimesFM-2.5	Periodic + q	-2.289	[-5.671, 0.690]	1.0000
TimesFM-2.5	q alone	-3.895	[-8.063, 1.359]	1.0000
TimesFM-2.5	Linear	14.440	[1.972, 29.216]	0.8672
TimesFM-2.5	Increment base	7.542	[-0.316, 19.235]	1.0000
TimesFM-2.5	Local + q	7.272	[1.698, 15.157]	0.0938
TimesFM-2.5	Fixed increment	5.378	[-5.575, 18.823]	1.0000
TimesFM-2.5	Uncorrected ($q^{(0)}$)	0.192	[-0.134, 0.737]	1.0000
Chronos-2	Value + q	-0.063	[-0.243, 0.106]	1.0000
Chronos-2	Weighted value + q	-0.013	[-0.154, 0.121]	1.0000
Chronos-2	Periodic + q	3.681	[-3.242, 11.209]	1.0000
Chronos-2	q alone	3.084	[-3.135, 9.578]	1.0000
Chronos-2	Linear	5.909	[-0.011, 12.305]	1.0000
Chronos-2	Increment base	2.059	[-2.698, 6.823]	1.0000
Chronos-2	Local + q	4.630	[1.643, 8.072]	0.2148
Chronos-2	Fixed increment	1.851	[-3.782, 7.604]	1.0000
Chronos-2	Uncorrected ($q^{(0)}$)	0.025	[0.000, 0.054]	1.0000

F.3 Periodic masks on real data

This comparison uses the unchanged values of the nineteen real series. The first, middle, and last seed-42 BM30 origins are selected without using errors. Seeds 3042–3044 determine the phase of up to six gaps of length $\lfloor p/4 \rfloor$ in the final periods. Aligned and staggered masks have the same missing-point count and final gap. All three predictors evaluate 342 histories. Because the gaps are shorter than the period, the same-phase branch is ineligible. Periodic connectivity alone does not yield a uniform forecasting advantage.

Table F3: Source-mean MSE reduction (%) with individual 95% intervals under controlled real masks. Positive favors SPImpute.

Predictor	Gaps	Weighted value	Periodic	Linear
Sundial	Aligned	-1.30 [-3.00, -0.29]	1.21 [-2.14, 5.18]	0.53 [-1.75, 3.05]
Sundial	Staggered	-0.01 [-0.03, 0.00]	-1.12 [-4.25, 1.11]	1.27 [0.22, 2.55]
TimesFM-2.5	Aligned	-1.37 [-4.92, 0.91]	4.31 [-1.96, 11.97]	4.70 [1.04, 8.90]
TimesFM-2.5	Staggered	-0.01 [-0.03, 0.00]	-0.01 [-4.85, 5.35]	0.26 [-3.60, 3.83]
Chronos-2	Aligned	0.36 [-3.28, 4.88]	1.99 [-3.71, 8.46]	6.34 [0.44, 12.67]
Chronos-2	Staggered	0.02 [0.00, 0.05]	-3.24 [-7.83, 1.08]	1.50 [-1.61, 5.45]

F.4 SHORT GAPS IN SYNTHETIC PULSES

Fourteen fixed conditions use 64 independent generation seeds (4000–4063), giving 896 histories. Twelve central conditions use $p = 48$, $n = 576$, and $x_t = \exp(-((t/p \bmod 1 - 0.4)/0.12)^2) + at/(4p) + \epsilon_t$, with $a \in \{0, 0.5, 2\}$ and $\epsilon_t \sim \mathcal{N}(0, 0.03^2)$. Gaps of length twelve occur in the last one, three, or six periods; the staggered condition shifts only the middle of three gaps by half a period. Two end-clipped controls are also retained. These short gaps leave the same-phase candidate ineligible, so SPImpute retains the same base reconstruction.

The following tables report every condition and comparator. The standalone same-phase baseline uses the uncorrected $q^{(0)}$. Holm correction covers all 210 primary comparisons, with individual paired intervals reported alongside the complete results. Appendix C.1 establishes equivalence for fixed weights when no edge joins two missing coordinates; this experiment also evaluates the effects of candidate selection.

Table F4: Sundial: complete pulse-replication MSE, averaged over 64 seeds. Among displayed candidates, shaded bold and underlined entries mark the two lowest distinct unrounded MSEs per row.

