SCALING GAINS HAVE SHAPE: STRUCTURAL RESPONSE IN TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling a time-series foundation model is usually summarized by the error reduction from a smaller checkpoint to a larger one. This number identifies whether an upgrade works, but collapses the data conditions that support its gain. We introduce *Structural Scaling Response* (SSR): the upgrade's log-risk gain on the original series minus its gain on a matched surrogate that preserves the power spectrum while randomizing Fourier phases. Positive SSR means that the original temporal phase structure amplifies the upgrade gain; negative SSR means that it attenuates the gain. Restoring the original phases from low to high frequencies produces an SSR curve that locates the timescales at which the difference emerges. We characterize when this response vanishes and bound the risk perturbation needed to change its sign. On untouched PEMS08, all four larger checkpoints improve accuracy, with channel-aggregated error reductions of 10.02–38.54%. Their SSRs nevertheless separate: the two Toto upgrades and the Chronos upgrade are positive (0.286, 0.209, and 0.113), while the Moirai upgrade is negative (-0.143). Matched surrogates, alternative losses, blocked resampling, mean aggregation, and channel-resampling analyses preserve all four signs. Across 16 upgrade–dataset combinations, SSR is specific to the pair: the same upgrade changes across datasets, and one dataset can contain both signs. On PEMS04 and PEMS08, each positive SSR reaches half its endpoint after phases at periods down to 8–15 hours have returned. Temporal phase structure is therefore a second coordinate of scaling. SSR provides a training-free protocol for measuring where scaling gains arise and comparing upgrades under deployment-relevant temporal structure.

1 INTRODUCTION

Larger checkpoints are now a standard upgrade path for time-series foundation models. Evaluations rank these upgrades by their error reduction on the original data. That scalar treats the upgrade gain as intrinsic to the checkpoint pair. It erases a second variable: the temporal phase structure of the deployment data.

Untouched PEMS08 exposes the consequence. Toto 4M→22M, Toto 22M→313M, Chronos tiny→base, and Moirai small→large improve forecasting error by 38.54%, 25.56%, 14.73%, and 10.02%. The leaderboard records four wins. A controlled phase intervention reveals a split: the Toto and Chronos gains increase as original phases return, while the Moirai gain increases after those phases are randomized. The same accuracy verdict can therefore hide opposite dependence on temporal phase structure.

Structural Scaling Response (SSR) measures this dependence. Endpoint SSR is the upgrade gain on the original series minus the gain on a power-spectrum-preserving, phase-randomized surrogate. Positive SSR means that original temporal phase structure amplifies the upgrade gain; negative SSR means that it attenuates the gain. Cumulatively restoring phases from low to high frequencies produces an SSR curve and locates the timescales at which the two gains diverge.

For a released upgrade, let $\mathcal{G}(q)$ be the log-risk gain of the larger checkpoint after retaining the lowest-frequency fraction q of original Fourier phases and randomizing the rest. Every value of q

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.

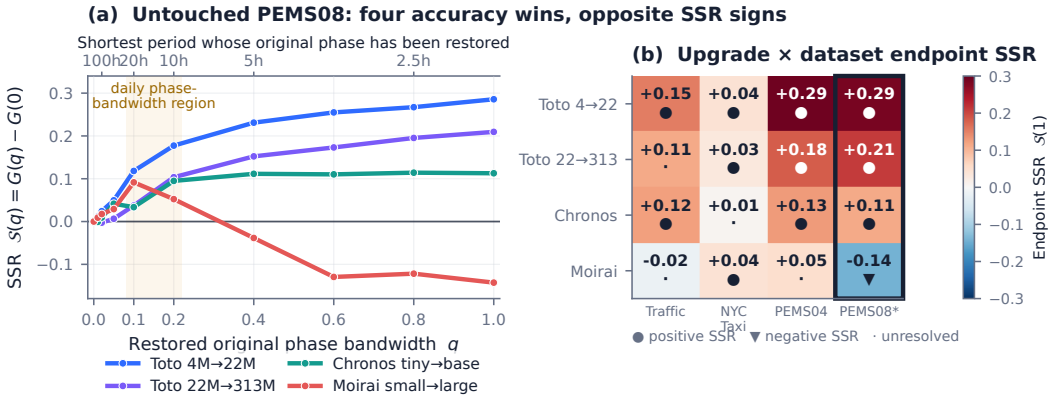


Figure 1: **Four accuracy wins, opposite SSR signs.** (a) PEMS08 contains three positive SSRs and one negative. (b) Endpoint SSR across 16 upgrade–dataset pairs; the outline marks untouched PEMS08.

preserves the per-channel power spectrum. The SSR curve is

$$\mathcal{S}(q) = \mathcal{G}(q) - \mathcal{G}(0), \qquad q \in [0, 1], \tag{1}$$

where $q = 0$ randomizes the nontrivial phases and $q = 1$ recovers the original series. At the endpoint,

$$\mathcal{G}(1) = \underbrace{\mathcal{G}(0)}_{\text{phase-randomized gain}} + \underbrace{\mathcal{S}(1)}_{\text{endpoint SSR}}, \tag{2}$$

Equation 2 separates the observed upgrade gain into a phase-randomized gain and endpoint SSR. A leaderboard observes only their sum, so it leaves the SSR sign undetermined. One controlled surrogate measurement reveals the missing coordinate.

SSR crosses two established experimental axes. Scaling studies vary resources and measure checkpoint loss. Phase-surrogate studies vary temporal structure and measure a single model. SSR varies temporal phase structure and measures the *upgrade gain*. It therefore distinguishes upgrades that share the same accuracy verdict but have opposite structural dependence.

Figure 1 summarizes the evidence. PEMS08 contains three positive SSRs and one negative SSR. The complete 16-combination map shows that SSR is specific to an upgrade–dataset pair. Iterative amplitude-adjusted Fourier transform (IAAFT), multivariate Fourier surrogates, a non-Fourier day shuffle, four loss definitions, and channel–time resampling preserve the PEMS08 signs. On PEMS04 and PEMS08, positive SSR reaches half of its endpoint after phases at periods down to 8–15 hours have returned.

We make three contributions:

- SSR exposes the structural dependence hidden by an original-data upgrade gain. A spectral zero-response benchmark and a sharp perturbation bound characterize when the response vanishes and when its sign is stable.
- Untouched PEMS08 establishes the central finding: four accuracy improvements split into three positive SSRs and one negative SSR, with every sign preserved by matched controls and sensitivity analyses.
- The 16-combination map identifies SSR as upgrade–dataset specific and localizes positive PEMS SSR to an 8–15 hour phase bandwidth. A separate endpoint audit finds no resolved positive contrast on Electricity or Weather.

Our claims concern named released upgrades on fixed evaluation datasets. Toto 4M→22M provides the cleanest capacity comparison because its published training exposure is matched. The remaining upgrades establish that SSR occurs across public model families. Within zero-shot hourly mobility forecasting, accuracy measures whether an upgrade wins; SSR measures how temporal phase structure supports its gain.

