Highlights

Global tree forecasters collapse at the hierarchical aggregate: a scale-aware, library-agnostic correction

Md Rezwanul Islam, Wael Mohammed

- Global boosted trees collapse 30–50× when scored at an out-of-support aggregate
- Curable on two axes: restore scale, or difference the target (recursion-robust)
- Both remedies are library-agnostic across LightGBM, XGBoost and CatBoost
- Collapse reproduces on synthetic data and three public sets (M5, Tourism, B2B)
- At the aggregate methods tie; per buyer 20 beat the baseline, FDR-controlled

Global tree forecasters collapse at the hierarchical aggregate: a scale-aware, library-agnostic correction

Md Rezwanul Islam[a,*], Wael Mohammed[a]

[a]*Field Nation LLC, Minneapolis, MN, USA*

Abstract

Global gradient-boosted-tree forecasters, trained by pooling series, are a workhorse of demand forecasting and reported accurate across hierarchy levels. We document a deployment-blocking exception. On a production panel of monthly gross transaction value from a business-to-business service marketplace, a global tree scored at the marketplace aggregate—far outside its per-buyer training support—collapses, under-predicting the total by an order of magnitude. We trace this to the constant extrapolation of trees beyond their training range: the failure is fundamentally one of scale, curable along either of two axes. One restores scale, via per-series scaling or a single-model cohort-weighted aggregate training row, both library-agnostic. The other renders the target stationary by seasonal differencing, which uniquely remains stable under recursive multi-step forecasting. The failure is structural—reproducing on synthetic data and three public datasets (M5, Tourism, B2B)—and seed-invariant, situated within a twenty-three-method benchmark documenting a unit-of-analysis dichotomy.

Keywords: gross transaction value, global forecasting models, gradient boosting, hierarchical forecasting, foundation models, two-sided marketplaces

1. Introduction

Project-based two-sided business-to-business (B2B) service marketplaces—on-demand IT field service, audio-visual installation, equipment-deployment

*Corresponding author. ORCID: https://orcid.org/0009-0002-4100-3796.

Email addresses: rezwanul.islam@fieldnation.com (Md Rezwanul Islam), wael.mohammed@fieldnation.com (Wael Mohammed)

URL: https://orcid.org/0009-0005-4211-8976 (Wael Mohammed)

dispatch, freelance professional work—pair enterprise service-buyer firms with independent providers under a transactional, non-subscription revenue model. Demand is dominated by discrete, project-shaped enterprise commitments (store technology refreshes, point-of-sale rollouts, OS-migration waves), so per-buyer monthly gross transaction value (GTV)—the pre-fee, pre-refund value of all completed work orders—is highly intermittent and heavy-tailed: a single enterprise customer can swing 10^5–10^6 USD month to month as upstream end-client capital schedules start and stop. The marketplace *aggregate* is comparatively smooth, because idiosyncratic per-buyer noise partially cancels in the sum. This gap between a lumpy per-buyer level and a smooth aggregate level—both consumed off the same panel, by account managers and by financial planning respectively—is the structural property around which our findings turn.

Global forecasting models (GFMs), which pool many series and learn one shared function, are the dominant applied approach to panels of this shape, and gradient-boosted trees (GBTs) are the most common GFM in practice (Januschowski et al., 2020; Makridakis et al., 2022). The prevailing evidence is favorable: Montero-Manso and Hyndman (2021) give theory for why a single global model can match or beat per-series local models with lower complexity, and recent work reports global LightGBM accurate across the levels of a hierarchy (Zhao and Abolghasemi, 2024). Against that backdrop we report a deployment-blocking exception that the favorable evidence does not anticipate.

The failure.. A global GBT with absolute lag features, trained on the per-buyer panel and then asked to forecast the marketplace aggregate, collapses. The aggregate lag-1 input sits roughly three orders of magnitude ($\approx 1{,}088\times$) above the median per-buyer lag-1 input, entirely outside the model's training support. Because trees extrapolate as constants beyond their observed leaves (Malistov and Trushin, 2019; März and Rasul, 2024), the model routes the out-of-support input to its largest leaf and under-predicts the aggregate by 30–50$\times$ ($> 90\%$ mean absolute percentage error (MAPE); equivalently seasonal mean absolute scaled error (MASE) $\approx$9.3 and root mean squared scaled error (RMSSE) $\approx$7.8 on the production aggregate, about nine times a seasonal-naive forecast)[1] before any correction. The general fact that trees

[1]We report MAPE first for interpretability—a $> 90\%$ aggregate MAPE maps directly to the order-of-magnitude under-forecast that is the phenomenon of interest—and confirm

extrapolate poorly is known; what is undocumented—and what this paper contributes—is the *aggregate-scale collapse* of a global GBT in a hierarchical setting, a working correction for it that reproduces on three public datasets, and the finding that the correction is *library-agnostic* once each library is given its load-bearing component.

Contributions..

(C1) The collapse is a training-support failure, curable on two axes (Section 5.2, Section 5.3). We diagnose the collapse and show it is fundamentally one of training-support scale: restoring the aggregate to the model's support cures it. The conventional remedy is *per-series scaling* (dividing each series by its own level), which prevents the collapse on every panel and is the more accurate of the two on one-step error; we also characterize a *single-model* alternative—a cohort-weighted aggregate training row under squared-error loss—for deployments that keep one un-normalized global model across levels (its edge is cold-start, where a new series has no level to scale by). A second, independent axis also cures it: rendering the target stationary by seasonal differencing, which alone avoids the recursive re-collapse the scale cures leave—it reconstructs each step by adding back the realized seasonal lag rather than asking a leaf to emit an out-of-range level. We recommend per-series scaling for one-step use and seasonal differencing wherever forecasts are rolled forward recursively. A leave-one-out ablation isolates the cohort-weighted row and the squared-error loss as the two load-bearing, interacting components of that single-model route (log scaling and a recursive guardrail are defensive refinements).

(C2) Both cures are library-agnostic (Section 5.4, Section 5.6, Section 5.7). Per-series scaling and the cohort-weighted row each prevent the collapse for all three major boosted-tree libraries (LightGBM, XGBoost, CatBoost) on every panel we test. We further show that a reported "CatBoost cannot forecast the aggregate" result—a finding from our

every principal result under the scale-free, competition-standard metrics seasonal MASE and RMSSE (Hyndman and Koehler, 2006; Makridakis et al., 2022), reported alongside MAPE throughout (Table 4). Because MAPE penalizes over-forecasts more heavily than under-forecasts (Hyndman and Koehler, 2006) and the collapse is a pure under-prediction, MAPE if anything understates it relative to a symmetric scaled metric.

own earlier deployment—is a deployment artifact of the single-model route: it arises only when CatBoost is trained without the load-bearing cohort weight; supplied it, CatBoost recovers across an exhaustive hyperparameter grid (14.9–18.5% one-step aggregate MAPE, tied with LightGBM).

(C3) The collapse is structural and reproduces on public data (Section 5.5, Section 5.6, Section 5.7, Section 5.8, Section 5.10). It reproduces on a synthetic hierarchical panel sharing only the scale gap (94.7% $\rightarrow$ 17.5% under the fix) and on three public datasets from different domains, including M5 retail (99.8% $\rightarrow$ 14.1%) and Australian Tourism (96.2% $\rightarrow$ 14.0%)—with restoration of training-support scale the decisive lever in every case; the collapse-versus-no-collapse outcome is also invariant across eight training seeds.

(C4) A reproducible benchmark in this sector, and the unit-of-analysis dichotomy (Section 5.11, Section 5.11). Twenty-three methods across four model families, including six 2025-era zero-shot foundation models, on five years of production data. At the aggregate, twelve methods tie within $\approx 8.25\%$ and the cohort-weighted trees are competitive but not leading; per buyer, twenty methods beat the baseline, nineteen of them significantly (paired Wilcoxon, n=1,420, Benjamini–Hochberg-controlled), a result that holds under both MAPE and scale-free MASE.

We are explicit about what we do not claim. We do not claim a corrected tree is the most accurate aggregate forecaster—at the aggregate everything ties and a simple decomposition method leads (Section 5.11)—nor that the single-model cohort-weighted row is the best cure: per-series scaling is simpler and more accurate on one-step error (Section 5.3), and we recommend it by default for one-step use, with seasonal differencing preferred under recursive multi-step. The contribution is the diagnosis—a deployment-blocking, scale-driven collapse—its generalization to public data, and the demonstration that restoring training-support scale, by either route, is the library-agnostic cure, not an aggregate-accuracy record. The reason a practitioner runs a global tree at all is its *per-buyer* value (twenty methods beat the baseline per buyer, Section 5.11); the aggregate is a *coherence constraint* the same model must satisfy when its forecasts are summed for financial planning. A collapse there is deployment-blocking regardless of aggregate accuracy, which is why

preventing it—returning the tree to the naive-competitive band—is the goal, not beating the seasonal-naive aggregate.

2. Related work

Global models and gradient-boosted trees.. The case for global over local models is well established: Montero-Manso and Hyndman (2021) show a single global model can match or exceed per-series local models with far lower complexity and without assuming series similarity, and GBT-based GFMs won or placed in the M5 competition (Makridakis et al., 2022; Bojer and Meldgaard, 2021). The favorable evidence extends to hierarchies: Zhao and Abolghasemi (2024) report global LightGBM accurate across hierarchy levels relative to local ETS/ARIMA. Our result is the boundary condition this literature does not surface—a global GBT can be accurate within the per-series support and yet collapse catastrophically when scored at an aggregate that lies orders of magnitude outside that support.

Extrapolation limits of trees, and remedies.. That trees cannot extrapolate beyond their training range is a known limitation, addressed by leaf-level linear extrapolation (Malistov and Trushin, 2019), parametric "hyper-tree" architectures that learn the parameters of an extrapolating model (März and Rasul, 2024), and, in the M5 setting, exponential-smoothing-anchored calibration of LightGBM to mitigate trend-extrapolation bias (Lainder and Wolfinger, 2022). These target *within-series* extrapolation along the trend. Our collapse is a distinct, *cross-scale* failure—the forecast target is at a different order of magnitude than any training series—and our remedy is a cohort-weighted training row plus log scaling rather than a new tree architecture or a post-hoc calibration anchor. We position the four-component fix as complementary to, not a replacement for, these.

Why the collapse has not been reported.. The failure is real on public data—it reproduces on M5 (Section 5.6)—yet the hierarchical-forecasting and M-competition literatures have not reported it, because their standard pipelines structurally avoid the operation that triggers it. Hierarchical methods forecast the bottom level and *reconcile* upward (Hyndman et al., 2011; Wickramasuriya et al., 2019), so the aggregate is a sum of base forecasts, never a single global model's direct output at aggregate scale; competition pipelines further fit level-aware models and routinely scale each series before training—both keep the model in support (Section 5.3 confirms it: per-series scaling applied to

the same global tree prevents the collapse). The collapse therefore requires a specific configuration: a single global model served across all aggregation levels on raw, unscaled features—a production convenience (one model, one pipeline) rather than a competition design. Only in that configuration does the known within-range extrapolation limit (above) manifest as a catastrophic, deployment-blocking collapse rather than mild bias. We report it because our production setting forces exactly this configuration; the contribution is naming and fixing the failure mode in the deployment where it is consequential—with a minimal, library-agnostic training-row remedy—not rediscovering that trees extrapolate poorly.

