FEDTINYTS: AN ULTRA-LOW-COMMUNICATION ONE-SHOT FEDERATED LEARNING FRAMEWORK FOR LONG-TERM TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

One-shot analytic federated learning in long-term time series forecasting remains challenging due to both high communication costs and limited local adaptation. These methods transmit dense moment matrices whose size grows quadratically with input length, while directly applying the collaboratively estimated knowledge to all clients can underperform local-only training by overlooking client-specific information. We propose FedTinyTS, a one-shot analytic protocol for a channel-independent linear forecasting backbone that removes that quadratic growth, exchanging compact summaries of temporal dependencies in a single round and fitting the model analytically rather than through repeated federated optimization. Because compression can reduce global forecasting performance, FedTinyTS uses reliability-aware aggregation to preserve the quality of the shared model. It further adapts the collaboratively learned knowledge to each client using locally retained information, improving personalized performance without further communication. Across five benchmarks and four horizons, FedTinyTS reduces the message payload by up to $8{,}548\times$ compared with FEDAVG and up to $90.7\times$ compared with dense FEDRIDGE. Averaged across horizons, the global model achieves lower mean squared error than FEDAVG on every benchmark but SolarEnergy, and personalization lowers it further on the same four, by up to 3.75%. The dense analytic baselines, which never adapt to a client, instead fall behind a purely local fit by up to 6.31%. Mean absolute error is less favorable, a gap the dense analytic baselines share. Compact temporal summaries therefore retain enough structure to fit a federated forecaster in one round. Code is available at https://anonymous.4open.science/r/FedTinyTS.

1 INTRODUCTION

Long-term time series forecasting (LTSF) supports applications such as energy management (Bahman & Zareipour, 2025; Zimmermann & Ziel, 2025; Kumar & Mandal, 2024), weather services (Zhao et al., 2024; Peings et al., 2026; Xue et al., 2025), and transportation planning (Kara et al., 2025; Shao et al., 2025; Zheng et al., 2025), where observations are often distributed across many clients. Federated learning (FL) lets those clients train a shared forecasting model without centralizing their raw series (McMahan et al., 2017; Li et al., 2020). What they exchange instead is high-dimensional model parameters, transmitted every round. For LTSF backbones, large parameter sizes combined with repeated communication rounds can incur substantial communication costs. Although linear forecasting models provide competitive backbones for LTSF (Zeng et al., 2023), federated collaboration still requires transmitting either model updates or client-side statistics.

To reduce the repeated communication required by iterative FL, one-shot analytic methods eliminate repeated model updates and thus avoid multiple rounds of exchange. However, their statistical summaries can still impose substantial communication costs. Existing analytic approaches communicate statistical summaries from which a global model is estimated directly (Alsulaimawi, 2026; Zhu et al., 2021). Although these approaches eliminate repeated optimization rounds, the required summaries can still be substantial. Their size grows quadratically with the history length, while cross-moment terms grow with both the history length and forecast horizon, making communication increasingly costly for long-horizon forecasting. This motivates compressing the transmitted information while

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

preserving the temporal dependencies required for accurate forecasting. Thus, the resulting global model is directly applied to all clients after aggregation, limiting its ability to capture client-specific temporal patterns. Effective federated LTSF therefore requires both communication-efficient temporal summaries and local adaptation that allows clients to recover information not captured by the shared representation.

A communication-efficient federated LTSF framework must satisfy three requirements in turn: **(1) Compact temporal summaries.** The transmitted information must be compressed far enough to matter, yet still carry the temporal relationships forecasting depends on. Because entries relating different time positions share the same lag, a summary can grow additively with history length and horizon. **(2) Reliable global aggregation.** Compression can reduce global forecasting performance by discarding information from the dense statistics, while differences in client sample support can further affect the reliability of the compressed summaries. Aggregation must therefore account for differences in client sample support and scale reliability to stabilize the global model. **(3) Communication-free personalization.** A single shared model causes limited local adaptation, as it cannot incorporate client-specific information that is absent from the shared summaries. Each client must therefore refine the shared model using locally retained information that is absent from the transmitted summaries, without further communication. This preserves client-specific temporal structure while still using knowledge learned across the federation.

Motivated by these requirements, we present FedTinyTS, a one-shot analytic protocol that reduces both the communication cost and the completion time of federated forecasting relative to iterative federated training. Its three phases take the three requirements in turn. **Phase 1: Robust Local Compression** builds the compact temporal summaries. Each client reduces the influence of large response residuals, retains its dense robust moments, and uploads two compact lag profiles with a spectral reliability scale and a training-row count. **Phase 2: Analytic Global Solve** makes aggregation reliable. Capping extreme precision scores, the server pools the profiles and sets the global ridge penalty from the pooled scale and row count before solving for the shared forecaster. **Phase 3: Communication-Free Personalization** adapts that shared model to each client. After one broadcast, each client combines the pooled profiles with its retained moments to obtain a personalized forecaster without further exchange. Our contributions are as follows:

- We identify two challenges in one-shot analytic FL for LTSF, namely dense analytic statistics that incur high communication costs and limited local adaptation of the shared model. Addressing both motivates a framework that compresses the temporal structure once, retaining what global forecasting requires and leaving each client what it needs to adapt.
- We introduce a lag-profile representation (Phase 1), an aggregation that reconstructs the solve's matrices from one pooled profile rather than a dense pair per client (Phase 2), and personalization from moments never sent (Phase 3). Under explicit assumptions we characterize the population target, its sampling convergence, and the compression error.
- Across five benchmarks and four forecasting horizons, FedTinyTS cuts payload by up to $8{,}548\times$ relative to FedAvg and $90.7\times$ relative to dense FedRidge. Averaged across horizons, global MSE improves over FedAvg on four of the five benchmarks, and communication-free personalization lowers that error further on the same four.

2 Related Work

General FL. FL methods divide into two broad approaches according to how they exchange client information, namely iterative model exchange and analytic statistic aggregation. The first approach learns iteratively, alternating local training with server aggregation through model averaging or local regularization (McMahan et al., 2017; Li et al., 2020). Its communication-oriented variants compress each message through quantization or layer-wise update recycling (Fang et al., 2026; Kim et al., 2025). These mechanisms address update size and training behavior. The repeated synchronization the schedule is built on goes untouched. The other approach fits models from aggregated statistics, using closed-form classifiers over pretrained representations (He et al., 2025), dual-stream estimation for personalization (Fan et al., 2025), or layer-wise residual learning for deeper representations (Tang et al., 2026). Analytic estimation can reduce synchronization, but layer-wise training need not conclude within a single communication round. Dense statistics carry a second cost, in the

size of each message rather than their number. Neither one-shot estimation nor analytic communication reduction is therefore unique to our work.

FL for TS forecasting. Within forecasting, a complementary line of research focuses on temporal structure and client heterogeneity. Fed-TREND constructs auxiliary synthetic data from client updates and global model trajectories to improve training across heterogeneous series (Yuan et al., 2026). FedTime instead combines client clustering, channel-independent patching, and parameter-efficient fine-tuning of a pretrained large language model (Abdel-Sater & Ben Hamza, 2024). Both retain a parameter-sharing backbone, so synthetic data and pretrained representations support iterative learning rather than replacing it with a one-shot analytic solve. Benchmarking frameworks such as FedProC separate forecasting backbones, client datasets, and federated strategies to support systematic comparison (Nguyen & Kim, 2026). That separation is what isolates a communication claim. Gains from a stronger backbone or a heterogeneity correction do not alone establish a smaller communication requirement. We therefore fix a linear backbone, which isolates the communication comparison from backbone capacity, and within that setting we target both exchange frequency and summary size. In this setting, DLSA (Zhu et al., 2021) and FedRidge (Alsulaimawi, 2026) provide direct analytic baselines, aggregating local least-squares approximations and sufficient statistics, respectively. Their dense formulations require large matrix payloads, although FedRidge also considers random projections to reduce dimension. Our method instead exchanges compact temporal lag summaries in place of dense matrices, while retaining local moments for personalization from the same broadcast. The claim concerns the representation and schedule of communication rather than any suggestion that analytic forecasting supplants general nonlinear FL.

3 Proposed Method

Our method considers N clients, each holding a local multivariate time series. Client i pairs an input window of length L with the following H observations from the same channel. Stacking these examples across channels gives $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times L}$ and $\boldsymbol{Y}_i \in \mathbb{R}^{n_i \times H}$, where n_i counts channel-window pairs. A channel-independent linear forecaster $\boldsymbol{\Theta} \in \mathbb{R}^{L \times H}$ predicts $\widehat{\boldsymbol{Y}}_i = \boldsymbol{X}_i \boldsymbol{\Theta}$ with shared coefficients across channels and zero bias. Its local Gram and cross-moments, $\mathcal{S}_{xx,i} = n_i^{-1} \boldsymbol{X}_i^\top \boldsymbol{X}_i$ and $\mathcal{S}_{xy,i} = n_i^{-1} \boldsymbol{X}_i^\top \boldsymbol{Y}_i$, have sizes $L \times L$ and $L \times H$ respectively. We use $\| \cdot \|_F$ for the Frobenius norm, $\| \cdot \|_2$ for the Euclidean vector or spectral matrix norm, and $\boldsymbol{I}$ for an identity matrix whose dimension is implied by context.

Unlike one-shot analytic methods that directly transmit dense moment matrices for global estimation, FedTinyTS first compresses, on each client, the information required for forecasting into compact temporal summaries (Phase 1). This compression reduces the communication payload but also discards position-specific information, which can reduce global forecasting performance. The subsequent reliability-aware aggregation on the server (Phase 2) therefore preserves the quality of the shared model obtained from the global analytic solve, while the final client-side phase (Phase 3) adapts that model to each client using locally retained information. In this way, FedTinyTS addresses both high communication costs and limited local adaptation within a single exchange. As illustrated in Figure 1, in Phase 1 each client uploads the lag profiles $(\boldsymbol{r}_i, \boldsymbol{s}_i)$, a spectral scale $\hat{\sigma}_i^2$ as a reliability proxy, and the training-row count n_i as sample support, totaling $L + H + 2$ values, while the dense moments $(\boldsymbol{G}_i, \boldsymbol{C}_i)$ stay on the client for later adaptation. In Phase 2 the server broadcasts the pooled profiles $(\boldsymbol{r}_g, \boldsymbol{s}_g)$ rather than the global coefficients $\boldsymbol{\Theta}_g$ it fits from them, totaling $L + H$ values per client, and Phase 3, run locally on each client, adds nothing further, so the position-specific detail omitted from the upload reaches the personalized forecaster $\boldsymbol{\Theta}_i$ without another exchange. The complete exchange contains $2L + 2H + 2$ values per client, independent of the number of local fitting updates. The following subsections detail the three phases.

3.1 Phase 1: Robust Local Compression

Phase 1 runs on each client and builds the compact temporal summaries, which must be compressed far enough to matter yet still carry the temporal relationships forecasting depends on. The client first reduces the influence of large response residuals through robust reweighting, so that its dense robust moments are not dominated by a few anomalies. It then compresses those moments into two compact lag profiles indexed by time separation rather than by position pair, so the summary grows

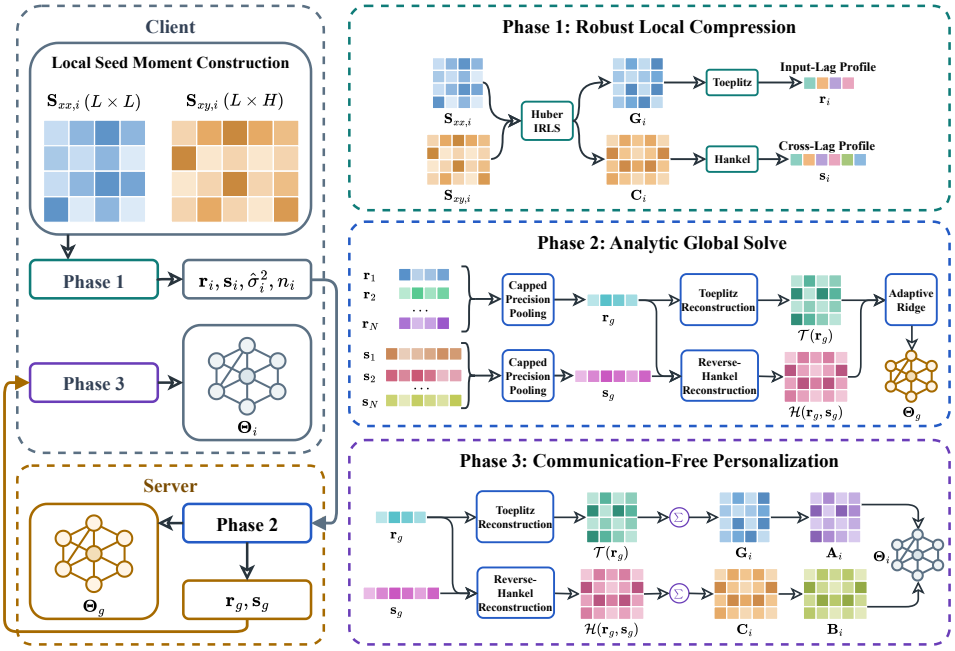


Figure 1: FedTinyTS in three phases. Phase 1 constructs robust local moments, extracts lag profiles, and estimates a spectral reliability scale. Phase 2 pools the profiles with capped precision weights and solves for a global model. Phase 3 combines the broadcast profiles with retained dense local moments for client-adaptive personalization. Each client uploads $L+H+2$ values once and receives $L+H$, with no further exchange.

additively with history length and horizon. It finally computes a spectral reliability scale and uploads it with the profiles and its training-row count, while retaining the dense robust moments for Phase 3.

Robustness comes first, because a single badly predicted window can distort many values that are later combined and shared. The client therefore uses Huber-type iteratively reweighted least squares (IRLS) (Huber, 1964), which fits a ridge-regularized local model, downweights windows with large response residuals, and refits. At each iterate, the client compares the size of each residual with a cutoff, set a fixed multiple of a robust scale above the median residual size. This robust scale is a floored median absolute deviation (Rousseeuw & Croux, 1993), so a few large residuals cannot inflate the cutoff meant to catch them. Windows at or below the cutoff keep full weight, while windows above it receive weights inversely proportional to their residual size. Writing $w_t^{(k)}$ for the weight of row t at iterate k, and $\boldsymbol{x}_t \in \mathbb{R}^L$ and $\boldsymbol{y}_t \in \mathbb{R}^H$ for its input window and target, the client recomputes two weighted moments over its n_i rows,

$$\boldsymbol{G}_i^{(k+1)} = \frac{1}{n_i} \sum_t w_t^{(k)} \boldsymbol{x}_t \boldsymbol{x}_t^\top, \qquad \boldsymbol{C}_i^{(k+1)} = \frac{1}{n_i} \sum_t w_t^{(k)} \boldsymbol{x}_t \boldsymbol{y}_t^\top. \tag{1}$$

The Gram moment $\boldsymbol{G}_i \in \mathbb{R}^{L \times L}$ records how each pair of input positions varies together, and the cross-moment $\boldsymbol{C}_i \in \mathbb{R}^{L \times H}$ records how each input position varies with each forecast step. The client refits from these moments and stops after a fixed number of updates, or once the relative change in the fit falls below a tolerance. The final dense robust moments $(\boldsymbol{G}_i, \boldsymbol{C}_i)$ are retained locally for Phase 3. Huber-type weighting serves as local preprocessing and is not the proposed communication reduction (Huber, 1964; 1981). For a fixed fit and cutoff with bounded inputs, the weighted response contributions are bounded (Proposition 2). This bound does not cover changes in the fit, the cutoff, or high-leverage inputs.

Compression then groups these position-by-position relationships by time separation. For example, two observations one step apart contribute to the same input-lag entry wherever they fall in the window. Under second-order stationarity, the population relationship depends on this separation

rather than on the positions themselves, which motivates the grouping. For input positions $a, b \in \{0, \dots, L-1\}$, forecast step $h \in \{0, \dots, H-1\}$, and input lag $\ell = |a - b|$, we extract

$$r_i[\ell] = \frac{1}{L-\ell} \sum_{a=0}^{L-\ell-1} \boldsymbol{G}_i[a, a+\ell], \qquad s_i[h] = \boldsymbol{C}_i[0, h]. \tag{2}$$

Each entry $r_i[\ell]$ averages the Gram entries whose two input positions are ℓ steps apart, and each entry $s_i[h]$ takes the cross-moment value between the earliest input position and forecast step h. The two compact lag profiles $\boldsymbol{r}_i \in \mathbb{R}^L$ and $\boldsymbol{s}_i \in \mathbb{R}^H$ replace the dense matrix pair with $L + H$ values, and $r_i[0]$ records the mean input energy. Writing $\boldsymbol{v}_i \in \mathbb{R}^{L+H}$ for the two profiles stacked, input profile first, the Toeplitz and reverse-Hankel reconstruction maps are

$$\mathcal{T}(\boldsymbol{r}_i)[a, b] = r_i[|a - b|], \qquad \mathcal{H}(\boldsymbol{r}_i, \boldsymbol{s}_i)[a, h] = v_i[L + h - a], \qquad 0 \le a, b < L, \quad 0 \le h < H. \tag{3}$$

Reconstruction fills every matrix entry with the profile value for its time separation. This gives the Toeplitz pattern for the Gram moment and the reverse-Hankel pattern for the cross-moment, shown in Figure B1. Because all clients order time separations in the same way, the server can average matching entries directly. Lemma 6 shows exact reconstruction for raw population moments of unweighted stationary windows, although finite samples and residual weighting can introduce reconstruction error. Instead of shortening the history or dropping the longest forecast steps, compression discards the position-specific information between positions at the same lag. Proposition 1 characterizes this discarded Gram component, which stays available locally for personalization.

The lag profiles describe temporal relationships but do not say how much each client should influence the pooled profiles, so the client also computes a spectral reliability scale. It uses the eigenvalues $\hat{\mu}_{i1} \le \cdots \le \hat{\mu}_{iL}$ of $\boldsymbol{G}_i$, which measure the weighted input energy along each eigenvector. The client averages only the eigenvalues at or below a cutoff $\hat{\mu}_{i,\mathrm{cut}}$, the Marchenko-Pastur (MP) bulk edge implied by its mean eigenvalue and its ratio of window length to training rows (Marčenko & Pastur, 1967), which gives

$$\hat{\sigma}_i^2 = \max\Big\{a_0,\ \mathrm{mean}\left\{\hat{\mu}_{i\ell} : \hat{\mu}_{i\ell} \le \hat{\mu}_{i,\mathrm{cut}}\right\}\Big\}, \tag{4}$$

where the floor $a_0 > 0$ keeps the result away from zero. Excluding the eigenvalues above this edge prevents a few high-energy directions from setting the scale alone. This scale is a reliability proxy rather than a calibrated noise estimate, and the server receives it instead of the full spectrum. The upload therefore carries lag structure, reliability, and sample support.

3.2 Phase 2: Analytic Global Solve

Phase 2 runs on the server and makes aggregation reliable, since compression can reduce global forecasting performance and clients differ in sample support. The server first scores each client from its spectral reliability scale and training-row count, then caps extreme precision scores against the cohort before pooling the lag profiles with them. It finally sets the global ridge penalty from the pooled scale and row count, reconstructs the shared matrices from the pooled profiles, and solves for the shared forecaster, so that neither an unusual client nor a poorly supported direction destabilizes the global model.

The server scores each client with two precisions. The sample-supported precision p_i is the training-row count divided by the spectral reliability scale and controls profile pooling, while the scale-only precision u_i is the reciprocal of that scale and controls the pooled scale. Without a limit, a client with an unusually small scale would receive almost all of the weight, and the shared forecaster would rely on that client alone. Each score is therefore capped at a multiple $\kappa \ge 1$ of a cohort reference, namely the lower medians $\tilde{p}$ and $\tilde{u}$ of the two scores across clients,

$$\begin{aligned} \check{p}_i &= \min\{p_i, \kappa\tilde{p}\}, \quad \check{u}_i = \min\{u_i, \kappa\tilde{u}\}, \quad \alpha_i = \frac{\check{p}_i}{\sum_j \check{p}_j}, \\ \hat{\sigma}_g^2 &= \frac{N}{\sum_j \check{u}_j}, \quad \boldsymbol{r}_g = \sum_i \alpha_i \boldsymbol{r}_i, \quad \boldsymbol{s}_g = \sum_i \alpha_i \boldsymbol{s}_i. \end{aligned} \tag{5}$$

Here $\check{p}_i$ and $\check{u}_i$ are the two capped scores, α_i is the pooling weight of client i, $\hat{\sigma}_g^2$ is the pooled scale, and $(\boldsymbol{r}_g, \boldsymbol{s}_g)$ are the pooled profiles. Scores below the cap keep their relative weights, while scores

above it stop growing, which limits extreme precision scores but still distinguishes between clients. Proposition 3 bounds the weight of each client by $\kappa/(\kappa + \lfloor N/2 \rfloor)$.

From the pooled profiles, the server then reconstructs the shared Gram moment and cross-moment and fits the shared forecaster. Because both reconstruction maps are linear, reconstruction and weighted averaging give the same result in either order, so the server needs only the pooled profiles rather than a dense matrix pair from every client. An adaptive ridge penalty reduces sensitivity in directions with little input energy, and the server sets it from the pooled scale $\hat{\sigma}_g^2$ and the total row count $\sum_i n_i$, both already uploaded, with a cap $\lambda_{\text{cap}} > 0$,

$$\lambda = \min\left\{\hat{\sigma}_g^2 \frac{L}{\sum_i n_i},\ \lambda_{\text{cap}}\right\}, \qquad \boldsymbol{\Theta}_g = \left(\mathcal{T}(\boldsymbol{r}_g) + \lambda \boldsymbol{I}\right)^{-1} \mathcal{H}(\boldsymbol{r}_g, \boldsymbol{s}_g). \tag{6}$$

The ratio in the first expression compares the input dimension with the total sample support, so, with the other held fixed, fewer rows or a larger pooled scale increases the penalty until it reaches the cap. Following ridge regression and covariance shrinkage (Hoerl & Kennard, 1970; Ledoit & Wolf, 2004), this rule sets the global ridge penalty from uploaded quantities without another exchange. Adding $\lambda \boldsymbol{I}$ shifts every eigenvalue of the reconstructed Gram moment by λ. However, diagonal averaging does not always preserve positive semidefiniteness, so a positive ridge alone does not guarantee that the reconstructed system is well conditioned. Theorem 2 bounds the effect of reconstruction error on the solve under explicit conditioning assumptions.

3.3 Phase 3: Communication-Free Personalization

Phase 3 runs on each client and adapts the shared model after one broadcast, because a single shared model limits local adaptation when it cannot incorporate client-specific information absent from the shared summaries. The client first sets its personalization coefficient, which decides how much shared structure to borrow from the reliability of its own summary and its agreement with the pooled lag profile. It then applies moment-level shrinkage, combining the pooled profiles with its retained moments into one personalized system, and solves it for a personalized forecaster without further communication. This preserves client-specific temporal structure while still using knowledge learned across the federation.