Condition	SPImpute	Value	Weighted	Periodic	$q^{(0)}$ only	Linear
$a = 0$, 1 aligned	0.001551	0.001552	0.001551	0.001308	**0.001306**	0.002825
$a = 0$, 3 aligned	0.002915	0.002948	0.00291	**0.001358**	0.002844	0.006944
$a = 0$, 6 aligned	0.006466	0.006463	0.006334	**0.001785**	0.008826	0.02086
$a = 0$, 3 staggered	0.001833	0.001833	0.001828	0.001314	**0.001311**	0.003201
$a = 0.5$, 1 aligned	1.927	1.927	1.927	**1.739**	1.8	2.386
$a = 0.5$, 3 aligned	2.038	2.035	2.039	**1.924**	2.277	2.655
$a = 0.5$, 6 aligned	**2.153**	2.154	2.161	2.529	2.711	2.773
$a = 0.5$, 3 staggered	2.12	2.12	2.12	1.79	**1.762**	2.695
$a = 2$, 1 aligned	34.35	34.35	34.35	35.32	**33.44**	35.56
$a = 2$, 3 aligned	36.33	36.15	36.13	40.38	**33.46**	39.66
$a = 2$, 6 aligned	**40.82**	40.98	40.92	44.31	44.98	47.78
$a = 2$, 3 staggered	35.42	35.42	35.42	35.5	**35.03**	38.46
End-clipped, $p = 24$	0.007276	0.007207	0.007276	**0.00167**	0.001682	0.01377
End-clipped, $p = 48$	0.003543	0.003518	0.003543	**0.001289**	0.001293	0.004631

Table F5: TimesFM-2.5: complete pulse-replication MSE, averaged over 64 seeds. Among displayed candidates, shaded bold and underlined entries mark the two lowest distinct unrounded MSEs per row.

Condition	SPImpute	Value	Weighted	Periodic	$q^{(0)}$ only	Linear
$a = 0$, 1 aligned	0.002692	0.002696	0.002692	0.001841	**0.00182**	0.002296
$a = 0$, 3 aligned	0.008917	0.009438	0.00919	0.00268	**0.002233**	0.02058
$a = 0$, 6 aligned	0.02423	0.026	0.02563	**0.008116**	0.03048	0.04553
$a = 0$, 3 staggered	0.003033	0.003049	0.00303	**0.001849**	0.001881	0.004459
$a = 0.5$, 1 aligned	0.2933	0.2934	0.2933	**0.1218**	0.1847	0.2661
$a = 0.5$, 3 aligned	**0.327**	0.349	0.3474	0.7032	0.8434	0.4415
$a = 0.5$, 6 aligned	**0.2068**	0.2392	0.2381	0.5554	0.5262	0.3079
$a = 0.5$, 3 staggered	0.1102	0.1102	0.1094	0.1149	**0.07237**	0.1328
$a = 2$, 1 aligned	3.027	3.042	3.027	3.473	**2.819**	5.39
$a = 2$, 3 aligned	3.538	3.836	3.858	2.777	**1.942**	8.67
$a = 2$, 6 aligned	**1.375**	1.753	1.746	2.374	3.683	4.987
$a = 2$, 3 staggered	0.7386	0.7417	0.7386	2.911	**0.1959**	1.042
End-clipped, $p = 24$	0.00454	0.004495	0.00454	0.001322	**0.001298**	0.009955
End-clipped, $p = 48$	0.03974	0.04074	0.03974	0.001894	**0.001803**	0.04112

Table F6: Chronos-2: complete pulse-replication MSE, averaged over 64 seeds. Among displayed candidates, shaded bold and underlined entries mark the two lowest distinct unrounded MSEs per row.

Condition	SPImpute	Value	Weighted	Periodic	$q^{(0)}$ only	Linear
$a = 0$, 1 aligned	0.002545	0.002548	0.002545	0.001009	**0.001002**	0.00445
$a = 0$, 3 aligned	0.009894	0.01032	0.01008	**0.00177**	0.004494	0.02967
$a = 0$, 6 aligned	0.01521	0.01608	0.01592	**0.006455**	0.0299	0.02522
$a = 0$, 3 staggered	0.001982	0.002005	0.001979	0.001022	**0.001013**	0.00185
$a = 0.5$, 1 aligned	0.5556	0.5533	0.5556	0.504	0.7042	**0.1462**
$a = 0.5$, 3 aligned	0.4237	0.4447	0.4169	0.9321	0.8798	**0.119**
$a = 0.5$, 6 aligned	**0.09256**	0.1028	0.1027	0.4892	0.1939	0.1455
$a = 0.5$, 3 staggered	0.2524	0.2554	0.2529	0.4429	0.09356	**0.03554**
$a = 2$, 1 aligned	18.24	18.25	18.24	18.58	**0.2954**	20.34
$a = 2$, 3 aligned	**1.85**	1.858	2.033	4.571	12.18	1.904
$a = 2$, 6 aligned	**0.4601**	0.5991	0.6025	7.233	0.5521	1.347
$a = 2$, 3 staggered	7.286	7.567	7.286	15.74	**0.3281**	15.44
End-clipped, $p = 24$	0.002792	0.002744	0.002792	0.001142	**0.001092**	0.00551
End-clipped, $p = 48$	0.02637	0.0239	0.02637	0.001021	**0.001003**	0.08459