The sector.. Published GTV/GMV benchmarks address mechanically distinct settings: business-to-consumer (B2C) e-commerce gross merchandise value (GMV) (Yu et al., 2022; Zhang et al., 2020), high-frequency spatial demand for ride-hailing and delivery (Han et al., 2023; Zhu and Laptev, 2017), and online-labor measurement. To our knowledge none benchmark monthly buyer-level GTV on a project-based B2B services marketplace, where upstream end-client capital is unobserved, there is no subscription revenue, and the aggregate is smooth while the per-buyer level is lumpy. A dated prior-art sweep (Scholar/web and patents, 2026-06-15) found no duplicate; we therefore hedge the claim as "to our knowledge, first."

Foundation models and the M-competition canon.. The M-competition lineage (Makridakis et al., 2020, 2022) established that simple statistical baselines are hard to beat at aggregated levels, and FFORMA (Montero-Manso et al., 2020) won M4 by feature-based per-series weighting. Zero-shot foundation time-series models—Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), TimesFM (Das et al., 2024), FlowState (IBM Research, 2025)—promise transfer from large pretraining corpora; independent benchmarks (Liu et al., 2024; Tan et al., 2024) show the gains are regime-dependent. We evaluate which of these conclusions transfer to this sector.

3. Data and setting

The empirical instantiation uses anonymized monthly GTV per buyer from the production transaction warehouse of one marketplace in the sector (an on-demand business-to-business field-service marketplace pairing several thousand enterprise service-buyer firms with tens of thousands of independent providers) over Jan 2021–Dec 2025 (T=60 training months; $\approx$ 5,400 active

buyers). All monetary quantities are reported in scale-free form—ratios, or values indexed to a base period—with absolute GTV levels withheld for commercial confidentiality; the analysis is invariant to this scaling, and the collapse is independently reproduced on public data (M5, Section 5.6; Australian Tourism, Section 5.7; Olist, Section 8). The training cutoff is the latest complete warehouse month; evaluation uses a strict five-month holdout, Jan–May 2026, with no holdout actuals informing selection, hyperparameters, or fitting. We evaluate at two levels with distinct constituents, and are explicit about which is which to avoid a selection ambiguity. The *aggregate* target is the full marketplace total—the sum over all active buyers, which is exactly what the global model's injected aggregate row represents (Section 5.2)—not a cohort subtotal. The *per-buyer* analysis covers the full active forecast population—every active buyer with at least fifteen months of history (n=1,659; n=1,420 of these have a finite baseline holdout MAPE)—rather than a revenue-weighted top cohort, with per-series model selection applied across the entire population. Buyers are stratified ex ante by coefficient of variation (CV) and seasonal strength (Wang et al., 2006) into *Steady* (CV<0.4), *Seasonal* (F_S>0.3), and *Variable* (residual) archetypes.

Structural panel statistics (load-bearing).. Three properties of the raw panel drive the central finding: (i) ≈ 73.5% of buyer-month cells are zero or absent; (ii) the median active-month count per buyer is seven months, so only a minority of buyers clear the longer-history minimums the deepest methods require; (iii) aggregate calendar-year GTV has not reversed in the window—every year posted higher GTV than the prior. The combination (sparse, short, seasonal with an upward trend) is what makes the drift-adjusted seasonal-naive baseline competitive at the aggregate and the foundation-model cluster competitive per buyer. The qualitative conclusions of this paper transfer only to panels sharing this sparsity-and-growth profile.

To avoid disclosing commercially sensitive level information, Figure 1 illustrates this profile on a seeded synthetic surrogate calibrated to the three statistics above (surrogate: 73.8% cell sparsity, seven-month median active history, a seasonal aggregate on a rising trend) rather than on the proprietary series. The surrogate ships in the replication package, so the orientation figure is fully reproducible, while the real-panel statistics reported here remain the load-bearing quantities.

The scale gap (the proximate cause).. The aggregate-series Jan-2026 lag-1 input is ≈ 1,088× the median per-buyer lag-1 input—entirely outside the

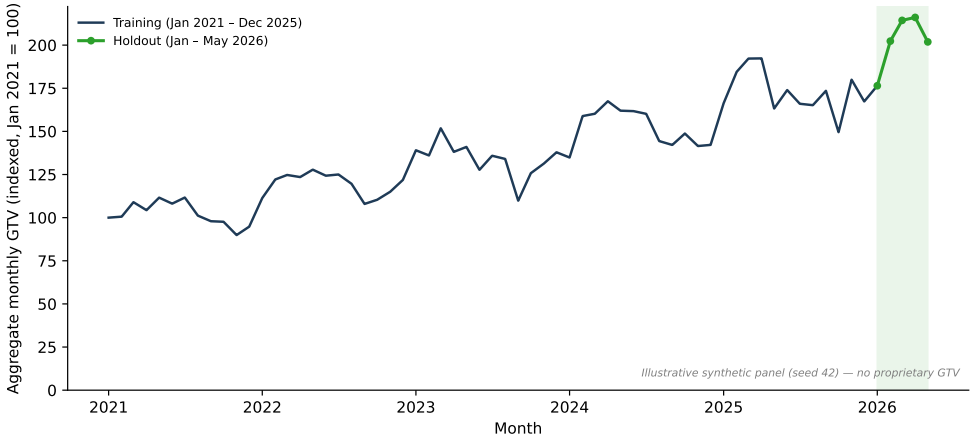

Figure 1: Full marketplace aggregate monthly GTV (summed over all buyers, indexed to Jan 2021=100), Jan 2021–May 2026, shown on a calibrated synthetic surrogate (Section 3; not the proprietary series). The aggregate trends upward with a recurring seasonal cycle (intra-year peaks and troughs) and irregular month-to-month variation; every calendar year posts higher GTV than the prior (no year-over-year reversal). The five-month held-out evaluation window (Jan–May 2026) is shaded. The per-buyer series feeding this total are lumpy, intermittent, heavy-tailed, and short—the gap this paper turns on.

per-buyer training support—the quantity that breaks an absolute-lag global tree. Figure 2 summarizes the mechanism and the two axes that return the target to support.

4. Methodology

Benchmark protocol.. We evaluate twenty-three methods across four families: *classical* (AutoETS, AutoTheta, MSTL, STL seasonal-trend, linear trend, Prophet); *foundation/zero-shot* (Chronos v1, Chronos-2, Chronos-Bolt, Moirai-2, FlowState, TimesFM)[2]; *intermittent* (Croston, Croston-SBA, TSB, IMAPA); *ML tree/neural* (LightGBM, XGBoost, CatBoost global; DeepAR, N-HiTS); and two *baseline* controls—a drift-adjusted seasonal-naive (the experimental control, not a contribution) and a three-month exponentially weighted moving average (EWMA). All twenty-three are scored at the aggre-

[2]Citations are to each model family's originating paper; the roster evaluates the latest public checkpoints—including the Chronos-2, Chronos-Bolt, and Moirai-2 releases—whose exact identifiers are pinned in the reproducibility code.

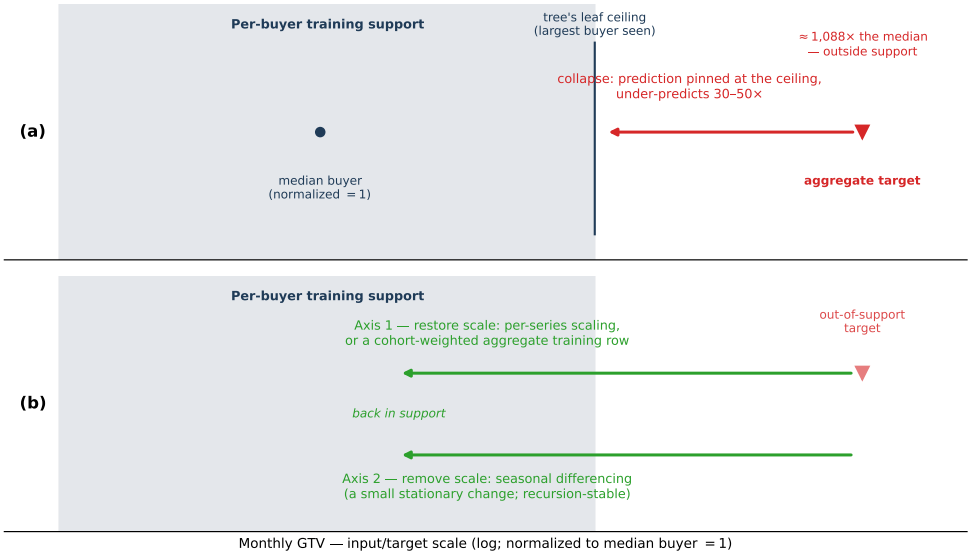

Figure 2: Why a global tree collapses at the aggregate, and the two cures. **(a)** A regression tree extrapolates as a constant: it cannot emit a value above its largest leaf—the largest buyer seen in training (the *ceiling*). The marketplace aggregate sits $\approx 1{,}088\times$ above the median buyer and outside the per-buyer training support, so the out-of-support input is routed to the ceiling and the forecast under-predicts the total by 30–50×. **(b)** Either axis returns the target to within support: *restoring scale*—per-series scaling, or a single cohort-weighted aggregate training row—or removing scale—seasonal differencing, the only route that remains stable under recursive multi-step forecasting. Conceptual schematic; the quantities are those reported in this section.

gate (Table 4). Classical seasonal ARIMA (SARIMA) was considered but excluded from the roster: on this heavy-tailed, intermittent monthly panel it behaved as a high-variance method whose occasional large errors offset its gains, adding no net accuracy over the automated ETS/Theta/MSTL classical baselines retained here. Methods are fit on the most recent 48 months per series (`TRAIN_WINDOW_MONTHS`); the five global cross-learning models train on the full panel. Per-series ensemble selection uses rolling-origin cross-validation; the operational and benchmark paths share one code path.

The baseline control, formally.. The seasonal-naive control—the reference every per-buyer comparison (Section 5.11) is measured against—is a damped-momentum seasonal-naive, not a plain one. For a series with last training month t and step $h{=}1, \ldots, H$, it scales the seasonal-naive base by a clipped trailing year-over-year (YoY) momentum multiplier damped toward 1 over the horizon (Gardner and McKenzie, 1985):

$$\begin{aligned} &\hat{y}_{t+h} = \mu_h \, y_{t+h-12}, \qquad \mu_h = 1 + (m-1)\,\varphi^{\,h-1}, \\ &m = \operatorname{clip}\!\left(\frac{\sum_{j=0}^{2} y_{t-j}}{\sum_{j=0}^{2} y_{t-12-j}},\ 0,\ 2 \right), \end{aligned} \tag{1}$$

with damping $\varphi{=}0.80$ per buyer—where trailing ratios mean-revert—and $\varphi{=}1$ (undamped, flat momentum) for the marketplace aggregate, whose growth is genuine. We call it a baseline because it cross-learns nothing, but it is a deliberately strong control: a method that cannot beat it per buyer is not competitive.