The retained moments $\boldsymbol{G}_i$ and $\boldsymbol{C}_i$ hold the position-specific information that compression removed, so the client only needs to decide how much shared structure to add to them. Two local signals make this decision. The normalized bulk scale η_i compares the lower-bulk variation with the total input energy and grows when the client summary is poorly determined. The nonnegative lag-profile alignment ρ_i measures agreement between the local and pooled input profiles and falls to zero when they are orthogonal or opposed. Their product is the personalization coefficient γ_i, which controls how much the client borrows, and the Phase 1 floor a_0 also protects the denominators,

$$\eta_i = \frac{\hat{\sigma}_i^2}{r_i[0] + a_0}, \qquad \rho_i = \max\left\{0, \frac{\boldsymbol{r}_i^\top \boldsymbol{r}_g}{\|\boldsymbol{r}_i\|_2 \|\boldsymbol{r}_g\|_2 + a_0}\right\}, \qquad \gamma_i = \eta_i^2 \rho_i. \tag{7}$$

Because the two signals multiply, a small value of either one reduces the shared contribution, so a high bulk ratio cannot make a client borrow a pattern pointing the opposite way, and agreement alone cannot override a small bulk ratio. Personalization uses the uncapped $\hat{\sigma}_i^2$, because the cap limits influence across clients and does not change local statistics, although capping still affects γ_i indirectly through $\boldsymbol{r}_g$ and its alignment with $\boldsymbol{r}_i$. Lemma 7 bounds γ_i between zero and one and shows its consistency under the stated assumptions.

With γ_i fixed, the client combines its retained moments with the shared reconstructions, following the principle of shrinkage, under which local estimation borrows shared information instead of relying only on a separate local fit (James & Stein, 1961; Efron & Morris, 1975). This moment-level shrinkage gives the personalized system

$$\boldsymbol{A}_i = \boldsymbol{G}_i + \gamma_i \mathcal{T}(\boldsymbol{r}_g) + \gamma_i \boldsymbol{I}, \qquad \boldsymbol{B}_i = \boldsymbol{C}_i + \gamma_i \mathcal{H}(\boldsymbol{r}_g, \boldsymbol{s}_g), \qquad \boldsymbol{\Theta}_i = \boldsymbol{A}_i^{-1} \boldsymbol{B}_i. \tag{8}$$

Here the personalized coefficient matrix $\boldsymbol{A}_i$ and its right-hand side $\boldsymbol{B}_i$ each add a γ_i-weighted shared reconstruction to a retained moment. At $\gamma_i = 0$ with nonsingular $\boldsymbol{G}_i$, the solution reduces to the local fit, while a positive γ_i adds shared structure and keeps the dense local information. Because the shared matrices enter both sides of the system with the same weight before any coefficient is fitted,

Table 1: Communication cost (MB) per round and over the full experiment, averaged over four horizons, with multiplicative increases relative to FedTinyTS shown in parentheses.

Method	ETH	Electricity	SolarEnergy	TW5K	Traffic	#1st	Avg Rank
Round							
FEDAVG	0.497 (↑ 75.3×)	7.957 (↑ 7.5×)	3.233 (↑ 7.1×)	4.973 (↑ 7.5×)	21.385 (↑ 7.5×)	0	4.00
QATFL	0.288 (↑ 43.5×)	4.600 (↑ 4.3×)	1.869 (↑ 4.1×)	2.875 (↑ 4.4×)	12.363 (↑ 4.3×)	0	2.00
FEDLUAR	0.320 (↑ 48.4×)	4.944 (↑ 4.7×)	2.010 (↑ 4.4×)	3.195 (↑ 4.8×)	13.518 (↑ 4.7×)	0	3.00
DLSA	0.565 (↑ 85.5×)	90.694 (↑ 85.5×)	38.707 (↑ 85.5×)	56.507 (↑ 85.5×)	243.546 (↑ 85.5×)	0	5.90
FEDRIDGE	0.565 (↑ 85.5×)	90.693 (↑ 85.5×)	38.707 (↑ 85.5×)	56.506 (↑ 85.5×)	243.542 (↑ 85.5×)	0	5.10
Ours	**0.006607** (1×)	**1.060** (1×)	**0.453** (1×)	**0.661** (1×)	**2.848** (1×)	5	1.00
Total							
FEDAVG	49.732 (↑ 7527×)	795.715 (↑ 750×)	323.259 (↑ 714×)	497.322 (↑ 753×)	2138.485 (↑ 751×)	0	6.00
QATFL	28.752 (↑ 4352×)	460.028 (↑ 434×)	186.886 (↑ 413×)	287.518 (↑ 435×)	1236.325 (↑ 434×)	0	4.00
FEDLUAR	32.008 (↑ 4845×)	494.409 (↑ 466×)	200.986 (↑ 444×)	319.501 (↑ 484×)	1351.787 (↑ 475×)	0	5.00
DLSA	0.565 (↑ 85.5×)	90.694 (↑ 85.5×)	38.707 (↑ 85.5×)	56.507 (↑ 85.5×)	243.546 (↑ 85.5×)	0	2.90
FEDRIDGE	0.565 (↑ 85.5×)	90.693 (↑ 85.5×)	38.707 (↑ 85.5×)	56.506 (↑ 85.5×)	243.542 (↑ 85.5×)	0	2.10
Ours	**0.006607** (1×)	**1.060** (1×)	**0.453** (1×)	**0.661** (1×)	**2.848** (1×)	5	1.00

the result is not an average of two finished models, so the personalized forecaster can combine client-specific information that was never transmitted with knowledge learned across the federation. The ridge $\gamma_i \boldsymbol{I}$ regularizes the system but does not by itself ensure nonsingularity, because it vanishes with γ_i and may not offset negative eigenvalues of the reconstructed Gram moment $\mathcal{T}(\boldsymbol{r}_g)$. Neither profile agreement nor moment-level shrinkage guarantees lower forecasting risk. Corollary 2 establishes convergence to the personalized population target under its conditioning assumptions.

4 PERFORMANCE EVALUATION

4.1 EVALUATION SETUP

We evaluate on five benchmarks, ETDatasetHour (Zhou et al., 2021) (ETH), Electricity and SolarEnergy (Lai et al., 2018), TinyWeather5K (Han et al., 2026) (TW5K), and Traffic (Lai et al., 2018), whose characteristics appear in Table C2. The baselines comprise four families, ordered by the communication each requires. FEDAVG (McMahan et al., 2017) represents unmodified iterative optimization, whereas QATFL (Fang et al., 2026) and FEDLUAR (Kim et al., 2025) retain its round schedule but compress individual messages through quantization and stale-update recycling. A single-round alternative is provided by DLSA (Zhu et al., 2021) and FEDRIDGE (Alsulaimawi, 2026), which exchange dense moments and admit an analytic solution. LOCALONLY and LOCALOLS complete the range by fitting each client independently, under gradient descent and ordinary least squares respectively, without any communication.

All methods share a chronological 80/20 train-test split applied before windowing, training-only per-channel standardization, and a channel-independent linear backbone with $L = 96$ and $H \in \{96, 192, 336, 720\}$. They differ in participation. FedAvg, QATFL, and FedLUAR execute 100 rounds with one local epoch each and draw a fixed cohort of about a tenth of the clients without replacement per round, except on ETH, whose two clients participate throughout, whereas DLSA, FedRidge, and FedTinyTS include all clients in their single round. We report MSE and mean absolute error (MAE), taking for each iterative seed the round of best test loss rather than the final round. Each iterative baseline is therefore credited with its best round, whereas the analytic methods run once and admit one value. Every configuration runs three seeds, and we report no significance test. Communication is measured as the declared message payload, counting transmitted fields rather than network latency or energy. Remaining configuration and table conventions appear in Section D.

Table 2: Long-term forecasting performance for generalization and personalization, averaged over four horizons.

Method	ETH		Electricity		SolarEnergy		TW5K		Traffic		#1st	Avg
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE		Rank
Generalization												
FEDAVG	0.3609	**0.4029**	0.2144	**0.2931**	**0.3229**	**0.3134**	0.6957	0.5933	0.5134	**0.3482**	5	2.20
QATFL	0.3896	0.4320	0.2164	0.2969	0.3249	0.3447	0.7024	0.5978	0.5142	0.3555	0	5.35
FEDLUAR	0.3619	0.4045	0.2157	0.2951	0.3248	0.3152	0.6999	**0.5932**	0.5127	0.3487	1	3.30
DLSA	0.3623	0.4122	0.2126	0.2943	0.3245	0.3957	0.6923	0.5961	0.5080	0.3705	0	3.75
FEDRIDGE	0.3623	0.4123	0.2114	0.2969	0.3242	0.3954	**0.6922**	0.5967	0.5074	0.3719	1	3.90
Ours	**0.3588**	0.4069	**0.2112**	0.2937	0.3240	0.3923	0.6930	0.5964	**0.5069**	0.3683	3	2.50
Personalization												
LocalOnly	0.3557	**0.3998**	0.2026	**0.2817**	0.3306	**0.3242**	0.6853	**0.5832**	0.5063	**0.3450**	5	1.80
LocalOLS	0.3580	0.4089	**0.1999**	0.2888	**0.3254**	0.3980	0.6827	0.5893	0.5022	0.3753	2	2.40
Ours	**0.3539**	0.4030	0.2033	0.2879	0.3255	0.3938	**0.6822**	0.5880	**0.4981**	0.3723	3	1.80

4.2 COMMUNICATION COST

FedTinyTS has the smallest per-round and total payload on all five benchmarks in Table 1. Averaged over horizons, FedRidge transmits $85.5\times$ as much on every benchmark, and up to $90.7\times$ at $H{=}720$. FedAvg transmits between $714\times$ and $7{,}527\times$ as much over its 100 rounds, and up to $8{,}548\times$ on ETH at $H{=}720$. These reductions reflect both smaller messages and fewer exchanges. Because both methods use one round with all clients, the FedRidge comparison isolates message size. FedRidge's messages carry dense moments and model parameters, ours two compact lag profiles and two scalars. The gap grows with the horizon because each additional forecast step adds two values to our exchange, compared with L moment entries and $L+1$ model parameters for FedRidge. Message size rather than repetition puts the iterative baselines above our per-round payload even where they sample a tenth of the clients, which is the case on four of the five benchmarks. ETH is the exception, where every method draws both clients, so its per-round figures compare message sizes alone. Quantization and update recycling reduce individual messages, but QATFL and FedLUAR still exceed the dense one-shot baselines in total payload. Personalization reuses the broadcast profiles and retained local moments without further communication, so the same $2L + 2H + 2$ values per client support both global and personalized forecasting. Tables E1 and E2 provide the payload dimensions and per-horizon measurements.

4.3 LONG-TERM FORECASTING PERFORMANCE

Compression discards moment detail, so Table 2 asks whether the compact summaries retain enough of it for forecasting. Relative to FedAvg, global horizon-averaged MSE falls by 0.38% to 1.49% on four benchmarks and rises on SolarEnergy. That MSE also stays within 1% of DLSA and FedRidge on every benchmark, so on this suite the lag profiles retain enough of that detail to match dense analytic aggregation at a far smaller payload. Compression by itself falls short of that parity, and the scale-weighted pooling and adaptive ridge of the global solve make up the difference. Each client holds a private split of its own series, so both blocks evaluate on the same per-client test data and differ only in the model applied to it. DLSA and FedRidge have no client-side adaptation step, so their generalization entries are also their personalized results, and those exceed the better local reference by 1.05% to 6.31% on four benchmarks, falling below it only on SolarEnergy. FedTinyTS instead improves on its own global model by 1.36% to 3.75%, most on Electricity, though SolarEnergy rises by 0.46%, and its personalized fit leads the local references on ETH, TW5K, and Traffic, with LocalOLS ahead on the other two. Combining retained dense moments with shared profiles restores position-specific information the upload omits, a mechanism consistent with these gains.

Global MAE is less favorable, exceeding FedAvg by 0.0789 on SolarEnergy and 0.0201 on Traffic while remaining within 0.0040 on the other three. The two metrics have different estimands. Squared error targets a conditional mean and absolute error a conditional median, and because our

Table 3: Progressive component ablation on M4Weekly. Performance columns report MSE.

Added operation	Generalization	Personalization	Uplink	Downlink
Dense OLS (FEDRIDGE)	2.2078		$L^2 + LH$	P
+ Compression (Equation 2)	3.3400		$L + H + 2$	$L + H$
+ MP bulk scale (Equation 4)	2.9671		$L + H + 2$	$L + H$
+ Huber-type IRLS (Equation 1)	2.1344		$L + H + 2$	$L + H$
+ Precision cap (Equation 5)	1.7975		$L + H + 2$	$L + H$
+ Adaptive ridge (Equation 6)	1.2659		$L + H + 2$	$L + H$
+ Personal., fixed γ_i (Equation 8)	1.2659	1.2975	$L + H + 2$	$L + H$
+ Personal., bulk-scale-adaptive γ_i (Equation 7)	1.2659	1.2142	$L + H + 2$	$L + H$
+ Personal., full $\gamma_i = \eta_i^2 \rho_i$ (Equation 7)	1.2659	1.2020	$L + H + 2$	$L + H$

analytic solves minimize weighted squared residuals, an MSE gain need not carry over. Both dense baselines pay a larger penalty, between 0.0820 and 0.0823 on SolarEnergy and between 0.0223 and 0.0237 on Traffic, a pattern consistent with the squared-error objective rather than with lag compression. The comparison does not isolate that cause, because the gradient baselines are trained separately for each reported loss while every analytic method solves squared-loss normal equations in both columns. The gap nonetheless bounds the performance claim, most clearly on SolarEnergy. Tables E5 and E6 give the per-horizon results.

4.4 REPRESENTATION EFFICIENCY

Table 3 adds the components of FedTinyTS one at a time to a dense analytic solve on M4Weekly (Makridakis et al., 2018). It runs at L=26 and the native weekly horizon H=13, where heterogeneous series lengths give the clients unequal row counts. Compression alone raises global MSE from 2.2078 to 3.3400, a 51.3% increase. Adding the MP bulk scale reduces it to 2.9671, followed by 2.1344 with Huber-type IRLS, 1.7975 with the precision cap, and 1.2659 with adaptive ridge. Personalized MSE falls from 1.2975 with a fixed coefficient to 1.2142 with bulk-scale adaptation and 1.2020 with the full coefficient. Together these additions therefore cover the 51.3% penalty and end 42.7% below the dense starting point. Robust weighting and adaptive regularization are not specific to a compressed representation, however, and a dense method could adopt them as well. What compression removes is position-specific moment detail, whereas these additions act on local residual weighting, client influence, and regularization inside the compact representation. The table shows component additions, not execution order. Since each row rests on a single execution, and since the dataset and window lengths differ from those of the main suite, none of these numbers compares directly with it. Nor does the sequence establish that every component is necessary, or that its gain survives a different ordering. Uplink and downlink stay at $L + H + 2$ and $L + H$ for every row after compression. The evidence therefore supports recovering performance inside a fixed communication budget, and leaves broader component-level conclusions to further validation.

5 CONCLUSION

One-shot analytic federated learning for LTSF faces high communication costs from dense moment matrices and limited local adaptation of the shared model. FedTinyTS addresses both within a single exchange, in which each client uploads compact temporal summaries, the server applies reliability-aware aggregation to preserve the quality of the shared model, and each client adapts that model using locally retained information without further communication. Across five benchmarks and four horizons, it reduces the message payload by up to $8{,}548\times$ relative to FedAvg and up to $90.7\times$ relative to dense FedRidge. The global model achieves lower horizon-averaged MSE than FedAvg on four of the five benchmarks while remaining comparable to the dense analytic baselines, and personalization lowers MSE further on the same four. Two limits remain, in that MAE is less favorable, particularly on SolarEnergy, and the channel-independent linear backbone leaves cross-channel structure unused. Future work will target MAE within the same communication budget and extend the compact temporal summaries to cross-channel dependencies.

AI USE STATEMENT

Generative AI tools supplied critical ingredients for proofs and assisted in writing them. They assisted with formulating and revising mathematical claims, drafting and polishing manuscript prose, organizing sections, checking references, and editing scientific figures. They also provided feedback on research methodology and experiments and assisted with result interpretation. We did not use these tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, or implement methods. They were not used for translation or for cleaning and reformatting datasets. Qualitative and thematic data analysis is not applicable to this work. Their coding role covered code review, bug finding, and bug fixing. We reviewed all AI-assisted work. AI-assisted proofs were checked against the implemented protocol. Closed-form derivations of the communication ratios were reconciled against raw run logs. Every reported number was traced to an archived experiment log. We take responsibility for the final content, including text, claims, and artifacts produced with generative AI assistance.

ETHICS STATEMENT

All experiments use public time series benchmarks that contain no personally identifying records. This work releases no new dataset. Consequently, the experiments raise no human-subjects or data-consent questions. A separate concern is that communication reduction could be mistaken for a privacy mechanism. Clients transmit second-order lag summaries instead of raw windows or dense Gram and cross-moments. Section C.2 gives distinct dense moments with identical full uploads. Neither fact provides a privacy guarantee. The summaries are deterministic functions of client data and retain temporal structure. No randomization is applied, and our analysis does not bound what an observing adversary can infer. This distinction matters because the electricity, solar, and traffic clients correspond to physical meters or sensors whose traces may be sensitive. Reducing the payload from $L^2 + LH$ to $L + H + 2$ changes the representation but does not quantify the change in disclosure risk. A deployment requiring formal privacy should combine this protocol with a mechanism that provides a privacy guarantee. This paper does not measure the resulting utility. Such a composition must account for the pooling weights because the server receives each n_i and $\hat{\sigma}_i^2$ in the clear. Hiding or perturbing those scalars would also perturb the reliability caps.

REPRODUCIBILITY STATEMENT

The protocol is specified as executable pseudocode in Appendix B, which gives the three client and server phases together with the constants that govern them. Appendix C states the assumptions and theoretical guarantees together with their required conditions. Appendix C records what the lag-summary compression discards. Dataset sources, the chronological split, the standardization fitted on training statistics only, and the full hyperparameter grid appear in Appendix D. Section 4.1 gives the reporting rule for iterative traces, loss weighting, participation rules, and the number of executions per configuration. Each choice affects the reported numbers. Reported communication is the declared wire size of the transmitted summaries rather than measured process-transport traffic. Appendix D lists the fields counted by the logger and distinguishes client-side personalization from the archived simulator's execution placement. A reimplementation can therefore reproduce the reported payloads without reproducing the simulator architecture. The structural diagnostics in Table C2 were recovered from archived logs of an earlier implementation version for the reason recorded in Appendix D. Per-horizon tables behind every aggregate in the main paper appear in Appendix E.

REFERENCES

Raed Abdel-Sater and A. Ben Hamza. A Federated Large Language Model for Long-Term Time Series Forecasting. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024)*, pp. 2452–2459. IOS Press, 2024. doi: 10.3233/faia240772.

Zahir Alsulaimawi. One-Shot Federated Ridge Regression: Exact Recovery via Sufficient Statistic Aggregation, 2026. URL https://arxiv.org/abs/2601.08216.

Donald W. K. Andrews. Empirical process methods in econometrics. In *Handbook of Econometrics*, volume 4, pp. 2247–2294. Elsevier, 1994. doi: 10.1016/S1573-4412(05)80006-6.

Shafie Bahman and Hamidreza Zareipour. Long-Term Multi-Resolution Probabilistic Load Forecasting Using Temporal Hierarchies. *Energies*, 18(11):2908, 2025. doi: 10.3390/en18112908.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Paul Doukhan. *Mixing: Properties and Examples*, volume 85 of *Lecture Notes in Statistics*. Springer, 1994. doi: 10.1007/978-1-4612-2642-0.

Bradley Efron and Carl Morris. Data Analysis Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975. doi: 10.2307/2285814.

Kejia Fan, Jianheng Tang, Zhirui Yang, Feijiang Han, Jiaxu Li, Run He, Yajiang Huang, Anfeng Liu, Houbing Herbert Song, Yunhuai Liu, and Huiping Zhuang. APFL: Analytic personalized federated learning via dual-stream least squares, 2025. URL https://arxiv.org/abs/2508.10732.

Xiang Fang, Li Chen, Huarui Yin, Xiaohui Chen, and Weidong Wang. Communication Efficient Federated Learning With Quantization-Aware Training Design. *IEEE Transactions on Machine Learning in Communications and Networking*, 4:45–59, 2026. ISSN 2831-316X. doi: 10.1109/tmlcn.2025.3635050.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York, 1986. doi: 10.1002/9781118186435.

Tao Han, Zhibin Wen, Zhenghao Chen, Dazhao Du, Song Guo, and Lei Bai. Benchmarking Physics-Informed Time-Series Models for Operational Global Station Weather Forecasting. Accepted at the 43rd International Conference on Machine Learning (ICML), 2026. URL https://arxiv.org/abs/2406.14399.

Run He, Kai Tong, Di Fang, Han Sun, Ziqian Zeng, Haoran Li, Tianyi Chen, and Huiping Zhuang. AFL: A Single-Round Analytic Approach for Federated Learning with Pre-trained Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4988–4998, June 2025.

Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Peter J. Huber. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Peter J. Huber. *Robust Statistics*. Wiley, New York, 1981. doi: 10.1002/0471725250.

William James and Charles Stein. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379. University of California Press, Berkeley, CA, 1961.

M. M. Kara, H. I. Turkmen, and M. A. Guvensan. Smart Organization of Imbalanced Traffic Datasets for Long-Term Traffic Forecasting. *Sensors*, 25(4):1225, 2025. doi: 10.3390/s25041225.

Jisoo Kim, Sungmin Kang, and Sunwoo Lee. Layer-wise Update Aggregation with Recycling for Communication-Efficient Federated Learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 86242–86273. Curran Associates, Inc., 2025. doi: 10.52202/085713-2888.

Vikash Kumar and Rajib Kumar Mandal. Long-term load forecasting for smart grid. *Engineering Research Express*, 6(4):045339, 2024. doi: 10.1088/2631-8695/ad8f92.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–104. ACM, 2018. doi: 10.1145/3209978.3210006.

Olivier Ledoit and Michael Wolf. A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2018.06.001.

V. A. Marčenko and L. A. Pastur. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994.

Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, Chichester, 2006. doi: 10.1002/0470010940.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Khoa Nguyen and Taehong Kim. FedProC: A benchmarking framework for federated time-series forecasting. *SoftwareX*, 34:102646, June 2026. ISSN 2352-7110. doi: 10.1016/j.softx.2026.102646.

Yannick Peings, Cameron Dong, Ankur Mahesh, Michael Pritchard, William Collins, and Gudrun Magnusdottir. Subseasonal Forecasting and MJO Teleconnections in Machine Learning Weather Prediction Models. *Journal of Geophysical Research: Atmospheres*, 131(3):e2025JD044910, 2026. doi: 10.1029/2025jd044910.

Peter J. Rousseeuw and Christophe Croux. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993. doi: 10.1080/01621459.1993.10476408.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, 1980. doi: 10.1002/9780470316481.

Zezhi Shao, Fei Wang, Tao Sun, Chengqing Yu, Yuchen Fang, Guangyin Jin, Zhulin An, Yang Liu, Xiaobo Qu, and Yongjun Xu. HUTFormer: Hierarchical U-Net transformer for long-term traffic forecasting. *Communications in Transportation Research*, 5:100218, 2025. doi: 10.1016/j.commtr.2025.100218.