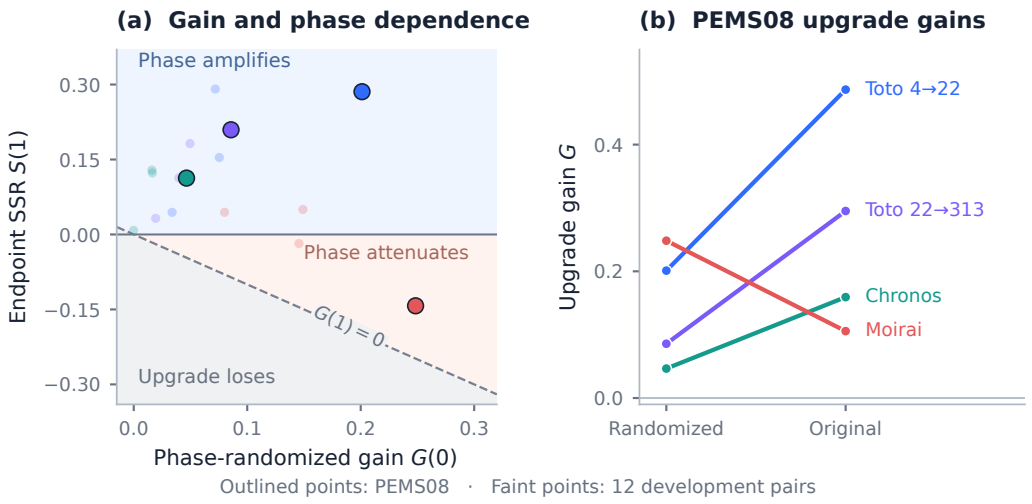


Figure 2: **Accuracy and phase dependence have different boundaries.** Left: all 16 estimates; outlined points are PEMS08. Right: PEMS08 endpoint gains, with straight lines joining estimates.

2 From a Scaling Gain to Structural Scaling Response

2.1 A paired comparison of upgrade gains

For an upgrade $u = (m_s, m_l)$ on dataset d, let c index channels, t forecast start times, and w a controlled input at restoration level q. We first average forecast loss within each channel and input, then compute the log-risk gain:

$$\mathcal{R}_{m,d,c,w} = \max\left\{ \frac{1}{|\mathcal{T}_d|} \sum_{t \in \mathcal{T}_d} \ell_{m,d,c,t,w}, 10^{-9} \right\}, \qquad g_{u,d,c,w} = \log \mathcal{R}_{m_s,d,c,w} - \log \mathcal{R}_{m_l,d,c,w}. \tag{3}$$

The median summarizes these paired gains over the fixed channel set $\mathcal{C}_d$ and the inputs $\mathcal{W}_q$ at that level:

$$\mathcal{G}_{u,d}(q) = \mathrm{median}_{c \in \mathcal{C}_d,\, w \in \mathcal{W}_q}\, g_{u,d,c,w}, \qquad \mathcal{S}_{u,d}(q) = \mathcal{G}_{u,d}(q) - \mathcal{G}_{u,d}(0). \tag{4}$$

The original series has one input; each synthetic level has seven. Positive $\mathcal{G}$ means that the larger checkpoint improves the channel-aggregated relative risk.

The decomposition in Equation 2 creates two distinct boundaries (Figure 2):

$$\underbrace{\mathcal{S}(1) = 0}_{\text{phase-dependence boundary}}, \qquad \underbrace{\mathcal{S}(1) = -\mathcal{G}(0)}_{\text{accuracy boundary}}. \tag{5}$$

When $\mathcal{G}(0) > 0$ and $-\mathcal{G}(0) < \mathcal{S}(1) < 0$, original temporal phase structure attenuates the upgrade gain while the larger checkpoint still improves accuracy. PEMS08 places Moirai in this region: $\mathcal{G}(0) = 0.248$, $\mathcal{S}(1) = -0.143$, and $\mathcal{G}(1) = 0.106$. The two Toto upgrades and Chronos instead have positive SSR. One original-series gain leaves both regimes possible; a surrogate comparison identifies the missing coordinate.

2.2 The cumulative phase-restoration path

For a channel $x \in \mathbb{R}^T$, write its discrete Fourier coefficients as $X_k = A_k e^{i\phi_k}$. At level q, retain the original phases through $k_q = \lfloor q \lfloor T/2 \rfloor \rfloor$ and randomize the higher nontrivial phases. The DC coefficient and, for even T, the Nyquist coefficient stay fixed. Conjugate symmetry keeps the reconstructed series real. Every intervention T_q preserves

$$|\mathcal{F}T_q(x)|^2 = |\mathcal{F}x|^2. \tag{6}$$

Truncated Fourier-transform surrogate analysis supplies this construction (Nakamura et al., 2006); SSR applies it to the *upgrade gain*. We use

$$q \in \{0, .01, .02, .05, .10, .20, .40, .60, .80, 1\},$$

with seven independent draws at each $q < 1$ and the original series at $q = 1$. The curve therefore summarizes successive restoration levels, with fresh random phases at each level.

For hourly data, the retained low-frequency prefix extends down to period T/k_q when $k_q \geq 1$. We report the convenient cutoff coordinate $2/q$ hours, its continuous-grid approximation. The low-to-high order localizes where upgrade gains diverge. It provides a cumulative bandwidth measurement rather than additive attribution to separate bands. Half recovery is the first crossing of half the positive endpoint SSR under an isotonic nondecreasing fit.

2.3 A SPECTRAL NULL AND A SHARP STABILITY BOUND

When the spectrum determines risk. Consider a length-T series with circular indexing. A fixed linear predictor has the form $\hat{x}_{t+h} = \sum_{\ell=0}^{L-1} a_{m,h,\ell} x_{t-\ell}$. Average its squared error over all T start times and H forecast steps. With $\omega_k = 2\pi k/T$ and $X_k = \sum_{t=0}^{T-1} x_t e^{-i\omega_k t}$, Parseval's identity gives

$$\mathcal{R}_m^{(2)}(x) = \frac{1}{HT^2} \sum_{h=1}^{H} \sum_{k=0}^{T-1} \left| e^{i\omega_k h} - \sum_{\ell=0}^{L-1} a_{m,h,\ell} e^{-i\omega_k \ell} \right|^2 |X_k|^2. \tag{7}$$

Only Fourier magnitudes enter the risk. Every magnitude-preserving phase surrogate has the same risk, so every upgrade between such predictors has $\mathcal{S}(q) = 0$. Applying the risk floor preserves this invariance. This supplies a concrete reference in which the retained spectrum determines risk. Our finite-start-time MAE evaluation measures how upgrade gains change outside that setting.

How stable is an SSR sign? Suppose every log risk changes by at most ε, for both checkpoints and $r \in \{0, q\}$, on the same evaluation units. Then

$$|\widetilde{\mathcal{G}}(r) - \mathcal{G}(r)| \leq 2\varepsilon, \qquad |\widetilde{\mathcal{S}}(q) - \mathcal{S}(q)| \leq 4\varepsilon. \tag{8}$$

The constant 4 is attainable under median aggregation. Thus $\varepsilon < |\mathcal{S}(q)|/4$ guarantees sign preservation. The four PEMS08 thresholds are 0.0714, 0.0524, 0.0282, and 0.0357 log-risk units. They measure sensitivity to a specified uniform perturbation; resampling assesses uncertainty under the observed evaluation design. Appendix A gives proofs and counterexamples that delineate the assumptions.