Pinned configuration and provenance.. Every result artifact was produced against one configuration on a short, strictly additive commit history—later analyses add new write-ups and figures without changing any forecasting recipe or the pipeline, so each lies on one linear history with the others. All share one configuration: input vintage `gtv_by_month_by_buyer_20260614`, training cutoff Dec 2025, holdout Jan–May 2026, and the fix recipe. Each artifact records its commit SHA, a clean-tree flag, the seed, and library versions in a provenance sidecar (Section 10). We make the configuration explicit because an earlier draft prominently reported a 5.4% aggregate MAPE for the corrected LightGBM that depended on a regime-step exogenous indicator since removed from the shipped code; that number is not reproducible from current code and we do not rely on it (see Section 10). It is distinct from the

per-series-scaling one-step MAPE of 5.4% in Table 3, a current result from a different recipe (per-series scaling, no exogenous indicator).

Metrics and significance.. We report MAPE at the aggregate and per-buyer levels and signed bias, and—for the collapse-and-fix result—also the scaled metrics seasonal MASE (Hyndman and Koehler, 2006) and an RMSSE (seasonal m=12 at the aggregate, one-step on the per-series public panels), evaluated across rolling origins, so the central finding does not rest on a single metric or split (Section 5.6). MAPE is retained for the aggregate benchmark—the marketplace total is strictly positive and smooth, the regime in which MAPE is well-behaved—and for the per-buyer leaderboard as a robustness check; the regime in which MAPE is fragile (intermittent, near-zero bottom series) is precisely where we report the scaled metrics, and the central collapse-and-fix result is confirmed under seasonal MASE and RMSSE on the intermittent, heavy-tailed public panels (M5, Tourism, and UCI Online Retail II). The aggregate collapse-and-fix gap is tested directly with the Diebold–Mariano test (Diebold and Mariano, 1995) under the Harvey–Leybourne–Newbold small-sample correction (Harvey et al., 1997), on the per-step absolute-percentage-error loss differential pooled across rolling origins, and cross-checked with a Wilcoxon signed-rank test across origins (Section 5.9, Table 8). Per-buyer comparisons against the baseline use the paired Wilcoxon signed-rank test (Wilcoxon, 1945) across the cross-section of buyers, with Benjamini–Hochberg false-discovery-rate (FDR) control across the twenty-two simultaneous comparisons, reported under both per-buyer MAPE and—as a scale-free cross-check—MASE. The scaled metrics divide by an in-sample naive mean absolute error whose season m we set by estimability rather than convention,

$$d_m(z) = \frac{1}{n-m} \sum_{s=m+1}^{n} \left| z_s - z_{s-m} \right|, \qquad m = \begin{cases} 12, & \text{aggregate } A, \\ 1, & \text{each buyer } i, \end{cases} \tag{2}$$

so MASE and RMSSE divide by the seasonal-naive error d_{12} at the long, seasonal aggregate and by the one-step error d_1 per buyer (Hyndman and Koehler, 2006). The seasonal denominator is not estimable for the median buyer's seven active months, whereas d_1 is defined for any buyer with two training observations and is robust to the intermittent, near-zero series on which MAPE is fragile. Stochastic methods are banded across seeds (Section 5.10); deterministic methods carry zero seed variance.

5. Results

5.1. The aggregate collapse

The mechanism, formally.. A regression tree is piecewise-constant: every prediction is a leaf value, and leaf values are averages of training targets, so the prediction is bounded by the training range,

$$\min_{i,t} y_{i,t} \;\le\; f(x) \;\le\; \max_{i,t} y_{i,t} \qquad \text{for every input } x, \tag{3}$$

with i indexing buyers and t months. Write the aggregate level as $A = \sum_i y_{i,t}$ and the tree's ceiling as $c = \max_{i,t} y_{i,t}$—the largest single-buyer level it ever saw. An aggregate query is routed to a near-ceiling leaf, so $\hat{A} \le c$ and the error is bounded below by a quantity that depends only on the *scale gap* $g \equiv A/c$:

$$\text{MAPE} = 1 - \frac{\hat{A}}{A} \;\ge\; 1 - \frac{c}{A} = 1 - \frac{1}{g}. \tag{4}$$

The collapse is therefore a structural ceiling, not a tuning failure: the observed 30–50× under-prediction is exactly $g \approx 30$–50 (MAPE $\ge 96.7\%$), and no leaf re-weighting can lift $\hat{A}$ above c. This ceiling gap is distinct from—and smaller than—the $\approx 1{,}088\times$ aggregate-to-median-buyer gap that puts the input out of support to begin with. Every cure that follows works by shrinking g to $O(1)$: raise c to aggregate scale (a cohort-weighted aggregate row, Section 5.2), rescale A down to per-buyer scale (per-series scaling, Section 5.3), or remove the level entirely (seasonal differencing, Section 5.3).

The unmodified global-tree-with-absolute-lags design under-predicts the aggregate by 30–50× across three independent libraries using disjoint codebases but identical feature engineering—evidence the failure is a property of the design pattern, not of any one library. DeepAR, a global ML method but with built-in per-series mean scaling (Salinas et al., 2020), *escapes the catastrophic collapse*—a median 12.5% aggregate MAPE across five seeds (range 6–15%), far from the absolute-lag trees' 30–50× under-prediction, though still not competitive (MASE 1.3, just above a seasonal-naive forecast)—isolating the absolute-lag tree structure, not global pooling itself, as the root cause of the catastrophic failure.

5.2. A single-model cure: the cohort-weighted aggregate row

Scale can be restored two ways. The conventional remedy—per-series scaling—we treat in Section 5.3; here we characterize a *single-model* route

that keeps one un-normalized global model spanning every aggregation level in training (the configuration our production system trains, and the one the leave-one-out dissects; at serving, production gates this tree from the aggregate and serves it per buyer, Section 7). It has four components, but a leave-one-out ablation (Table 2) shows only two are individually load-bearing—the cohort-weighted aggregate row (c) and the RMSE loss (b), which interact—while the log transform (a) and the recursive guardrail (d) are defensive refinements we retain for robustness:

(a) **Log target and lag transform.** Replace y_t with $\tilde{y}_t = \log(1+y_t)$ for the target and lags, compressing the multi-order-of-magnitude scale gap into a contiguous range ($\tilde{y} \in [10, 18]$) inside which trees interpolate.

(b) **RMSE loss on the log target.** L_1/MAE is mis-specified for intermittent monthly data whose median is often zero (trees collapse to a near-zero constant leaf); RMSE on $\log(1+y)$ is approximately error-symmetric and recovers a coherent gradient.

(c) **Cohort-weighted aggregate training row.** Insert the cross-buyer total as one extra series (`buyer_id`$=-1$) so trees observe the high-lag region in training, weighted by $w_{\text{agg}} = n_{\text{buyer}}/n_{\text{agg}}$ so its rows reach parity with the $\approx$ 211,000 per-buyer rows.

(d) **Recursive-step guardrail.** Cap each predicted step at $2\times$ the trailing-12-month buffer maximum, preventing rare leakage of aggregate-scale predictions into the largest per-buyer queries. The cap is dormant on the aggregate by construction.

The cumulative ablation (Table 1) shows the *construction path*: recovery happens entirely when the cohort-weighted aggregate row (c) is added (95% $\rightarrow$ 17.5%), with (a) and (b) doing nothing in that order because there is no aggregate-scale row for them to act on yet. A cumulative table cannot, however, establish necessity. We therefore run a leave-one-out ablation—start from the full fix and remove one component at a time—on both the synthetic and the real production panel (Table 2). Two components are load-bearing, and they interact: the aggregate row (c) supplies the only aggregate-scale training signal (removing it reverts the collapse on both panels), and the RMSE loss (b) is what lets that single high-magnitude row move the global fit—under MAE the lone aggregate row is washed out among $\approx 10^5$ per-buyer

Table 1: Cumulative ablation of the four-component fix on a synthetic hierarchical panel (one-step aggregate MAPE, mean over six data seeds; Section 5.5). Component (c)—the cohort-weighted aggregate row—is the load-bearing lever; (a) and (b) do nothing alone because the collapse is a training-support problem, not a loss/transform problem.

Configuration	One-step aggregate MAPE
Baseline: absolute lags, MAE, per-buyer only	94.7%
+ (a) $\log(1+y)$ target/lags	94.5%
+ (b) RMSE loss on log scale	95.4%
+ (c) cohort-weighted aggregate row	**17.5%**
+ (d) 2× recursive guardrail (full fix)	**17.5%**

Table 2: Leave-one-out ablation: one-step aggregate MAPE when each component is removed from the full fix, on the synthetic panel (mean over six seeds) and the real production panel (pinned Jan–May 2026). Two components are load-bearing on both panels and interact; the log transform and the recursive cap are defensive refinements, not individually necessary.

Configuration	Synthetic	Production	Role
Full fix (reference)	17.5%	15.1%	—
− (c) cohort-weighted aggregate row	95.4%	97.8%	load-bearing
− (b) RMSE loss (use MAE)	202%	99.6%	load-bearing
− cohort weight only (row kept)	31.1%	79.6%	decisive at extreme gap
− (a) $\log(1+y)$ transform	15.7%	18.6%	refinement
− (d) 2× recursive guardrail	17.5%	15.1%	dormant guardrail

rows, so removing (b) while keeping the row also reverts the collapse. Cohort-weighting the row is decisive specifically at production's extreme scale gap (removing the weight: 15% → 80%). By contrast the log transform (a) and the recursive guardrail (d) are not individually necessary in either panel: (a) changes one-step MAPE by ≤ 4 percentage points (pp) and (d) never binds at these horizons. We report them as defensive refinements retained for the heavier-tailed, longer-horizon production tail, not as load-bearing levers.

5.3. Two further cures: per-series scaling and a stationary target

The single-model route is not the only—or the simplest—cure. The standard way global models handle disparate scales is *per-series scaling*: divide each series by its own level before fitting and multiply back afterwards, as DeepAR does by construction (Salinas et al., 2020). Applied to the same

global tree—each series divided by one plus its training mean, the aggregate scaled by its own mean at inference—this restores the aggregate to the model's support and prevents the collapse on every panel (Table 3):

$$\tilde{y}_t^{(i)} = \frac{y_t^{(i)}}{1+\bar{y}^{(i)}}, \quad \bar{y}^{(i)} = \frac{1}{n_i}\sum_t y_t^{(i)}, \qquad \hat{y}_{t+h}^{(i)} = \left(1+\bar{y}^{(i)}\right)\hat{\tilde{y}}_{t+h}^{(i)}, \tag{5}$$

mapping every series, the aggregate included, to $O(1)$—that is, $g \to O(1)$. It is in fact the more accurate of the two routes on one-step aggregate error (production 5.4% vs. 15.1%; M5 8.2% vs. 14.7%, five-seed means), and it is library-agnostic. We therefore recommend per-series scaling as the default cure for one-step forecasts. The single-model cohort-weighted row (Section 5.2) is preferable in two narrower situations: when one un-normalized global model must serve every level from a shared (log) feature space without per-series scale bookkeeping, and at *cold start*, where a new series has no history from which to estimate a scale yet the aggregate row and log space still place its queries in range. Both routes confirm the central claim—the collapse is a training-support scale failure, curable by restoring scale, not a property of trees that resists remedy.

