Do Time Series Foundation Model Benchmarks Hide Regime-Dependent Failures? Evidence from Traffic Speed Forecasting

Yingshuo Wang [1] **Xian Sun** [2] **Lingdong Kong** [3] **Wei Gao** [4] **Yanhang Li** [4] **Zhichao Fan** [5] **Zexin Zhuang** [6]

Abstract

Standard benchmarks evaluate time series foundation models (TSFMs) using aggregate metrics, but these can mask severe failures in critical operating regimes. We introduce *regime-stratified evaluation* and apply it to three TSFMs on two standard traffic speed benchmarks. Traffic exhibits abrupt regime switching between free-flow and congested states, producing bimodal speed distributions during transitions. When we stratify by traffic regime, both accuracy and prediction-interval coverage degrade sharply during transitions: transition-regime MAE reaches 11 mph (versus 3 mph overall), and empirical coverage of 90% prediction intervals drops as low as 55%. These failures are invisible in aggregate metrics because free-flow observations dominate the sample. A simple historical conditional baseline (sampling from per-sensor training distributions) achieves better transition coverage than any TSFM, but has far worse overall accuracy. We propose *bimodal mixture augmentation* (BMA), a post-hoc method that combines TSFM forecasts with historical distributional knowledge, approaching the historical baseline's transition coverage while preserving the TSFM's accuracy. Our results suggest that TSFM benchmarks should incorporate regime-aware evaluation to surface failures that aggregate metrics hide.

[1]University of California, Berkeley, CA, USA [2]Duke University, Durham, NC, USA [3]National University of Singapore, Singapore [4]Northeastern University, Boston, MA, USA [5]University of Illinois Urbana-Champaign, IL, USA [6]Southern Methodist University, Dallas, TX, USA. Correspondence to: Yingshuo Wang <yingshuow@berkeley.edu>.

Accepted at the Workshop on Forecasting as a New Frontier of Intelligence, ICML 2026, Seoul, South Korea.

1. Introduction

Traffic speed forecasting matters most during transitions between free-flow and congestion. In stable conditions, real-time data suffices; during onset or dissipation of congestion, traffic can break down or recover, and systems that consume future speeds (estimated times of arrival (ETAs), freight windows, emergency dispatch) need forecasts with reliable uncertainty. This is where forecasting earns its value.

Time series foundation models (TSFMs) are positioned as general-purpose probabilistic forecasters (Ansari et al., 2024; Woo et al., 2024). TSFM-Bench (Li et al., 2025) and GIFT-Eval (Aksu et al., 2025) include METR-LA (207 loop detectors, Los Angeles) and PEMS-BAY (325 detectors, San Francisco Bay Area) (Li et al., 2018), the standard benchmarks used by DCRNN (Li et al., 2018), STGCN (Yu et al., 2018), and subsequent graph neural network models. These leaderboards stratify by domain and frequency but not by operating regime within a domain. A model can appear accurate on average while failing in the regimes that matter most. Adler et al. (2025) studied TSFM prediction-interval coverage on six general datasets and found foundation models "consistently better calibrated," but did not include high-frequency traffic data. Traffic at 5-minute resolution exhibits *abrupt regime switching* between free-flow ($\sim$65 mph) and congested ($\sim$10–20 mph) states (Greenshields, 1935), driven by sharp capacity thresholds. This is a property of traffic physics, not of any particular sensor. During transitions, the true future speed distribution is bimodal (Figure 1): speed will either remain high or drop sharply. TSFMs produce unimodal intervals centered between the modes ($\sim$30–45 mph), a range that is transient and rarely sustained.

Related work. Conformal methods have been extended to non-stationary series: Gibbs & Candès (2021) introduced adaptive conformal inference (ACI); recent work addresses regime switching (Lu et al., 2025) and correlated series (Cini et al., 2025). All adjust interval *width* but cannot correct distributional *shape*. On the traffic side, Wu et al. (2023) and Zheng et al. (2025) proposed uncertainty quantification for deep traffic models, but train domain-specific architectures

from scratch and do not evaluate zero-shot TSFMs. Widening a unimodal interval centered on 40 mph still misses outcomes near 15 or 65 mph; a different approach is needed.