The relative-risk interpretation also survives aggregation: if $Q(q) = \exp[-\mathcal{G}(q)]$, then $\exp[\mathcal{S}(q)] = Q(0)/Q(q)$. Above the numerical floor, a model-independent multiplicative risk scale for each channel cancels inside every log-risk gain. Raw MAE and fixed-global normalization consequently yield the same SSR, whereas start-time-dependent normalization can change it.

2.4 INFERENCE

Positive SSR must satisfy three pre-specified conditions: positive one-sided 95% lower bounds for both $\mathcal{G}(1)$ and $\mathcal{S}(1)$, and Spearman correlation $\rho(q, \mathcal{G}(q)) \geq 0.5$. Negative SSR requires the one-sided 95% upper bound for $\mathcal{S}(1)$ to fall below zero. The 5th–95th bootstrap percentiles form a central 90% interval and also supply these directional bounds. Primary uncertainty is conditional on the fixed channel sets and released checkpoints; channel resampling is reported separately as sensitivity analysis.

3 SSR CROSSES SCALING AND SURROGATE ANALYSIS

Neural scaling laws vary parameters, data, or compute and observe checkpoint loss (Kaplan et al., 2020; Hoffmann et al., 2022). Time-series scaling studies extend this analysis across model size, data scale, architecture, and distribution shift (Yao et al., 2025; Shi et al., 2025; Khwaja et al., 2026). They establish whether larger models improve. SSR supplies the missing interaction between upgrade gain and temporal phase structure.

Surrogate analysis instead varies temporal structure. Truncated Fourier-transform surrogates restore low-frequency phases (Nakamura et al., 2006). SINE measures how the absolute value of context mechanisms or forecasters changes between observed and spectrum-controlled series (Cakiroglu et al., 2026). SSR uses this controlled axis to make the small-to-large upgrade gain the measured variable. It predicts that two upgrades can both improve the original forecast yet have opposite SSR signs on the same dataset (Table 1).

Table 1: **SSR crosses two experimental axes.** Scaling varies resources; surrogate analysis varies temporal structure. SSR measures how temporal phase structure changes an upgrade gain.

Work	Measured object	Controlled axis	Decisive observable
Scaling laws	Checkpoint loss	Model, data, and compute	Scaling slope
TFTS	Nonlinearity statistic	Phase cutoff	Surrogate rejection
SINE	Absolute context/forecaster value	Spectrum-matched endpoint	Beyond-spectrum value
SSR	**Checkpoint-upgrade gain**	**Cumulative phase bandwidth**	**Upgrade–dataset SSR curve**

4 Experimental Design

4.1 Released upgrades and mobility datasets

We evaluate seven released checkpoints through four upgrades: Toto 4M→22M, Toto 22M→313M, Chronos tiny→base, and Moirai small→large (Khwaja et al., 2026; Ansari et al., 2024; Woo et al., 2024). Each named upgrade is an empirical object available to a model user. Toto 4M→22M provides the cleanest capacity comparison. Both checkpoints share the published 400k-step, 3.40T-point training exposure, global batch 64, context 4096, and corpus proportions. Toto 22M→313M also changes published exposure. Chronos and Moirai extend the study to other public model families. Our claims therefore attach to released upgrades, with Toto 4M→22M supplying the matched-exposure capacity comparison.

The evaluation suite contains four hourly mobility datasets:

- **Traffic**: eight frozen channels and 16 forecast start times from the canonical Traffic matrix (Lai et al., 2018);
- **NYC Taxi**: 24 frozen demand zones and 24 forecast start times (New York City Taxi and Limousine Commission, 2025);
- **PEMS04**: 24 hash-selected sensors and 24 forecast start times, with twelve 5-minute flow observations aggregated by hourly mean (Guo et al., 2019; California Department of Transportation, 2026);
- **PEMS08**: a disjoint set of 24 sensors and 24 forecast start times, reserved as untouched confirmation data before the model results were inspected.

All checkpoints receive the same information: 512 observations of context, a 96-step horizon, and no fine-tuning. We evaluate zero-shot median forecasts with context-normalized MAE,

$$\ell(\hat{y}, y; x) = \frac{96^{-1} \sum_{h=1}^{96} |\hat{y}_h - y_h|}{\mathrm{std}(x)}, \tag{9}$$

with a mean-absolute fallback only for constant contexts. Reported channel-aggregated relative error reductions are $1 - \exp[-\mathcal{G}(1)]$; SSR remains on the additive log-risk scale.

4.2 Controlled inputs and matched surrogates

For each dataset, we evaluate the original series and nine phase-restoration levels with seven random draws each. We add seven iterative amplitude-adjusted Fourier transform (IAAFT) surrogates, seven multivariate Fourier-transform (MVFT) surrogates, and seven weekday-stratified day shuffles. This

gives 85 controlled inputs per dataset. The empirical base contains 16 upgrade–dataset pairs, 80 fixed channels, and 88 forecast start times. All surrogate draws are repeated interventions on these fixed units.

The controls retain complementary information. IAAFT preserves each channel's empirical marginal and approximately matches its global power spectrum (Schreiber & Schmitz, 2000). MVFT applies one frequency-wise phase rotation across channels. It preserves every auto-spectrum and the full complex cross-spectrum (Prichard & Theiler, 1994). The day shuffle preserves complete within-day trajectories, weekday class, and the contemporaneous cross-section while removing cross-week order. Agreement across these constructions ties SSR to temporal phase structure instead of a single surrogate generator.

The design separates four explanations. If an upgrade's gain depends only on spectrum-preserved information, its SSR should be near zero. If marginal distortion creates SSR, IAAFT should eliminate it. If disrupted cross-channel second-order structure creates SSR, MVFT should eliminate it. If within-day phase structure contributes, shuffling whole days while keeping their interiors intact should retain a substantial fraction of SSR. Untouched PEMS08 tests the four sign predictions fixed from the development datasets.

We use 1,000 bootstrap repetitions with non-circular blocks of two consecutive forecast start times. Each repetition draws one common surrogate at every restoration level. This preserves channel comovement and short-range temporal dependence. We report all 12 development pairs and evaluate PEMS08 once as the four-pair confirmation dataset. The intervals quantify surrogate and start-time uncertainty conditional on the frozen channel sets. Training-run and sensor-population variation lie outside this estimand.

Finally, we test the PEMS08 signs across loss and sampling choices. Saved predictions yield raw MAE and two losses with denominators fixed across original and surrogate inputs. We repeat the endpoint analysis with mean aggregation and blocks of 4 and 8 forecast start times. We also remove each of the 24 channels in turn and jointly bootstrap channels and forecast start times. Context-normalized MAE remains the pre-specified primary loss.

5 Results

5.1 One leaderboard, opposite SSR signs

Untouched PEMS08 provides the decisive contrast (Table 2). Every larger checkpoint improves nMAE on the original series: 38.54%, 25.56%, 14.73%, and 10.02% across the four upgrades. A conventional leaderboard places all four on the same continuum of successful scaling. SSR separates those successes by how they depend on temporal phase structure.