A second axis: a stationary target.. Restoring scale is not the only way back into support. A complementary axis treats the target: predict a seasonal difference $y_t - y_{t-12}$ rather than the level and reconstruct each step by adding back the realized seasonal lag:

$$d_t = y_t - y_{t-12}, \qquad \hat{y}_{t+h} = y_{t+h-12} + \hat{d}_{t+h}. \tag{6}$$

A seasonal difference is scale-free—a marketplace and a single buyer each growing 8% year-over-year present the same target—so the aggregate is no longer out of support, and differencing cures the one-step collapse on every panel (Table 3). It has one advantage the scale cures lack: it is the only cure that does not recursively re-collapse. The single-model cohort row, accurate one step ahead, re-collapses under recursive multi-step (production 15% → 27%, M5 15% → 34% recursive mean) because each step still asks a leaf to emit a larger absolute level than any it grew; seasonal differencing instead reconstructs each step by adding back the realized seasonal lag, inherits no leaf ceiling, and so its one-step and recursive errors coincide (production 11.0%/10.9%, Tourism 4.6%/4.5%). We therefore prefer it wherever forecasts are rolled forward. One caution: a purely multiplicative target (the ratio

Table 3: Three cures across two axes: aggregate MAPE (five-seed mean±sd), one-step and recursive, identical recipe. The naive global tree collapses on every panel. *Scale axis:* per-series scaling and the single-model cohort-weighted row both restore scale (scaling the more accurate one-step). *Trend axis:* seasonal differencing makes the target stationary. Only seasonal differencing avoids the recursive re-collapse, and its seed variance is the lowest of the cures; the cohort row's recursive error is seed-fragile (M5 ±30.8). Values are from the five-seed cures-comparison run; the synthetic baseline (93.2%) is thus ≈1.5 points below the six-seed cumulative-ablation baseline (94.7%, Table 1), run-to-run variation immaterial to the collapse.

Dataset	Baseline	Per-series	Cohort row	Seasonal diff.
One-step aggregate MAPE				
Synthetic	93.2 ± 0.0	11.7 ± 0.7	14.9 ± 0.9	$\mathbf{6.9 \pm 0.1}$
Production	97.8 ± 0.0	$\mathbf{5.4 \pm 0.3}$	15.1 ± 0.3	11.0 ± 0.0
M5	99.8 ± 0.0	$\mathbf{8.2 \pm 0.1}$	14.7 ± 1.5	10.0 ± 0.0
Tourism	96.1 ± 0.1	9.6 ± 0.6	14.0 ± 0.1	$\mathbf{4.6 \pm 0.0}$
Recursive aggregate MAPE				
Synthetic	93.5 ± 0.1	17.2 ± 0.5	61.4 ± 4.5	$\mathbf{6.6 \pm 0.1}$
Production	98.0 ± 0.0	8.2 ± 1.3	27.1 ± 10.6	$\mathbf{10.9 \pm 0.0}$
M5	99.8 ± 0.0	17.9 ± 0.2	33.6 ± 30.8	$\mathbf{10.0 \pm 0.0}$
Tourism	96.1 ± 0.1	10.4 ± 0.1	15.0 ± 0.2	$\mathbf{4.5 \pm 0.0}$

y_t/y_{t-12}) compounds under recursion and is unstable on heavy-tailed panels (production recursive MAPE diverges); the additive seasonal difference is stable.

5.4. Library-agnostic transfer of both cures

Given the identical single-model fix—crucially including the load-bearing cohort weight—the collapse is prevented for all three major boosted-tree libraries (Ke et al., 2017; Chen and Guestrin, 2016; Prokhorenkova et al., 2018) on every panel we test. On the production panel, a hyperparameter sweep of full-fix CatBoost (depth $\in \{4, 6, 8, 10\}\times$ learning rate $\in \{0.03, 0.05, 0.1\}\times$ $L_2 \in \{1, 3, 9\}$, 36 configurations) recovers the aggregate in every configuration: 14.9–18.5% one-step MAPE (median 15.9%), statistically tied with LightGBM (15.1%) on the identical panel.

The CatBoost non-transfer as a deployment artifact.. The shipped production benchmark (Table 4) lists CatBoost at 92.9% MAPE, collapsed. That figure does not reflect a library limitation: the production CatBoost path is trained without the cohort-balanced sample weight (on the prior belief that its

symmetric trees were brittle under the weight), and the leave-one-out (Table 2) shows withholding that weight reverts the collapse for any library—it sends the full-fix LightGBM from 15% to 80% on the same panel. Supplied the full fix including the weight, CatBoost recovers (above). The 92.9% is therefore an artifact of withholding the fix's load-bearing component from one library, not evidence that the library cannot do it; we revise this figure here. A controlled isolation confirms the weight—not the dataset or the library—is the variable: CatBoost recovers with the cohort-weighted row and collapses with an unweighted row on both M5 (16.8% vs. 94.4%) and production (15.7% vs. 87.0%), the weight being decisive whenever the single aggregate row competes with many per-buyer rows.

The one residual difference is second-order.. CatBoost is somewhat more fragile under recursive multi-step inference at the extreme production gap (recursive 12-step MAPE 37–70% across the grid versus LightGBM's ≈37%), but this is tunable, not categorical: at depth 6/learning-rate 0.1 recursive MAPE is 18%, better than LightGBM. Its maximum emittable (leaf-ceiling) value under the baseline recipe is marginally lower than LightGBM's (52× versus 43× below the true aggregate), consistent with its oblivious-tree structure—but both ceilings are catastrophically below the aggregate, so this difference does not distinguish the libraries at the level that matters. On a synthetic panel where we sweep the scale gap from 38× to 1,568× under the full fix, CatBoost tracks LightGBM throughout (both in the 10–24% band, CatBoost within a few points of LightGBM at every gap and indistinguishable at the largest); there is no scale-gap threshold beyond which CatBoost alone fails. We thus find no genuine cross-library boundary—the fix transfers across all three libraries—and report the earlier production-only "non-transfer" as the withheld-weight artifact above.

The ranking is not a MAPE artifact.. The MASE and RMSSE columns (both seasonal, m=12) confirm the ordering directly: across all 23 methods, the Spearman rank correlation between the MAPE ordering and the MASE and RMSSE orderings is 0.98 and 0.93. The collapse is just as stark under the scaled metrics—the shipped no-weight CatBoost sits at MASE 9.3, more than nine times a seasonal-naive forecast, while every method in the tie band is below MASE 0.85 and the cohort-weighted XGBoost recovers to MASE 1.2—so the benchmark conclusions do not depend on MAPE (Section 5.9 reports the formal significance test).

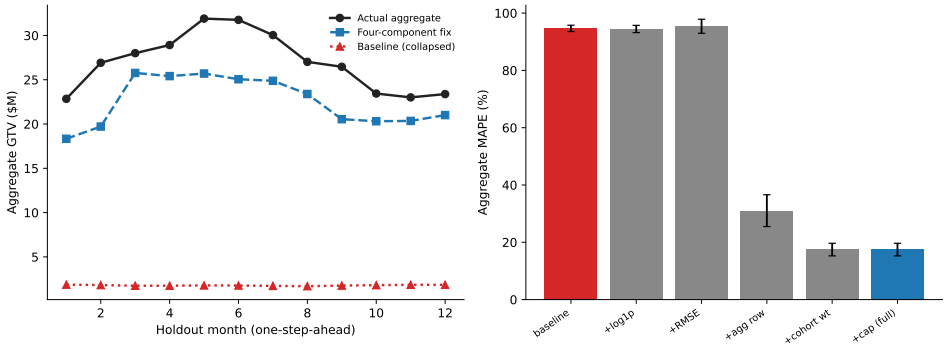

Figure 3: Synthetic reproduction of the collapse and fix. The baseline global tree (flat line) under-predicts the smooth synthetic aggregate; the cohort-weighted fix tracks it. Mechanism confirmation, not MAPE replication—synthetic per-buyer series carry no learnable cross-buyer structure, so recovery stops at 17.5% rather than production's single digits.

5.5. Synthetic reproduction of the collapse

To show the failure is a property of the design pattern and not of the proprietary data, we build a synthetic hierarchical panel (300 buyers $\times$ 60 months) that carries no exploitable cross-buyer structure yet reproduces the production conditions that drive the collapse: a scale gap, idiosyncratic per-buyer seasonal phases (so the aggregate is smooth by cancellation), and $\approx 7\%$/yr growth. The synthetic aggregate is 11.5$\times$ the largest single-buyer series—out of per-buyer support. The baseline tree collapses to 94.7% one-step aggregate MAPE across six data seeds (it routes the out-of-support input to its largest leaf and predicts a flat $\approx$ \$2M for a $\approx$ \$30M series); the four-component fix recovers it to 17.5% (Table 1, Figure 3). The residual $\approx 17\%$ is the same leaf-ceiling mechanism in milder form—a growing aggregate climbs above its highest training leaf—and is why production pairs the tree with log scaling and an ensemble rather than trusting a bare tree to extrapolate.

5.6. Generalization to the public M5 benchmark

To establish that the collapse and the fix are not artifacts of one proprietary panel, we replicate both on the public **M5** competition dataset (Makridakis et al., 2022)—the canonical hierarchical forecasting benchmark—using the identical feature, fix, and recursive-forecast code. M5's 30,490 bottom item-store series aggregate to one grand total; aggregated to monthly, the total is 113$\times$ the largest single series in training, the same out-of-support condition.

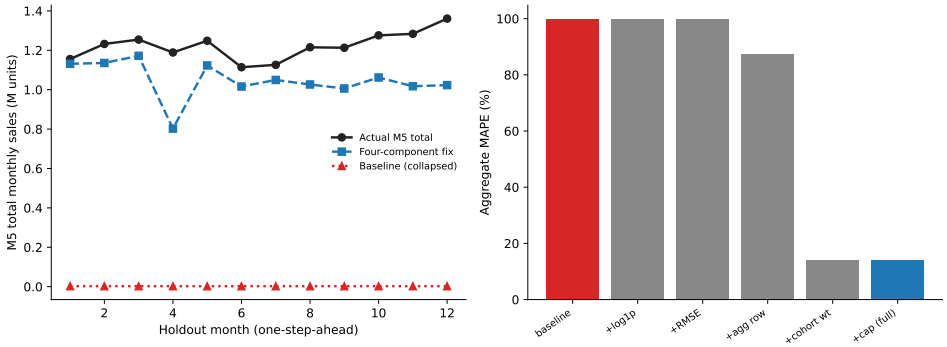

Figure 4: Collapse and fix replicated on public M5. Left: the baseline global tree collapses on the M5 grand total while the four-component fix tracks it. Right: the cumulative ablation, with the cohort-weighted aggregate row as the decisive lever—the same structure as the synthetic (Figure 3) and production panels.