We present the first *regime-stratified* evaluation of TSFMs on standard traffic speed benchmarks. Our contributions are: **(1)** We show that aggregate metrics mask severe regime-dependent failures in *both* accuracy and prediction-interval coverage during traffic regime transitions. **(2)** We show that a simple historical conditional baseline achieves better transition coverage than any TSFM but worse overall accuracy, revealing complementary strengths. **(3)** We propose *bimodal mixture augmentation* (BMA), which combines TSFM forecasts with historical distributional knowledge, improving transition coverage by 3–19 pp without retraining.

2. Experimental Setup

Datasets. We use METR-LA (207 sensors, Los Angeles freeways, Mar–Jun 2012) and PEMS-BAY (325 sensors, San Francisco Bay Area, Jan–May 2017) (Li et al., 2018), the standard benchmarks for spatio-temporal traffic forecasting. Both record speed at 5-minute intervals. We use the standard 70/10/20 train/validation/test split.

Sensor selection. To focus on the regimes where coverage failures matter most, we rank sensors by their training-set congestion frequency (fraction of readings below 25 mph) and form a pool of the top 50 per dataset. Each random seed then subsamples 30 sensors from this pool, introducing variability across runs. This uses only training data, avoiding test-set leakage.

Models. We evaluate three TSFMs in zero-shot mode: **Chronos-T5-Base** (Ansari et al., 2024), a T5-based encoder-decoder that tokenizes time series values and generates 100 distributional samples; **Chronos-Bolt-Small** (Ansari et al., 2024), an efficient quantile-based variant from which we draw pseudo-samples via interpolation; and **Moirai-1.1-R-Base** (Woo et al., 2024), which uses any-variate attention with mixture outputs and generates 100 samples. Baselines include **ACI-LR** (linear regression wrapped with adaptive conformal inference (Gibbs & Candès, 2021)) and a **historical conditional** baseline that draws 100 samples directly from the per-sensor empirical distribution $P(\text{speed}_{t+h} \mid \text{speed}_t)$ computed from training data, with no forecasting model involved. The historical conditional tests whether BMA's TSFM anchor adds value beyond the historical lookup it uses.

Evaluation protocol. From the pool of 50 congestion-prone sensors, we draw three different random subsets of 30 sensors each (using seeds 42, 43, 44) so that means and standard deviations reflect variability across sensor selections. For each subset, we evaluate on 50 test windows at horizons $h \in \{3, 6, 12\}$ steps (15, 30, 60 minutes ahead), with 14 hours of context for foundation models. We report *mean absolute error* (MAE) and *empirical coverage*: the fraction of true observations that fall within the model's 90% prediction interval. Differences between coverage rates are reported in percentage points (pp).

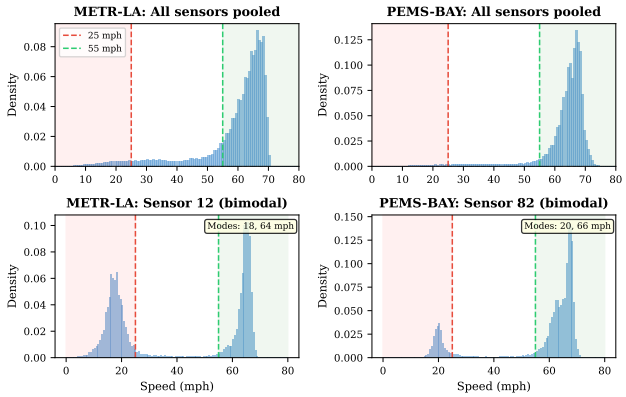

Figure 1. Speed distributions from training data. **Top:** pooled across all sensors, the distribution appears unimodal near 60–65 mph. **Bottom:** individual congestion-prone sensors reveal strong bimodality with modes near ∼18 mph and ∼64–66 mph. Red and green lines mark the regime thresholds (25 and 55 mph); shaded regions indicate the congested (red) and free-flow (green) regimes.

Regime detection. We classify each (window, sensor) target period using Highway Capacity Manual (Transportation Research Board, 2022) thresholds: *free-flow* (all steps >55 mph, LOS A/B), *congested* (all steps <25 mph, LOS E/F), or *transition* (mixed/intermediate, LOS C/D). These thresholds align with the bimodal speed distributions in Figure 1: individual congestion-prone sensors show modes near ∼18 and ∼65 mph with a valley near 30–40 mph.