For Toto 4M→22M, the log-risk gain rises from $\mathcal{G}(0) = 0.201$ on phase-randomized data to $\mathcal{G}(1) = 0.487$ on the original series. Its endpoint SSR is 0.286, with a one-sided lower bound of 0.254 and a monotone SSR curve. Toto 22M→313M and Chronos gain an additional 0.209 and 0.113 as original phases return. Moirai moves in the opposite direction. Its gain falls from 0.248 to 0.106, giving an endpoint SSR of -0.143 with a one-sided upper bound of -0.070. Phase randomization strengthens the Moirai upgrade gain even though the larger checkpoint remains more accurate on the original data. Temporal phase structure therefore moves four successful upgrades in opposite directions.

Table 2: **Untouched PEMS08.** Gain is channel-aggregated nMAE reduction on original data. Endpoint SSR compares original and randomized log-risk gains. Brackets give central 90% intervals; P_{50} is the half-recovery period.

Released upgrade	Gain	Endpoint SSR	90% interval	ρ	P_{50}
Toto 4M→22M	38.54%	0.286	[0.254, 0.342]	1.000	14.1h
Toto 22M→313M	25.56%	0.209	[0.168, 0.260]	0.964	9.8h
Chronos tiny→base	14.73%	0.113	[0.087, 0.159]	0.952	15.0h
Moirai small→large	10.02%	−0.143	[−0.192, −0.070]	−0.576	negative

5.2 THE SSR SIGNS SURVIVE MATCHED SURROGATES AND RESAMPLING

The phase-restoration path fixes each channel's power spectrum while changing its marginal distribution and cross-channel structure. Two matched surrogates restore these features separately. Across all 16 upgrade–dataset pairs, random-phase endpoint SSR correlates at $r = 0.994$ with IAAFT and $r = 0.998$ with MVFT. Mean absolute discrepancies are 0.0116 and 0.0060 log risk. Signs agree in 15/16 and 16/16 pairs. On PEMS08, Moirai remains negative under all three constructions (-0.143, -0.148, and -0.152); both Toto upgrades and Chronos remain positive.

Analysis choices preserve the result. Raw MAE and both fixed-denominator losses reproduce the primary endpoint SSR values within 0.003 log risk. Removing any one of the 24 channels preserves all four signs. A joint channel–time bootstrap gives central 90% intervals of $[0.200, 0.385]$, $[0.119, 0.306]$, $[0.081, 0.200]$, and $[-0.205, -0.046]$. Block sizes of 2, 4, and 8 consecutive start times preserve the three-positive, one-negative pattern. Mean aggregation does as well.

Figures 5 and 6 in Appendix C show the complete sensitivity comparison. Raw MAE and fixed-global normalization coincide by scale invariance; context-dependent normalization provides the distinct weighting test.

Absolute risk summaries align with this contrast. Every checkpoint has lower median raw MAE on the original series. Original-to-randomized risk ratios are .805, .604, and .469 for Toto 4M, 22M, and 313M; .584 and .513 for Chronos tiny and base; and .476 and .538 for Moirai small and large. These checkpoint-level summaries show a larger relative reduction for the larger Toto and Chronos checkpoints, but for the smaller Moirai checkpoint.

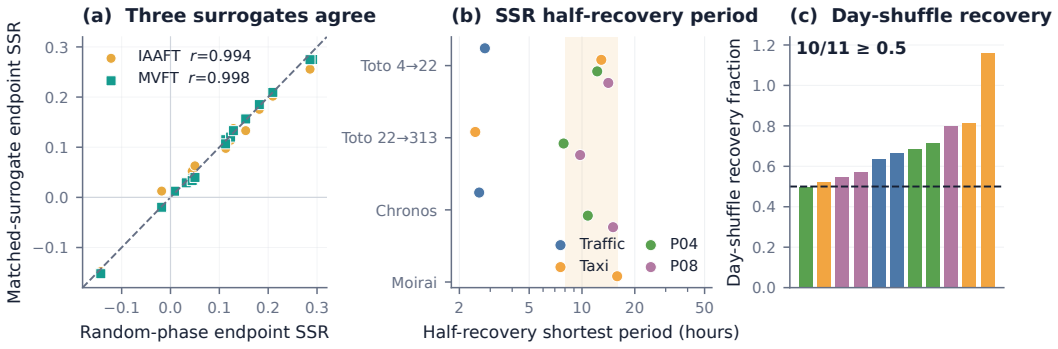


Figure 3: **Independent controls preserve SSR.** (a) Matched surrogates agree. (b) Positive PEMS curves reach half recovery at 8–15h. (c) Day shuffling retains at least half in 10/11 positive pairs.

5.3 SSR IS SPECIFIC TO AN UPGRADE–DATASET PAIR

Figure 4 expands the PEMS08 finding to all 16 SSR curves. Eleven pairs satisfy the positive-SSR criteria, yet the signs vary across both upgrades and datasets. Toto 4M→22M is positive on every dataset. Toto 22M→313M is positive on Taxi, PEMS04, and PEMS08. On Traffic, its gain appears mainly after $q > 0.8$ and misses the monotonicity criterion. Chronos is positive on Traffic and both PEMS datasets; Taxi offers little original-series headroom. Moirai is positive on Taxi, unresolved on PEMS04, trends negative on Traffic, and is decisively negative on PEMS08.

SSR adds information beyond the original-series gain. On Traffic, Toto 22M→313M and Moirai small→large improve by a similar 14.18% and 11.97%. Their correlations between restored phase bandwidth and upgrade gain are $+0.455$ and -0.758. On PEMS08, Chronos and Moirai again have comparable gains (14.73% versus 10.02%) but opposite endpoint SSR values ($+0.113$ versus -0.143). The SSR curve separates upgrades that a leaderboard places together.

5.4 POSITIVE PEMS SSR EMERGES AT AN 8–15 HOUR PHASE BANDWIDTH

The SSR curve locates where original and phase-randomized upgrade gains diverge. On both PEMS datasets, the positive Toto and Chronos curves reach half of endpoint SSR after phases at periods