(M5 is canonically a daily, 28-day-horizon, RMSSE-scored competition; we re-aggregate to monthly to match the monthly production cadence and to test the cross-scale mechanism, not competition-ranking accuracy; 113× is the top-level grand-total-to-largest-series ratio.) The naive global tree collapses to 99.8% aggregate MAPE (496× under-prediction), and the cumulative ablation reproduces the production structure exactly: the log transform and RMSE loss do nothing on their own (99.8% → 99.8%), the aggregate training row drops it to 87.5%, and the *cohort-weighted* aggregate row is again the decisive lever (→ 14.1%). The fix recovers LightGBM (14.1% one-step / 45.1% recursive-12mo) and XGBoost (6.3% / 10.3%) (Table 5, Figure 4); the 14.1% is the single 12-month holdout reported here, while the ten-origin rolling-mean LightGBM fix is 12.6% (Table 8).

Robust to metric and origin.. The M5 result is not an artifact of MAPE or of a single train/test split. Across three rolling origins, evaluated with the M-competition's own scaled metrics (seasonal MASE and one-step RMSSE), the baseline collapses on every origin and metric (MASE 8.5–12.4, i.e., ≈10× worse than a seasonal-naive forecast) and the four-component fix recovers on every origin (MASE 1.0–2.9). Crucially, the fix lands in the *naive-competitive band* (MASE ≈1–3), not below the naive baseline—confirming, in scaled-metric terms, that the contribution is *collapse-prevention*, not an aggregate-accuracy win (Section 5.11). The same holds on the production aggregate: under MASE, the shipped no-weight CatBoost stays collapsed (MASE 9.3)

while the cohort-weighted XGBoost (MASE 1.2) and LightGBM (MASE 3.8) recover—and, given the cohort weight, CatBoost recovers as well (Section 5.4).

5.7. A second public benchmark from another domain: Australian Tourism

To guard against M5 being a retail-specific coincidence, we replicate on a second public hierarchical benchmark from an unrelated domain—the Australian **Tourism** dataset of monthly visitor nights (Wickramasuriya et al., 2019), the canonical hierarchical set of the reconciliation literature. Its 304 bottom region× purpose-of-travel series sum to a national total that is ≈598× the largest single series in training—a cross-domain scale gap larger than M5's 113× and second only to UCI Online Retail II's 2,062× (Table 7). Using the identical code, the naive global tree collapses (96.2% one-step aggregate MAPE, seasonal MASE 16.4) and recovery again happens entirely at the cohort-weighted aggregate row (95.9% → 13.5% in the cumulative ablation; log and RMSE alone do nothing). The full fix recovers all three libraries (Table 6), and the collapse-and-fix holds across three rolling origins (baseline MASE 15.5–16.4 every origin; fix MASE 1.55–2.30).

5.8. A third public domain: a B2B buyer panel

Because the production panel is a field-service marketplace, we test whether the collapse is sector-specific by replicating on a public business-to-business buyer panel from an unrelated industry, country, and decade—UCI Online Retail II (Chen et al., 2012), a UK online retailer whose customers are largely wholesale businesses (5,878 business customers, monthly revenue, a ≈2,062× scale gap). The baseline collapses (98.7% one-step aggregate MAPE, MASE 9.5) and the four-component fix prevents it (36.8%, MASE 3.1), recovery again landing at the cohort-weighted aggregate row. The panel spans only 25 months—too short for strong seasonal modeling—so the fix prevents the collapse (a 2.7× error reduction) rather than reaching single digits; we report the one-step scale effect, not seasonal accuracy. The failure is thus a property of heavy-tailed business-buyer panels scored out of support, not of one sector.

The collapse and the library-agnostic fix thus reproduce on five datasets—synthetic (Section 5.5), production (Section 5.2), and three public datasets spanning three domains (M5 retail, Australian Tourism, and a B2B wholesale buyer panel; Table 7)—and under both percentage and scaled error metrics across rolling origins, so contribution C1 is a property of global tree forecasters scored out of support, not a single-panel, single-domain, or single-metric artifact, and the gap is statistically significant on every dataset (Section 5.9).

5.9. Statistical significance of the collapse–fix gap

The recovery is not confined to a few favorable origins. The fix wins on every rolling origin of every dataset, so the assumption-free Wilcoxon signed-rank test across origins rejects at its all-same-sign floor ($p \leq 10^{-3}$; Table 8). A forecasting-specific test agrees: pooling the per-step absolute-percentage-error loss differential across origins, the Diebold–Mariano statistic (Harvey–Leybourne–Newbold corrected) for the four-component fix against the uncorrected (collapsed) baseline tree is 57–251 one-step and 7–199 recursive. These magnitudes are large by construction: the collapsed baseline's per-step loss is near-constant at ≈98–100%, so the loss differential has little variance; we therefore read the statistic as an effect-size sanity check rather than a precise p-value, and rely on the all-origins sign result for inference. The recursive Diebold–Mariano statistic is smaller on M5 (7.1) than its one-step value (57.6) because on M5's steep grand total the fix's recursive forecast carries the leaf-ceiling re-collapse residual (Section 5.10), inflating its mean error to 39%—still far below the baseline's 99.8%. The all-origins sign result and the distribution-free Wilcoxon test thus carry the inference, with the large Diebold–Mariano magnitudes as a corroborating effect size; both agree that the collapse-and-fix gap is real.

5.10. Seed invariance of the collapse outcome

Of the twenty-three methods, eighteen are deterministic given the input (classical, foundation zero-shot point forecasts, intermittent, baseline) and carry exactly zero seed variance. The five stochastic methods are the three global trees and two neural models. Banding the three trees over eight seeds at the shipped configurations (Table 9), the *collapse-versus-no-collapse outcome is invariant*: the shipped no-weight CatBoost collapses on every seed ($91.9\% \pm 1.2$, CV 1.3%)—confirming the collapse is a stable property of withholding the cohort weight, not a seed draw—while the cohort-weighted XGBoost ($13.9\% \pm 2.6$) and LightGBM ($31.9\% \pm 12.0$) forecast the aggregate on every seed. Given the cohort weight, CatBoost likewise recovers on every configuration of the Section 5.4 sweep. LightGBM's point MAPE is seed-sensitive (range 15.7–49.8%), which—together with the configuration fragility noted in Section 10—is why we emphasize the mechanism rather than any single corrected-tree number.

5.11. The benchmark and the unit-of-analysis dichotomy

Table 4 reports the aggregate benchmark. Twelve methods cluster within $\approx 8.25\%$; with five holdout months no pairwise difference inside the band is statistically resolvable. The corrected trees are competitive-not-leading at the aggregate, and a simple decomposition method (Prophet, 4.16%) is nominally first, with the drift-adjusted baseline mid-pack (5.93%). This is the more forgiving level of the dichotomy: at the aggregate, the smoothness of the series renders the methods statistically indistinguishable.

The per-buyer level is the opposite. Twenty methods beat the baseline on per-buyer holdout MAPE (paired Wilcoxon, n=1,420); nineteen are statistically significant, and *all nineteen survive Benjamini–Hochberg FDR control across the twenty-two comparisons* (largest adjusted $p \approx 0.035$). The same robust leg holds under scale-free MASE—twenty methods beat the baseline on per-buyer holdout MASE and all twenty survive Benjamini–Hochberg control (n=1,659, largest adjusted $p \approx 2 \times 10^{-4}$)—so the dichotomy is not an artifact of the percentage metric, and is in fact sharper under the scale-free metric: Prophet, the *aggregate* leader (Section 5.11), is significantly worse than the baseline per buyer (Figure 5). The operational rule follows directly: choose the model class at the aggregation level the forecast is consumed on—a smooth aggregate forgives method choice, a lumpy per-buyer series does not.

6. Robustness

Table 10 maps each stress test to its outcome; the paragraphs that follow expand each row. The pattern is uniform: none of the single-axis M5-style interventions substitutes for restoring aggregate scale, and the collapse-and-fix result survives every probe.

Training-window upper bound.. Across ten rolling cutoffs, the aggregate optimum is 36–48 months (no failures); a 60-month window fails ($> 20\%$ aggregate MAPE) on 2/10 cutoffs (max 39.3%), and 67% of the failing ensemble slots are seasonal-differencing methods refitting over the pre-2025 low-growth regime. Production caps the window at 48 months. More history is a concentrated, intermittent tail risk here, not a monotone gain. Window length is moreover not a cure for the collapse: holding the holdout fixed and growing the training window from 36 to (where data allow) 120 months, the baseline collapses at every window length on production, M5, and Tourism alike (the collapse is an out-of-support scale problem, to which per-buyer history adds

no aggregate-scale rows), while the fixed tree's accuracy does not improve and on Tourism degrades monotonically (6.7% at 36 months to 12.0% at 120). Production's 60-month span precludes longer windows there; Tourism's 228 months confirm the pattern out to 120.

Foundation-model growth-regime bias.. On both an H2-2025 and a Jan–May 2026 origin, the six zero-shot foundation models under-forecast the growing aggregate (-11.3% and -3.3% mean signed bias)—they regress toward the mean and lag an accelerating series. The sharper "foundation-negative while intermittent/classical-positive" contrast holds only on the milder 2026 anchor; we scope it as regime-specific.

Entity identifiers, hyperparameters, and feature richness.. The collapse is not an artifact of an under-specified model. Adding the series identity (`buyer_id`) as a categorical feature—standard in M5-style global models—leaves the baseline collapsed (one-step 98–100% across panels): with no aggregate training row the aggregate is an unseen category the identifier cannot place in support, and with the row it adds nothing beyond the cohort weight. A sweep over leaves, depth, and boosting rounds (to 255 leaves, depth 12, 3,000 rounds) recovers in no configuration, and an enriched feature set (additional lags, a six-month rolling mean) does not help. A leaf cannot emit above the largest value it grew; the collapse is a property of training support, not of capacity.

Non-recursive forecasting and scale segmentation do not substitute for restoring scale.. Two structural choices from the M5 winners were tested directly. *Direct, non-recursive* per-horizon models—one model per step, the M5-winning structure—remove the recursive re-collapse (cohort-row M5 $45\% \to 10\%$, production $37\% \to 16\%$) but still require a scale cure to forecast the aggregate at all: non-recursion addresses multi-step drift, not the support gap, consistent with the trend-axis result (Section 5.3). *Segmenting by scale*—training only on the largest-decile buyers, those closest to the aggregate—still collapses (92–100%), because the aggregate exceeds the largest single buyer by 11.5 to 2,062×; no choice of training buyers places it in support.

Tweedie loss and post-hoc multipliers.. The M5-standard Tweedie objective does not cure the collapse on its own—its log link compresses scale like `log1p` but leaves the support gap, so bare-Tweedie one-step MAPE stays at the collapse level—and with the cohort row it does not improve on squared error,

the aggregate being a smooth high-level series rather than the zero-inflated count demand Tweedie targets. A leak-free bias multiplier (estimated on training-period one-step errors and applied to the holdout, in the M5 manner) corrects the cohort row's one-step level bias but barely changes its recursive error—locating the residual re-collapse as a slope, not a level, problem, which is exactly what seasonal differencing addresses (Section 5.3).