Jianheng Tang, Yajiang Huang, Kejia Fan, Feijiang Han, Jiaxu Li, Jinfeng Xu, Run He, Anfeng Liu, Houbing Herbert Song, Huiping Zhuang, and Yunhuai Liu. DeepAFL: Deep Analytic Federated Learning. International Conference on Learning Representations (ICLR), OpenReview, 2026. URL https://openreview.net/forum?id=ve3EzAvMGe.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996. doi: 10.1007/978-1-4757-2545-2.

Xianghai Xue, Xinguang He, Binrui Liu, Wenkai Lyu, and Mengjie Yang. Subseasonal Forecasting and Analysis of China Temperature Using an Explainable Deep Learning Model. *International Journal of Climatology*, 45(13):e70063, 2025. doi: 10.1002/joc.70063.

Wei Yuan, Chaoqun Yang, Xiangyu Zhao, Quoc Viet Hung Nguyen, Yang Cao, Tieke He, and Hongzhi Yin. Tackling Data Heterogeneity in Federated Time Series Forecasting. *Science China Information Sciences*, 69(5):152102, 2026. doi: 10.1007/s11432-025-4553-x.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, June 2023. ISSN 2159-5399. doi: 10.1609/aaai.v37i9.26317.

Shan Zhao, Zhitong Xiong, and Xiao Xiang Zhu. Efficient Subseasonal Weather Forecast Using Teleconnection-Informed Transformers. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2847–2852, 2024. doi: 10.1109/IGARSS53475.2024.10641568.

Xiao Zheng, Saeed Asadi Bagloee, and Majid Sarvi. Landformer: years ahead long-term traffic forecasting with land use information. *Transportmetrica A: Transport Science*, pp. 1–30, 2025. doi: 10.1080/23249935.2025.2565379.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

Xuening Zhu, Feng Li, and Hansheng Wang. Least-Square Approximation for a Distributed System. *Journal of Computational and Graphical Statistics*, 30(4):1004–1018, June 2021. ISSN 1537-2715. doi: 10.1080/10618600.2021.1923517.

Monika Zimmermann and Florian Ziel. Efficient mid-term forecasting of hourly electricity load using generalized additive models. *Applied Energy*, 388:125444, 2025. doi: 10.1016/j.apenergy.2025.125444.

A MATHEMATICAL NOTATION

Table A1 defines the mathematical notation used throughout the main paper and appendix. The entries distinguish dense client moments, transmitted lag profiles, and reconstructed shared quantities, so objects with different roles can be located without retracing the derivations. The communication symbols at the end describe message dimensions and participation rather than forecasting error.

Table A1: Mathematical notation used throughout the paper.

Notation	Meaning
General conventions and operators	
$\mathbb{R}$, $\mathbb{R}^{a\times b}$	Real numbers and real matrices with a rows and b columns
$\mathbf{0}_H$, $\mathbf{0}$, $\boldsymbol{I}$	Length-H zero vector, zero matrix, and identity matrix of the dimension implied by context
$(\cdot)^\top$, $(\cdot)^{-1}$	Matrix transpose and inverse
$\mathrm{Tr}(\cdot)$, $\mathrm{diag}(\cdot)$	Matrix trace; diagonal extraction for a matrix or diagonal-matrix construction from entries
$\min$, $\max$	Minimum and maximum operators
$\|\cdot\|_2$, $\|\cdot\|_\infty$	Euclidean vector norm or spectral matrix norm; maximum absolute vector entry
$\|\cdot\|_F$	Frobenius norm for matrices
$\langle\cdot,\cdot\rangle_F$	Frobenius inner product for matrices
$\succeq$, $\succ$, $\perp$	Positive-semidefinite order, positive-definite order, and orthogonality
$\lambda_{\min}(\cdot)$, $\dim(\cdot)$	Smallest eigenvalue and subspace dimension
$\lceil x\rceil$, $\lfloor x\rfloor$	Ceiling and floor
$\mathbb{E}$, $\Pr$, $\sup$	Expectation, probability, and supremum
$\mathrm{med}_{\mathrm{lo}}$, $\mathrm{med}_{\mathbb{P}}$	Lower sample median and population median
$\mathcal{T}(\cdot)$, $\mathcal{H}(\cdot,\cdot)$	Toeplitz and reverse-Hankel reconstruction operators
$\Pi_{\mathcal{T}}$	Linear diagonal-averaging projection onto the Toeplitz reconstruction
$\mathcal{T}_L$, $\mathcal{T}_L^\perp$	Symmetric Toeplitz subspace in $\mathbb{R}^{L\times L}$ and its Frobenius orthogonal complement within the symmetric matrices
$\mathcal{D}_0$, $\boldsymbol{D}\in\mathcal{D}_0$	Subspace of diagonal $L\times L$ matrices with zero trace and one of its perturbations
$\boldsymbol{T}_\ell$	Toeplitz basis matrix with ones where $\|a-b\|=\ell$ and zeros elsewhere
$\mathbf{1}\{\cdot\}$	Indicator function
$\hat{\cdot}$, $(\cdot)^*$	Hat-marked empirical fit, quantile-based scale, spectral statistic, or stopping quantity; star-marked population counterpart. Empirical moments and profiles are unstarred without hats
$(\cdot)^{(0)}$, $(\cdot)^{(k)}$	Initial fit or unweighted diagnostic, and IRLS iterate k
$\bar{\cdot}$, $\tilde{\cdot}$, $\check{\cdot}$	Arithmetic average, median-based quantity, and median-capped quantity
$\xrightarrow{p}$, $\Rightarrow$	Convergence in probability and convergence in distribution
$O_p(a_n)$, $o_p(1)$, $O(a_n)$	Stochastic order, convergence to zero in probability, and deterministic order
$\approx$, $\uparrow$	Approximate equality and the multiplicative increase relative to FedTinyTS reported in the communication tables
Indices, dimensions, and data	
$i,j\in\{1,\dots,N\}$	Client indices and total client count N
t	Channel-window row index
a,b	Matrix row and column indices
ℓ,h	Input-lag or eigenvalue index (with range specified locally), and forecast-step index
$k\in\{0,\dots,T\}$	IRLS update index and maximum update count T
n_i, $n=\sum_i n_i$	Client i row count and total row count
c_i, $c_{i,n}$	Population row proportion and finite-sample proportion n_i/n
L, H, C	Input length, forecast horizon, and number of data channels, with L and H fixed in client-wise asymptotics unless stated otherwise
$\boldsymbol{X}_i$, $\boldsymbol{Y}_i$	Client i input and target window matrices in $\mathbb{R}^{n_i\times L}$ and $\mathbb{R}^{n_i\times H}$
$\boldsymbol{x}_t$, $\boldsymbol{y}_t$, $\boldsymbol{e}_t$	Transposed input and target rows, and column-vector forecast residual for window t

Continued on the next page

Table A1 continued

Notation	Meaning
$\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{e}$	Generic population input row, target row, and residual
$\mathbb{P}_i$, $\mathbb{P}_{n_i}$, $\mathbb{P}$	Client population law, client empirical measure, and a generic row distribution
Q_1, $\tilde{y}$, Q_3, $\bar{y}$	First quartile, median, third quartile, and mean of the raw target
$\bar{n}$, $\overline{\mathrm{CV}}$	Across-client mean row count $N^{-1}\sum_i n_i$ and across-client mean coefficient of variation
Forecasters, objectives, and moments	
$\boldsymbol{\Theta}$, $\widehat{\boldsymbol{Y}}_i$	Shared linear-forecaster parameter and client prediction $\boldsymbol{X}_i\boldsymbol{\Theta}$
$\boldsymbol{\Theta}_g$, $\boldsymbol{\Theta}_i$, $\boldsymbol{\Theta}_i^*$	Proposed global, finite-sample personalized, and population personalized forecasters
$\boldsymbol{b}_g$, $\boldsymbol{b}_i$	Zero global and personalized analytic bias vectors
λ, ϵ	Global ridge coefficient and local IRLS ridge
$\mathcal{S}_{xx,i}$, $\mathcal{S}_{xy,i}$	Unweighted empirical Gram moment and cross-moment, each normalized by n_i
$\boldsymbol{G}_i$, $\boldsymbol{C}_i$	Terminal empirical moments formed with Huber residual weights
$\boldsymbol{G}_i^*$, $\boldsymbol{C}_i^*$	Exact population counterparts of client i's finite, ridged, early-stopped terminal IRLS moments
$\mathcal{S}_{xx,i}^*$, $\mathcal{S}_{xy,i}^*$	Raw unweighted population moments
$\varrho_i(k)$	Client i's uncentered lag-k second moment $\mathbb{E}[X_tX_{t+k}]$ in the stationarity argument
Lag compression	
$\boldsymbol{r}$, $\boldsymbol{s}$, $\boldsymbol{\Sigma}$	Generic input-lag summary, cross-lag summary, and matrix argument in the reconstruction subroutines
$\boldsymbol{r}_i$, $r_i[\ell]$	Client input-lag summary vector and its lag-ℓ entry
$\boldsymbol{s}_i$, $s_i[h]$	Client cross-lag summary vector and its forecast-step-h entry
$\boldsymbol{v}_i$, $\boldsymbol{v}$	Concatenation of the input-lag and cross-lag vectors used by reverse-Hankel indexing
$\boldsymbol{r}_g$, $\boldsymbol{s}_g$	Finite-sample pooled lag summaries broadcast by the server
$\boldsymbol{r}_i^*$, $\boldsymbol{s}_i^*$	Population client lag summaries
$\varepsilon_{\mathcal{T},i}$, $\varepsilon_{\mathcal{H},i}$	Relative terminal weighted Toeplitz and Hankel reconstruction errors
$\varepsilon_{\mathcal{T},i}^{(0)}$, $\varepsilon_{\mathcal{H},i}^{(0)}$	Corresponding ancillary errors from unweighted training windows
$\bar{\varepsilon}_{\mathcal{T}}$, $\bar{\varepsilon}_{\mathcal{H}}$, $\tilde{\varepsilon}_{\mathcal{T}}^{(0)}$, $\tilde{\varepsilon}_{\mathcal{H}}^{(0)}$	Reported mean weighted errors and median ancillary unweighted errors
$\boldsymbol{\Delta}_{xx,i}$, $\boldsymbol{\Delta}_{xy,i}$	Finite-sample Gram and cross-moment residuals discarded by compression
$\boldsymbol{\Delta}_{xx,i}^*$, $\boldsymbol{\Delta}_{xy,i}^*$	Population compression residuals
Robust local fitting and spectral scale	
$z_t^{(k)}$, $\tilde{z}_i^{(k)}$	Empirical row residual norm and its lower sample median at iterate k
$\hat{\sigma}_{e,i}^{(k)}$, $\delta_i^{(k)}$	Empirical MAD-based residual scale and Huber cutoff
c_{H}, b_{MAD}	Huber cutoff multiplier and MAD normalizer
$w_t^{(k)}$, $w(z,u)$	Empirical row weight and generic weight map with residual magnitude z and cutoff u
τ, a_0, b_0, d_0	IRLS stopping tolerance and positive scale or denominator safeguards
q_i	Client aspect proxy L/n_i
$\hat{\mu}_{i\ell}$, $\hat{\bar{\mu}}_i$, $\hat{\mu}_{i,\mathrm{cut}}$	Ordered empirical eigenvalue, arithmetic mean of the client spectrum, and MP-inspired cutoff
$\hat{\mathcal{I}}_i$, $\lvert\hat{\mathcal{I}}_i\rvert$	Selected empirical eigenvalue index set and its cardinality, not a training row count
$\mu_{i\ell}^*$, $\bar{\mu}_i^*$, $\mathcal{I}_i^*$	Ordered population eigenvalue, mean spectrum, and below-mean selected index set
$\hat{\sigma}_i^2$, $\sigma_i^{2,*}$	Empirical client bulk scale and its fixed-dimensional population target
$\hat{\sigma}_g^2$, $\sigma_g^{2,*}$	Empirical capped pooled bulk scale and its population limit
Reliability pooling and personalization	
p_i, u_i	Sample-weighted inverse bulk-scale proxy $n_i/\hat{\sigma}_i^2$ and inverse bulk-scale proxy $1/\hat{\sigma}_i^2$

Continued on the next page

Table A1 continued

Notation	Meaning
$\tilde{p}$, $\tilde{u}$, $\check{p}_i$, $\check{u}_i$	Cohort lower medians and their median-relative capped values
κ, α_i	Precision-cap multiplier and finite-sample aggregation weight
q_{pool}, λ_{cap}	Pooled aspect proxy $L/\sum_i n_i$ and global ridge cap
η_i, ρ_i, γ_i	Normalized client bulk scale, nonnegative lag-profile alignment, and personalization coefficient
$\boldsymbol{A}_i$, $\boldsymbol{B}_i$	Finite-sample personalized coefficient matrix and right-hand-side matrix
η_i^*, ρ_i^*, γ_i^*	Population limits of the personalization quantities
$\boldsymbol{A}_i^*$, $\boldsymbol{B}_i^*$	Population personalized coefficient matrix and right-hand-side matrix
Convergence targets and assumptions	
$\mathcal{G}_i$	Finite-recursion empirical-process function class for client i
$z_i(\boldsymbol{\Theta})$, $m_i(\boldsymbol{\Theta})$, $D_i(\boldsymbol{\Theta}, m)$	Residual-norm map for a random row from client i, its population median, and the absolute-deviation median with supplied center m
$\boldsymbol{M}_{xx,i}(\boldsymbol{\Theta}, u)$, $\boldsymbol{M}_{xy,i}(\boldsymbol{\Theta}, u)$	Population weighted-moment maps at a supplied fit and cutoff
U_{i,n_i}	Supremum empirical error over the local function class $\mathcal{G}_i$
$\mathcal{N}_{ik}$, C_{ac}, ζ, ζ_0	Neighborhood, anti-concentration constant, deviation radius, and local radius limit in the cutoff margin condition
m_w, M_w	Uniform lower and upper eigenvalue bounds for terminal dense and Toeplitz moments
$\sigma^2_{\min}$, $\sigma^2_{\max}$	Uniform lower and upper bounds on population bulk scales
p_i^*, $\tilde{p}^*$, $\check{p}_i^*$	Limiting normalized precision, its lower median, and its capped value
α_i^*, $\boldsymbol{r}_g^*$, $\boldsymbol{s}_g^*$	Limiting aggregation weight and pooled population lag summaries
$\boldsymbol{\Theta}_g^*$	Implemented compressed robust population target
$\boldsymbol{G}_{g,\mathrm{dense}}^*$, $\boldsymbol{C}_{g,\mathrm{dense}}^*$, $\boldsymbol{\Theta}_{g,\mathrm{dense}}^*$	Dense robust population moments and their counterfactual pooled target under the same weights
$\boldsymbol{A}_g$, $\boldsymbol{B}_g$, $\boldsymbol{A}_g^*$, $\boldsymbol{B}_g^*$	Empirical and population coefficient and right-hand-side matrices for the global solve
$\boldsymbol{\Delta}_{xx,g}^*$, $\boldsymbol{\Delta}_{xy,g}^*$	Population compression residuals pooled with α_i^*
$\boldsymbol{\Theta}_{\mathrm{OLS}}^*$	Raw sample-size-weighted pooled OLS population target
B_x	Uniform bound on input-vector norm in the fixed-fit sensitivity statement
ξ_i	Arbitrary Toeplitz-lemma weight satisfying $\xi_i \geq 0$ and $\sum_i \xi_i = 1$
$p_{i,n} = p_i/n$	Normalized empirical precision used to derive the population weights
u_i^*, $\tilde{u}^*$, $\check{u}_i^*$	Population inverse bulk scale, its cohort lower median, and its capped value
S_n, S^*	Sums of normalized capped empirical precisions and their population limits
Population IRLS recursion and asymptotic distributions	
$\hat{\boldsymbol{\Theta}}_i^{(k)}$, $\boldsymbol{\Theta}_i^{(k),*}$	Empirical and population client fits at iterate k
$\boldsymbol{G}_i^{(k)}$, $\boldsymbol{C}_i^{(k)}$, $\boldsymbol{G}_i^{(k),*}$, $\boldsymbol{C}_i^{(k),*}$	Empirical and population weighted moments at update $k \geq 1$
$\tilde{z}_i^{(k),*}$, $\sigma_{e,i}^{(k),*}$, $\delta_i^{(k),*}$, $w_i^{(k),*}$	Population residual median, MAD-based scale, cutoff, and row-weight function
$\hat{K}_i$, K_i^*	Empirical and population stopping indices
$\hat{d}_i^{(k)}$, $d_i^{(k),*}$	Empirical and population relative fit changes
g_i^{stop}, g_i^{spec}	Minimum population gaps from the stopping threshold and spectral-selection boundary
$\mathcal{N}(0, \sigma^2)$, σ, Φ, Φ^{-1}	Centered Gaussian law, its scale, standard-normal cumulative distribution function, and its quantile function
Z_H, m_H, D_H, δ_H, π_H	Gaussian residual norm, its population median and MAD, population cutoff, and downweighting probability
S_H, ξ_h	Sum of squared independent standard normals and its standard-normal coordinates
Communication, optimization, and reporting	
$P = LH + H = H(L+1)$, M, R	Implemented parameter count, iterative participants per round, and communication rounds
q_s, $\bar{u}$, $\bar{\iota}$	QATFL quantized upload and FedLUAR mean omitted-parameter and index overhead

Continued on the next page

Table A1 continued

Notation	Meaning
ν, β_1, β_2, M_{qat}	Adam learning rate, first- and second-moment decay coefficients, and QATFL local update budget

B Full Protocol and Algorithm

The protocol separates client-side compression and personalization from server-side aggregation. Each phase below pairs the unchanged algorithm with its numerical definitions and derivation, using the line numbers printed in that algorithm. The single upload contains $(\boldsymbol{r}_i, \boldsymbol{s}_i, \hat{\sigma}_i^2, n_i)$, and the single broadcast contains $(\boldsymbol{r}_g, \boldsymbol{s}_g)$. Appendix D distinguishes this protocol placement from the archived simulator used for timing measurements.

B.1 Phase 1: Robust Local Compression

Algorithm 1 Phase 1 robust local compression at client i

Require: input length L, horizon H, updates $T = 5$, ridge $\epsilon = 10^{-6}$, tolerance $\tau = 10^{-4}$, constants $c_{\mathrm{H}} = 1.345$, $b_{\mathrm{MAD}} = 0.6745$, $a_0 = d_0 = 10^{-8}$, $b_0 = 10^{-12}$

1: Flatten channels into $n_i \geq 1$ rows $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times L}$, $\boldsymbol{Y}_i \in \mathbb{R}^{n_i \times H}$
2: $\mathcal{S}_{xx,i} \leftarrow n_i^{-1} \boldsymbol{X}_i^\top \boldsymbol{X}_i, \quad \mathcal{S}_{xy,i} \leftarrow n_i^{-1} \boldsymbol{X}_i^\top \boldsymbol{Y}_i$
3: $\hat{\boldsymbol{\Theta}}_i^{(0)} \leftarrow (\mathcal{S}_{xx,i} + \epsilon \boldsymbol{I})^{-1} \mathcal{S}_{xy,i}$
4: **for** $k = 0, \ldots, T-1$ **do**
5: $\quad \boldsymbol{e}_t^{(k)} \leftarrow \boldsymbol{y}_t - (\hat{\boldsymbol{\Theta}}_i^{(k)})^\top \boldsymbol{x}_t, \quad z_t^{(k)} \leftarrow \|\boldsymbol{e}_t^{(k)}\|_2$ for every row t
6: $\quad \tilde{z}_i^{(k)} \leftarrow \mathrm{med}_{\mathrm{lo}}\{z_t^{(k)}\}_{t=1}^{n_i}$
7: $\quad \hat{\sigma}_{e,i}^{(k)} \leftarrow \max\{\mathrm{med}_{\mathrm{lo}}\{|z_t^{(k)} - \tilde{z}_i^{(k)}|\}_{t=1}^{n_i} / b_{\mathrm{MAD}},\ a_0\}$
8: $\quad \delta_i^{(k)} \leftarrow \tilde{z}_i^{(k)} + c_{\mathrm{H}} \hat{\sigma}_{e,i}^{(k)}$
9: $\quad w_t^{(k)} \leftarrow 1$ if $z_t^{(k)} \leq \delta_i^{(k)}$, else $w_t^{(k)} \leftarrow \delta_i^{(k)} / (z_t^{(k)} + b_0)$
10: $\quad \boldsymbol{G}_i^{(k+1)} \leftarrow n_i^{-1} \sum_t w_t^{(k)} \boldsymbol{x}_t \boldsymbol{x}_t^\top$
11: $\quad \boldsymbol{C}_i^{(k+1)} \leftarrow n_i^{-1} \sum_t w_t^{(k)} \boldsymbol{x}_t \boldsymbol{y}_t^\top$
12: $\quad \hat{\boldsymbol{\Theta}}_i^{(k+1)} \leftarrow (\boldsymbol{G}_i^{(k+1)} + \epsilon \boldsymbol{I})^{-1} \boldsymbol{C}_i^{(k+1)}$
13: $\quad \hat{d}_i^{(k+1)} \leftarrow \|\hat{\boldsymbol{\Theta}}_i^{(k+1)} - \hat{\boldsymbol{\Theta}}_i^{(k)}\|_F / (\|\hat{\boldsymbol{\Theta}}_i^{(k)}\|_F + d_0)$
14: $\quad \hat{K}_i \leftarrow k + 1$
15: $\quad$ **if** $\hat{d}_i^{(k+1)} < \tau$ **then**
16: $\quad\quad$ **break**
17: $\quad$ **end if**
18: **end for**
19: $\boldsymbol{G}_i \leftarrow \boldsymbol{G}_i^{(\hat{K}_i)}, \quad \boldsymbol{C}_i \leftarrow \boldsymbol{C}_i^{(\hat{K}_i)}$
20: $\boldsymbol{r}_i \leftarrow \textsc{ExtractAutocorr}(\boldsymbol{G}_i), \quad \boldsymbol{s}_i \leftarrow \boldsymbol{C}_i[0, :]$
21: $(\hat{\mu}_{i1}, \ldots, \hat{\mu}_{iL}) \leftarrow \mathrm{eigvalsh}(\boldsymbol{G}_i)$ ▷ *in increasing order*
22: $\hat{\bar{\mu}}_i \leftarrow L^{-1} \sum_{\ell=1}^{L} \hat{\mu}_{i\ell}, \quad q_i \leftarrow L / n_i$
23: $\hat{\mu}_{i,\mathrm{cut}} \leftarrow \hat{\bar{\mu}}_i (1 + \sqrt{q_i})^2$
24: $\hat{\mathcal{I}}_i \leftarrow \{\ell \in \{1, \ldots, L\} : \hat{\mu}_{i\ell} \leq \hat{\mu}_{i,\mathrm{cut}}\}$
25: $\hat{\sigma}_i^2 \leftarrow \max\{a_0,\ |\hat{\mathcal{I}}_i|^{-1} \sum_{\ell \in \hat{\mathcal{I}}_i} \hat{\mu}_{i\ell}\}$ ▷ $|\hat{\mathcal{I}}_i| \geq 1$ *is the selected count*
26: **logically upload** $(\boldsymbol{r}_i, \boldsymbol{s}_i, \hat{\sigma}_i^2, n_i)$
27: Retain $\boldsymbol{G}_i, \boldsymbol{C}_i$ locally

Client i first constructs moments that reduce the contribution of windows poorly explained by its local fit. Lines 1-3 of Algorithm 1 lay out the rows, take their unweighted moments, and seed the iteration with the ridge fit

$$\hat{\boldsymbol{\Theta}}_i^{(0)} = (\mathcal{S}_{xx,i} + \epsilon \boldsymbol{I})^{-1} \mathcal{S}_{xy,i}, \tag{B1}$$

where $\epsilon > 0$ is the local ridge coefficient that keeps the initial solve well posed.