F.5 Long-gap confirmation

The confirmation study uses the 1,152 fixed synthetic inputs described in Appendix D.3. Figure F2 reports every near-origin condition; the corresponding interior conditions all preserve the forecasts from the uncorrected procedure. Each MSE is averaged over sixteen generation seeds. Two families of 216 comparisons test SPImpute against the procedure using $q^{(0)}$ and against the increment base, respectively. The main confirmation figure weights the two gap lengths equally within each waveform, period, and drift group; its intervals resample the common generation seeds.

All 576 interior-gap inputs preserve the uncorrected output exactly, as specified by the candidate-equality rule. Near-origin effects vary by predictor and condition. The identity for corrected coordinates does not guarantee lower forecasting error.

F.6 Confirmation on additional real data

The 225 confirmation tasks use five source collections and disjoint future intervals. Figure 7 shows four comparisons for each source, and Table F7 reports the twelve prespecified tests across sources. None passes the 12-test Holm correction. With five sources, the smallest attainable two-sided sign-flip value is 0.0625; individual bootstrap intervals do not change this resolution. These results describe the tested inputs and do not establish an overall advantage over matched weighted-value reconstruction.

Table F7: Five-source confirmation: MSE reductions with individual 95% intervals. All twelve tests share Holm correction.

Predictor	Comparator	Reduction (%)	95% CI	Raw p	Holm p
Sundial	Uncorrected ($q^{(0)}$)	-0.058	[-0.165, 0.000]	0.5000	1.0000
Sundial	Weighted value + q	-0.405	[-0.920, -0.062]	0.0625	0.7500
Sundial	Periodic + q	-0.242	[-3.112, 2.618]	0.8125	1.0000
Sundial	Linear	16.610	[2.023, 31.197]	0.1250	1.0000
TimesFM-2.5	Uncorrected ($q^{(0)}$)	0.107	[0.000, 0.282]	0.5000	1.0000
TimesFM-2.5	Weighted value + q	-0.090	[-0.191, 0.008]	0.2500	1.0000
TimesFM-2.5	Periodic + q	2.718	[-0.185, 6.142]	0.2500	1.0000
TimesFM-2.5	Linear	19.861	[-0.240, 42.600]	0.1875	1.0000
Chronos-2	Uncorrected ($q^{(0)}$)	-0.010	[-0.028, 0.000]	0.5000	1.0000
Chronos-2	Weighted value + q	-0.319	[-0.738, 0.029]	0.2500	1.0000
Chronos-2	Periodic + q	4.977	[-0.644, 10.957]	0.3125	1.0000
Chronos-2	Linear	12.929	[-2.788, 29.734]	0.5000	1.0000