Prediction intervals are calibrated and proper-scored.. The benchmark is point-only, but the deployed system ships split-conformal prediction bands (Lei et al., 2018; Stankeviciute et al., 2021); we score them on the same pinned holdout with both empirical coverage (calibration) and the MSIS, the strictly proper interval score of the M4/M5 competitions (Gneiting and Raftery, 2007; Makridakis et al., 2020)—which rewards calibration and sharpness jointly, so a needlessly wide band cannot game it. Of the 1,444 scored buyers in the active forecast population, 1,181 (82%) carry a calibrated band; the 263 without one are short-history, low-volume buyers (fewer than ten conformal calibration residuals, $\approx 13\%$ of recent-year GTV), so the figures below describe the calibratable majority and are mildly optimistic for that tail. Across the 1,181, mean coverage is 0.82 at the nominal-80% band and 0.91 at 95%; the median buyer covers fully at both bands—the band is mildly conservative on the typical buyer—with a 10th percentile of 0.33/0.6 where the low-volume tail under-covers. This conservative direction is what the finite-sample quantile inflation at small calibration n (the $\lceil (n{+}1)\alpha \rceil / n$ correction) produces; we do not invoke the exact marginal-coverage guarantee, which assumes an exchangeability that this growing, regime-shifting aggregate violates (Tibshirani et al., 2019; Barber et al., 2023). The proper score is modest (median MSIS 4.0 and 7.1 at 80%/95%); its mean is inflated by the same low-volume tail where the seasonal-naive scale is near zero—a known fragility of scaled scores on intermittent series—so we summarize by the median. On the single aggregate series (n=5 monthly holdout points, hence noisy) the 95% band covers fully and the 80% band covers 3/5, with a tight MSIS of 2.1/2.3; at five points, 3/5 is the nearest attainable value below nominal and reflects discretization, not miscalibration. Intervals are thus measured and proper-scored, not asserted: the per-buyer band covers at or above nominal for the median buyer and is mildly conservative at the 80% band (0.82), with a slight shortfall at 95% (0.91) confined to the low-volume tail.

7. Reconciliation and forecast combination

Two standard alternatives to forecasting the aggregate as a single series are available here—hierarchical *reconciliation* (forecast each buyer, then enforce coherence with the total) (Hyndman et al., 2011; Wickramasuriya et al., 2019; Panagiotelis et al., 2023) and cross-method *combination* (average a pool of forecasts)—and neither beats the direct aggregate forecast. For all four production base methods, the direct (top-down) aggregate forecast beats both the bottom-up sum of independent per-buyer forecasts and minimum-trace shrinkage (MinT-shrink) reconciliation (Wickramasuriya et al., 2019) on the Jan–May 2026 holdout (Table 11): MinT pulls the incoherent bottom-up sum back toward the aggregate but never recovers the direct forecast, and the bottom-up sum is 2–3× worse.

Cross-sectional MinT and temporal THieF (Athanasopoulos et al., 2017) reconciliation fail for a reason measurable a priori: the off-diagonal error covariance these methods exploit is ≈ 0. On the forecast cohort the buyer×buyer base-forecast error correlation centers at +0.010 (95% CI $[+0.008, +0.012]$); a permutation null that destroys cross-correlation reproduces the observed magnitude almost exactly (real mean $|r| = 0.164$ vs. null 0.135), and an identical audit on a synthetic compound-symmetric panel with true ρ=0.40 recovers 0.39—so the audit detects reconcilable structure when it exists, and finds essentially none here (Figure 7). We therefore recommend running this buyer-covariance audit before investing in reconciliation.

The same near-diagonal error structure predicts that cross-method *combination*—whose equal-weight form is the classic hard-to-beat forecasting baseline (Bates and Granger, 1969; Claeskens et al., 2016)—is null here too, and it is. An equal-weight average of the eighteen methods the production system considers at the aggregate (the roster minus the five aggregate-gated learners) scores 8.17% MAPE (MASE 0.80, RMSSE 0.68) on the holdout—barely improving on its members' mean individual error (8.23%, a 0.06 pp gain) and far from the best single method (Prophet, 4.16%; Table 4). With per-series errors essentially uncorrelated the simple average has almost no diversification to exploit: combination is the within-level analogue of the cross-series reconciliation null, and the smooth aggregate—already forgiving of method choice (Section 5.11)—gains nothing from either.

8. Limitations

The benchmark and the unit-of-analysis dichotomy rest on a single proprietary marketplace with a five-month aggregate holdout; the per-buyer paired Wilcoxon (n=1,420) is the cross-sectionally robust leg, and aggregate claims rest on a short window. A second marketplace would generalize the sector benchmark and the dichotomy and remains future work—a public attempt is instructive but insufficient: on the only public two-sided marketplace we could obtain (the Olist Brazilian e-commerce dataset (Olist and Sionek, 2018)), the collapse-and-fix mechanism reconfirms, but its short, fast-growing aggregate does not exhibit the variance-cancellation smoothness the dichotomy turns on, indicating the dichotomy is a property of a mature marketplace and that generalizing these two legs requires a second mature marketplace, which is not public. The *collapse-and-fix mechanism* (C1) is on firmer ground: it reproduces on synthetic data and on three public datasets across three domains—M5 retail, Australian Tourism, and a B2B wholesale buyer panel (Section 5.6, Section 5.7, Section 5.8)—so it is not a single-panel or single-domain artifact, and a leave-one-out ablation isolates the load-bearing components (Section 5.2). The fix is library-agnostic (Section 5.4); we found no genuine cross-library boundary. The findings are empirical rather than formal theorems, but they rest on one stated mechanism—a piecewise-constant tree cannot emit a value above its largest training leaf (Section 9)—and are confirmed on public data, not a single-panel engineering anecdote. The qualitative conclusions transfer to panels sharing the sparse, short, growing profile of Section 3.

9. Conclusion and sector guidance

A global gradient-boosted-tree forecaster trained on a lumpy per-buyer panel and scored at a smooth aggregate orders of magnitude outside its training support collapses—a deployment-blocking failure the favorable global-model literature does not anticipate. The failure is fundamentally one of training-support scale, and is cured along either of two axes. The first restores scale, by two library-agnostic routes: per-series scaling—the conventional remedy, and the more accurate on one-step error—or a single-model cohort-weighted aggregate training row under squared-error loss, for deployments that keep one un-normalized model across levels. The second removes scale from the target by seasonal differencing, the only cure that does not recursively re-collapse.

The failure is structural—it reproduces on a synthetic panel and three public datasets (M5, Australian Tourism, and a B2B panel)—and seed-invariant.

Every result here reduces to one fact: a regression tree is a bounded, piecewise-constant function and cannot emit a value above the largest it saw in training, so the cure is to ensure the quantity it must predict—the level after scaling, or the change after differencing—always lies inside that training range. From this the practitioner procedure follows. (i) *Diagnose* cheaply: compute the ratio of the target aggregate to the largest single training series; a ratio far above one (here 11.5 to 2,062×) flags an out-of-support query, and a one-step teacher-forced check at the aggregate confirms a collapse in minutes. (ii) *Treat* along the axis the deployment needs: for one-step use, per-series scaling is the simplest and most accurate cure; where forecasts are rolled forward recursively, predict a seasonal difference, the only cure that stays stable; where one unnormalized model must serve every level, add a cohort-weighted aggregate row under squared-error loss; or adopt a natively scale-aware architecture (DeepAR-style RNN, N-HiTS, a foundation model). (iii) *Do not* expect the usual reflexes to substitute: a richer feature set, the series identifier, deeper or longer-trained trees, the Tweedie objective, a per-horizon (non-recursive) structure, scale-based segmentation, or a post-hoc multiplier each leaves the collapse or its recursive residual in place (Section 6). Finally, choose the model class at the aggregation level the forecast is consumed on: a smooth aggregate forgives method choice while a lumpy per-buyer series does not.

10. Reproducibility

Every result artifact was produced against one configuration: input vintage `gtv_by_month_by_buyer_20260614`, training cutoff Dec 2025, holdout Jan–May 2026, and the fix recipe. All artifacts lie on one linear, strictly additive commit history—later analyses add new write-ups and figures with no change to any forecasting recipe or the pipeline—and each embeds its commit SHA, a clean-tree flag, library versions, and its seed in a provenance sidecar. The synthetic-panel generator (Section 5.5) reproduces the collapse and fix without the proprietary data, alongside the M5 and Tourism replication scripts (both public, auth-free); each public number is reproducible from its recorded commit and seed, and a tagged snapshot of the analysis code accompanies the release. The public analysis code and the synthetic-panel generator accompany this manuscript as a public supplementary code capsule (the public and synthetic subset never touches the proprietary panel), and are

released publicly, archived with a Zenodo DOI, on publication. We note for honesty that an earlier draft reported a 5.4% aggregate MAPE for the corrected LightGBM ranked first in the benchmark; that figure depends on a regime-step exogenous indicator since removed from the shipped code and is not reproducible from current code. We therefore emphasize the collapse-prevention mechanism and the library-agnostic transfer rather than any single corrected-tree aggregate number.

Leakage scope.. The collapse-and-fix results (C1–C3) are computed by retraining each tree on the training window only (train $\leq$ Dec 2025) and scoring the held-out window, on the synthetic, M5, Tourism, and production panels alike; they do not depend on the operational cross-validation method-selection path. The per-buyer dichotomy itself (C4, Section 5.11) shares this leak-free footing: it pairs each method's forward forecast—fit on the same train $\leq$ Dec 2025 window and scored on the Jan–May 2026 holdout—against the baseline per buyer, and never consults the cross-validation winner, so the selection-path leakage described next does not reach it. The per-buyer critical-difference diagram (Figure 5) ranks these same forward holdout forecasts and inherits the same footing: its global and neural rows score each model's forecast of the Jan–May 2026 holdout, fit on data through Dec 2025 and therefore temporally out-of-sample (the full-panel fit below is cross-sectional—all buyers, training months only), so the diagram's rankings are untouched by the selection-path leak. A separate known leakage in that selection path (the fitted global/neural models are fit once on the full panel—including the validation folds the selection later scores them on—during cross-validation-winner ranking) therefore does not affect the paper's central results; it would, if uncorrected, bias only the per-buyer selection comparison, which we flag for the benchmark leg and is the subject of a separate, eval-gated fix.

Declaration of competing interest

The authors are employed by Field Nation LLC, the marketplace operator whose anonymized data is used in this study. The authors declare no other competing financial or personal interests. Company approval was obtained for publication.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Md Rezwanul Islam: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Wael Mohammed:** Conceptualization, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used a large language model (Claude, Anthropic) to assist with drafting and editing the prose and LaTeX. After using this tool the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

The underlying transaction data are proprietary and cannot be released. The core collapse-and-fix result is nonetheless independently verifiable on fully public data: it replicates on the M5 competition dataset (Section 5.6) and the Australian Tourism hierarchical dataset (Section 5.7)—both openly available through the `datasetsforecast` package—on the UCI Online Retail II B2B wholesale panel (Chen et al., 2012), and on a synthetic hierarchical-panel generator (Section 5.5) that reproduces the collapse and the four-component fix with no proprietary data; the mechanism additionally reconfirms on the public Olist marketplace dataset (Olist and Sionek, 2018). The Diebold–Mariano significance test (Section 5.9) likewise runs on these public datasets. The analysis code, the synthetic generator, and the provenance-stamped result artifacts are released at a pinned commit; see Section 10.