Lines 5-9 then run Huber-type IRLS (Huber, 1964), which alternates residual-based weights with a weighted ridge fit for at most T updates. Write $\boldsymbol{x}_t \in \mathbb{R}^L$ and $\boldsymbol{y}_t \in \mathbb{R}^H$ for the transposed input and target rows of client i, suppressing the client index within this local calculation. At iterate k, let $\boldsymbol{e}_t^{(k)} = \boldsymbol{y}_t - (\hat{\boldsymbol{\Theta}}_i^{(k)})^\top \boldsymbol{x}_t$ be the residual of row t and $z_t^{(k)} = \|\boldsymbol{e}_t^{(k)}\|_2$ its size. The residual cutoff is set from the median and median absolute deviation of these sizes (Rousseeuw & Croux, 1993),

$$\begin{aligned} \tilde{z}_i^{(k)} &= \mathrm{med}_{\mathrm{lo}}\{z_t^{(k)}\}_{t=1}^{n_i}, & \delta_i^{(k)} &= \tilde{z}_i^{(k)} + c_{\mathrm{H}} \hat{\sigma}_{e,i}^{(k)}, \\ \hat{\sigma}_{e,i}^{(k)} &= \max\Big\{ \tfrac{1}{b_{\mathrm{MAD}}} \mathrm{med}_{\mathrm{lo}}\{|z_t^{(k)} - \tilde{z}_i^{(k)}|\}_{t=1}^{n_i},\, a_0 \Big\}. \end{aligned} \tag{B2}$$

Here c_{H} controls the cutoff above the median and b_{MAD} normalizes the median absolute deviation. The positive constant a_0 is a scale floor. The operator $\mathrm{med}_{\mathrm{lo}}$ takes the lower middle value for an even count. Rows at or below the cutoff keep weight $w_t^{(k)} = 1$, and rows above it receive $w_t^{(k)} = \delta_i^{(k)}/(z_t^{(k)} + b_0)$ with denominator safeguard $b_0 > 0$.

Lines 10-12 recompute the weighted moments $\boldsymbol{G}_i^{(k+1)}$ and $\boldsymbol{C}_i^{(k+1)}$ by Equation 1, from which the updated fit is $\hat{\boldsymbol{\Theta}}_i^{(k+1)} = (\boldsymbol{G}_i^{(k+1)} + \epsilon \boldsymbol{I})^{-1} \boldsymbol{C}_i^{(k+1)}$. Normalizing by n_i preserves the reduction in total contribution caused by downweighting, rather than inflating the remaining weights to sum to one.

Iteration stops at lines 13-19, once the relative change $\|\hat{\boldsymbol{\Theta}}_i^{(k+1)} - \hat{\boldsymbol{\Theta}}_i^{(k)}\|_F / (\|\hat{\boldsymbol{\Theta}}_i^{(k)}\|_F + d_0)$ falls below the tolerance τ, or after T updates. The positive safeguard d_0 stabilizes this relative-change denominator. Algorithm 1 fixes $T = 5$, $\epsilon = 10^{-6}$, $\tau = 10^{-4}$, $c_{\mathrm{H}} = 1.345$, $b_{\mathrm{MAD}} = 0.6745$, $a_0 = d_0 = 10^{-8}$, and $b_0 = 10^{-12}$. The terminal moments $(\boldsymbol{G}_i, \boldsymbol{C}_i)$ are retained locally for Phase 3. Huber-type weighting supplies the local preprocessing rule rather than the proposed communication reduction (Huber, 1964; 1981). At a fixed fit and cutoff with bounded inputs, weighted response contributions are bounded (Proposition 2). This fixed-fit bound does not control the full estimator's sensitivity to changes in its fit, cutoff, or high-leverage inputs.

Line 20 evaluates Equation 2, giving $r_i[\ell] = (L-\ell)^{-1} \sum_{a=0}^{L-\ell-1} \boldsymbol{G}_i[a, a+\ell]$ and $s_i[h] = \boldsymbol{C}_i[0, h]$. The reconstructions in Equation 3 use $\mathcal{T}(\boldsymbol{r}_i)[a, b] = r_i[|a-b|]$ and $\mathcal{H}(\boldsymbol{r}_i, \boldsymbol{s}_i)[a, h] = v_i[L+h-a]$. The Toeplitz reconstruction gives every input pair at lag ℓ the same value, replacing position-specific differences by their lag average. The reverse-Hankel reconstruction places $\boldsymbol{s}_i$ in its first row and shifts it one entry to the right per lower row, filling the exposed leading entries with the tail of $\boldsymbol{r}_i$. The index $L + h - a$ runs from 1 to $L + H - 1$, so $r_i[0]$ enters only the Toeplitz diagonal. The first row of $\boldsymbol{C}_i$ pairs the earliest input position with every forecast step, covering separations from L to $L + H - 1$. The tail of $\boldsymbol{r}_i$ supplies the shorter separations. Reversing the input-row order gives a Hankel matrix with constant anti-diagonals, so the profiles specify an $L \times H$ cross-moment reconstruction. The two profiles have complementary roles. The input profile covers separations shorter than the input window, while the cross-moment anchor extends the representation through the forecast horizon. Together they specify both matrices needed by an analytic forecaster without transmitting either dense matrix. The representation preserves the diagonal averages of $\boldsymbol{G}_i$ and the first row of $\boldsymbol{C}_i$ while discarding position-specific deviations. Both reconstructions are exact for raw population moments of unweighted contiguous windows from a second-order stationary stream (Lemma 6). Finite samples and residual-dependent weighting can break this structure even under stationarity. Unlike truncating the input history, this representation retains coordinates for both short and long lags. Every client uses the same window-defined coordinates, allowing the server to average corresponding temporal relations directly. The approximation concerns variation between positions at a shared lag, rather than the removal of a selected lag range. The discarded component of $\boldsymbol{G}_i$ lies in the orthogonal complement of the symmetric Toeplitz subspace, of dimension $L(L-1)/2$ (Proposition 1). Phase 3 retains access to this position-specific information through the dense local moments.

The lag profiles describe temporal relations but do not specify how strongly each client should influence the pool. A spectral bulk scale supplies a compact reliability proxy for that decision. Lines 21-25 compute it from the eigenvalues $\hat{\mu}_{i1} \le \cdots \le \hat{\mu}_{iL}$ of $\boldsymbol{G}_i$, which measure weighted input energy along its eigenvectors. With mean $\hat{\bar{\mu}}_i = L^{-1} \sum_\ell \hat{\mu}_{i\ell}$ and aspect ratio $q_i = L/n_i$, the client averages the lower spectral bulk by Equation 4. The cutoff is inspired by the Marchenko-Pastur edge factor $(1 + \sqrt{q_i})^2$ (Marčenko & Pastur, 1967). Eigenvalues above the cutoff do not enter the

final average, so a few large spectral directions are excluded from the reported bulk scale. Here $|\hat{\mathcal{I}}_i|$ denotes the number of eigenvalues selected by the cutoff. For dependent, weighted windows, this bulk statistic is a spectral-scale proxy, not a calibrated estimate of observation noise or prediction uncertainty. Unlike the residual scale inside IRLS, which is computed from prediction errors and used only locally, $\hat{\sigma}_i^2$ is computed from the terminal input moment and used for cross-client pooling. At a fixed row count, a smaller $\hat{\sigma}_i^2$ increases the client's precision score in Phase 2, subject to the server's cap. Uploading this scalar in place of the eigenspectrum keeps the reliability signal at one value per client.

Line 26 sends the compact tuple

$$(\boldsymbol{r}_i, \boldsymbol{s}_i, \hat{\sigma}_i^2, n_i). \tag{B3}$$

It contains $L + H + 2$ values, while line 27 holds $(\boldsymbol{G}_i, \boldsymbol{C}_i)$ back for personalization.

B.2 Phase 2: Analytic Global Solve

Algorithm 2 Phase 2 one-shot aggregation and global solve at the server

Require: uploads $\{(\boldsymbol{r}_i, \boldsymbol{s}_i, \hat{\sigma}_i^2, n_i)\}_{i=1}^N$, ridge cap $\lambda_{\text{cap}} = 0.01$, precision cap $\kappa = 10$

1: **for** $i = 1, \dots, N$ **do**
2: $\quad p_i \leftarrow n_i/\hat{\sigma}_i^2, \quad u_i \leftarrow 1/\hat{\sigma}_i^2$
3: **end for**
4: $\tilde{p} \leftarrow \text{med}_{\text{lo}}\{p_j\}_{j=1}^N, \quad \tilde{u} \leftarrow \text{med}_{\text{lo}}\{u_j\}_{j=1}^N$
5: **for** $i = 1, \dots, N$ **do**
6: $\quad \check{p}_i \leftarrow \min(p_i,\ \kappa\,\tilde{p}), \quad \check{u}_i \leftarrow \min(u_i,\ \kappa\,\tilde{u})$ ▷ *median-relative precision cap*
7: **end for**
8: $\alpha_i \leftarrow \check{p}_i \big/ \sum_j \check{p}_j$ for $i = 1, \dots, N$
9: $\boldsymbol{r}_g \leftarrow \sum_i \alpha_i \boldsymbol{r}_i, \quad \boldsymbol{s}_g \leftarrow \sum_i \alpha_i \boldsymbol{s}_i$
10: $\hat{\sigma}_g^2 \leftarrow N \big/ \sum_i \check{u}_i$
11: $q_{\text{pool}} \leftarrow L \big/ \sum_i n_i$
12: $\lambda \leftarrow \min\big(\hat{\sigma}_g^2\, q_{\text{pool}},\ \lambda_{\text{cap}}\big)$
13: $\boldsymbol{\Theta}_g \leftarrow \big(\mathcal{T}(\boldsymbol{r}_g) + \lambda \boldsymbol{I}\big)^{-1} \mathcal{H}(\boldsymbol{r}_g, \boldsymbol{s}_g)$ ▷ *server-side only, for centralized evaluation*
14: $\boldsymbol{b}_g \leftarrow \boldsymbol{0}_H$ ▷ *analytic linear-layer bias*
15: **broadcast** $(\boldsymbol{r}_g, \boldsymbol{s}_g)$ to all clients

Algorithm 2 computes the two precision scores in line 2, their lower medians in line 4, and their capped values in line 6. Lines 8-10 then normalize the profile weights, pool both profiles, and compute the harmonic-type scale. The server receives profiles in a common lag coordinate system, but their sample support and spectral scales can differ across clients. It uses both quantities to determine each client's contribution to the shared representation. Unrestricted inverse-scale weighting, however, would let one client with a tiny $\hat{\sigma}_i^2$ dominate both the pooled profiles and the global scale. The sample-supported precision $p_i = n_i/\hat{\sigma}_i^2$ controls profile pooling, while the scale-only precision $u_i = 1/\hat{\sigma}_i^2$ controls the pooled scale. For cap multiplier $\kappa \geq 1$, their reference values are the cohort lower medians $\tilde{p} = \text{med}_{\text{lo}}\{p_j\}_{j=1}^N$ and $\tilde{u} = \text{med}_{\text{lo}}\{u_j\}_{j=1}^N$. The server caps and pools these scores by Equation 5. The normalized weights α_i are nonnegative and sum to one. Using the same weights for both profile averages keeps the reconstructed input and cross-moments associated with the same client mixture. The capped u_i instead define a harmonic-type global scale $\hat{\sigma}_g^2$ without an explicit row-count multiplier. The distinction prevents sample support from being counted in both that scale and the pooled aspect ratio. The local spectral proxy can still depend on n_i through its cutoff. Capping both scores limits extreme contributions to both the profiles and the global scale. Once a score lies above the lower median, increasing that score alone does not change the median reference. Proposition 3 bounds each weight by $\kappa/(\kappa + \lfloor N/2 \rfloor)$, which still permits a large individual share in small cohorts. Capping changes cross-client pooling only, leaving each client's own $\hat{\sigma}_i^2$ and dense moments untouched.

The pooled profiles now supply the two sides of a global normal equation. Both reconstruction maps are linear, so reconstructing after pooling gives the same matrices as pooling the client reconstructions. The server therefore forms only one reconstructed Gram moment and one reconstructed

cross-moment, rather than a dense pair for every client. Lines 11-13 set the ridge from two quantities already uploaded, the pooled scale and the total row count, with upper ridge bound $\lambda_{\text{cap}} > 0$, and solve Equation 6. The ratio q_{pool} compares input dimension with total sample support. Fewer rows or a larger pooled scale increase the ridge, with the other quantity held fixed, until the cap is reached. This rule uses information already in the upload, avoiding a separate exchange to select the regularization strength. Its motivation follows ridge regularization and covariance shrinkage, which control sensitivity through the coefficient matrix (Hoerl & Kennard, 1970; Ledoit & Wolf, 2004). This scale-based rule does not guarantee minimum prediction risk.

Adding $\lambda \boldsymbol{I}$ shifts every eigenvalue of the reconstructed Gram moment by λ. However, diagonal averaging need not preserve positive semidefiniteness, so a positive ridge alone does not guarantee that this reconstructed system is well conditioned. Theorem 2 bounds the effect of compression through reconstruction residuals under explicit conditioning assumptions. The implementation obtains $\boldsymbol{\Theta}_g \in \mathbb{R}^{L \times H}$ by a direct linear solve.

The server retains $\boldsymbol{\Theta}_g$ for generalization, but broadcasts only the pooled profiles $(\boldsymbol{r}_g, \boldsymbol{s}_g)$. This sends $L+H$ values instead of the $L \times H$ coefficient matrix. The profiles also let each client reconstruct both shared moment terms and combine them with its own dense statistics. The next phase determines their contribution separately for each client. Line 14 sets the analytic layer bias to $\boldsymbol{b}_g = \boldsymbol{0}_H$, which adds no trained intercept and no transmitted coefficient. Line 15 broadcasts the profiles alone, while $\boldsymbol{\Theta}_g$ stays on the server for centralized evaluation.

B.3 Phase 3: Communication-Free Personalization

Algorithm 3 Phase 3 communication-free personalization at client i

Require: local $(\boldsymbol{r}_i, \hat{\sigma}_i^2, \boldsymbol{G}_i, \boldsymbol{C}_i)$, broadcast $(\boldsymbol{r}_g, \boldsymbol{s}_g)$

1: $\eta_i \leftarrow \hat{\sigma}_i^2 / \big(r_i[0] + a_0\big)$

2: $\rho_i \leftarrow \max\Big(0,\ \boldsymbol{r}_i^\top \boldsymbol{r}_g / \big(\|\boldsymbol{r}_i\|_2 \, \|\boldsymbol{r}_g\|_2 + a_0\big)\Big)$

3: $\gamma_i \leftarrow \eta_i^2 \, \rho_i$

4: $\boldsymbol{A}_i \leftarrow \boldsymbol{G}_i + \gamma_i \, \mathcal{T}(\boldsymbol{r}_g) + \gamma_i \boldsymbol{I}$

5: $\boldsymbol{B}_i \leftarrow \boldsymbol{C}_i + \gamma_i \, \mathcal{H}(\boldsymbol{r}_g, \boldsymbol{s}_g)$

6: $\boldsymbol{\Theta}_i \leftarrow \boldsymbol{A}_i^{-1} \boldsymbol{B}_i$

7: $\boldsymbol{b}_i \leftarrow \boldsymbol{0}_H$

Each client must decide how much of the shared structure to include in its local solve. The strength should distinguish the amount of lower-bulk variation from whether the shared temporal pattern agrees with the client. Two signals implement this distinction using quantities already available after the broadcast. Line 1 computes the normalized bulk scale, line 2 the nonnegative lag-profile alignment, and line 3 combines them as in Equation 7,

$$\eta_i = \frac{\hat{\sigma}_i^2}{r_i[0] + a_0}, \qquad \rho_i = \max\left\{0, \frac{\boldsymbol{r}_i^\top \boldsymbol{r}_g}{\|\boldsymbol{r}_i\|_2 \|\boldsymbol{r}_g\|_2 + a_0}\right\}, \qquad \gamma_i = \eta_i^2 \rho_i.$$

Since $0 < \eta_i \leq 1$ (Lemma 7), squaring suppresses small ratios while leaving values near one intact. Clipping ρ_i at zero removes the global terms when the local profile is orthogonal or opposed to the pool. Thus a large γ_i requires both a large normalized bulk scale and strong alignment with the pooled lag profile. Neither signal alone is sufficient. A high bulk ratio produces little shared contribution when profile agreement is weak, while strong agreement does not override a small bulk ratio. Personalization uses the uncapped $\hat{\sigma}_i^2$, because the cap limits cross-client influence rather than modifying local statistics, although capping still reaches γ_i indirectly through $\boldsymbol{r}_g$ and its alignment with $\boldsymbol{r}_i$.

Lines 4 and 5 form the coefficient matrix and right-hand side, and line 6 solves the system in Equation 8,

$$\boldsymbol{A}_i = \boldsymbol{G}_i + \gamma_i \mathcal{T}(\boldsymbol{r}_g) + \gamma_i \boldsymbol{I}, \qquad \boldsymbol{B}_i = \boldsymbol{C}_i + \gamma_i \mathcal{H}(\boldsymbol{r}_g, \boldsymbol{s}_g), \qquad \boldsymbol{\Theta}_i = \boldsymbol{A}_i^{-1} \boldsymbol{B}_i.$$

Here $\boldsymbol{A}_i \in \mathbb{R}^{L \times L}$ is the combined coefficient matrix and $\boldsymbol{B}_i \in \mathbb{R}^{L \times H}$ is its cross-moment right-hand side. The construction is motivated by shrinkage, which combines local estimation with shared

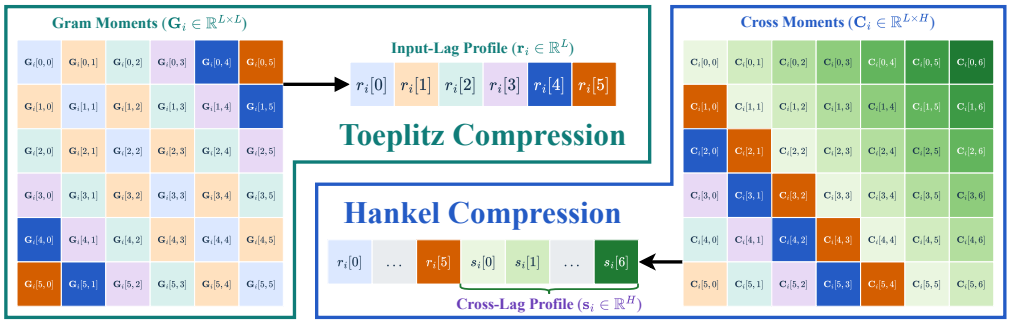


Figure B1: Toeplitz and reverse-Hankel compression of the dense client moments. Entries sharing a color share a lag.

information (James & Stein, 1961; Efron & Morris, 1975). Here that combination takes place in the normal equations, not through interpolation of parameter matrices. The shared Gram and cross-moments enter the coefficient matrix and right-hand side with the same γ_i. Their effect on the final coefficients also depends on the client's dense Gram moment, so γ_i is not a direct mixture weight on predictions.

At $\gamma_i = 0$ with nonsingular $\boldsymbol{G}_i$, the solution reduces to the local fit $\boldsymbol{G}_i^{-1}\boldsymbol{C}_i$ based on the terminal robust moments. For positive γ_i, the client incorporates the shared structure while retaining the position-specific relations in $\boldsymbol{G}_i, \boldsymbol{C}_i$. Every quantity needed by Equation 8 is either retained locally or supplied by the single broadcast, so adaptation requires no further communication.

The identity term regularizes the system, but does not by itself ensure nonsingularity when γ_i vanishes or the reconstructed Gram moment is indefinite. Corollary 2 establishes eventual nonsingularity under its conditioning assumptions. Neither profile agreement nor the moment-level shrinkage construction guarantees lower forecasting risk. We therefore evaluate personalization separately to assess its effect on local predictions. Line 7 sets $\boldsymbol{b}_i = \mathbf{0}_H$, matching the zero-bias global model and the common linear backbone.

B.4 Structural Subroutines

Algorithm 4 implements the extraction and reconstruction maps in Equation 2 and Equation 3. Extraction converts a dense input moment into one value per time separation; reconstruction places those values into the matrix shapes required by the analytic solves. The routines use zero-based indices and the same input-window ordering on every client.

Figure B1 illustrates how lag grouping changes the two dense moments. Matching colors identify entries assigned to the same time separation, making the repeated diagonals in the input moment and the shifted pattern in the cross-moment visible. The compact profiles preserve these shared relationships, while differences between positions at the same lag remain only in the dense moments retained locally.

Algorithm 4 Subroutines

function TOEPLITZ($\boldsymbol{r} \in \mathbb{R}^L$)
  return $\big[r[|a-b|]\big]_{0\le a,b<L} \in \mathbb{R}^{L\times L}$
end function
function HANKEL($\boldsymbol{r} \in \mathbb{R}^L$, $\boldsymbol{s} \in \mathbb{R}^H$)
  $\boldsymbol{v} \leftarrow (r[0],\dots,r[L-1],s[0],\dots,s[H-1])$
  return $\big[v[L+h-a]\big]_{\substack{0\le a<L\\ 0\le h<H}} \in \mathbb{R}^{L\times H}$ ▷ *rows are stored in reverse-Hankel order*
end function
function EXTRACTAUTOCORR($\boldsymbol{\Sigma} \in \mathbb{R}^{L\times L}$)
  for $\ell = 0,\dots,L-1$ **do**
    $r[\ell] \leftarrow (L-\ell)^{-1}\sum_{a=0}^{L-\ell-1} \boldsymbol{\Sigma}_{a,a+\ell}$
  end for
  return $\boldsymbol{r}$
end function

Extracting the input-lag profile. For a symmetric input moment matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{L\times L}$, the entries $\boldsymbol{\Sigma}[a, a+\ell]$ all describe input positions separated by ℓ steps. ExtractAutocorr averages these entries along the upper diagonal at that separation,

$$r[\ell] = \frac{1}{L-\ell}\sum_{a=0}^{L-\ell-1} \boldsymbol{\Sigma}[a, a+\ell], \qquad 0 \le \ell < L.$$

The denominator is the number of entries on that diagonal, so longer lags are not made smaller merely because fewer position pairs contribute. At zero lag, $r[0] = \mathrm{Tr}(\boldsymbol{\Sigma})/L$ is the mean input energy; at the longest lag, $r[L-1] = \boldsymbol{\Sigma}[0, L-1]$ contains a single entry. Symmetry makes the lower diagonals redundant. Despite the routine's name, these are weighted raw second moments, not correlation coefficients normalized to unit variance.