Post-hoc correction methods. We compare four post-hoc correction strategies, all applied to the cached TSFM samples without retraining:

Global ACI (G-ACI): Adaptive conformal inference (Gibbs & Candès, 2021) tracks a miscoverage rate α_t updated sequentially: when recent observations fall outside the interval, α_t shrinks, widening future intervals. Because α_t is shared across regimes, free-flow observations (near-nominal) dilute the transition-regime signal.

Regime-Conditional ACI (R-ACI): Maintains separate α parameters per regime, so transition-regime errors drive wider transition intervals without affecting free-flow. Both ACI variants adjust interval *width* only and cannot shift probability mass toward a missing mode.

Table 1. MAE (mph) at h=12 stratified by traffic regime. TSFMs achieve much lower overall MAE than the historical conditional baseline, but all methods fail similarly during transitions.

	Method	Overall	Free	Trans.	Cong.
METR-LA	Hist. cond.	12.90	—	9.47	—
	Chronos-T5	5.77	2.13	9.83	1.19
	Chronos-Bolt	6.20	2.63	9.99	2.34
	Moirai	5.84	2.16	9.72	1.89
P-BAY	Hist. cond.	3.11	—	11.05	—
	Chronos-T5	3.07	1.35	11.04	2.70
	Chronos-Bolt	3.04	1.29	11.11	2.86
	Moirai	3.09	1.19	11.33	4.89

Bimodal Mixture Augmentation (BMA): Our proposed method. From training data, we precompute per-sensor historical transition probabilities: $P(\text{speed}_{t+h} \in R \mid \text{speed}_t)$ for each regime R and horizon h. At test time, we replace a fraction of the TSFM's 100 forecast samples with draws from this historical conditional distribution, scaled by the transition probability. This injects the missing mode (e.g., congested speeds ∼15 mph) when historical data indicates a non-trivial chance of regime switching. The mixing weight $w \in [0.1, 0.5]$ is selected on 10 held-out windows and evaluated on all remaining data.

BMA + ACI: BMA fixes the distributional *shape* by injecting the missing mode, but may leave residual over- or undercoverage because the mixing weight w is fixed per configuration. Global ACI then adjusts the interval *width* sequentially to close any remaining coverage gap. The two corrections are complementary: shape first, then width.

3. Results

3.1. Aggregate Metrics Hide Regime-Dependent Failures

Table 1 reports MAE stratified by traffic regime at h=12. TSFMs achieve much lower overall MAE than the historical conditional baseline (5.8 vs 12.9 mph on METR-LA), confirming that the TSFM provides genuine forecasting value. But during transitions, all methods converge to similar error (∼10–11 mph) because neither the TSFM nor the historical lookup can predict which direction traffic will move. Free-flow dominates the sample, so the aggregate is pulled toward the easy regime.

3.2. Calibration Failures Concentrate in Transitions

We construct 90% prediction intervals for each model (the standard operating point in TSFM benchmarks (Adler et al., 2025; Aksu et al., 2025)), meaning the interval should contain the true speed 90% of the time, and measure how often it actually does (empirical coverage). We then stratify by

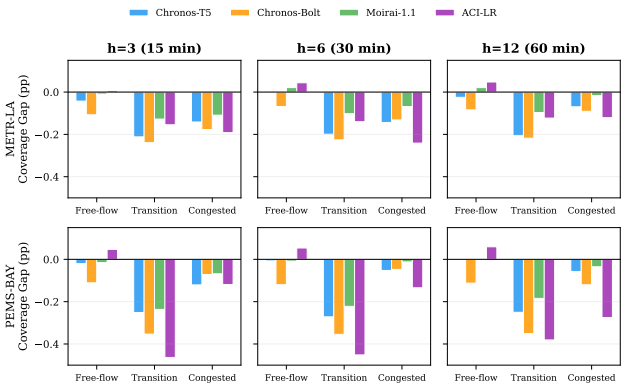

Figure 2. Coverage gap by traffic regime at three horizons. Each bar shows empirical coverage minus the 90% nominal target; the zero line means the model's intervals achieve exactly 90% coverage. Negative bars indicate undercoverage (intervals too narrow). All models fail during **transitions**, with gaps reaching −35 pp (Chronos-Bolt) and −48 pp (ACI-LR) on PEMS-BAY. ACI-LR overcorrects in free-flow (bars above zero) but has the worst transition coverage, illustrating that widening intervals alone cannot fix the shape mismatch.