References

Ansari, A.F., et al., 2024. Chronos: Learning the language of time series. Transactions on Machine Learning Research ArXiv:2403.07815.

Athanasopoulos, G., Hyndman, R.J., Kourentzes, N., Petropoulos, F., 2017. Forecasting with temporal hierarchies (THieF). Eur. J. Oper. Res. 262, 60–74.

Barber, R.F., Candès, E.J., Ramdas, A., Tibshirani, R.J., 2023. Conformal prediction beyond exchangeability. Ann. Statist. 51, 816–845.

Bates, J.M., Granger, C.W.J., 1969. The combination of forecasts. J. Oper. Res. Soc. 20, 451–468.

Bojer, C.S., Meldgaard, J.P., 2021. Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting 37, 587–603. doi:10.1016/j.ijforecast.2020.07.007.

Chen, D., Sain, S.L., Guo, K., 2012. Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management 19, 197–208. doi:10.1057/dbm.2012.17. source publication for the UCI Online Retail / Online Retail II dataset (UK wholesale online retailer).

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system, in: Proc. ACM SIGKDD, pp. 785–794.

Claeskens, G., Magnus, J.R., Vasnev, A.L., Wang, W., 2016. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting 32, 754–762. doi:10.1016/j.ijforecast.2015.12.005.

Das, A., Kong, W., Sen, R., Zhou, Y., 2024. A decoder-only foundation model for time-series forecasting (TimesFM), in: Proc. ICML.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13, 253–263.

Gardner, E.S., McKenzie, E., 1985. Forecasting trends in time series. Management Science 31, 1237–1246. doi:10.1287/mnsc.31.10.1237.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102, 359–378.

Han, Y., Ma, R., Zhang, Q., et al., 2023. Fast forecasting of unstable data streams for on-demand service platforms. arXiv preprint arXiv:2303.01887 .

Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. International Journal of Forecasting 13, 281–291. doi:10.1016/S0169-2070(96)00719-4.

Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L., 2011. Optimal combination forecasts for hierarchical time series. Computational Statistics & Data Analysis 55, 2579–2589. doi:10.1016/j.csda.2011.03.006.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecasting 22, 679–688.

IBM Research, 2025. Granite Time Series — FlowState R1 model card. Hugging Face. https://huggingface.co/ibm-granite/granite-timeseries-flowstate-r1.

Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., Callot, L., 2020. Criteria for classifying forecasting methods. International Journal of Forecasting 36, 167–177. doi:10.1016/j.ijforecast.2019.05.008.

Ke, G., et al., 2017. LightGBM: A highly efficient gradient boosting decision tree, in: Adv. Neural Inf. Process. Syst. (NeurIPS).

Koning, A.J., Franses, P.H., Hibon, M., Stekler, H.O., 2005. The M3 competition: Statistical tests of the results. International Journal of Forecasting 21, 397–409. doi:10.1016/j.ijforecast.2004.10.003.

Lainder, A.D., Wolfinger, R.D., 2022. Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies. International Journal of Forecasting 38, 1426–1433. doi:10.1016/j.ijforecast.2021.12.003.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L., 2018. Distribution-free predictive inference for regression. Journal of the American Statistical Association 113, 1094–1111. doi:10.1080/01621459.2017.1307116.

Liu, S., et al., 2024. TimerBench: A comprehensive benchmark of time-series foundation models. arXiv preprint arXiv:2410.10393 .

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020. The M4 competition: 100,000 time series and 61 forecasting methods. Int. J. Forecasting 36, 54–74.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2022. The M5 accuracy competition: results, findings and conclusions. Int. J. Forecasting 38, 1346–1364.

Malistov, A., Trushin, A., 2019. Gradient boosted trees with extrapolation, in: 18th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE. pp. 783–789. doi:10.1109/ICMLA.2019.00138.

März, A., Rasul, K., 2024. Forecasting with hyper-trees. arXiv preprint arXiv:2405.07836. doi:10.48550/arXiv.2405.07836.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R.J., Talagala, T.S., 2020. FFORMA: Feature-based forecast model averaging. Int. J. Forecasting 36, 86–92.

Montero-Manso, P., Hyndman, R.J., 2021. Principles and algorithms for forecasting groups of time series: Locality and globality. International Journal of Forecasting 37, 1632–1653. doi:10.1016/j.ijforecast.2021.03.004.

Olist, Sionek, A., 2018. Brazilian e-commerce public dataset by Olist. Kaggle. doi:10.34740/KAGGLE/DSV/195341. public two-sided e-commerce marketplace orders, 2016–2018.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., Hyndman, R.J., 2023. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. European Journal of Operational Research 306, 693–706. doi:10.1016/j.ejor.2022.07.040.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features, in: Adv. Neural Inf. Process. Syst. (NeurIPS).

Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T., 2020. DeepAR: probabilistic forecasting with autoregressive recurrent networks. Int. J. Forecasting 36, 1181–1191.

Stankeviciute, K., Alaa, A.M., van der Schaar, M., 2021. Conformal time-series forecasting, in: Adv. Neural Inf. Process. Syst. (NeurIPS).

Tan, M., Merrill, M., Gupta, V., Althoff, T., Hartvigsen, T., 2024. Are language models actually useful for time series forecasting?, in: Adv. Neural Inf. Process. Syst. (NeurIPS).

Tibshirani, R.J., Barber, R.F., Candès, E.J., Ramdas, A., 2019. Conformal prediction under covariate shift, in: Adv. Neural Inf. Process. Syst. (NeurIPS).

Wang, X., Smith, K., Hyndman, R.J., 2006. Characteristic-based clustering for time series data. Data Min. Knowl. Discov. 13, 335–364.

Wickramasuriya, S.L., Athanasopoulos, G., Hyndman, R.J., 2019. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. J. Amer. Statist. Assoc. 114, 804–819.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bull. 1, 80–83.

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., Sahoo, D., 2024. Unified training of universal time series forecasting transformers (Moirai), in: Proc. ICML.

Yu, B., Ma, S., Tang, R., Yang, J., Li, Z., 2022. GAIA: Graph neural network with temporal shift aware attention for gross merchandise value forecast in e-commerce, in: Proc. IEEE Int. Conf. Big Data, pp. 3262–3271.

Zhang, C., Tian, Y., Zhao, J., et al., 2020. How much can a retailer sell? Sales forecasting on Tmall. arXiv preprint arXiv:2002.11940 .

Zhao, Y., Abolghasemi, M., 2024. Local vs. global models for hierarchical forecasting. arXiv preprint arXiv:2411.06394. doi:10.48550/arXiv.2411.06394.

Zhu, L., Laptev, N., 2017. Deep and confident prediction for time series at Uber, in: Proc. IEEE Int. Conf. Data Mining Workshops, pp. 103–110.

Table 4: Aggregate benchmark on the Jan–May 2026 holdout (five-month, multi-step; production benchmark at git `313ffa6`, vintage 20260614). All twenty-three roster methods are scored at the aggregate. At the aggregate everything ties and a decomposition method (Prophet) is nominally first—the corrected trees are competitive-not-leading. The column is MAPE over the five-month holdout, a multi-step forecast for the recursive ML methods: the cohort-weighted LightGBM holdout value (36.8%) carries the recursive re-collapse of Section 5.3; its one-step aggregate MAPE is 15.1% (Table 2). The CatBoost row is the shipped configuration, which omits the cohort weight; supplying it recovers CatBoost to ≈15% (Section 5.4)—the 92.9% is a deployment artifact, not a library limit.

Rank	Method	Family	Agg. MAPE	MASE	RMSSE
1	Prophet	classical	**4.16%**	**0.41**	**0.46**
2	AutoETS	classical	5.63%	0.59	0.58
3	MSTL	classical	5.74%	0.56	0.51
4	Naive YoY + Mom. (baseline)	baseline	5.93%	0.58	0.51
5	AutoTheta	classical	5.94%	0.61	0.60
6	FlowState	foundation	6.38%	0.65	0.59
7	TimesFM	foundation	6.78%	0.69	0.61
8	Moirai-2	foundation	6.94%	0.66	0.58
9	Chronos-2	foundation	6.95%	0.72	0.68
10	STL seasonal-trend	classical	7.23%	0.73	0.64
11	Chronos-Bolt	foundation	7.42%	0.81	0.83
12	Chronos v1	foundation	8.25%	0.84	0.76
13	Linear trend	classical	10.8%	0.96	1.05
14	**XGBoost** (cohort-weighted)	ml-tree	**11.5%**	**1.23**	**1.22**
15	IMAPA	intermittent	11.7%	1.06	1.06
16	EWMA-3M	baseline	11.9%	1.11	1.00
17	N-HiTS	ml-neural	12.0%	1.25	1.18
18	Croston classic	intermittent	12.1%	1.16	0.99
19	TSB	intermittent	12.1%	1.16	0.99
20	Croston-SBA	intermittent	12.5%	1.26	1.08
21	DeepAR[†]	ml-neural	12.5%	1.31	1.25
22	**LightGBM** (cohort-weighted)	ml-tree	**36.8%**	**3.81**	**3.41**
23	**CatBoost** (no cohort weight)	ml-tree	**92.9%**	**9.33**	**7.80**

[†] DeepAR is reported as the median of five seeds (MAPE range 6–15%): it is the one row we band, because its sampled-path head produced material aggregate seed variance and a single draw is unreliable. N-HiTS, a point forecaster without that sampling head, is reported at a single fixed seed (unbanded), a minor asymmetry we flag; the remaining rows are deterministic or fixed-seed. N-HiTS and DeepAR forecast the aggregate series directly—their built-in per-series scaling is the escape mechanism (Section 5.1)—and are re-scored here on the same pinned config as every other row.

Table 5: Public replication on M5 (30,490 monthly item-store series, 12-month holdout). One-step aggregate MAPE. The collapse and the four-component fix reproduce; the cohort-weighted aggregate row is the load-bearing lever, exactly as on the synthetic and production panels. The fix rescues CatBoost here too (16.8%); given the cohort weight, CatBoost recovers on production as well (Section 5.4).

Configuration / library	Baseline	Four-component fix
LightGBM (ablation)	99.8%	**14.1%**
XGBoost	99.9%	**6.3%**
CatBoost	99.8%	**16.8%**

Table 6: Second public replication on Australian Tourism (304 monthly series, 12-month holdout, ≈598× gap). One-step aggregate MAPE / seasonal MASE. The collapse and the library-agnostic fix reproduce in a non-retail domain.

Configuration / library	MAPE	MASE
Baseline (LightGBM, no fix)	96.2%	16.4
Full fix — LightGBM	**14.0%**	2.30
Full fix — XGBoost	**5.2%**	0.90
Full fix — CatBoost	**11.0%**	1.84

Table 7: Generalization of the collapse and the four-component fix across all five panels, ordered by scale gap. One-step aggregate MAPE, baseline (global tree, no fix) versus the four-component fix; the *scale gap* is the aggregate target divided by the largest single training series (for the production panel, the aggregate-to-largest-buyer ceiling gap g of Eq. (4); its aggregate-to-median-buyer ratio is the larger ≈1,088× out-of-support gap). The collapse (baseline $\gtrsim$90%, an order-of-magnitude under-prediction) and its recovery hold across a ∼180-fold range of scale gaps (11.5× to 2,062×) spanning five domains. Production, M5, and Tourism are rolling-origin means (Table 8)—so the production fix here (9.5%) is the twelve-origin mean, distinct from the 15.1% single pinned-holdout value carried by the ablation tables (Table 2); Synthetic is a six-seed mean and UCI Online Retail II a single 25-month holdout (no rolling origins available). Scaled-metric (seasonal MASE / RMSSE) and significance confirmation are in Section 5.9 and the per-dataset tables; the mechanism additionally reconfirms on the public Olist marketplace (Section 8).