Reconstructing the input moment. Toeplitz assigns $r[|a-b|]$ to entry (a, b), giving every pair with the same separation the same value. This produces a symmetric matrix with constant diagonals and preserves the extracted lag averages. It cannot recover differences between positions within a diagonal, because extraction has replaced them by their average. For symmetric input matrices, this averaging is the Frobenius projection onto the symmetric Toeplitz subspace characterized in Proposition 1. Projection onto that subspace does not in general preserve positive semidefiniteness, which is why the solve analysis states separate conditioning assumptions.

Reconstructing the cross-moment. Hankel also uses the anchor $\boldsymbol{s} = \boldsymbol{C}[0,:]$, whose entries relate the earliest input position to each future target. In a contiguous window, input position a and target h are separated by $L + h - a$ steps. Concatenating the two profiles therefore provides the required lookup,

$$\boldsymbol{v} = (r[0],\dots,r[L-1],s[0],\dots,s[H-1]), \qquad \mathcal{H}(\boldsymbol{r},\boldsymbol{s})[a,h] = v[L+h-a].$$

Separations below L use the input profile; separations from L onward use the anchor. Consequently, the first reconstructed row equals $\boldsymbol{s}$, and each subsequent row shifts the lookup one position toward shorter lags. The smallest queried separation is one, so $r[0]$ is used only in the Gram-moment diagonal, not in the cross-moment reconstruction. Reversing the input-row order makes the entries constant along anti-diagonals, explaining the reverse-Hankel ordering in the algorithm.

A small example. Let $L = 3$ and $H = 2$, with the input moment and cross-moment anchor

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 3 \\ 1 & 4 & 5 \\ 3 & 5 & 6 \end{pmatrix}, \qquad \boldsymbol{s} = (7, 8).$$

Extraction gives $\boldsymbol{r} = ((2+4+6)/3, (1+5)/2, 3) = (4, 3, 3)$, so $\boldsymbol{v} = (4, 3, 3, 7, 8)$ and

$$\mathcal{T}(\boldsymbol{r}) = \begin{pmatrix} 4 & 3 & 3 \\ 3 & 4 & 3 \\ 3 & 3 & 4 \end{pmatrix}, \qquad \mathcal{H}(\boldsymbol{r},\boldsymbol{s}) = \begin{pmatrix} 7 & 8 \\ 3 & 7 \\ 3 & 3 \end{pmatrix}.$$

The input reconstruction replaces the unequal diagonal and one-step entries by their averages. The cross-moment reconstruction keeps the anchor row and fills the remaining positions from the same lag lookup; it does not recover untransmitted cross-moment entries. Exact recovery requires the lag structure specified in Lemma 6, whereas the finite weighted moments generally yield an approximation. All three routines are linear in their inputs, allowing the server to pool profiles before reconstruction while clients retain the original dense moments for personalization.

C Theoretical Analysis

C.1 Assumptions

The guarantees concern the displayed equations in exact arithmetic as every client's row count increases, with L, H, N fixed. Communication-round convergence and unlimited-iteration IRLS are outside this regime. A star denotes a population counterpart. Quantities computed from the sample remain unstarred. Hats mark empirical fits, quantile-based scales, spectral statistics, and stopping quantities, while moments, profiles, and derived weights retain their algorithmic notation. In particular, $(\boldsymbol{G}_i^*, \boldsymbol{C}_i^*, \boldsymbol{r}_i^*, \boldsymbol{s}_i^*, \sigma_i^{2,*})$ is obtained from the same finite local procedure with population expectations and quantiles. The complete population recursion is defined in Section C.3. The precision-cap multiplier remains $\kappa = 10$.

Assumption 1 (Dependent-row regularity)**.** *For every client, $\mathbb{E}\|\boldsymbol{x}\|_2^2 + \mathbb{E}[\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2] < \infty$. Write $\mathbb{P}_{n_i} g = n_i^{-1} \sum_{t=1}^{n_i} g(\boldsymbol{x}_t, \boldsymbol{y}_t)$ and $\mathbb{P}_i g = \mathbb{E}_i[g(\boldsymbol{x}, \boldsymbol{y})]$. The local function class $\mathcal{G}_i$ specified in Section C.3 satisfies*

$$U_{i,n_i} := \sup_{g \in \mathcal{G}_i} |\mathbb{P}_{n_i} g - \mathbb{P}_i g| \xrightarrow{p} 0.$$

The class contains the raw and weighted moment coordinates and the indicators needed to estimate residual medians and median absolute deviations. Window overlap alone does not establish this uniform law.

Assumption 2 (Residual-quantile regularity)**.** *Let $z_i(\boldsymbol{\Theta}) = \|\boldsymbol{y} - \boldsymbol{\Theta}^\top \boldsymbol{x}\|_2$ for a random row $(\boldsymbol{x}, \boldsymbol{y})$ from client i. Its value at observed row t and fit $\hat{\boldsymbol{\Theta}}_i^{(k)}$ is the residual norm $z_t^{(k)}$ used in Phase 1. For $k = 0, \ldots, T-1$, the residual norm at $\boldsymbol{\Theta}_i^{(k),*}$ and its absolute deviation from its median have unique medians. Their densities are continuous and strictly positive near those medians. For a neighborhood $\mathcal{N}_{ik}$ of the population fit and cutoff $(\boldsymbol{\Theta}_i^{(k),*}, \delta_i^{(k),*})$, assume*

$$\sup_{(\boldsymbol{\Theta}, u) \in \mathcal{N}_{ik}} \mathbb{P}_i\{|z_i(\boldsymbol{\Theta}) - u| \le \zeta\} \le C_{\mathrm{ac}} \zeta, \qquad 0 < \zeta < \zeta_0,$$

where $C_{\mathrm{ac}}, \zeta_0 > 0$. In particular, the population residual distribution has no atom at a Huber cutoff.

Assumption 3 (Finite recursion and separated stopping decisions)**.** *The population procedure uses the same $\epsilon, T, \tau, a_0, b_0, d_0, c_{\mathrm{H}}, b_{\mathrm{MAD}}$ as Algorithm 1. Its relative changes $d_i^{(k),*}$ satisfy $d_i^{(k),*} \neq \tau$ for $k = 1, \ldots, T-1$. Both empirical and population recursions can be continued to T for analysis, even when the implemented procedure stops earlier. The target is the output selected by that finite stopping rule, not a converged Huber M-estimator (Huber, 1981; Hampel et al., 1986; Maronna et al., 2006).*

Assumption 4 (Terminal conditioning)**.** *For every client, both $\boldsymbol{G}_i^*$ and $\mathcal{T}(\boldsymbol{r}_i^*)$ have eigenvalues in $[m_w, M_w]$, where $0 < m_w \le M_w < \infty$. The Toeplitz condition is separate because diagonal averaging need not preserve positive semidefiniteness.*

Assumption 5 (Spectral separation and client proportions)**.** *For $n = \sum_j n_j$, the client proportions satisfy $n_i/n \to c_i > 0$, with $\sum_i c_i = 1$. The population scales satisfy $0 < \sigma_{\min}^2 \le \sigma_i^{2,*} \le \sigma_{\max}^2 < \infty$. When $L \ge 2$, no eigenvalue of $\boldsymbol{G}_i^*$ equals its mean $\bar{\mu}_i^*$. This separation stabilizes the selected eigenvalue set as the empirical cutoff tends to the mean. At $L = 1$, the selected set is always $\{1\}$, without a separation condition.*

We state the uniform laws directly, with mixing and empirical-process methods providing the relevant background (Doukhan, 1994; Andrews, 1994). The stronger rate assumptions below are additional conditions, not consequences of finite second moments alone.

Table C1: Information retained under lag-summary compression.

Quantity	Fate under compression
Preserved exactly	
$r_i[\ell]$, $s_i[h]$, n_i	Transmitted diagonal averages, first-row cross-moments, and row count
$\hat{\sigma}_i^2$	Computed from $\boldsymbol{G}_i$ before compression and transmitted
$\mathrm{DFT}(\boldsymbol{r}_i)$	Determined by $\boldsymbol{r}_i$
Replaced by the reconstruction	
$\boldsymbol{G}_i[a, b]$	Replaced by its lag-diagonal average $r_i[\lvert a-b\rvert]$
$\boldsymbol{C}_i[a, h]$	First row retained; other rows reconstructed from both profiles
$\mathrm{diag}(\boldsymbol{G}_i)$	Reconstructed as the constant diagonal $r_i[0]$
$\boldsymbol{G}_i^{-1}$	Replaced by $\mathcal{T}(\boldsymbol{r}_i)^{-1}$ when both inverses exist
Eigenvalues and eigenvectors of $\boldsymbol{G}_i$	Full eigensystem not generally determined; the transmitted bulk scale retains partial spectral information
$\boldsymbol{G}_i^{-1}\boldsymbol{C}_i$	Reconstructed solve agrees if both moment residuals vanish and the inverses exist
Not generally recoverable from the upload	
$\boldsymbol{\Delta}_{xx,i}$, $\boldsymbol{\Delta}_{xy,i}$	Position-specific reconstruction residuals
Not generally recoverable from either representation	
$\boldsymbol{x}_t$, $\boldsymbol{y}_t$	Individual rows are aggregated when moments are formed

C.2 Information Content of the Compression

Table C1 distinguishes the terminal dense moments retained locally from the quantities determined by the logical upload. The server receives the profiles, row count, and spectral scale, but cannot generally recover the position-specific reconstruction residuals. This distinction explains why personalization can use information unavailable to the global solve, without implying that the summaries provide a privacy guarantee.

The ridge-free reconstruction uses both profiles because $\boldsymbol{r}_i$ determines the Toeplitz factor, while $(\boldsymbol{r}_i, \boldsymbol{s}_i)$ determine the cross-moment factor. When the inverses exist, $\boldsymbol{G}_i = \mathcal{T}(\boldsymbol{r}_i)$ and $\boldsymbol{C}_i = \mathcal{H}(\boldsymbol{r}_i, \boldsymbol{s}_i)$ are sufficient for equality of the two solves. They are not necessary because nonzero moment residuals can cancel in the solve. Table C2 reports relative finite-sample moment reconstruction errors, not inverse-matrix or parameter errors.

Let $\mathcal{T}_L$ be the symmetric Toeplitz subspace of the symmetric $L \times L$ matrices, equipped with the Frobenius inner product. Let

$$\mathcal{D}_0 = \{\mathrm{diag}(d_1, \ldots, d_L) : \textstyle\sum_{a=1}^{L} d_a = 0\}, \qquad \Pi_{\mathcal{T}}(\boldsymbol{G}_i) = \mathcal{T}(\boldsymbol{r}_i).$$

Here $\Pi_{\mathcal{T}}$ denotes diagonal averaging followed by reconstruction. Proposition 1 characterizes the information removed by this lag map alone, without conditioning on the additional transmitted spectral scale.

Proposition 1 (The withheld Gram information is the Toeplitz orthogonal complement)**.** *The zero-trace diagonal subspace satisfies $\mathcal{D}_0 \subseteq \mathcal{T}_L^{\perp}$ and has dimension $L-1$. Adding an element of $\mathcal{D}_0$ to a symmetric matrix leaves $\Pi_{\mathcal{T}}$ unchanged. Across the symmetric preimage of $\mathcal{T}(\boldsymbol{r}_i)$, the Gram residual $\boldsymbol{\Delta}_{xx,i}$ ranges over $\mathcal{T}_L^{\perp}$, which has dimension $L(L-1)/2$. If $\mathcal{T}(\boldsymbol{r}_i) \succ 0$, its positive-definite preimage contains a relatively open neighborhood in every direction of that complement.*

This result identifies dense Gram information that the compressed profile cannot recover. It is not a privacy guarantee.

Proof. The matrices $\boldsymbol{T}_\ell$, with ones at positions $\lvert a-b\rvert = \ell$ and zeros elsewhere, form an orthogonal basis of $\mathcal{T}_L$. For $\boldsymbol{D} \in \mathcal{D}_0$,

$$\langle \boldsymbol{D}, \boldsymbol{T}_0 \rangle_F = \mathrm{Tr}(\boldsymbol{D}) = 0, \qquad \langle \boldsymbol{D}, \boldsymbol{T}_\ell \rangle_F = 0 \quad (\ell \geq 1),$$

since $\boldsymbol{T}_0 = \boldsymbol{I}$ and $\boldsymbol{D}$ has no off-diagonal entries. Thus $\mathcal{D}_0 \subseteq \mathcal{T}_L^{\perp}$. One linear trace constraint on L diagonal entries gives $\dim(\mathcal{D}_0) = L-1$.

For a symmetric matrix $\boldsymbol{G}$, the projection coefficients are

$$\frac{\langle \boldsymbol{G}, \boldsymbol{T}_0 \rangle_F}{\|\boldsymbol{T}_0\|_F^2} = \frac{\mathrm{Tr}(\boldsymbol{G})}{L}, \qquad \frac{\langle \boldsymbol{G}, \boldsymbol{T}_\ell \rangle_F}{\|\boldsymbol{T}_\ell\|_F^2} = \frac{2\sum_{a=0}^{L-\ell-1} \boldsymbol{G}[a, a+\ell]}{2(L-\ell)} \quad (\ell \geq 1).$$

These are exactly the diagonal averages defining $\Pi_{\mathcal{T}}$. Its kernel is therefore $\mathcal{T}_L^\perp$, with dimension

$$\frac{L(L+1)}{2} - \dim(\mathcal{T}_L) = \frac{L(L+1)}{2} - L = \frac{L(L-1)}{2}.$$

Every matrix in the preimage is uniquely $\mathcal{T}(\boldsymbol{r}_i) + \boldsymbol{D}$ for some $\boldsymbol{D} \in \mathcal{T}_L^\perp$. If $\mathcal{T}(\boldsymbol{r}_i) \succ 0$, Weyl's inequality gives

$$\lambda_{\min}(\mathcal{T}(\boldsymbol{r}_i) + \boldsymbol{D}) \geq \lambda_{\min}(\mathcal{T}(\boldsymbol{r}_i)) - \|\boldsymbol{D}\|_2 > 0$$

whenever $\|\boldsymbol{D}\|_2 < \lambda_{\min}(\mathcal{T}(\boldsymbol{r}_i))$. Hence all sufficiently small perturbations in the complement preserve positive definiteness. □

The spectral scalar can further restrict this preimage without generally identifying the dense moments. For example, at $L = 2$, $\mathrm{diag}(1, 3)$ and $\mathrm{diag}(3, 1)$ share $\boldsymbol{r}_i = (2, 0)^\top$ and the same eigenspectrum. With equal n_i and $\boldsymbol{C}_i = \boldsymbol{0}$, they also share $\boldsymbol{s}_i = \boldsymbol{0}_H$ and $\hat{\sigma}_i^2$, yielding identical full uploads but different Gram residuals. For the cross-moment, holding $\boldsymbol{G}_i$ and the first row of $\boldsymbol{C}_i$ fixed leaves the upload unchanged when the remaining rows of $\boldsymbol{C}_i$ vary. These examples establish nonuniqueness, not a privacy guarantee.

C.3 Population Target

The population counterpart of Algorithm 1 replaces empirical moments and lower sample medians with expectations and population quantiles. Throughout, L, H, N are fixed and every $n_i \to \infty$, unless the horizon limit is explicitly considered.

For client law $\mathbb{P}_i$, use the population median

$$\mathrm{med}_{\mathbb{P}_i} Z = \inf\{z : \mathbb{P}_i(Z \leq z) \geq \tfrac{1}{2}\}.$$

Define the residual and absolute-deviation maps by

$$\begin{aligned} z_i(\boldsymbol{\Theta}) &= \|\boldsymbol{y} - \boldsymbol{\Theta}^\top \boldsymbol{x}\|_2, \\ m_i(\boldsymbol{\Theta}) &= \mathrm{med}_{\mathbb{P}_i} z_i(\boldsymbol{\Theta}), \\ D_i(\boldsymbol{\Theta}, m) &= \mathrm{med}_{\mathbb{P}_i} |z_i(\boldsymbol{\Theta}) - m|. \end{aligned}$$

Here m_i is the population residual median, while D_i is the absolute-deviation median about a supplied center m. For $z \geq 0$ and $u > 0$, define the implemented weight by

$$w(z, u) = \begin{cases} 1, & z \leq u, \\ u/(z + b_0), & z > u. \end{cases}$$

This map satisfies $0 < w(z, u) \leq 1$. Because $b_0 > 0$, it has a small jump at $z = u$. The no-atom condition in Assumption 2 is therefore relevant even though the safeguard is numerically small.

Start with

$$\mathcal{S}_{xx,i}^* = \mathbb{P}_i[\boldsymbol{x}\boldsymbol{x}^\top], \qquad \mathcal{S}_{xy,i}^* = \mathbb{P}_i[\boldsymbol{x}\boldsymbol{y}^\top], \qquad \boldsymbol{\Theta}_i^{(0),*} = (\mathcal{S}_{xx,i}^* + \epsilon \boldsymbol{I})^{-1} \mathcal{S}_{xy,i}^*.$$

For $k = 0, \ldots, T-1$, compute

$$\begin{aligned} \tilde{z}_i^{(k),*} &= m_i(\boldsymbol{\Theta}_i^{(k),*}), \\ \sigma_{e,i}^{(k),*} &= \max\left\{ \frac{D_i(\boldsymbol{\Theta}_i^{(k),*}, \tilde{z}_i^{(k),*})}{b_{\mathrm{MAD}}}, a_0 \right\}, \\ \delta_i^{(k),*} &= \tilde{z}_i^{(k),*} + c_{\mathrm{H}} \sigma_{e,i}^{(k),*}, \\ w_i^{(k),*} &= w\left(z_i(\boldsymbol{\Theta}_i^{(k),*}), \delta_i^{(k),*} \right), \\ \boldsymbol{G}_i^{(k+1),*} &= \mathbb{P}_i[w_i^{(k),*} \boldsymbol{x}\boldsymbol{x}^\top], \qquad \boldsymbol{C}_i^{(k+1),*} = \mathbb{P}_i[w_i^{(k),*} \boldsymbol{x}\boldsymbol{y}^\top], \\ \boldsymbol{\Theta}_i^{(k+1),*} &= (\boldsymbol{G}_i^{(k+1),*} + \epsilon \boldsymbol{I})^{-1} \boldsymbol{C}_i^{(k+1),*}. \end{aligned} \tag{C1}$$

The constants are exactly those in Algorithm 1. The empirical recursion uses $\mathbb{P}_{n_i}$ and lower sample medians in place of $\mathbb{P}_i$ and its quantiles. Its fits are $\hat{\boldsymbol{\Theta}}_i^{(k)}$, and its weighted moments are $\boldsymbol{G}_i^{(k)}$ and $\boldsymbol{C}_i^{(k)}$, as in the main method.

For $k = 1, \dots, T$, let

$$d_i^{(k),*} = \frac{\|\boldsymbol{\Theta}_i^{(k),*} - \boldsymbol{\Theta}_i^{(k-1),*}\|_F}{\|\boldsymbol{\Theta}_i^{(k-1),*}\|_F + d_0}, \qquad K_i^* = \min\Big(\{k \in \{1, \dots, T-1\} : d_i^{(k),*} < \tau\} \cup \{T\}\Big).$$

Define $\hat{d}_i^{(k)}$ and $\hat{K}_i$ by the same expressions using empirical fits. The terminal moments are

$$\boldsymbol{G}_i^* = \boldsymbol{G}_i^{(K_i^*),*}, \qquad \boldsymbol{C}_i^* = \boldsymbol{C}_i^{(K_i^*),*}, \qquad \boldsymbol{G}_i = \boldsymbol{G}_i^{(\hat{K}_i)}, \qquad \boldsymbol{C}_i = \boldsymbol{C}_i^{(\hat{K}_i)}.$$

Continuing the recursion after its first stopping decision is only a device for defining all candidate iterates on one probability space. The actual procedure still stops at $\hat{K}_i$. The terminal moments use weights computed from the preceding fit and also determine the terminal ridged fit.

Apply Equation 2 to $(\boldsymbol{G}_i^*, \boldsymbol{C}_i^*)$ to obtain $(\boldsymbol{r}_i^*, \boldsymbol{s}_i^*)$. Write the ordered eigenvalues of $\boldsymbol{G}_i^*$ as $\mu_{i1}^* \le \cdots \le \mu_{iL}^*$, and set

$$\bar{\mu}_i^* = \frac{1}{L}\sum_{\ell=1}^{L} \mu_{i\ell}^*, \qquad \mathcal{I}_i^* = \{\ell \in \{1, \dots, L\} : \mu_{i\ell}^* \le \bar{\mu}_i^*\}, \qquad \sigma_i^{2,*} = \max\left\{a_0, \frac{1}{|\mathcal{I}_i^*|}\sum_{\ell \in \mathcal{I}_i^*} \mu_{i\ell}^*\right\}.$$

Here $|\mathcal{I}_i^*|$ denotes the cardinality of the selected index set. The sample eigenvalues $\hat{\mu}_{i\ell}$ use the same increasing order. Their mean, cutoff, selected set, and scale are precisely those of Equation 4, as lines 21-25 of Algorithm 1 compute them. At fixed L, the factor $(1 + \sqrt{L/n_i})^2$ tends to one, which explains why the population cutoff is $\bar{\mu}_i^*$. Neither selected set is empty because the smallest nonnegative eigenvalue is no larger than the mean, and the sample cutoff is at least the sample mean. This is the fixed-dimensional target of the spectral heuristic, not a claim that the dependent weighted Gram matrix follows the classical MP law.

Function classes and rate conditions. For later use, define the population weighted-moment maps

$$\boldsymbol{M}_{xx,i}(\boldsymbol{\Theta}, u) = \mathbb{P}_i[w(z_i(\boldsymbol{\Theta}), u)\boldsymbol{x}\boldsymbol{x}^\top], \qquad \boldsymbol{M}_{xy,i}(\boldsymbol{\Theta}, u) = \mathbb{P}_i[w(z_i(\boldsymbol{\Theta}), u)\boldsymbol{x}\boldsymbol{y}^\top].$$

The class $\mathcal{G}_i$ in Assumption 1 contains each raw moment coordinate and each coordinate of these weighted integrands. It also contains the indicators $\mathbf{1}\{z_i(\boldsymbol{\Theta}) \le v\}$ and $\mathbf{1}\{|z_i(\boldsymbol{\Theta}) - m| \le v\}$. The parameters range over fixed neighborhoods of all population fits, centers, quantiles, and cutoffs needed by the finite recursion.

C.3.1 Additional conditions for the root-n rate

For rate statements, strengthen the uniform law to

$$U_{i,n_i} = O_p(n_i^{-1/2}).$$

The residual and absolute-deviation densities must be bounded below by a positive constant near their medians, uniformly over the parameter neighborhoods just specified. In addition, require the deterministic population maps

$$m_i(\boldsymbol{\Theta}), \qquad D_i(\boldsymbol{\Theta}, m), \qquad \boldsymbol{M}_{xx,i}(\boldsymbol{\Theta}, u), \qquad \boldsymbol{M}_{xy,i}(\boldsymbol{\Theta}, u)$$

to be locally Lipschitz in their arguments. For example, the weighted maps must satisfy

$$\begin{aligned}&\|\boldsymbol{M}_{xx,i}(\boldsymbol{\Theta}, u) - \boldsymbol{M}_{xx,i}(\boldsymbol{\Theta}', u')\|_F \\ &\quad + \|\boldsymbol{M}_{xy,i}(\boldsymbol{\Theta}, u) - \boldsymbol{M}_{xy,i}(\boldsymbol{\Theta}', u')\|_F \le C_i\big(\|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|_F + |u - u'|\big)\end{aligned}$$

for a finite C_i on each relevant neighborhood. Finally, require $n_i/n - c_i = O(n^{-1/2})$. An unweighted cutoff margin and integrable moments give continuity of the weighted maps, but do not by themselves give these Lipschitz bounds. The rate conditions are additional assumptions and have not been verified for the benchmark data-generating processes (Andrews, 1994; van der Vaart & Wellner, 1996).