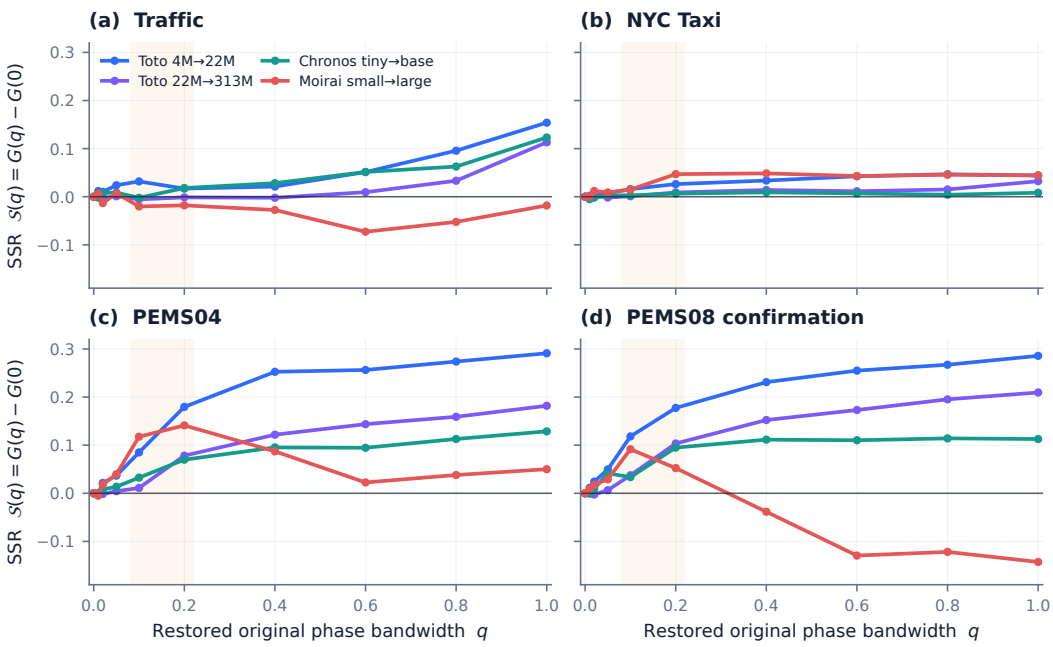


Figure 4: **SSR is upgrade–dataset specific.** Each plot traces four upgrades along the same restoration path. The gold band marks shortest restored periods of roughly 10–25h.

down to roughly 8–15 hours have returned. Toto 4M→22M crosses at 12.2h on PEMS04 and 14.1h on PEMS08. Toto 22M→313M crosses at 7.9h and 9.8h; Chronos crosses at 10.8h and 15.0h. Untouched PEMS08 reproduces both the SSR signs and the crossing region.

The day shuffle supplies a non-Fourier control. It preserves complete within-day trajectories, weekday type, and the contemporaneous cross-section while removing cross-week order. This control retains at least half of endpoint SSR in 10 of 11 positive pairs. Together, the two interventions place much of the positive SSR in within-day and weekday-aligned temporal phase structure. The 8–15 hour crossing localizes this dependence along a cumulative restoration path; it is not an additive band attribution.

5.5 An external audit suggests a mobility boundary

A separate endpoint audit compares original and low-frequency-preserving surrogate inputs using pooled-risk contrasts. Positive one-sided lower bounds appear in four of six mobility pairs and in none of six Electricity or Weather pairs (Appendix D). Its distinct surrogate and aggregation provide an exploratory domain contrast alongside the full mobility SSR curves.

6 Discussion

Upgrade gain is dataset-conditional. The original-series gain shows whether a released upgrade improves one dataset. SSR shows how temporal phase structure changes that gain. PEMS08 makes the distinction visible: all four upgrades improve accuracy, but their SSR signs are $+, +, +, -$. Across the complete map, one upgrade changes across datasets and one dataset contains multiple signs. The practical unit of scaling analysis is therefore an upgrade–dataset pair.

Matched surrogates isolate temporal phase structure. Random-phase, IAAFT, and MVFT surrogates preserve different combinations of marginal, spectral, and cross-channel information. The day shuffle preserves natural within-day trajectories without Fourier synthesis. All four PEMS08 signs remain stable across these controls, alternative losses, and channel–time resampling. The

spectral null identifies a setting where the preserved spectrum fully determines risk; the observed SSRs quantify departures from that reference under the evaluated forecast loss.

Implications for scaling evaluation. SSR adds two coordinates to a scaling curve: the upgrade gain that survives phase randomization and the restored phase bandwidth at which the remaining difference emerges. These coordinates separate upgrades with similar average accuracy and generate testable hypotheses about architectures and training regimes. Scaling evaluations should therefore report the deployment structure that supports an upgrade gain.

Supported scope. SSR identifies behavioral dependence for named released upgrades. Toto 4M→22M isolates capacity most closely under matched published exposure; the other upgrades establish the phenomenon across public model families. The full evidence supports zero-shot hourly mobility forecasting. A separate endpoint audit finds no resolved positive contrast on Electricity or Weather. The next scientific axes are internal mechanism, sampling frequency, probabilistic scores, restoration order, and independently trained checkpoints.

7 Conclusion

Scaling gains have shape. SSR measures that shape by comparing an upgrade gain on original and phase-randomized data, then tracing it as phases return. On untouched PEMS08, four accuracy improvements split into three positive SSRs and one negative SSR; matched surrogates and sensitivity analyses preserve every sign. Across 16 pairs, SSR is upgrade–dataset specific, and positive PEMS SSR concentrates at an 8–15 hour restoration scale. Accuracy reports whether an upgrade works; SSR reveals the temporal phase structure that supports its gain.

Reproducibility statement. Appendices A and B provide proofs, assumptions, and the complete evaluation recipe. Appendix C specifies every loss and resampling variant. The accompanying numerical supplement contains the exact curve points, uncertainty summaries, and plotting inputs.

REFERENCES

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

Mert Onur Cakiroglu, Mehmet Dalkilic, and Hasan Kurban. The spectrum is not enough: When context helps time-series forecasting. *arXiv preprint arXiv:2607.13006*, 2026.

California Department of Transportation. Performance measurement system. `https://pems.dot.ca.gov/`, 2026.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 922–929, 2019.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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:2605.20119*, 2026.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Han Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–104, 2018.

Tomomichi Nakamura, Michael Small, and Yoshito Hirata. Testing for nonlinearity in irregular fluctuations with long-term trends. *Physical Review E*, 74(2):026205, 2006. doi: 10.1103/PhysRevE.74.026205.

New York City Taxi and Limousine Commission. NYC Taxi & Limousine Commission trip record data. `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`, 2025.

Dean Prichard and James Theiler. Generating surrogate data for time series with several simultaneously measured variables. *Physical Review Letters*, 73(7):951–954, 1994. doi: 10.1103/PhysRevLett.73.951.

Thomas Schreiber and Andreas Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142(3–4):346–382, 2000. doi: 10.1016/S0167-2789(00)00043-9.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, 2025.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 53140–53164. PMLR, 2024.

Qingren Yao, Chao-Han Huck Yang, Renhe Jiang, Yuxuan Liang, Ming Jin, and Shirui Pan. Towards neural scaling laws for time series foundation models. In *International Conference on Learning Representations*, 2025.

A Theoretical Results and Proofs

A.1 Spectral zero-response benchmark

For a fixed horizon, let $e_t = x_{t+h} - \sum_\ell a_{m,h,\ell} x_{t-\ell}$. Under circular indexing and the DFT convention in Section 2, time shifts give

$$E_k = \left(e^{i\omega_k h} - \sum_\ell a_{m,h,\ell} e^{-i\omega_k \ell} \right) X_k.$$

Parseval's identity states $\sum_t |e_t|^2 = T^{-1} \sum_k |E_k|^2$. Dividing by T and averaging over the H horizons proves Equation 7. Preserving $|X_k|$ preserves each checkpoint risk, including after the numerical risk floor. The result holds separately for every channel and hence after median aggregation. Fixed positive per-channel normalization also preserves it.