Panel	Domain	Scale gap	Baseline	Full fix
Synthetic	hierarchical sim.	11.5×	94.7%	**17.5**%
Production	B2B field service	≈48×	97.9%	**9.5**%
M5	retail (Walmart)	113×	99.8%	**12.6**%
Australian Tourism	visitor nights	598×	95.9%	**11.1**%
UCI Online Retail II	B2B wholesale	2,062×	98.7%	**36.8**%

Table 8: The aggregate collapse–fix gap is statistically significant on every dataset. Diebold–Mariano statistic (Harvey–Leybourne–Newbold small-sample correction) for the four-component fix versus the uncorrected (collapsed) baseline tree, on the per-step absolute-percentage-error loss differential pooled across rolling origins; a positive value favors the fix (read as an effect size—see Section 5.9). The Wilcoxon column is the distribution-free signed-rank test across origins and carries the inference. Mean one-step aggregate MAPE (baseline→fix), averaged over the listed rolling origins, and the origins on which the fix wins are shown for scale.

Dataset	Rolling origins	Diebold–Mariano one-step	Diebold–Mariano recursive	Wilcoxon p	Mean MAPE base→fix
Production	12	179.5	36.4	2×10^{-4}	97.9% → 9.5%
Tourism	24	250.5	199.2	6×10^{-8}	95.9% → 11.1%
M5	10	57.6	7.1	1×10^{-3}	99.8% → 12.6%

Table 9: Eight-seed variance band on the stochastic tree models, pinned Jan–May 2026 aggregate MAPE. Mechanism (collapse vs. not) is seed-invariant; LightGBM's point MAPE is seed-sensitive.

Model	Mean±sd	Range	Reading
CatBoost (no cohort wt.)	91.9 ± 1.2	89.1–92.9	collapses every seed
XGBoost (cohort-wt.)	13.9 ± 2.6	11.1–18.3	stable, never collapses
LightGBM (cohort-wt.)	31.9 ± 12.0	15.7–49.8	forecasts agg. every seed

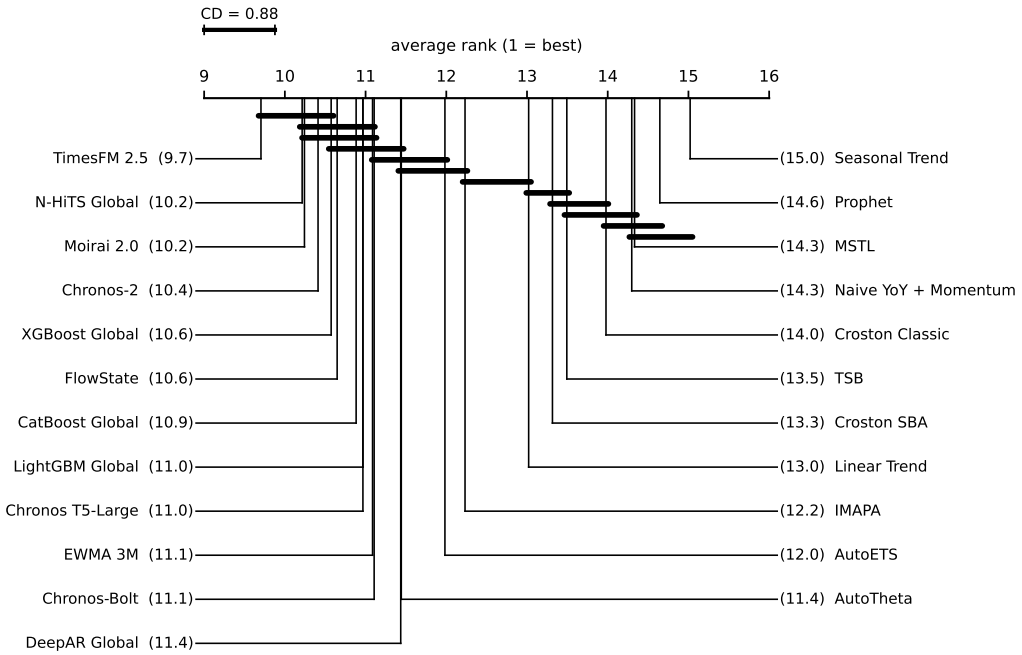

Figure 5: Per-buyer critical-difference diagram over all twenty-three methods (Demšar, 2006; Koning et al., 2005), using scale-free per-buyer MASE ranks on the n=1,564 buyers scored by every method (Friedman χ^2_{22}=2023, p numerically 0). Each method is placed at its mean rank (1= best, i.e., lowest MASE); a horizontal bar joins methods whose mean-rank gap falls below the Nemenyi critical distance (CD=0.88, α=0.05) and are therefore not statistically separable. The cross-learning tree and neural learners and the zero-shot foundation models hold the best per-buyer ranks, while the drift-adjusted baseline sits near the bottom (20th of 23) and Prophet—the *aggregate* leader (Table 4, Section 5.11)—is among the worst per buyer. This is the per-buyer half of the dichotomy: methods that tie at the smooth aggregate separate sharply on the lumpy per-buyer series.

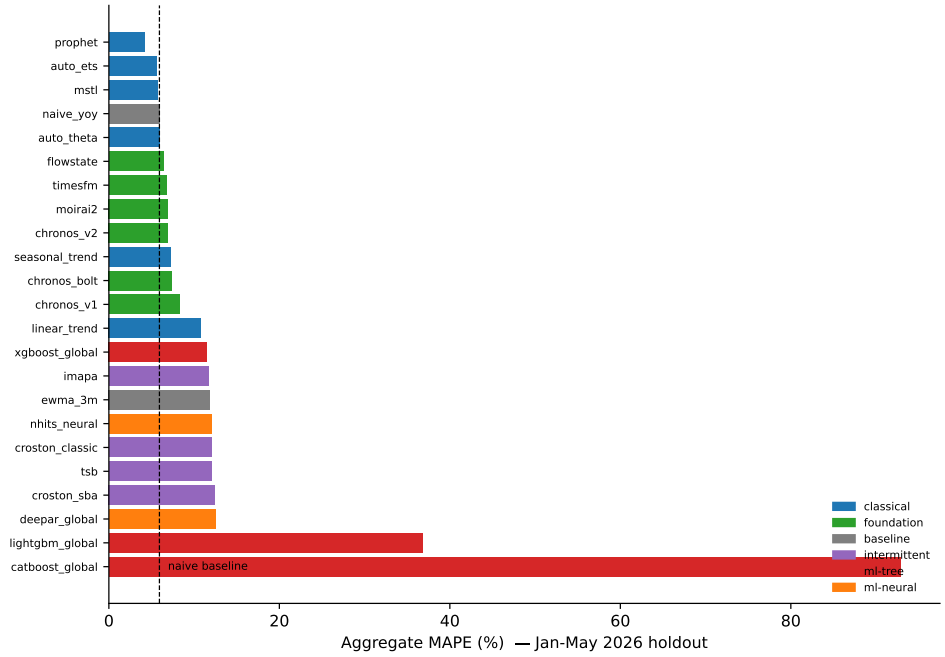

Figure 6: Pinned aggregate benchmark (Jan–May 2026). The shipped no-weight CatBoost bar (collapsed) beside the cohort-weighted trees and the tied competitive cluster is the visual statement of the dichotomy; supplying CatBoost the cohort weight recovers it to ≈15% (Section 5.4).

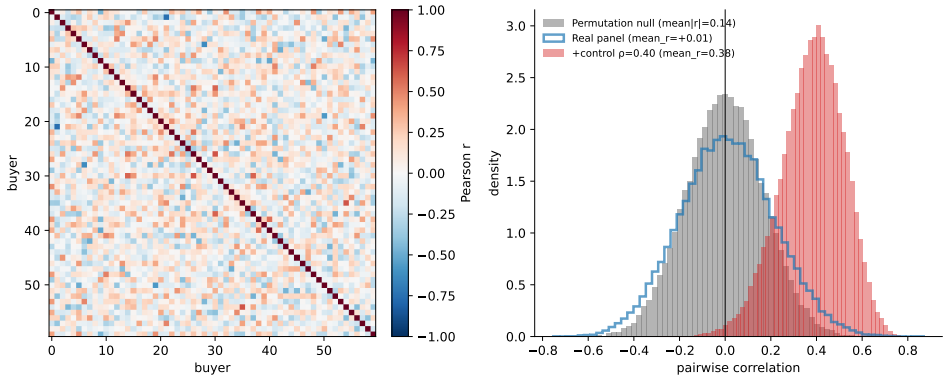

Figure 7: Off-diagonal error-correlation distribution (real vs. permutation null vs. ρ=0.40 positive control). The real panel sits on the null—no exploitable cross-series structure—explaining the measured reconciliation null.

Table 10: Robustness map: each probe and its outcome. The collapse and the four-component fix survive every stress test; no single-axis intervention substitutes for restoring aggregate scale. The paragraphs below expand each row in order.

Probe / intervention	Outcome
Training-window length (36–120 mo)	Not a cure: the baseline collapses at every length; 60 mo adds intermittent tail risk.
Foundation models, growth regime	Under-forecast the accelerating aggregate ($-11.3\%/-3.3\%$ signed bias); scoped as regime-specific.
Entity id + hyperparameter sweep + richer features	Still collapsed (98–100%): a training-support property, not a capacity one.
Direct (non-recursive) horizons; largest-decile segmentation	Remove recursive drift but still need a scale cure; the aggregate exceeds the largest buyer 11.5–2,062×.
Tweedie loss; leak-free bias multiplier	Neither cures: the residual re-collapse is a slope, not a level, problem (seasonal differencing).
Prediction intervals (split-conformal)	Calibrated (0.82/0.91 coverage at 80/95%) and proper-scored (mean scaled interval score, MSIS), not asserted.

Table 11: Reconciliation is null at the aggregate. For each production base method, the direct (top-down) aggregate forecast beats both the bottom-up sum of independent per-buyer forecasts and cross-sectional MinT-shrink reconciliation on the Jan–May 2026 holdout (MAPE). MinT improves on the incoherent bottom-up sum but never recovers the direct forecast. Production benchmark, vintage 20260614, cutoff Dec 2025; direct and MinT at git `313ffa6`, the bottom-up column an additive descendant.

Base method	Direct	MinT	Bottom-up
AutoETS	**5.63%**	8.29%	13.82%
MSTL	**5.74%**	7.00%	15.17%
Naive YoY	**5.93%**	7.98%	17.54%
FlowState	**6.38%**	10.15%	19.70%