C.4 Existence and Consistency of the Local IRLS Statistic

Lemma 1 (Existence of the finite recursion). *The moment conditions in Assumption 1 and the quantile conditions in Assumption 2 make Equation C1 well defined through T.*

Thus the population reference for every permitted local update is well defined under the stated moment and quantile conditions.

Proof. For any weight $0 \le w \le 1$,

$$\|\mathbb{P}_i[w\boldsymbol{x}\boldsymbol{x}^\top]\|_2 \le \mathbb{E}_i\|\boldsymbol{x}\|_2^2 < \infty, \qquad \|\mathbb{P}_i[w\boldsymbol{x}\boldsymbol{y}^\top]\|_F \le \mathbb{E}_i[\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2] < \infty.$$

These bounds apply to the seed and every weighted update.

For every vector $\boldsymbol{v}$,

$$\boldsymbol{v}^\top\mathbb{P}_i[w\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{v} = \mathbb{P}_i[w(\boldsymbol{v}^\top\boldsymbol{x})^2] \ge 0.$$

Consequently,

$$\boldsymbol{G}_i^{(k),*} + \epsilon\boldsymbol{I} \succeq \epsilon\boldsymbol{I}, \qquad \|(\boldsymbol{G}_i^{(k),*} + \epsilon\boldsymbol{I})^{-1}\|_2 \le \epsilon^{-1}.$$

The same argument applies to $\mathcal{S}_{xx,i}^* + \epsilon\boldsymbol{I}$.

A finite fit gives a finite residual norm almost surely. Finite population medians and the scale floor give $\delta_i^{(k),*} \ge c_{\mathrm{H}}a_0 > 0$. Together with the moment and inverse bounds, this defines the next weight and ridged fit. Since $d_0 > 0$, every stopping comparison is also defined. Induction from the seed establishes existence through T. □

Lemma 2 (Consistency of the implemented local statistic). *Under Assumptions 1-5,*

$$\hat{\boldsymbol{\Theta}}_i^{(k)} \xrightarrow{p} \boldsymbol{\Theta}_i^{(k),*} \quad (k = 0, \dots, T), \qquad \Pr(\hat{K}_i = K_i^*) \to 1,$$

and

$$(\boldsymbol{G}_i, \boldsymbol{C}_i, \boldsymbol{r}_i, \boldsymbol{s}_i, \hat{\sigma}_i^2) \xrightarrow{p} (\boldsymbol{G}_i^*, \boldsymbol{C}_i^*, \boldsymbol{r}_i^*, \boldsymbol{s}_i^*, \sigma_i^{2,*}).$$

Under these assumptions, data-dependent reweighting and stopping preserve consistency of the local summaries uploaded to the server.

Proof. Lemma 1 ensures that the population fits are defined. The uniform law includes the raw moment coordinates, so $\mathcal{S}_{xx,i} \xrightarrow{p} \mathcal{S}_{xx,i}^*$ and $\mathcal{S}_{xy,i} \xrightarrow{p} \mathcal{S}_{xy,i}^*$. Subtracting the two seed normal equations gives

$$\hat{\boldsymbol{\Theta}}_i^{(0)} - \boldsymbol{\Theta}_i^{(0),*} = (\mathcal{S}_{xx,i} + \epsilon\boldsymbol{I})^{-1}\left[(\mathcal{S}_{xy,i} - \mathcal{S}_{xy,i}^*) - (\mathcal{S}_{xx,i} - \mathcal{S}_{xx,i}^*)\boldsymbol{\Theta}_i^{(0),*}\right].$$

The inverse has norm at most ϵ^{-1}, and both differences on the right tend to zero. Hence the seed fit converges.

Suppose the claim holds at iterate k. The residual norm satisfies

$$|z_i(\hat{\boldsymbol{\Theta}}_i^{(k)}) - z_i(\boldsymbol{\Theta}_i^{(k),*})| \le \|\boldsymbol{x}\|_2\|\hat{\boldsymbol{\Theta}}_i^{(k)} - \boldsymbol{\Theta}_i^{(k),*}\|_F.$$

For deterministic fits tending to the population fit, the right side tends to zero almost surely under $\mathbb{P}_i$. Thus the corresponding population residual distributions converge at their continuity points. The uniform empirical law controls the extra error from evaluating the residual indicators at the random empirical fit. For the limiting residual cumulative distribution function (CDF) F and its unique median m, continuity and positive local density give, for sufficiently small $h > 0$,

$$F(m - h) < \tfrac{1}{2} < F(m + h).$$

The empirical distribution obeys the same inequalities with probability tending to one, placing its lower median between $m - h$ and $m + h$. This proves $\tilde{z}_i^{(k)} \xrightarrow{p} \tilde{z}_i^{(k),*}$. Applying the same argument to absolute deviations, using

$$\big||z - \tilde{z}| - |z^* - \tilde{z}^*|\big| \le |z - z^*| + |\tilde{z} - \tilde{z}^*|,$$

proves convergence of the MAD. The maximum with a_0 and the threshold addition are continuous, so

$$\hat{\sigma}_{e,i}^{(k)} \xrightarrow{p} \sigma_{e,i}^{(k),*}, \qquad \delta_i^{(k)} \xrightarrow{p} \delta_i^{(k),*}.$$

For a raw moment coordinate f, write $w_{i,n}^{(k)} = w(z_i(\hat{\boldsymbol{\Theta}}_i^{(k)}), \delta_i^{(k)})$. Then

$$\mathbb{P}_{n_i}[w_{i,n}^{(k)} f] - \mathbb{P}_i[w_i^{(k),*} f] = (\mathbb{P}_{n_i} - \mathbb{P}_i)[w_{i,n}^{(k)} f] + \mathbb{P}_i[(w_{i,n}^{(k)} - w_i^{(k),*}) f].$$

The first term is bounded in absolute value by U_{i,n_i} once the parameters lie in the specified neighborhood. For the second term, the weight map is continuous away from the cutoff boundary. Assumption 2 makes that boundary a null set. Moreover, its absolute integrand is bounded by $2(\|\boldsymbol{x}\|_2^2 + \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2)$. Dominated convergence therefore makes the population weighted-moment maps continuous at the limiting fit and cutoff. Evaluation at consistent random parameters now gives

$$\boldsymbol{G}_i^{(k+1)} \xrightarrow{p} \boldsymbol{G}_i^{(k+1),*}, \qquad \boldsymbol{C}_i^{(k+1)} \xrightarrow{p} \boldsymbol{C}_i^{(k+1),*}.$$

This argument accounts for the actual discontinuous safeguard branch rather than replacing it with an idealized continuous Huber weight.

The weighted normal equations give the exact identity

$$\hat{\boldsymbol{\Theta}}_i^{(k+1)} - \boldsymbol{\Theta}_i^{(k+1),*} = (\boldsymbol{G}_i^{(k+1)} + \epsilon \boldsymbol{I})^{-1} \times \left[(\boldsymbol{C}_i^{(k+1)} - \boldsymbol{C}_i^{(k+1),*}) - (\boldsymbol{G}_i^{(k+1)} - \boldsymbol{G}_i^{(k+1),*}) \boldsymbol{\Theta}_i^{(k+1),*} \right].$$

The inverse norm is at most ϵ^{-1}. The vanishing weighted-moment differences therefore imply convergence of the fit. Induction proves convergence at every candidate iterate.

Continuity of the norm and the positive denominator imply $\hat{d}_i^{(k)} \xrightarrow{p} d_i^{(k),*}$. When $T = 1$, both procedures terminate at the sole update, so their stopping indices agree identically. For $T \geq 2$, finiteness of the recursion and separation from the threshold give

$$g_i^{\text{stop}} := \min_{1 \leq k < T} |d_i^{(k),*} - \tau| > 0, \qquad \Pr\left\{ \max_{1 \leq k < T} |\hat{d}_i^{(k)} - d_i^{(k),*}| < \tfrac{1}{2} g_i^{\text{stop}} \right\} \to 1.$$

On this event every relevant comparison with τ agrees, so $\hat{K}_i = K_i^*$. The terminal empirical moments therefore converge to the terminal population moments.

At each fixed coordinate,

$$|r_i[\ell] - r_i^*[\ell]| \leq \frac{1}{L-\ell} \sum_{a=0}^{L-\ell-1} |\boldsymbol{G}_i[a, a+\ell] - \boldsymbol{G}_i^*[a, a+\ell]|, \qquad s_i[h] - s_i^*[h] = \boldsymbol{C}_i[0,h] - \boldsymbol{C}_i^*[0,h].$$

There are only $L + H$ coordinates, so both vector differences tend to zero.

Weyl's inequality and the mean definition give

$$\max_\ell |\hat{\mu}_{i\ell} - \mu_{i\ell}^*| \leq \|\boldsymbol{G}_i - \boldsymbol{G}_i^*\|_2 \xrightarrow{p} 0, \qquad |\hat{\bar{\mu}}_i - \bar{\mu}_i^*| \leq \max_\ell |\hat{\mu}_{i\ell} - \mu_{i\ell}^*|.$$

The moving cutoff satisfies

$$\hat{\mu}_{i,\text{cut}} - \bar{\mu}_i^* = (\hat{\bar{\mu}}_i - \bar{\mu}_i^*) + \hat{\bar{\mu}}_i \left(2\sqrt{L/n_i} + L/n_i \right) \xrightarrow{p} 0.$$

For $L \geq 2$, let $g_i^{\text{spec}} = \min_\ell |\mu_{i\ell}^* - \bar{\mu}_i^*| > 0$. Whenever

$$\max_\ell |\hat{\mu}_{i\ell} - \mu_{i\ell}^*| + |\hat{\mu}_{i,\text{cut}} - \bar{\mu}_i^*| < g_i^{\text{spec}},$$

every eigenvalue remains on its population side of the cutoff. This event has probability tending to one, and hence $\Pr(\hat{\mathcal{I}}_i = \mathcal{I}_i^*) \to 1$. For $L = 1$, equality of the selected sets holds identically. On either equality event, the scale floor is 1-Lipschitz and

$$|\hat{\sigma}_i^2 - \sigma_i^{2,*}| \leq \frac{1}{|\mathcal{I}_i^*|} \sum_{\ell \in \mathcal{I}_i^*} |\hat{\mu}_{i\ell} - \mu_{i\ell}^*| \leq \max_\ell |\hat{\mu}_{i\ell} - \mu_{i\ell}^*|.$$

This proves the final convergence. □

C.5 Rate of Convergence of the Local Summaries

Lemma 3 (Root-n_i rate for the local statistic)**.** *Under Assumptions 1-5 and the additional local rate conditions in Section C.3.1, each fit, terminal moment, lag profile, and spectral scale in Lemma 2 has error* $O_p(n_i^{-1/2})$.

The stronger assumptions give a sampling-error rate for the finite local pipeline, not a convergence rate in the number of IRLS updates.

Proof. The raw moments have error $O_p(n_i^{-1/2})$ because their coordinates belong to $\mathcal{G}_i$. The seed identity in Lemma 2 gives the same rate for $\hat{\boldsymbol{\Theta}}_i^{(0)}$. For any relevant population CDF F with median q, the uniform positive density bound $f_0 > 0$ implies

$$F(q+h) \geq \tfrac{1}{2} + f_0 h, \qquad F(q-h) \leq \tfrac{1}{2} - f_0 h$$

for small $h > 0$. If the uniform empirical CDF error is at most U_{i,n_i}, choosing $h > U_{i,n_i}/f_0$ brackets the empirical lower median. Thus the empirical median error at a supplied parameter is uniformly $O_p(n_i^{-1/2})$. The same argument applies to the supplied-center absolute-deviation quantile.

By Lemma 2 and finiteness of the recursion, the estimated parameters and quantile centers lie simultaneously in the neighborhoods from Section C.3.1 with probability tending to one. The following local bounds hold on this event. Its complement contributes a vanishing probability to each tail bound and therefore does not affect the stated stochastic order.

Use the deterministic local Lipschitz bounds to split the residual-median error into

$$\begin{aligned}|\tilde{z}_i^{(k)} - \tilde{z}_i^{(k),*}| &\leq |\tilde{z}_i^{(k)} - m_i(\hat{\boldsymbol{\Theta}}_i^{(k)})| + |m_i(\hat{\boldsymbol{\Theta}}_i^{(k)}) - m_i(\boldsymbol{\Theta}_i^{(k),*})| \\ &\leq |\tilde{z}_i^{(k)} - m_i(\hat{\boldsymbol{\Theta}}_i^{(k)})| + C_i \|\hat{\boldsymbol{\Theta}}_i^{(k)} - \boldsymbol{\Theta}_i^{(k),*}\|_F.\end{aligned}$$

The uniform quantile bound makes the first term $O_p(n_i^{-1/2})$. For the MAD, apply the same split to $D_i(\boldsymbol{\Theta}, m)$, including the center difference just bounded. The floor and threshold addition are Lipschitz. Consequently, a root-n_i fit error implies a root-n_i cutoff error.

The empirical/population decomposition from the consistency proof now yields

$$\begin{aligned}&\|\boldsymbol{G}_i^{(k+1)} - \boldsymbol{G}_i^{(k+1),*}\|_F + \|\boldsymbol{C}_i^{(k+1)} - \boldsymbol{C}_i^{(k+1),*}\|_F \\ &\quad \leq C_i U_{i,n_i} + C_i \left(\|\hat{\boldsymbol{\Theta}}_i^{(k)} - \boldsymbol{\Theta}_i^{(k),*}\|_F + |\delta_i^{(k)} - \delta_i^{(k),*}| \right) = O_p(n_i^{-1/2}).\end{aligned}$$

Constants may depend on the fixed dimensions, client law, and finite iterate, but not on n_i. Substituting these errors into the weighted normal-equation identity gives

$$\begin{aligned}\|\hat{\boldsymbol{\Theta}}_i^{(k+1)} - \boldsymbol{\Theta}_i^{(k+1),*}\|_F \leq \epsilon^{-1} \big(&\|\boldsymbol{C}_i^{(k+1)} - \boldsymbol{C}_i^{(k+1),*}\|_F \\ &+ \|\boldsymbol{G}_i^{(k+1)} - \boldsymbol{G}_i^{(k+1),*}\|_2 \|\boldsymbol{\Theta}_i^{(k+1),*}\|_F \big) = O_p(n_i^{-1/2}).\end{aligned}$$

Finite induction through T proves the iterate rates without assuming a rate for a plug-in weighted moment.

For any $M > 0$, the terminal Gram error obeys

$$\Pr\{\sqrt{n_i}\|\boldsymbol{G}_i - \boldsymbol{G}_i^*\|_F > M\} \leq \Pr(\hat{K}_i \neq K_i^*) + \Pr\{\sqrt{n_i}\|\boldsymbol{G}_i^{(K_i^*)} - \boldsymbol{G}_i^{(K_i^*),*}\|_F > M\}.$$

The stopping-error probability vanishes, while the fixed-iterate rate bounds the second term uniformly for large M. No rate for the stopping-error probability is needed. This argument also covers the terminal cross-moment and, by linearity in fixed dimensions, both lag profiles. Weyl's inequality gives root-n_i eigenvalue errors, and $\hat{\mu}_{i,\mathrm{cut}} - \bar{\mu}_i^* = O_p(n_i^{-1/2})$. The final selected-average bound from Lemma 2, together with its set-equality event, proves the scale rate. □

C.6 Spectral Well-Posedness of the Sample Operators

Lemma 4 (Eventual positive definiteness)**.** *Under Assumptions 1-5, with probability tending to one, every $\boldsymbol{G}_i$ and $\mathcal{T}(\boldsymbol{r}_i)$ has smallest eigenvalue at least $m_w/2$. For any $\xi_i \geq 0$ with $\sum_i \xi_i = 1$,*

$$\sum_i \xi_i \mathcal{T}(\boldsymbol{r}_i^*) \succeq m_w \boldsymbol{I}.$$

The corresponding empirical convex combination is bounded below by $(m_w/2)\boldsymbol{I}$ on the same event.

Population conditioning therefore transfers to the empirical pooled system with probability tending to one, without guaranteeing positive definiteness at every finite sample size.

Proof. The local consistency lemma and linearity give

$$\|\boldsymbol{G}_i - \boldsymbol{G}_i^*\|_2 \xrightarrow{p} 0, \qquad \|\mathcal{T}(\boldsymbol{r}_i) - \mathcal{T}(\boldsymbol{r}_i^*)\|_2 \xrightarrow{p} 0.$$

Weyl's inequality therefore yields

$$\lambda_{\min}(\mathcal{T}(\boldsymbol{r}_i)) \geq m_w - \|\mathcal{T}(\boldsymbol{r}_i) - \mathcal{T}(\boldsymbol{r}_i^*)\|_2,$$

and the same inequality holds for the dense Gram matrix. Each error is below $m_w/2$ with probability tending to one. A union bound over fixed N gives a simultaneous event for all clients. For the population pool, the matrix-order calculation is

$$\sum_i \xi_i \mathcal{T}(\boldsymbol{r}_i^*) \succeq \sum_i \xi_i m_w \boldsymbol{I} = m_w \boldsymbol{I}.$$

Replacing m_w by $m_w/2$ proves the empirical assertion on that event, even if the convex weights are data-dependent. □

For the probability statements, assign the estimator zero whenever its coefficient matrix is singular. The probability of this event tends to zero, leaving consistency and O_p rates unchanged. This convention is confined to the analysis. The implementation remains a direct solve.

C.7 Compression Algebra and Approximation Error

Lemma 5 (Compression commutes with pooling)**.** *Let $\boldsymbol{G}_i$ be symmetric matrices, and extract $\boldsymbol{r}_i$ from their upper-diagonal averages and $\boldsymbol{s}_i$ from the first rows of $\boldsymbol{C}_i$. For any scalar weights α_i,*

$$\mathcal{T}\left(\sum_i \alpha_i \boldsymbol{r}_i\right) = \Pi_{\mathcal{T}}\left(\sum_i \alpha_i \boldsymbol{G}_i\right), \qquad \mathcal{H}\left(\sum_i \alpha_i \boldsymbol{r}_i, \sum_i \alpha_i \boldsymbol{s}_i\right) = \sum_i \alpha_i \mathcal{H}(\boldsymbol{r}_i, \boldsymbol{s}_i).$$

The server can pool compact profiles before reconstruction without changing the resulting structured moments.

Proof. At every lag ℓ, interchange the finite sums to obtain

$$\frac{1}{L-\ell} \sum_{a=0}^{L-\ell-1} \sum_i \alpha_i \boldsymbol{G}_i[a, a+\ell] = \sum_i \alpha_i r_i[\ell].$$

Repeating this average on the corresponding diagonals gives the Toeplitz identity. For the cross-moment, concatenate the pooled profiles into $\boldsymbol{v} = \sum_i \alpha_i \boldsymbol{v}_i$. Then at every (a, h),

$$v[L + h - a] = \sum_i \alpha_i v_i[L + h - a] = \sum_i \alpha_i \mathcal{H}(\boldsymbol{r}_i, \boldsymbol{s}_i)[a, h].$$

This proves the second identity. It uses both profiles and does not assert that $\mathcal{H}$ is an orthogonal projection of $\boldsymbol{C}_i$. □

The lag grouping in Figure B1 underlies the following compression residuals.

Lemma 6 (Consistency and scope of the compression residuals)**.** *Under Assumptions 1-5, define*

$$\boldsymbol{\Delta}_{xx,i} = \boldsymbol{G}_i - \mathcal{T}(\boldsymbol{r}_i), \qquad \boldsymbol{\Delta}_{xy,i} = \boldsymbol{C}_i - \mathcal{H}(\boldsymbol{r}_i, \boldsymbol{s}_i).$$

Then $(\boldsymbol{\Delta}_{xx,i}, \boldsymbol{\Delta}_{xy,i}) \xrightarrow{p} (\boldsymbol{\Delta}^*_{xx,i}, \boldsymbol{\Delta}^*_{xy,i})$*. The corresponding raw population residuals vanish for unweighted contiguous windows from a second-order stationary stream. Residual-dependent weights do not generally preserve this exact structure.*

More local data consistently estimates the compression residual, but need not eliminate the structural approximation error.

Proof. For the weighted residuals,

$$\boldsymbol{\Delta}_{xx,i} - \boldsymbol{\Delta}^*_{xx,i} = (\boldsymbol{G}_i - \boldsymbol{G}^*_i) - \mathcal{T}(\boldsymbol{r}_i - \boldsymbol{r}^*_i),$$
$$\boldsymbol{\Delta}_{xy,i} - \boldsymbol{\Delta}^*_{xy,i} = (\boldsymbol{C}_i - \boldsymbol{C}^*_i) - \mathcal{H}(\boldsymbol{r}_i - \boldsymbol{r}^*_i, \boldsymbol{s}_i - \boldsymbol{s}^*_i).$$

Each right side tends to zero by local consistency and linearity.