Fixed linear coefficients, squared loss, and complete circular evaluation are sufficient conditions for this null. The loss assumption has a direct counterexample. Let $T = 64$ and

$$x_t = \cos(6\pi t/64) + 0.7\cos(14\pi t/64 + 0.2) + 0.5\cos(22\pi t/64 + 0.6).$$

Compare the one-step predictors $f_s = 0$ and $f_l = 0.8x_t$. Draw the nontrivial rFFT phases independently from $\mathrm{Uniform}[-\pi, \pi]$, using NumPy's default generator with seed 123, and preserve the DC and Nyquist phases. Full-origin squared risk gives zero SSR. Under MAE, the same construction gives SSR -0.01471. Thus predictor linearity alone does not imply zero SSR under the loss used in our experiments.

A.2 A sharp bound under median aggregation

If $\|a - \widetilde{a}\|_\infty \le \delta$, then $a - \delta\mathbf{1} \le \widetilde{a} \le a + \delta\mathbf{1}$ coordinatewise. Monotonicity and translation equivariance of the median give

$$|\operatorname{median}(\widetilde{a}) - \operatorname{median}(a)| \le \delta.$$

This holds for odd sample sizes and for the average of the middle two order statistics when the sample size is even. A perturbation of at most ε to each log risk changes each paired log-risk gain by at most 2ε. The median at each endpoint therefore changes by at most 2ε. The triangle inequality yields

$$|\widetilde{\mathcal{S}}(q) - \mathcal{S}(q)| \le |\widetilde{\mathcal{G}}(q) - \mathcal{G}(q)| + |\widetilde{\mathcal{G}}(0) - \mathcal{G}(0)| \le 4\varepsilon.$$

The two endpoints may have different sample sizes; each endpoint must retain its own evaluation units under perturbation. If the four log-risk terms have different uniform error bounds, their sum replaces 4ε.

To attain the bound, start with every risk equal to e^{ε}. At q, multiply small-checkpoint risks by e^{ε} and large-checkpoint risks by $e^{-\varepsilon}$. At 0, reverse the multipliers. All risks remain above the floor. The gains change by $+2\varepsilon$ and -2ε, so SSR changes by exactly 4ε. Consequently $\varepsilon < |\mathcal{S}(q)|/4$ preserves a nonzero sign; equality can produce zero. This is a deterministic guarantee for the empirical risks under the stated uniform perturbation, separate from the bootstrap uncertainty intervals.

A.3 Non-identifiability, units, and checkpoint chains

Fix an observed original-series gain g. For any $s \in \mathbb{R}$, setting $\mathcal{G}(0) = g - s$ and $\mathcal{S}(1) = s$ yields the same original gain. Thus g alone leaves the sign and magnitude of SSR unidentified. The surrogate comparison measures the additional coordinate.

For a fixed channel and matched input, log-risk gains telescope across checkpoints. Setting the two checkpoints equal gives zero; reversing them negates the gain and SSR. Multiplying both checkpoint risks by a common positive scale adds the same log constant and therefore cancels. Zero response, antisymmetry, and common-scale invariance survive median aggregation. When scaling losses before applying the numerical floor, the scale invariance holds provided neither version activates the floor.

Checkpoint-chain additivity requires more care. Let two adjacent upgrades have channel gains $x = (0, 0, 10)$ and $y = (0, 10, 0)$. Their medians are zero, but $\mathrm{median}(x + y) = 10$. These gains arise from positive risks with log risks $L_k = 0$, $L_j = y$, and $L_i = x + y$. Taking all randomized-input risks equal gives the same counterexample for SSR. Additivity holds pointwise or under a common linear averaging operator, not in general after separately taking medians.

Likewise, $\mathcal{G}$ aggregates log-risk ratios rather than pooled risks. Small-checkpoint risks $(1, 4, 9)$ and large-checkpoint risks $(1, 2, 100)$ give median log gain zero, while the log-ratio of their median risks is $\log 2$. The exact aggregate interpretation is $Q(q) = \exp[-\mathcal{G}(q)]$ and $\exp[\mathcal{S}(q)] = Q(0)/Q(q)$.

A.4 Upgrade-gain ranking boundaries

For two upgrades, write $D_r = \mathcal{G}_u(r) - \mathcal{G}_v(r)$. Then

$$D_1 = D_0 + \mathcal{S}_u(1) - \mathcal{S}_v(1), \qquad \text{strict rank reversal} \iff D_0 D_1 < 0.$$

These are relative gains over each upgrade's own smaller checkpoint, rather than absolute accuracy across families. The log-risk perturbation assumption gives $|\widetilde{D}_r - D_r| \le 4\varepsilon$. Both endpoint rankings are preserved if $\varepsilon < \min(|D_0|, |D_1|)/4$.

PEMS08 provides an exploratory illustration. The Chronos-minus-Moirai point difference changes from -0.2018 on randomized inputs to $+0.0537$ on original data. A post-hoc paired bootstrap of saved losses shares origin blocks and surrogate draws across all models and keeps channels fixed. With 10,000 replicates and the primary block length of two, the original-data difference has a central 90% interval of $[-0.0082, 0.1232]$. Block lengths four and eight give $[0.0351, 0.0889]$ and $[0.0512, 0.0837]$. This comparison illustrates the ranking identity; the confirmed finding remains the four-way SSR sign pattern.

B Evaluation Protocol

B.1 Data, checkpoints, and controlled inputs

Table 3: Frozen evaluation datasets. PEMS04 and PEMS08 aggregate twelve 5-minute flow observations into each hourly mean.

Dataset	Length	Channels	Start times	Role
Traffic	17,544	8	16	development
NYC Taxi	1,464	24	24	development
PEMS04	1,416	24	24	development
PEMS08	1,488	24	24	confirmation

The model pool is Toto 2.0 4M, 22M, and 313M; Chronos tiny and base; and Moirai small and large. Every call uses context 512, horizon 96, and a median forecast without fine-tuning. Moirai uses 20 predictive samples to estimate its median. The matched Toto 4M and 22M training exposure is described in the main text.

The original input has index zero. Each $q < 1$ has seven inputs with deterministic dataset-level seeds and independent per-channel phase draws. The DC and Nyquist coefficients remain unchanged; conjugate symmetry preserves real-valued output. Amplitudes remain fixed, so both the periodogram and circular autocovariance are preserved. The original series, rather than a reconstructed surrogate, supplies $q = 1$.

IAAFT uses 120 iterations and a fixed global spectrum-residual ceiling of 0.08. Multivariate Fourier surrogates apply one phase rotation per frequency across all channels, preserving the full complex cross-spectrum. Weekday-stratified day shuffles divide each hourly dataset into complete days, permute only within weekday class, and retain each 24-hour multivariate block intact.

B.2 Aggregation and resampling recipe

1. Fix dataset, channel indices, forecast start times, context length, horizon, checkpoint pair, loss, restoration grid, and seeds. PEMS08 upgrade-level sign tests were fixed before inspecting its checkpoint results.
2. Construct the controlled inputs, then extract identical context and target indices for every checkpoint. Model-visible information and forecast horizons are matched.
3. Store the loss tensor with axes input, channel, and forecast start time. Average over start times within each input and channel, apply the risk floor, take the paired log-risk difference, and finally take the median over channels and same-level surrogate inputs.
4. Subtract the randomized-input gain to obtain the SSR curve. Compute endpoint SSR, Spearman direction, and the isotonic half-recovery crossing on the complete grid.
5. Draw non-circular blocks of forecast start times with replacement. At every synthetic q, select one surrogate shared across channels and both checkpoints. Recompute all summaries for 1,000 replicates. The primary block length is two; channels remain fixed.
6. Report the 5th and 95th percentiles as a central 90% interval. Apply the pre-specified sign rules from the main text. Keep unresolved combinations in the map. Evaluate IAAFT, multivariate Fourier, and day-shuffle controls separately.