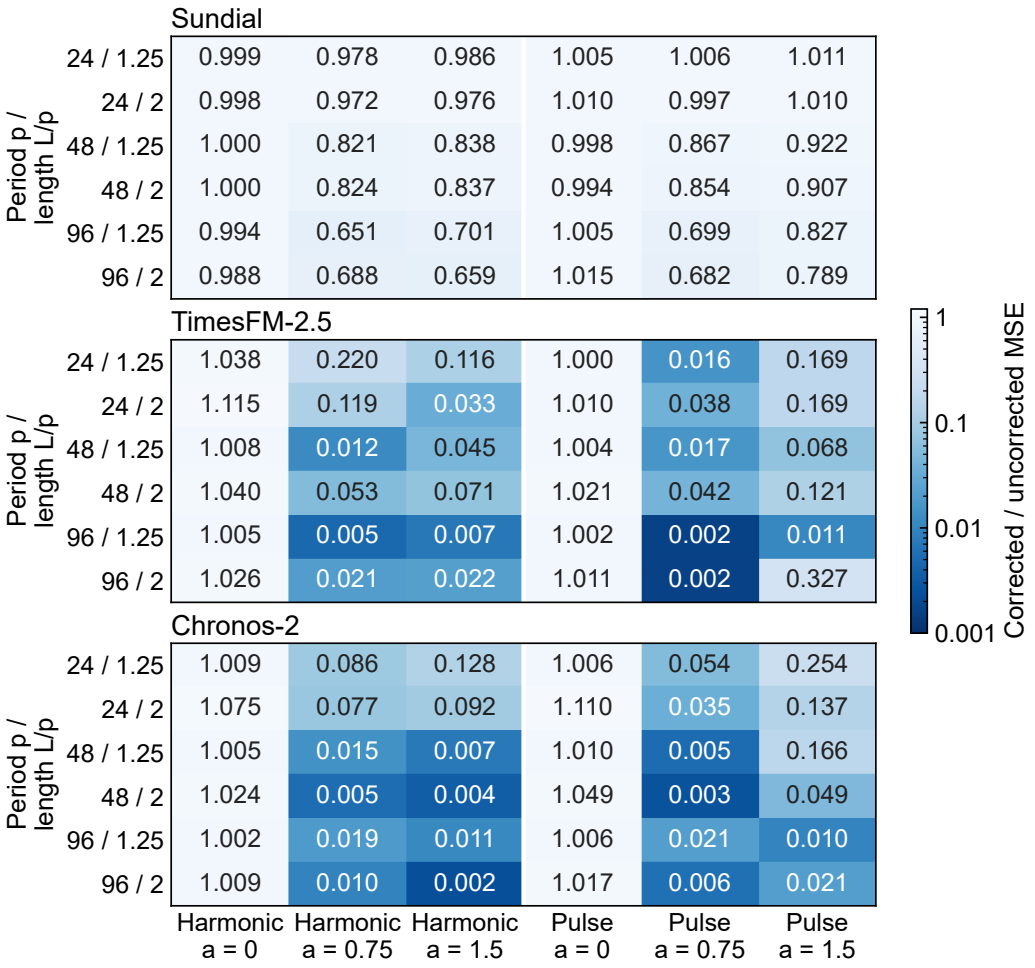


Figure F2: All 36 near-origin conditions per predictor. Cells divide corrected by uncorrected MSE, each averaged over sixteen seeds. Rows specify period and normalized gap length; columns specify waveform and drift. Darker blue indicates lower ratios on the logarithmic scale. The 36 corresponding interior conditions have ratio 1.000 for every predictor: all 576 inputs and forecasts are unchanged.

F.7 Ranks including base objectives

Table F8 adds local smoothing, value fidelity, periodic consistency, and increment reconstruction with q disabled. These larger pools compare base objectives with external methods. Table F2 separately compares objectives under matched same-phase selection.

Table F8: Expanded source-weighted MSE ranks: pools of 25, 25, and nine methods. Shaded bold and underlined entries mark the two lowest distinct unrounded values per column.

Method	Sundial	TimesFM-2.5	Chronos-2
Periodic consistency	**6.772**	**7.560**	4.733
SPImpute	7.028	7.886	3.857
Value fidelity	7.777	9.023	**3.823**
Increment base	8.082	9.721	4.067
BiSmooth-Kalman	10.210	7.997	5.305
SAITS	10.927	8.808	5.818
Nearest	11.080	10.223	–
Linear	11.393	13.147	5.408
LQDA	11.753	13.330	–
Backward	11.867	13.200	–
GPT4TS	11.977	9.828	–

Table F8 (continued)

Method	Sundial	TimesFM-2.5	Chronos-2
STL-Kalman	12.023	14.225	–
FIM	12.033	12.843	5.403
Forward	12.832	11.608	–
Local smoothing	13.047	16.193	6.585
Mean	13.405	11.545	–
TimeMixer (PyPOTS)	13.605	10.287	–
Kalman-ARIMA	13.942	11.898	–
HELIX	14.150	12.037	–
TOTEM	14.782	12.652	–
Polynomial	16.533	17.663	–
Kalman-Struct	17.175	19.187	–
GP-RBF	18.238	19.895	–
Spline	20.398	20.747	–
Zero	23.972	23.497	–

G Runtime and Training Budgets

G.1 Sparse solution and caching

Each difference operator and Gram matrix requires $\mathcal{O}(n)$ storage in the sparse implementation. Eliminating observations leaves $m = |M|$ unknowns. Direct factorization cost depends on bandwidth and fill-in (Golub & Van Loan, 2013, Chapter 11); conjugate gradients cost approximately $\mathcal{O}(K \operatorname{nnz}(A))$ for K iterations (Saad, 2003, Chapter 6). Operator caches depend only on length, order, and period and hold up to 64 entries. Each validation task constructs its references once, and each weight triple requires one base solve. The same-phase candidate leaves the quadratic system unchanged.