To check the raw stationary case, order inputs as $(X_t, \dots, X_{t+L-1})$ and targets as $(X_{t+L}, \dots, X_{t+L+H-1})$. Let $\varrho_i(k) = \mathbb{E}[X_t X_{t+k}]$ be the uncentered lag-k second moment. Stationarity gives

$$\mathcal{S}^*_{xx,i}[a,b] = \varrho_i(|a-b|), \qquad \mathcal{S}^*_{xy,i}[a,h] = \varrho_i(L+h-a).$$

In this paragraph only, let $\boldsymbol{r}, \boldsymbol{s}$ be extracted from these raw moments. Then $r[\ell] = \varrho_i(\ell)$ and $s[h] = \varrho_i(L+h)$. For the concatenated vector $\boldsymbol{v}$,

$$v[k] = \begin{cases} r[k] = \varrho_i(k), & 0 \le k < L, \\ s[k-L] = \varrho_i(k), & L \le k < L+H. \end{cases}$$

Substituting $k = L + h - a$ proves $\mathcal{H}(\boldsymbol{r}, \boldsymbol{s})[a,h] = \mathcal{S}^*_{xy,i}[a,h]$. The Gram identity follows directly from the lag difference. With estimated residual weights, however, $\mathbb{E}[w\ x_a x_b]$ and $\mathbb{E}[w\ x_a y_h]$ can depend on absolute window positions as well as their lag. The stationary raw identities therefore do not establish zero terminal weighted residuals. □

C.8 Convergence of the One-Shot Global Solve

Dividing every empirical precision p_i by the common total n leaves the normalized capped weights unchanged. The limiting precisions therefore use client proportions,

$$p^*_i = \frac{c_i}{\sigma^{2,*}_i}, \qquad \tilde{p}^* = \mathrm{med}_{\mathrm{lo}}\{p^*_j\}_{j=1}^N, \qquad \check{p}^*_i = \min\{p^*_i, \kappa\tilde{p}^*\}, \qquad \alpha^*_i = \frac{\check{p}^*_i}{\sum_j \check{p}^*_j}.$$

Define the pooled population profiles and their compressed target by

$$\boldsymbol{r}^*_g = \sum_i \alpha^*_i \boldsymbol{r}^*_i, \qquad \boldsymbol{s}^*_g = \sum_i \alpha^*_i \boldsymbol{s}^*_i, \qquad \boldsymbol{\Theta}^*_g = \mathcal{T}(\boldsymbol{r}^*_g)^{-1}\mathcal{H}(\boldsymbol{r}^*_g, \boldsymbol{s}^*_g).$$

Theorem 1 (Consistency of the one-shot global solve)**.** *Under Assumptions 1-5, the global coefficient matrix is nonsingular with probability tending to one and*

$$\boldsymbol{r}_g \xrightarrow{p} \boldsymbol{r}^*_g, \qquad \boldsymbol{s}_g \xrightarrow{p} \boldsymbol{s}^*_g, \qquad \boldsymbol{\Theta}_g \xrightarrow{p} \boldsymbol{\Theta}^*_g \qquad \text{as every } n_i \to \infty.$$

Under the additional uniform empirical-process, quantile, and local Lipschitz conditions in Section C.3.1, together with $n_i/n - c_i = O(n^{-1/2})$*, all three differences have order* $O_p(n^{-1/2})$*. The target uses the implemented finite robust recursion, lag compression, and capped precision weights.*

Proof. Set

$$c_{i,n} = \frac{n_i}{n}, \qquad p_{i,n} = \frac{p_i}{n} = \frac{c_{i,n}}{\hat{\sigma}^2_i}, \qquad p^*_i = \frac{c_i}{\sigma^{2,*}_i}.$$

For any positive common multiplier n, the lower median and cap are homogeneous. In particular,

$$\frac{\check{p}_i}{n} = \min\left\{p_{i,n}, \kappa\, \mathrm{med}_{\mathrm{lo}}\{p_{j,n}\}_{j=1}^N\right\}.$$

Dividing numerator and denominator of α_i by n therefore leaves α_i unchanged. Moreover,

$$|p_{i,n} - p_i^*| \le \frac{|c_{i,n} - c_i|}{\hat{\sigma}_i^2} + \frac{c_i|\hat{\sigma}_i^2 - \sigma_i^{2,*}|}{\hat{\sigma}_i^2 \sigma_i^{2,*}} \xrightarrow{p} 0.$$

The positive scale floor and the population scale bound control the denominators. Under the rate conditions, Lemma 3 makes this difference $O_p(n^{-1/2})$, since $n_i/n \to c_i > 0$.

For two vectors $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^N$, their order statistics satisfy

$$|\operatorname{med}_{\text{lo}}(\boldsymbol{p}) - \operatorname{med}_{\text{lo}}(\boldsymbol{q})| \le \|\boldsymbol{p} - \boldsymbol{q}\|_\infty.$$

Coordinatewise minima consequently make the cap map Lipschitz, including at median ties and at the cap threshold. Let $S_n = \sum_j \check{p}_j/n$ and $S^* = \sum_j \check{p}_j^* > 0$. Then $S_n \xrightarrow{p} S^*$, and

$$\alpha_i - \alpha_i^* = \frac{\check{p}_i/n - \check{p}_i^*}{S_n} + \frac{\check{p}_i^*(S^* - S_n)}{S_n S^*}.$$

Both terms tend to zero. Under the rate conditions, each is $O_p(n^{-1/2})$. Unlike the discontinuous spectral selection rule, the continuous cap map requires no assumption excluding median or cap ties.

Subtract the population pool and add back $\sum_i \alpha_i \boldsymbol{r}_i^*$ to obtain

$$\boldsymbol{r}_g - \boldsymbol{r}_g^* = \sum_i \alpha_i(\boldsymbol{r}_i - \boldsymbol{r}_i^*) + \sum_i (\alpha_i - \alpha_i^*)\boldsymbol{r}_i^*.$$

The first sum uses consistent local profiles and weights in $[0, 1]$. In the second, the weight errors vanish while the population profiles are fixed. Both vanish because N is fixed. The identical decomposition holds for $\boldsymbol{s}_g - \boldsymbol{s}_g^*$. Under the stronger conditions, these finite sums have order $O_p(n^{-1/2})$. A union bound combines the client results without cross-client independence.

Define

$$u_i^* = \frac{1}{\sigma_i^{2,*}}, \qquad \tilde{u}^* = \operatorname{med}_{\text{lo}}\{u_j^*\}_{j=1}^N, \qquad \check{u}_i^* = \min\{u_i^*, \kappa\tilde{u}^*\}, \qquad \sigma_g^{2,*} = \frac{N}{\sum_i \check{u}_i^*}.$$

The same continuous-map argument gives $\hat{\sigma}_g^2 \xrightarrow{p} \sigma_g^{2,*} \in (0, \infty)$. Therefore,

$$0 \le \lambda = \min\{\hat{\sigma}_g^2 L/n, \lambda_{\text{cap}}\} \le \hat{\sigma}_g^2 L/n = O_p(n^{-1}).$$

The global ridge consequently disappears from the population target.

Write

$$\boldsymbol{A}_g = \mathcal{T}(\boldsymbol{r}_g) + \lambda \boldsymbol{I}, \quad \boldsymbol{B}_g = \mathcal{H}(\boldsymbol{r}_g, \boldsymbol{s}_g), \quad \boldsymbol{A}_g^* = \mathcal{T}(\boldsymbol{r}_g^*), \quad \boldsymbol{B}_g^* = \mathcal{H}(\boldsymbol{r}_g^*, \boldsymbol{s}_g^*).$$

The reconstructed differences satisfy

$$\boldsymbol{A}_g - \boldsymbol{A}_g^* = \mathcal{T}(\boldsymbol{r}_g - \boldsymbol{r}_g^*) + \lambda \boldsymbol{I}, \qquad \boldsymbol{B}_g - \boldsymbol{B}_g^* = \mathcal{H}(\boldsymbol{r}_g - \boldsymbol{r}_g^*, \boldsymbol{s}_g - \boldsymbol{s}_g^*).$$

On the simultaneous positive-definiteness event in Lemma 4, $\|\boldsymbol{A}_g^{-1}\|_2 \le 2/m_w$. Using $\boldsymbol{A}_g\boldsymbol{\Theta}_g = \boldsymbol{B}_g$ and $\boldsymbol{A}_g^*\boldsymbol{\Theta}_g^* = \boldsymbol{B}_g^*$, compute

$$\begin{aligned}
\boldsymbol{A}_g(\boldsymbol{\Theta}_g - \boldsymbol{\Theta}_g^*) &= \boldsymbol{B}_g - \boldsymbol{A}_g\boldsymbol{\Theta}_g^* \\
&= (\boldsymbol{B}_g - \boldsymbol{B}_g^*) - (\boldsymbol{A}_g - \boldsymbol{A}_g^*)\boldsymbol{\Theta}_g^*, \\
\boldsymbol{\Theta}_g - \boldsymbol{\Theta}_g^* &= \boldsymbol{A}_g^{-1}\left[(\boldsymbol{B}_g - \boldsymbol{B}_g^*) - (\boldsymbol{A}_g - \boldsymbol{A}_g^*)\boldsymbol{\Theta}_g^*\right].
\end{aligned}$$

It follows that

$$\|\boldsymbol{\Theta}_g - \boldsymbol{\Theta}_g^*\|_F \le \frac{2}{m_w}\left(\|\boldsymbol{B}_g - \boldsymbol{B}_g^*\|_F + \|\boldsymbol{A}_g - \boldsymbol{A}_g^*\|_2\|\boldsymbol{\Theta}_g^*\|_F\right).$$

The profile differences and vanishing ridge make the right side $o_p(1)$, proving consistency. Under the additional conditions, its profile terms are $O_p(n^{-1/2})$ and its ridge term is $O_p(n^{-1})$. This proves the stated root-n rate. □

Sampling error therefore vanishes around the specified population version of the implemented procedure. Under the stronger assumptions, its rate is $n^{-1/2}$. Neither statement identifies that target with raw pooled least squares. The derivation above uses the local quantile and empirical-process bounds (Serfling, 1980; van der Vaart & Wellner, 1996). It does not establish a finite-sample concentration bound (Boucheron et al., 2013).

Precision weighting and the dense comparison target. With equal population scales and an inactive precision cap, the limiting weights equal c_i. Equal scales alone do not prevent sample-size imbalance from activating that cap. Even if $\alpha_i^* = c_i$, terminal residual-weighted moments can differ from raw moments. Thus the dense comparison estimator $\boldsymbol{\Theta}^*_{g,\mathrm{dense}}$ need not equal $\boldsymbol{\Theta}^*_{\mathrm{OLS}}$. No minimum-variance or centralized-efficiency conclusion follows from the convergence theorem.

C.9 Proof of the Compression-Error Bound

For the limiting weights of Section C.8, define the dense comparison moments and estimator by

$$\boldsymbol{G}^*_{g,\mathrm{dense}} = \sum_i \alpha_i^* \boldsymbol{G}_i^*, \qquad \boldsymbol{C}^*_{g,\mathrm{dense}} = \sum_i \alpha_i^* \boldsymbol{C}_i^*, \qquad \boldsymbol{\Theta}^*_{g,\mathrm{dense}} = (\boldsymbol{G}^*_{g,\mathrm{dense}})^{-1} \boldsymbol{C}^*_{g,\mathrm{dense}}.$$

The subscript dense distinguishes this estimator from the compressed target. It is a solve from pooled moments, not an arithmetic mean of client estimators. Lemma 6 distinguishes their sampling fluctuations from population approximation error. The following proof holds the limiting weights fixed and isolates only the latter error.

Theorem 2 (Lag-compression perturbation bound)**.** *Let*

$$\boldsymbol{\Delta}^*_{xx,i} = \boldsymbol{G}_i^* - \mathcal{T}(\boldsymbol{r}_i^*), \qquad \boldsymbol{\Delta}^*_{xy,i} = \boldsymbol{C}_i^* - \mathcal{H}(\boldsymbol{r}_i^*, \boldsymbol{s}_i^*),$$

and set $\boldsymbol{\Delta}^*_{xx,g} = \sum_i \alpha_i^* \boldsymbol{\Delta}^*_{xx,i}$ *and* $\boldsymbol{\Delta}^*_{xy,g} = \sum_i \alpha_i^* \boldsymbol{\Delta}^*_{xy,i}$. *Under Assumptions 4-5,*

$$\left\|\boldsymbol{\Theta}^*_g - \boldsymbol{\Theta}^*_{g,\mathrm{dense}}\right\|_F \le \frac{\|\boldsymbol{\Delta}^*_{xx,g}\|_F \|\boldsymbol{\Theta}^*_{g,\mathrm{dense}}\|_F + \|\boldsymbol{\Delta}^*_{xy,g}\|_F}{m_w}.$$

Proof. Joint linearity from Lemma 5 gives

$$\mathcal{T}(\boldsymbol{r}_g^*) = \sum_i \alpha_i^* \mathcal{T}(\boldsymbol{r}_i^*) = \sum_i \alpha_i^* (\boldsymbol{G}_i^* - \boldsymbol{\Delta}^*_{xx,i}) = \boldsymbol{G}^*_{g,\mathrm{dense}} - \boldsymbol{\Delta}^*_{xx,g},$$
$$\mathcal{H}(\boldsymbol{r}_g^*, \boldsymbol{s}_g^*) = \sum_i \alpha_i^* \mathcal{H}(\boldsymbol{r}_i^*, \boldsymbol{s}_i^*) = \boldsymbol{C}^*_{g,\mathrm{dense}} - \boldsymbol{\Delta}^*_{xy,g}.$$

The dense normal equation is $\boldsymbol{G}^*_{g,\mathrm{dense}} \boldsymbol{\Theta}^*_{g,\mathrm{dense}} = \boldsymbol{C}^*_{g,\mathrm{dense}}$. Substituting these moment identities gives

$$\begin{aligned}
\mathcal{T}(\boldsymbol{r}_g^*)(\boldsymbol{\Theta}^*_g - \boldsymbol{\Theta}^*_{g,\mathrm{dense}}) &= \mathcal{H}(\boldsymbol{r}_g^*, \boldsymbol{s}_g^*) - \mathcal{T}(\boldsymbol{r}_g^*)\boldsymbol{\Theta}^*_{g,\mathrm{dense}} \\
&= (\boldsymbol{C}^*_{g,\mathrm{dense}} - \boldsymbol{\Delta}^*_{xy,g}) - (\boldsymbol{G}^*_{g,\mathrm{dense}} - \boldsymbol{\Delta}^*_{xx,g})\boldsymbol{\Theta}^*_{g,\mathrm{dense}} \\
&= \boldsymbol{\Delta}^*_{xx,g}\boldsymbol{\Theta}^*_{g,\mathrm{dense}} - \boldsymbol{\Delta}^*_{xy,g}.
\end{aligned}$$

The population matrix-order argument gives $\mathcal{T}(\boldsymbol{r}_g^*) \succeq m_w \boldsymbol{I}$. Multiplication by its inverse and the inequalities $\|\boldsymbol{AB}\|_F \le \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F$ and $\|\boldsymbol{A}\|_2 \le \|\boldsymbol{A}\|_F$ yield

$$\begin{aligned}
\|\boldsymbol{\Theta}^*_g - \boldsymbol{\Theta}^*_{g,\mathrm{dense}}\|_F &\le \|\mathcal{T}(\boldsymbol{r}_g^*)^{-1}\|_2 \left(\|\boldsymbol{\Delta}^*_{xx,g}\|_2 \|\boldsymbol{\Theta}^*_{g,\mathrm{dense}}\|_F + \|\boldsymbol{\Delta}^*_{xy,g}\|_F\right) \\
&\le \frac{\|\boldsymbol{\Delta}^*_{xx,g}\|_F \|\boldsymbol{\Theta}^*_{g,\mathrm{dense}}\|_F + \|\boldsymbol{\Delta}^*_{xy,g}\|_F}{m_w}.
\end{aligned}$$

No division by a cross-moment norm or target norm is involved. □

This bound isolates the deterministic effect of lag compression while holding robustification and aggregation weights fixed. Small pooled reconstruction residuals imply a small parameter perturbation when the Toeplitz operator remains well conditioned. The absolute bound is also defined when a cross-moment or target is zero.

For comparison, let $\mathcal{S}^*_{xx,i} = \mathbb{E}_i[\boldsymbol{x}\boldsymbol{x}^\top]$ and $\mathcal{S}^*_{xy,i} = \mathbb{E}_i[\boldsymbol{x}\boldsymbol{y}^\top]$. When the raw pooled Gram matrix is nonsingular, define

$$\boldsymbol{\Theta}^*_{\mathrm{OLS}} = \left(\sum_i c_i \mathcal{S}^*_{xx,i}\right)^{-1} \sum_i c_i \mathcal{S}^*_{xy,i}.$$

Corollary 1 (Total error decomposition)**.** *Under the conditions of Theorems 1 and 2, and when* $\boldsymbol{\Theta}^*_{\mathrm{OLS}}$ *is defined,*

$$\|\boldsymbol{\Theta}_g - \boldsymbol{\Theta}^*_{\mathrm{OLS}}\|_F \le \underbrace{\|\boldsymbol{\Theta}_g - \boldsymbol{\Theta}^*_g\|_F}_{\textit{estimation}} + \underbrace{\|\boldsymbol{\Theta}^*_g - \boldsymbol{\Theta}^*_{g,\mathrm{dense}}\|_F}_{\textit{compression}} + \underbrace{\|\boldsymbol{\Theta}^*_{g,\mathrm{dense}} - \boldsymbol{\Theta}^*_{\mathrm{OLS}}\|_F}_{\textit{estimand shift}}.$$

Only the estimation term is guaranteed to vanish as every n_i *grows.*

Proof. Add and subtract the two intermediate targets before taking norms,

$$\boldsymbol{\Theta}_g - \boldsymbol{\Theta}^*_{\mathrm{OLS}} = (\boldsymbol{\Theta}_g - \boldsymbol{\Theta}^*_g) + (\boldsymbol{\Theta}^*_g - \boldsymbol{\Theta}^*_{g,\mathrm{dense}}) + (\boldsymbol{\Theta}^*_{g,\mathrm{dense}} - \boldsymbol{\Theta}^*_{\mathrm{OLS}}),$$
$$\|\boldsymbol{\Theta}_g - \boldsymbol{\Theta}^*_{\mathrm{OLS}}\|_F \le \|\boldsymbol{\Theta}_g - \boldsymbol{\Theta}^*_g\|_F + \|\boldsymbol{\Theta}^*_g - \boldsymbol{\Theta}^*_{g,\mathrm{dense}}\|_F + \|\boldsymbol{\Theta}^*_{g,\mathrm{dense}} - \boldsymbol{\Theta}^*_{\mathrm{OLS}}\|_F.$$

Theorem 1 controls the first term, and Theorem 2 controls the second. The third compares different population objectives and is not a sampling error. It can be zero, but increasing row counts alone need not remove it. □

The decomposition separates sampling error from the two changes made to the population target. Lag compression accounts for the second term, while finite-step robustification and capped precision weighting account for the third. The latter terms are zero only when the corresponding targets coincide.

C.10 Conditional Response Sensitivity and Capped Client Weights

Per-window response contribution.

Proposition 2 (Fixed-fit response clipping)**.** *At fixed* $(\boldsymbol{\Theta}, \delta)$*, the implemented Huber-type weight satisfies* $w_t\|\boldsymbol{e}_t\|_2 \le \delta$*. If* $\|\boldsymbol{x}_t\|_2 \le B_x$*, then*

$$\|w_t \boldsymbol{x}_t \boldsymbol{y}_t^\top\|_F \le B_x^2\|\boldsymbol{\Theta}\|_2 + B_x\delta.$$

Proof of Proposition 2. Fix $(\boldsymbol{\Theta}, \delta)$, set $z = \|\boldsymbol{e}\|_2$, and write $\boldsymbol{y} = \boldsymbol{\Theta}^\top\boldsymbol{x} + \boldsymbol{e}$. The two weight branches give

$$wz = \begin{cases} z \le \delta, & z \le \delta, \\ \delta z/(z + b_0) \le \delta, & z > \delta. \end{cases}$$

Using the rank-one identity $\|\boldsymbol{x}\boldsymbol{y}^\top\|_F = \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2$, expand and bound

$$\begin{aligned} \|w\boldsymbol{x}\boldsymbol{y}^\top\|_F &= \|w\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\Theta} + w\boldsymbol{x}\boldsymbol{e}^\top\|_F \\ &\le w\|\boldsymbol{x}\|_2\|\boldsymbol{\Theta}^\top\boldsymbol{x}\|_2 + w\|\boldsymbol{x}\|_2\|\boldsymbol{e}\|_2 \\ &\le \|\boldsymbol{x}\|_2^2\|\boldsymbol{\Theta}\|_2 + \|\boldsymbol{x}\|_2\delta \\ &\le B_x^2\|\boldsymbol{\Theta}\|_2 + B_x\delta. \end{aligned}$$

Similarly, $\|w\boldsymbol{x}\boldsymbol{x}^\top\|_F \le B_x^2$. Each normalized moment contribution is divided by n_i. The fitted parameters and cutoff are held fixed in this argument. Changing one raw observation can also change the fitted parameters, the cutoff, and overlapping windows. The bound therefore does not establish full-estimator influence or breakdown guarantees (Hampel et al., 1986; Huber, 1981). □

For a fixed fit and cutoff with bounded inputs, an arbitrarily large response residual has a bounded weighted cross-moment contribution. This conditional statement does not establish robustness to unbounded input leverage or to changes in the data-dependent fit and cutoff.

Finite-cohort bound on a client weight.

Proposition 3 (Median-cap weight bound)**.** *For finite positive* $p_1, \dots, p_N$*, lower median* $\tilde{p}$*, and* $\kappa \ge 1$*,*

$$\alpha_i \le \frac{\kappa}{\kappa + \lfloor N/2 \rfloor} \qquad \textit{for every } i.$$

Median capping limits a client's normalized pooling share, although that share can remain large in a small federation.

Proof. The lower median is the order statistic with index $\lceil N/2 \rceil$. At least $\lfloor N/2 \rfloor + 1$ scores are at least $\tilde{p}$. After excluding client i, at least $\lfloor N/2 \rfloor$ other scores still meet this threshold. For each such score, capping leaves at least $\tilde{p}$, since $\kappa \geq 1$. Therefore,

$$\sum_j \check{p}_j \geq \check{p}_i + \lfloor N/2 \rfloor \tilde{p}.$$

The function $x/(x+a)$ is nondecreasing for $x > 0$ and $a \geq 0$. Using $\check{p}_i \leq \kappa \tilde{p}$, obtain

$$\alpha_i \leq \frac{\check{p}_i}{\check{p}_i + \lfloor N/2 \rfloor \tilde{p}} \leq \frac{\kappa \tilde{p}}{\kappa \tilde{p} + \lfloor N/2 \rfloor \tilde{p}} = \frac{\kappa}{\kappa + \lfloor N/2 \rfloor}.$$

□

Capping bounds the final weight without enforcing equal participation. For example, at $N = 2$ and $\kappa = 10$, a client can still receive weight $10/11$. A score that is already strictly above the median cannot move that median by increasing further. A client whose score determines the median can change the threshold. The reported diagnostics record capped precisions for 3 of 321 Electricity clients and 1 of 862 Traffic clients, and none for ETH, SolarEnergy, or TW5K.

C.11 Stability and Limit of the Personalization Coefficient

Lemma 7 (Bounds and consistency of the personalization coefficient)**.** *For the exact-arithmetic empirical moments,*

$$0 < \eta_i \leq 1, \qquad 0 \leq \rho_i \leq 1, \qquad 0 \leq \gamma_i \leq 1.$$

Under Assumptions 1-5, as every client's row count increases,

$$\eta_i \xrightarrow{p} \eta_i^* = \frac{\sigma_i^{2,*}}{r_i^*[0] + a_0},$$

$$\rho_i \xrightarrow{p} \rho_i^* = \max\left\{0, \frac{(\boldsymbol{r}_i^*)^\top \boldsymbol{r}_g^*}{\|\boldsymbol{r}_i^*\|_2 \|\boldsymbol{r}_g^*\|_2 + a_0}\right\},$$

$$\gamma_i \xrightarrow{p} \gamma_i^* = (\eta_i^*)^2 \rho_i^*.$$

Adaptive strength remains bounded and approaches a population value under the stated assumptions. This does not establish an optimal regularization choice.

Proof. The empirical weighted Gram matrix is positive semidefinite, and

$$r_i[0] = \frac{\mathrm{Tr}(\boldsymbol{G}_i)}{L} = \frac{1}{L} \sum_{\ell=1}^{L} \hat{\mu}_{i\ell} \geq 0.$$

The selected spectrum is a nonempty initial segment of the ordered eigenvalues. Its average is no larger than the average of all eigenvalues. Consequently,

$$0 < \hat{\sigma}_i^2 \leq \max\{a_0, r_i[0]\} \leq a_0 + r_i[0], \qquad 0 < \eta_i \leq 1.$$

Cauchy-Schwarz gives

$$0 \leq \rho_i \leq \frac{\|\boldsymbol{r}_i\|_2 \|\boldsymbol{r}_g\|_2}{\|\boldsymbol{r}_i\|_2 \|\boldsymbol{r}_g\|_2 + a_0} \leq 1.$$

The stated bound on $\gamma_i = \eta_i^2 \rho_i$ follows.