traffic regime. Figure 2 reveals a striking pattern. During **free-flow** (>55 mph), TSFMs achieve coverage near the 90% target (bars near zero). During **congestion** (<25 mph), coverage drops moderately. The **transition** regime (25–55 mph) is the worst: Chronos-Bolt achieves only 54.9% on PEMS-BAY at h=12, a 35 pp gap. The ACI-LR baseline (linear regression with adaptive conformal inference) is even worse during transitions on PEMS-BAY despite overcovering in free-flow, confirming that width-only adjustment cannot fix this problem.

The transition regime's severe undercoverage matters most when operators need to act. Consider a freeway segment at 55 mph at 6:45 AM as morning demand builds: it will either hold near free-flow or collapse to 15 mph once a downstream bottleneck activates. During stable conditions real-time observations suffice, but at the edge of breakdown only a forward-looking prediction can distinguish a 4-minute from a 16-minute traversal. The root cause is a shape mismatch: the true future speed distribution during transitions is bimodal (Figure 1), but in zero-shot mode all three TSFMs produce effectively unimodal prediction intervals centered between the modes. The congested regime is itself unimodal (concentrated near 10–20 mph), so the problem there is merely one of interval width; transitions are fundamentally bimodal, making this a problem of distributional *shape*.

3.3. Post-Hoc Coverage Correction

Table 2 compares the four post-hoc methods on transition-regime coverage at h=12.

Global ACI barely moves the needle: +0.5–1.2 pp on tran-

Table 2. Transition-regime coverage (%) at h=12, 90% nominal. The historical conditional baseline (Hist.) achieves high transition coverage by sampling from training data. BMA approaches this level while preserving the TSFM's superior point accuracy.

	Model	Hist.	Native	G-ACI	R-ACI	BMA	+ACI
METR-LA	Chr-T5		68.2	71.3	70.7	78.3	**81.9**
	Chr-Bolt	81.6	68.6	69.7	70.2	78.7	**80.3**
	Moirai		80.3	80.9	81.6	84.1	**84.5**
P-BAY	Chr-T5		65.0	65.5	67.9	76.7	**77.2**
	Chr-Bolt	81.7	54.9	56.1	57.5	71.2	**73.6**
	Moirai		71.6	72.1	74.3	74.2	**74.9**

sition coverage. This is expected: widening a unimodal interval centered at the wrong location cannot recover the missing mode.

Regime-Conditional ACI does slightly better (+2.3–2.9 pp on PEMS-BAY) because the transition-specific α adapts more aggressively. But it still only adjusts width, not shape.

BMA produces the largest improvements, ranging from +2.6 pp (Moirai, PEMS-BAY) to +16.3 pp (Chronos-Bolt, PEMS-BAY). The improvement is largest for Chronos-Bolt, which has the worst native coverage: BMA lifts its PEMS-BAY transition coverage from 54.9% to 71.2%. **BMA + ACI** adds a further 1–2 pp on top of BMA by adjusting residual width, achieving the best overall results. The two methods are complementary: BMA corrects distributional *shape* (injecting the missing mode) while ACI corrects residual *width* (scaling the interval to account for remaining miscoverage). We use global ACI rather than regime-conditional ACI in the combination because once BMA has corrected the shape mismatch, the residual coverage error is uniform enough across regimes that per-regime width adjustment offers no additional benefit.

The historical conditional baseline (Table 2, "Hist.") achieves 81–82% transition coverage by construction, since it samples directly from the bimodal historical distribution. But it has poor overall MAE (Table 1), because it is a lookup table, not a forecaster. TSFMs provide the opposite: good overall accuracy but poor transition coverage. BMA combines both: it keeps the TSFM's point predictions while borrowing the historical distribution's shape, approaching the historical baseline's transition coverage without sacrificing overall accuracy.