G.2 Runtime comparison

The runtime study covers all ten methods displayed in Figure 3. Each method uses the same 57 BM30 histories: first, middle, and last origins from all nineteen series under mask seed 42, with the original float32-to-float64 conversion and imputation seeds. One warmup per method and history precedes three repetitions, yielding 171 timed calls per method. The first batch measured Linear, Nearest, BiSmooth-Kalman, SPImpute, SAITS, and FIM; a second batch added GPT4TS, LQDA, TimeMixer, and TOTEM on the same host. Histories and methods are randomly interleaved within each batch. The two batches were measured on consecutive days.

Linear, Nearest, LQDA, SPImpute, and BiSmooth-Kalman run on the Intel Core i7-10700, with BLAS and PyTorch each limited to two CPU threads. SAITS, FIM, GPT4TS, TimeMixer, and TOTEM use the RTX 3060. All calls execute serially. Timing covers normalization, gap or window construction, fitting or training, solving or inference, and observation restoration. GPU timings include transfers and synchronization. Input DataFrame construction, experiment-file I/O, process communication, output checks, and forecasting are excluded. Figure 8 uses one common time axis; the measurements retain these device and workload differences.

SAITS uses the forecasting comparison's PyPOTS 1.5 configuration: length-96 windows with stride 48, two layers of width 64, four heads, batch size at most 32, and at most ten epochs with patience three. Each call initializes and trains a new model. FIM retains the temporal-base and pointwise-base weights, float32 inference, length-256 adapter, and linear fallback. Loading FIM and checking checkpoint identity took 6.89 s once, outside per-call timing. FIM repairs 48,000 missing points directly and 518 through fallback; 19 of the 57 histories include fallback.

GPT4TS, TimeMixer, and TOTEM use the original length-96 adapters, batch size at most 32, and three training epochs per history. GPT4TS reloads the locally cached GPT-2 weights on each call; this loading cost is included, with remote metadata requests disabled. TimeMixer and TOTEM initialize new models on each call. GPT4TS requires no fallback on this timing panel. LQDA uses its original local quantile-alignment implementation.

All 570 warmup outputs match the saved predictor inputs. All 1,710 timed outputs equal their warmup outputs and preserve observations. Unrelated system activity can affect timings on the shared host and GPU; the separately measured batches do not control for changes in this activity.

G.3 SAITS TRAINING BUDGET

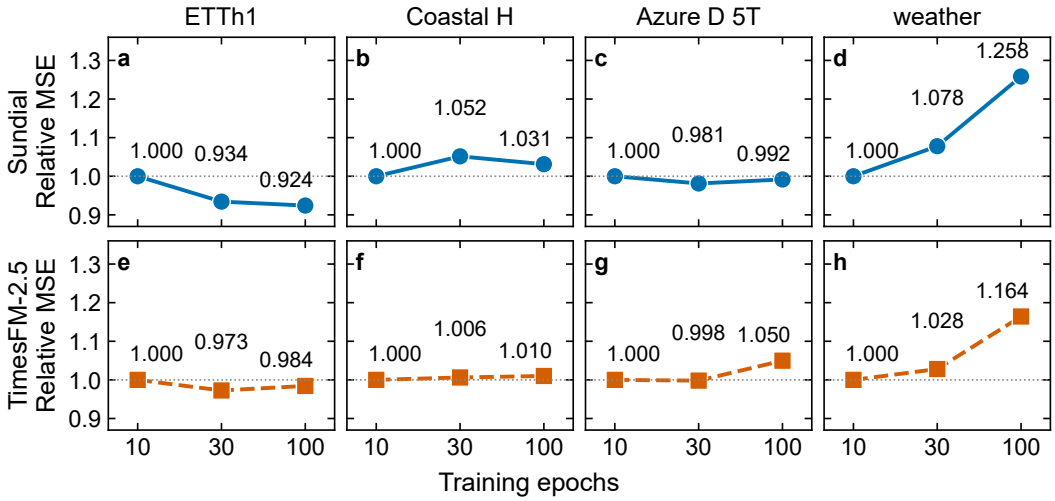


Figure G1: SAITS training budgets by series (columns) and predictor (rows). Curves show forecast MSE at the best validation checkpoint within 10, 30, or 100 epochs, relative to the 10-epoch MSE; lower is better. Main-benchmark adapter settings remain unchanged.

This separate baseline study uses 108 BM30 histories from four main sources. Independent validation blocks are excluded from overlapping training windows and normalization. SAITS uses two layers, width 64, four heads, and length-96 windows. A single 100-epoch training run per history provides the checkpoints with lowest validation RMSE within the first 10, 30, and 100 epochs. The main comparison uses the adapter configuration in Appendix D.2.