Forecast start times are repeated decisions along a stream, handled with blocks. Channels form a fixed evaluation set in the primary analysis. Surrogates are repeated transformations of the same series, rather than additional datasets or training realizations. Trajectory counts describe computational scale, while inference uses the declared channels, start times, and surrogate draws.

C Sensitivity to Loss, Sampling, and Aggregation

C.1 Loss definitions

The analysis reuses saved losses and inputs. If $d_{w,c,t}$ is the context scale used in primary nMAE, the raw error is $\mathrm{MAE}_{w,c,t} = \ell_{w,c,t} d_{w,c,t}$ up to saved precision. Fixed-context normalization divides it by the original-input scale $d_{0,c,t}$. Fixed-global normalization divides it by the original dataset's channel standard deviation. No forecasts are rerun.

Figure 5 shows all 64 endpoint values and the PEMS08 curve differences. Raw and fixed-global losses share every channel-wise log-risk gain because their constant scale cancels; their overlapping curves check this invariance. Input-dependent and fixed-context normalization instead reweight errors across forecast start times. Across all four definitions, the PEMS08 three-positive, one-negative pattern persists.

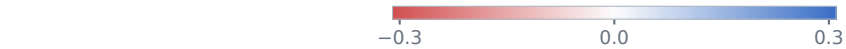
C.2 Sampling and aggregation choices

Time-only analyses repeat the original 1,000-replicate procedure with blocks of two, four, and eight consecutive start times. Leave-one-channel-out analysis removes each of the 24 PEMS08 channels once. The joint bootstrap resamples 24 channels with replacement, resamples non-circular blocks of four start times, and selects one phase-randomized surrogate for 2,000 replicates. Mean aggregation replaces the channel-and-surrogate median of log gains by its arithmetic mean. These are post-hoc sensitivity analyses of the fixed dataset.

Figure 6 shows all 96 metric–analysis combinations. Bootstrap intervals and leave-one-channel-out ranges retain all four signs, as does mean aggregation. The numerical supplement also supplies the per-checkpoint absolute risks and all primary inference summaries.

(a) Endpoint SSR across all four loss definitions

	Context nMAE	Raw MAE	Fixed context	Fixed global
Toto 4→22 · Traffic	0.154	0.154	0.155	0.154
Toto 4→22 · Taxi	0.044	0.044	0.044	0.044
Toto 4→22 · PEMS04	0.291	0.291	0.292	0.291
Toto 4→22 · PEMS08	0.286	0.286	0.286	0.286
Toto 22→313 · Traffic	0.113	0.120	0.113	0.120
Toto 22→313 · Taxi	0.032	0.031	0.032	0.031
Toto 22→313 · PEMS04	0.182	0.180	0.182	0.180
Toto 22→313 · PEMS08	0.209	0.212	0.209	0.212
Chronos · Traffic	0.123	0.117	0.123	0.117
Chronos · Taxi	0.008	0.008	0.008	0.008
Chronos · PEMS04	0.129	0.129	0.129	0.129
Chronos · PEMS08	0.113	0.113	0.113	0.113
Moirai · Traffic	-0.018	-0.027	-0.019	-0.027
Moirai · Taxi	0.044	0.042	0.043	0.042
Moirai · PEMS04	0.050	0.050	0.050	0.050
Moirai · PEMS08	-0.143	-0.142	-0.142	-0.142



(b) PEMS08 curve change from primary nMAE

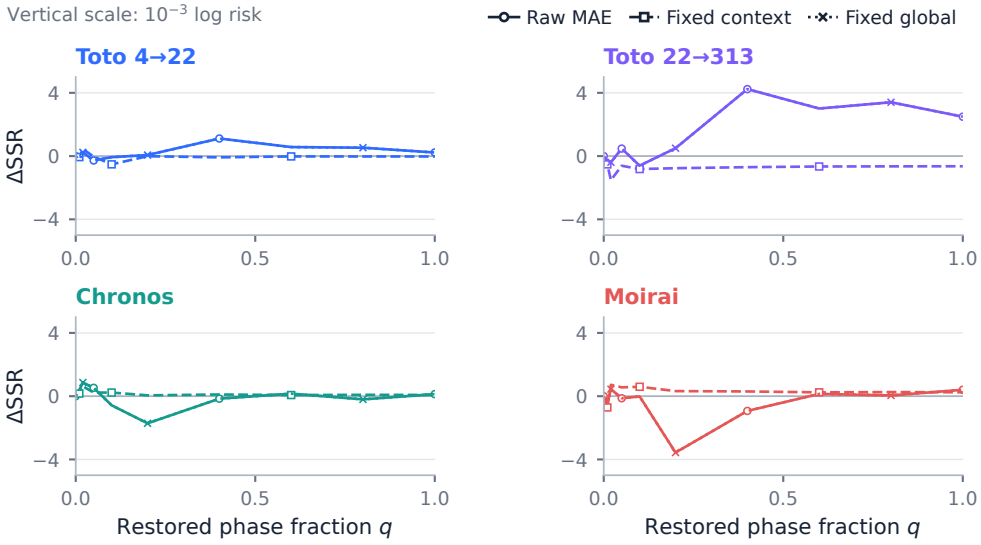


Figure 5: **Loss sensitivity across 64 endpoints.** Lower plots show PEMS08 changes from primary nMAE (10^{-3} log-risk units). Raw and fixed-global normalization coincide.