BMA uses per-sensor historical data, just as ACI uses past prediction residuals. The TSFM itself remains zero-shot; BMA is a post-hoc correction applied to its outputs, analogous to how conformal methods adjust intervals using domain data. BMA does not degrade coverage in other regimes because the replacement fraction is modulated by the transition probability: in stable free-flow ($P(\text{congested}) \approx 0$), no samples are replaced. Results are stable across mixing weights ($w \in [0.2, 0.5]$) and degrade only below $w = 0.1$.

4. Discussion and Limitations

Implications for benchmarks. Our results suggest that aggregate evaluation metrics are insufficient for domains with regime switching. Regime-stratified evaluation is straightforward to implement and could be applied to any domain where physical thresholds produce distinct operating regimes: electricity spot prices switch between normal and spike regimes (Huisman & Mahieu, 2003; Lehna et al., 2025), and wind power near turbine cut-in speeds exhibits similar bimodality.

BMA and the zero-shot boundary. BMA uses per-sensor historical data from the training set, just as conformal methods use past residuals. The TSFM itself remains zero-shot; BMA is applied to its outputs. An alternative is to fine-tune the TSFM on traffic data directly, which may teach it to produce bimodal predictions natively. Whether fine-tuning closes the gap is an open empirical question. BMA remains relevant when fine-tuning is impractical or when using closed-source models.

Other limitations. Our evaluation uses 30 congestion-prone sensors per dataset with univariate forecasts; spatial information from upstream and downstream sensors could improve both accuracy and coverage. BMA's mixing weight is tuned on 10 held-out windows; a validation-set approach would be more principled. The interval width increase ($\sim$50–80%) trades sharpness for coverage; whether this is acceptable depends on the application.

References

Adler, C., Chang, Y., Draxler, F., Abdi, S., and Smyth, P. Beyond accuracy: Are time series foundation models well-calibrated? *arXiv preprint arXiv:2510.16060*, 2025.

Aksu, T., Woo, G., Liu, J., Liu, X., Liu, C., Savarese, S., Xiong, C., and Sahoo, D. GIFT-Eval: A benchmark for general time series forecasting model evaluation. In *International Conference on Learning Representations*, 2025.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Sez, H., Shchur, O., Rangapuram, S. S., Arango, S. P., Kapoor, S., et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Cini, A., Jenkins, A., Mandic, D., Alippi, C., and Bianchi, F. M. Relational conformal prediction for correlated time series. In *International Conference on Machine Learning*, 2025.

Gibbs, I. and Candès, E. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021.

Greenshields, B. D. A study of traffic capacity. *Highway Research Board Proceedings*, 14:448–477, 1935.

Huisman, R. and Mahieu, R. Regime jumps in electricity prices. *Energy Economics*, 25(5):425–434, 2003.

Lehna, M., Sieberichs, C., and Scholz, C. Assessing time series foundation models for probabilistic electricity price forecasting. *Energies*, 18(23):6269, 2025.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Li, Z., Qiu, X., Chen, P., Wang, Y., Cheng, H., Shu, Y., Hu, J., Guo, C., Zhou, A., Wen, Q., Jensen, C. S., and Yang, B. TSFM-Bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

Lu, E. D., Findling, C., Clausel, M., Leite, A., Gong, W., and Kersaudy, P. Adaptive regime-switching forecasts with distribution-free uncertainty: Deep switching state-space models meet conformal prediction. *arXiv preprint arXiv:2512.03298*, 2025.

Transportation Research Board. *Highway Capacity Manual*. National Academies of Sciences, Engineering, and Medicine, Washington, DC, 7th edition, 2022.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., and Sahoo, D. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, 2024.

Wu, Y., Ye, Y., Zeb, A., Yu, J. J., and Wang, Z. Adaptive modeling of uncertainties for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

Yu, B., Yin, H., and Zhu, Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *International Joint Conference on Artificial Intelligence*, pp. 3634–3640, 2018.

Zheng, V. Z., Choi, S., and Sun, L. Probabilistic traffic forecasting with dynamic regression. *Transportation Science*, 59(4):689–707, 2025.