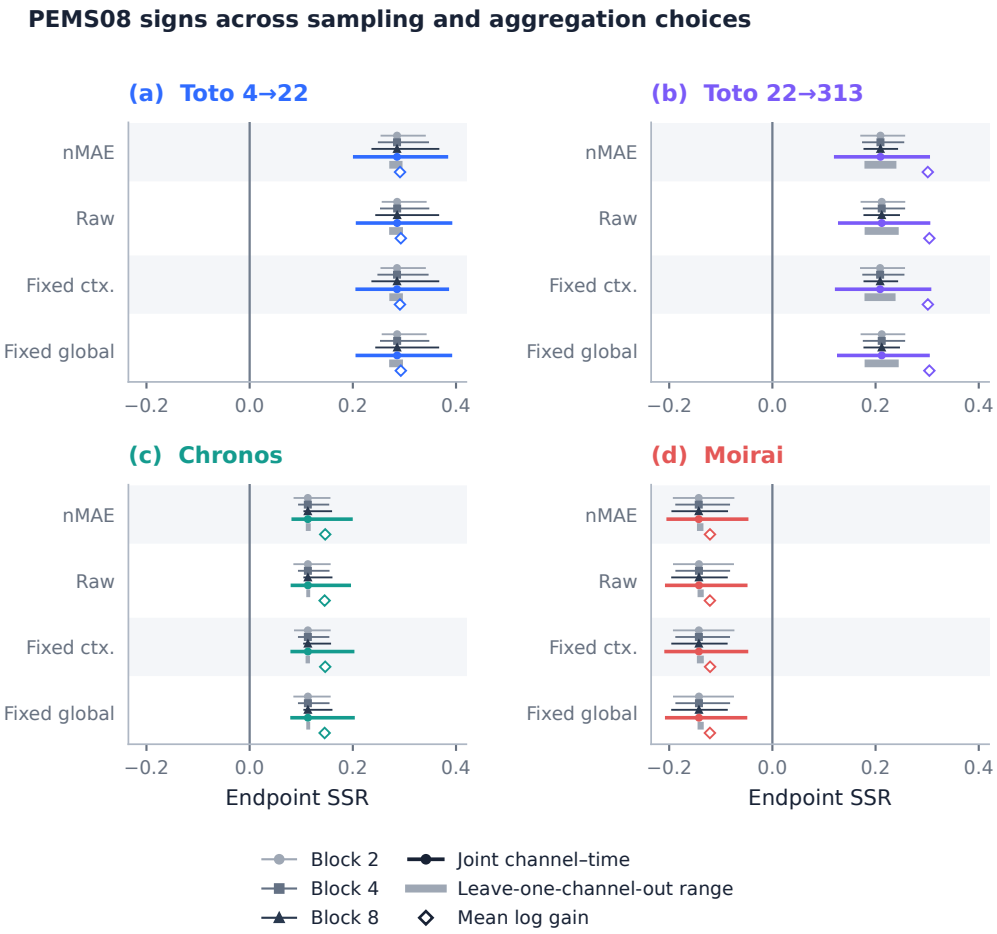


Figure 6: **PEMS08 signs persist across specifications.** Bars: central 90% bootstrap intervals. Thick gray segments: leave-one-channel-out ranges. Diamonds: mean log gains. All four loss definitions are shown.

D Exploratory Domain Contrast

A separate development audit used low-frequency-phase-preserving surrogates on Traffic, NYC Taxi, Electricity, and Weather. Its upgrades are Toto 4M→313M, Chronos tiny→base, and Moirai small→large. The audit uses a distinct endpoint construction and aggregates errors before taking logs. For each resampled original or surrogate input, let $\mathcal{R}_m^{\text{pool}}$ be the mean loss over sampled channels and start times. Its contrast is

$$\left[\log \mathcal{R}_s^{\text{pool}}(\text{original}) - \log \mathcal{R}_l^{\text{pool}}(\text{original})\right] - \left[\log \mathcal{R}_s^{\text{pool}}(\text{surrogate}) - \log \mathcal{R}_l^{\text{pool}}(\text{surrogate})\right].$$

Each of 2,000 bootstrap replicates samples channels and individual start times with replacement and selects one surrogate. Figure 7 reports the resulting distribution's median and central 90% interval. Positive one-sided lower bounds occur in four of six mobility comparisons and none of six Electricity or Weather comparisons. This pooled-risk audit provides an exploratory domain contrast; the primary SSR-curve analysis uses the median of channel-wise log gains.

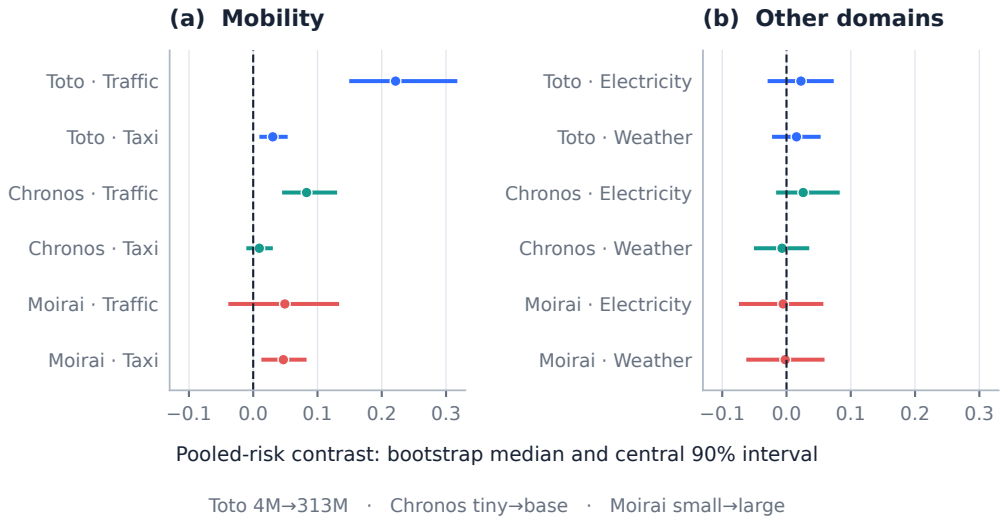


Figure 7: **A separate low-frequency-preserving audit.** Points and bars show bootstrap medians and central 90% intervals of pooled-risk contrasts, a different aggregation from primary SSR.

E Exploratory Downstream Model Choice

A separate analysis uses saved Traffic and NYC Taxi predictions from the same seven checkpoints. At start time t, an earlier outcome enters model choice only after its complete 96-step horizon has been observed. A top-2 policy averages the two checkpoints with the best matured channel history. A cost-aware policy admits checkpoints within 2% of the best matured score and selects the smallest. These policies use prediction history, rather than an estimated SSR, and are evaluated separately from the structural claims.

Table 4: Exploratory seven-checkpoint model choice. nMAE weights Traffic and Taxi equally; parameter-calls sum the sizes of invoked checkpoints.

Policy	nMAE ↓	Error vs. best fixed	Parameter-calls
Best fixed checkpoint	0.29052	reference	205.29M
Cost-aware choice	0.28624	−1.47%	187.81M (−8.52%)
Top-2 ensemble	**0.28141**	**−3.13%**	381.69M
Uniform seven-model ensemble	0.29235	+0.63%	877.49M

F Numerical Supplement and Compute

The accompanying numerical supplement contains all 640 metric-specific curve points, the 16 primary endpoint and shape summaries, matched-surrogate controls, 96 PEMS08 sensitivity records, absolute-risk summaries, and the separate domain audit. These files retain exact values for lookup; the manuscript figures organize their comparisons visually. Plotting inputs preserve unresolved and negative results alongside positive ones.

The structural experiment produced 1,104,320 forecast trajectories and 106,014,720 horizon values. Input construction took 27.65 CPU seconds. Inference took 3,484.17 seconds (58.07 minutes) on one NVIDIA RTX A6000, with mean GPU utilization 94.10% and peak observed memory 13,681MiB. The evaluation is inference-only. Per-checkpoint timings and the exact evaluation configuration are included in the numerical supplement.

G Broader Impact

This work uses aggregate, non-biological public time series and released forecasting checkpoints. It supports more informative allocation of inference capacity across deployment streams. Any operational deployment remains responsible for its own reliability and latency requirements.

AI Use Statement

We used generative AI tools to clean and reformat the dataset. We did not use them for any other research tasks requiring disclosure (e.g., ideation, experimental design, theory, method implementation, or result interpretation). Additionally, we used AI tools to help debug code and to polish the writing. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for this work.