The denominator $r_i[0] + a_0$ is bounded below by a_0 on this feasible domain. Its error can be written exactly as

$$\eta_i - \eta_i^* = \frac{\hat{\sigma}_i^2 - \sigma_i^{2,*}}{r_i[0] + a_0} - \frac{\sigma_i^{2,*}(r_i[0] - r_i^*[0])}{(r_i[0] + a_0)(r_i^*[0] + a_0)}.$$

Both numerators vanish by Lemma 2, proving the first limit.

Local and pooled profile convergence give convergence of their inner product and norms. The alignment denominator is at least a_0. Continuity of $x \mapsto \max\{0, x\}$, including at zero, then gives the alignment limit. Thus $\rho_i \xrightarrow{p} \rho_i^*$. Finally,

$$|\gamma_i - \gamma_i^*| \leq |\eta_i^2 - (\eta_i^*)^2| \rho_i + (\eta_i^*)^2 |\rho_i - \rho_i^*| \leq 2|\eta_i - \eta_i^*| + |\rho_i - \rho_i^*| \xrightarrow{p} 0.$$

□

C.12 Convergence of the Personalized Solve

Corollary 2 (Consistency of the personalized forecaster)**.** *Under Assumptions 1-5, as every* $n_j \to \infty$,

$$\boldsymbol{\Theta}_i \xrightarrow{p} \boldsymbol{\Theta}_i^* = (\boldsymbol{A}_i^*)^{-1}\boldsymbol{B}_i^*,$$

where

$$\boldsymbol{A}_i^* = \boldsymbol{G}_i^* + \gamma_i^*\mathcal{T}(\boldsymbol{r}_g^*) + \gamma_i^*\boldsymbol{I}, \qquad \boldsymbol{B}_i^* = \boldsymbol{C}_i^* + \gamma_i^*\mathcal{H}(\boldsymbol{r}_g^*, \boldsymbol{s}_g^*).$$

The personalized forecaster therefore approaches its population target as all clients accumulate data, without a guarantee of improved prediction risk.

Proof. Using joint linearity of the reconstruction maps,

$$\boldsymbol{A}_i - \boldsymbol{A}_i^* = (\boldsymbol{G}_i - \boldsymbol{G}_i^*) + \gamma_i\mathcal{T}(\boldsymbol{r}_g - \boldsymbol{r}_g^*) + (\gamma_i - \gamma_i^*)\big(\mathcal{T}(\boldsymbol{r}_g^*) + \boldsymbol{I}\big),$$
$$\boldsymbol{B}_i - \boldsymbol{B}_i^* = (\boldsymbol{C}_i - \boldsymbol{C}_i^*) + \gamma_i\mathcal{H}(\boldsymbol{r}_g - \boldsymbol{r}_g^*, \boldsymbol{s}_g - \boldsymbol{s}_g^*) + (\gamma_i - \gamma_i^*)\mathcal{H}(\boldsymbol{r}_g^*, \boldsymbol{s}_g^*).$$

Local consistency, global-profile consistency, and Lemma 7 make every term tend to zero.

The dense and Toeplitz conditioning assumptions imply

$$\boldsymbol{A}_i^* = \boldsymbol{G}_i^* + \gamma_i^*\big(\mathcal{T}(\boldsymbol{r}_g^*) + \boldsymbol{I}\big) \succeq \big(m_w + \gamma_i^*(m_w + 1)\big)\boldsymbol{I} \succeq m_w\boldsymbol{I}.$$

Weyl's inequality then makes $\boldsymbol{A}_i$ invertible with probability tending to one and bounds $\|\boldsymbol{A}_i^{-1}\|_2$ by $2/m_w$ on that event.

Since $\boldsymbol{A}_i\boldsymbol{\Theta}_i = \boldsymbol{B}_i$ and $\boldsymbol{A}_i^*\boldsymbol{\Theta}_i^* = \boldsymbol{B}_i^*$,

$$\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^* = \boldsymbol{A}_i^{-1}\left[(\boldsymbol{B}_i - \boldsymbol{B}_i^*) - (\boldsymbol{A}_i - \boldsymbol{A}_i^*)\boldsymbol{\Theta}_i^*\right].$$

Therefore,

$$\|\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i^*\|_F \le \frac{2}{m_w}\left(\|\boldsymbol{B}_i - \boldsymbol{B}_i^*\|_F + \|\boldsymbol{A}_i - \boldsymbol{A}_i^*\|_2\|\boldsymbol{\Theta}_i^*\|_F\right) \xrightarrow{p} 0.$$

□

Corollary 3 (Stability of local adaptation)**.** *Under Assumptions 1-5,*

$$(\eta_i, \rho_i, \gamma_i) \xrightarrow{p} (\eta_i^*, \rho_i^*, \gamma_i^*), \qquad \gamma_i^* \ge 0.$$

The personalized coefficient matrix in Equation 8 is nonsingular with probability tending to one, and $\boldsymbol{\Theta}_i \xrightarrow{p} \boldsymbol{\Theta}_i^*$. *Here* $\boldsymbol{\Theta}_i^*$ *is the same personalized solve evaluated at the population moments and coefficient.*

Proof. Lemma 7 gives the coefficient limit, and Corollary 2 gives the forecaster limit. □

This result establishes statistical stability of both the adaptive coefficient and the resulting local forecaster. It does not guarantee that personalization improves prediction risk. If $\rho_i^* = 0$, then $\gamma_i^* = 0$, so direct substitution gives

$$\boldsymbol{A}_i^* = \boldsymbol{G}_i^*, \qquad \boldsymbol{B}_i^* = \boldsymbol{C}_i^*, \qquad \boldsymbol{\Theta}_i^* = (\boldsymbol{G}_i^*)^{-1}\boldsymbol{C}_i^*.$$

This is the ridge-free terminal-moment solve, not the ϵ-ridged terminal IRLS fit.

Scope of the personalization heuristic. The coefficient controls the global structured-moment contribution to both sides of the personalized normal equations. It does not interpolate directly between local and global parameter matrices. Input-lag alignment does not ensure that clients share the same forecasting target, so no monotone prediction-risk improvement follows. Its empirical effect is evaluated in Section 4.3.

C.13 Horizon Scaling

Cross-moment estimation. The previous results hold at each fixed H. They do not give a joint limit as H and n_i grow. The following Gaussian calculation concerns a population fraction only.

Population fraction of downweighted rows.

Proposition 4 (Gaussian horizon limit with population quantiles)**.** *Suppose a residual vector has independent $\mathcal{N}(0, \sigma^2)$ coordinates, with fixed $\sigma > 0$. Let Z_H be its norm, $m_H = \mathrm{med}(Z_H)$, $D_H = \mathrm{med}\,|Z_H - m_H|$, and*

$$\delta_H = m_H + c_{\mathrm{H}} \max\{D_H/b_{\mathrm{MAD}}, a_0\}, \qquad \pi_H = \Pr(Z_H > \delta_H).$$

Then

$$\pi_H \longrightarrow 1 - \Phi\left(c_{\mathrm{H}} \max\left\{\frac{\Phi^{-1}(0.75)}{b_{\mathrm{MAD}}}, \frac{\sqrt{2}a_0}{\sigma}\right\}\right) \quad \text{as } H \to \infty.$$

If $\sigma\Phi^{-1}(0.75)/(\sqrt{2}\, b_{\mathrm{MAD}}) \geq a_0$, this limit is approximately 0.0893.

In this independent Gaussian example, centering the cutoff at the residual median prevents the fraction of downweighted rows from tending to one as the horizon grows.

Proof. Write $Z_H = \sigma\sqrt{S_H}$, where $S_H = \sum_{h=1}^{H} \xi_h^2$ and the ξ_h are independent standard normals. The central limit theorem and law of large numbers give

$$\frac{S_H - H}{\sqrt{2H}} \Rightarrow \mathcal{N}(0, 1), \qquad \frac{S_H}{H} \xrightarrow{p} 1.$$

Rationalizing the square-root difference yields

$$\frac{Z_H - \sigma\sqrt{H}}{\sigma/\sqrt{2}} = \frac{S_H - H}{\sqrt{2H}} \frac{2\sqrt{H}}{\sqrt{S_H} + \sqrt{H}} \Rightarrow \mathcal{N}(0, 1),$$

because the second factor tends to one.

The limiting normal distribution has a unique median and positive density there. The quantile-bracketing argument therefore gives

$$\frac{m_H - \sigma\sqrt{H}}{\sigma/\sqrt{2}} \to 0.$$

After subtracting this median offset, $(Z_H - m_H)/(\sigma/\sqrt{2}) \Rightarrow \mathcal{N}(0, 1)$. For $Z \sim \mathcal{N}(0, 1)$, the median of $|Z|$ solves

$$\Pr(|Z| \leq d) = 2\Phi(d) - 1 = \tfrac{1}{2} \quad \Longrightarrow \quad d = \Phi^{-1}(0.75).$$

Consequently,

$$D_H \longrightarrow \frac{\sigma}{\sqrt{2}}\Phi^{-1}(0.75).$$

Substitute the two quantile limits into

$$\frac{\delta_H - \sigma\sqrt{H}}{\sigma/\sqrt{2}} = \frac{m_H - \sigma\sqrt{H}}{\sigma/\sqrt{2}} + c_{\mathrm{H}} \max\left\{\frac{\sqrt{2}D_H}{\sigma b_{\mathrm{MAD}}}, \frac{\sqrt{2}a_0}{\sigma}\right\}.$$

The right side tends to the argument of Φ in the proposition. Since the limiting normal CDF is continuous there, the probability of exceeding this moving cutoff tends to the stated upper tail. When the floor is inactive, $\Phi^{-1}(0.75)/b_{\mathrm{MAD}}$ is close to one because $b_{\mathrm{MAD}} = 0.6745$ is a rounded normal quantile. Substitution gives the reported approximate value. □

Proposition 4 uses population medians and MADs. It can describe a sequential limit in which consistent row-level quantile estimation is taken first at each fixed horizon. It does not justify the same constant for an empirical threshold based on a fixed number of rows as $H \to \infty$. Nor does the marginal downweighting fraction determine which inputs are reweighted, the resulting weighted Gram matrix, or prediction risk.

Table C2: Benchmark characteristics and weighted Toeplitz/Hankel reconstruction errors, reported at $H = 720$ except for M4Weekly, which uses its own L=26, H=13 configuration and has no logged reconstruction errors.

Dataset	N	$\bar{n}$	C	$\overline{\mathrm{CV}}$	$\bar{\varepsilon}_{\mathcal{T}}$	$\bar{\varepsilon}_{\mathcal{H}}$	Q_1	$\tilde{y}$	Q_3	$\bar{y}$
ETH	2	91,847	7	0.74	0.00358	0.01077	7.29	10.84	14.42	10.79
Electricity	321	20,228	1	0.41	0.00214	0.00411	1,498.11	2,519.07	3,217.18	2,538.79
SolarEnergy	137	41,233	1	1.46	0.00138	0.00448	0.00	0.00	12.20	6.35
TW5K	200	346,515	5	0.57	0.00251	0.00354	225.89	244.91	263.10	244.58
Traffic	862	13,220	1	0.81	0.00321	0.00607	0.022	0.052	0.073	0.057
M4Weekly	294	956	1	0.49	n/a	n/a	2,322.46	3,311.01	4,777.43	3,742.50

D Implementation Details

All experiments were run in FedProC, a benchmarking framework for federated time series forecasting (Nguyen & Kim, 2026). Every strategy uses its shared dataset loaders, client partitioning, training loop, and wire-size logger. The client-side protocol retains $\boldsymbol{G}_i, \boldsymbol{C}_i$ locally and uses the broadcast profiles for Phase 3. The archived simulator underlying the reported timings instead computes that solve at the server using dense fields returned in memory. The declared wire schema counts only $(\boldsymbol{r}_i, \boldsymbol{s}_i, \hat{\sigma}_i^2, n_i)$ and the corresponding summary downlink. Communication accounting sums the wire size of these fields, billing each tensor at its element size and each scalar at four bytes, while omitting package containers and the undeclared dense moments. Thus the reported communication is the declared wire size of Algorithms 1-3, not measured process-transport traffic. The two placements use the same personalized normal equations, but the archived timings are not measurements of client-side deployment. They include the archived server-side personalization cost.

The terminal weighted diagnostics in Table C2 were recovered from an earlier internal diagnostic run because the reconstruction-error logger was later removed from the reference implementation. That run used the same terminal weighted-moment construction as the current implementation. A multivariate row-count defect present in that earlier run affects the bulk-scale cutoff but not the weighted moments or their reconstruction residuals. Electricity uses one deterministic diagnostic execution, while each other dataset uses three. The separate $\varepsilon_{\mathcal{T}}^{(0)}$ and $\varepsilon_{\mathcal{H}}^{(0)}$ plots use unweighted training windows from the archived configurations. An audit of all five H=720 configurations found identical dataset paths, split ratios, standardization, L, H, batch size, sample ratio, and client count.

In the five-benchmark study, gradient methods share a fixed three-seed schedule, and analytic losses are likewise averaged over three archived executions despite having no stochastic optimizer. Each gradient-method round draws a fixed cohort of $\max(1, \lfloor 0.1N \rfloor)$ clients uniformly without replacement rather than sampling each client independently, giving 32, 13, 20, and 86 participants on Electricity, SolarEnergy, TW5K, and Traffic, so the inclusion indicators within a round are dependent. Both ETDatasetHour clients participate every round, and one-shot analytic methods use all available clients in their sole round. The gradient-based methods (FedAvg, QATFL, FedLUAR) run for 100 rounds with one local epoch per round, while the one-shot analytic methods (DLSA, FedRidge, FedTinyTS) run for a single round. Gradient methods use Adam with learning rate $\nu = 0.0005$ and first- and second-moment decay coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.999$, respectively. One-shot analytic methods have no learning rate. The configured loss determines the gradient-training objective and reported metric. Analytic methods retain their squared-loss normal equations in both loss-configured run families. The evaluated QATFL configuration uses 16 quantization levels and M_{qat}=2, capped by the shared one-local-epoch budget (Fang et al., 2026). The evaluated FedLUAR configuration recycles one layer per round (Kim et al., 2025). DLSA is unregularized, while FedRidge uses ridge 0.1. For FedTinyTS, the local ridge is ϵ=10^{-6} and the Huber coefficient is c_{H}=1.345. IRLS runs for at most five updates with tolerance 10^{-4}. The precision cap is κ=10, and the adaptive global ridge is capped at 0.01.

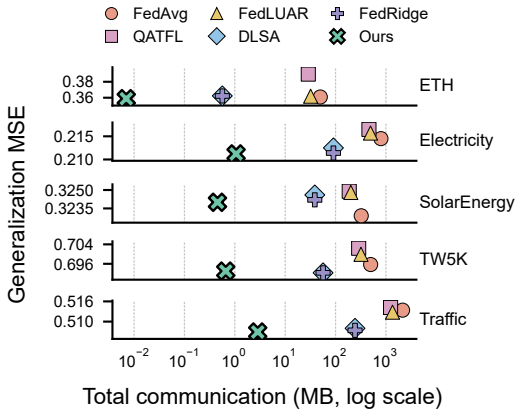


Figure E1: Generalization MSE and message payload.

E Additional Experimental Results

This section provides the supplementary empirical results omitted from the main paper by the page limit. Section E.1 visualizes forecasting performance, communication, personalization, and structural diagnostics. Section E.2 reports the dataset-horizon values underlying the aggregate comparisons in Section 4.

Table C2 lists client count N, mean row count $\bar{n}$, channel count C, the mean coefficient of variation $\overline{\mathrm{CV}}$ (signal standard deviation over mean, averaged across clients), and the across-client means of the relative reconstruction errors $\varepsilon_{\mathcal{T},i} = \|\boldsymbol{G}_i - \mathcal{T}(\boldsymbol{r}_i)\|_F / \|\boldsymbol{G}_i\|_F$ and $\varepsilon_{\mathcal{H},i} = \|\boldsymbol{C}_i - \mathcal{H}(\boldsymbol{r}_i, \boldsymbol{s}_i)\|_F / \|\boldsymbol{C}_i\|_F$ of the terminal weighted moments at $H = 720$, together with the across-client means of the per-client first quartile, median, third quartile, and mean of the raw targets $Q_1, \tilde{y}, Q_3, \bar{y}$. The final row characterizes M4Weekly, used in Table 3. Its 294 clients are the series of the 359 in the collection that admit at least one training window at L=26, H=13, and their raw lengths span 260 to 2,610 points, which is the length heterogeneity the five evaluation benchmarks lack. Its $\bar{n}$ is therefore counted at that configuration rather than at $H = 720$, and its reconstruction-error columns are unavailable because those runs predate the archived logger.

E.1 Result Visualizations

Figures E1, E2, E3, E4, E5, E6, and E7 summarize the aggregate, horizon-wise, personalized, client-level, and structural diagnostics. Figures E1, E2, and E4 average over the four forecast horizons. Figure E3 instead adds individual dataset-horizon settings sequentially. The client-level and structural diagnostics use $L = 96$ and $H = 720$ only. Per-client test MSE is averaged over three seeded executions under the manuscript best-round reporting rule. Figure E1 averages generalization MSE and total message payload over the four forecast horizons for each dataset. Across every dataset, FedTinyTS occupies the smallest-payload position while remaining within the narrow MSE range occupied by the comparison methods. Read off the horizon-averaged totals in Table 1, the payload reduction is $85.5\times$ relative to FedRidge on every dataset and $714\times$-$7{,}536\times$ relative to FedAvg. FedRidge and DLSA are also one-shot baselines, but they transmit dense moments. FedTinyTS matches or improves on FedAvg's horizon-averaged MSE on four datasets and trails it by 0.34% on SolarEnergy. The principal empirical advantage is therefore payload reduction with comparable MSE, rather than a uniform gain in forecasting performance.

Figure E2 adds datasets in the fixed order used by the manuscript tables. Each point averages MSE and MAE ranks over four horizons and all datasets included up to that point. Lower ranks are better, intermediate points are order-dependent, and the final Traffic point equals the full-table average rank. FedTinyTS and FedAvg occupy the top two positions at every prefix, trading first place within a narrow band while QATFL remains last throughout, and DLSA and FedRidge cross once the SolarEnergy prefix is included. At the final point, FedAvg and FedTinyTS have the two lowest aggregate ranks, at 2.16 and 2.60 respectively, matching the Avg Rank row of Table E5.

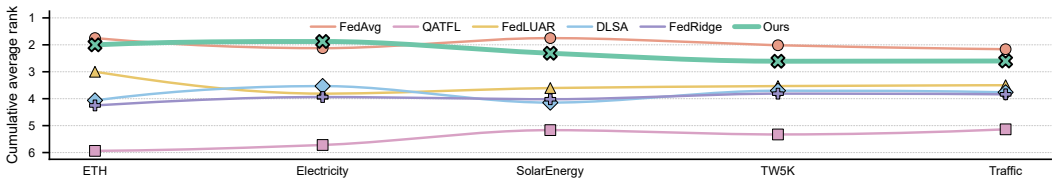


Figure E2: Cumulative generalization rank by dataset.

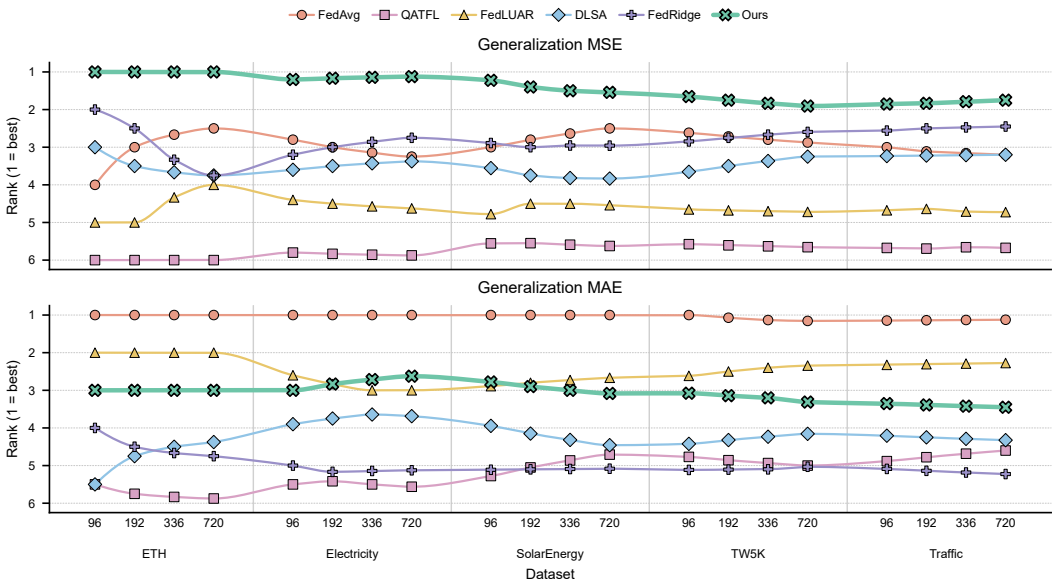


Figure E3: Cumulative generalization rank by dataset and forecast horizon.

Figure E3 orders all 20 dataset-horizon settings, with $H \in \{96, 192, 336, 720\}$ inside each dataset block. Each marker averages the ranks over all settings included up to that point, while the shape-preserving curves only guide the eye. The upper panel reports MSE and the lower panel reports MAE. Ties use the displayed means and share the average occupied rank. At ETH H=96, FedTinyTS ranks first on MSE but third on MAE behind FedAvg and FedLUAR. As settings are added, its cumulative MSE rank approaches roughly 2.2 and its cumulative MAE rank approaches 3. These ranks describe relative method ordering, not the intrinsic difficulty of the datasets. Across all prefixes, FedTinyTS retains the lowest cumulative MSE rank, while FedAvg retains the lowest cumulative MAE rank, consistent with FedAvg posting the best displayed MAE in most individual rows of Table E5.

Figure E4 averages MSE over the four horizons and normalizes each global result to 100%. Each arrow ends at the corresponding personalized result, and its label gives the relative MSE reduction. Personalization lowers MSE on four datasets and slightly raises it on SolarEnergy, where the arrow points right. Using the displayed means in Table 2, the MSE reduction ranges from 1.37% on ETH to 3.74% on Electricity. TW5K and Traffic improve by 1.56% and 1.74%, respectively, whereas SolarEnergy increases by 0.46%. These percentages use the same displayed-table calculation as the main text rather than reproducing the figure's rounded annotations. Comparing FedTinyTS across Tables E5 and E6, personalization raises SolarEnergy MSE at all four horizons. The shortfall against LocalOLS at $H \in \{336, 720\}$ is a separate comparison, not the global-to-personalized change plotted here.

Figure E5 compares LocalOnly, LocalOLS, and personalized FedTinyTS. Each point is one client's test MSE. Violin envelopes contain embedded boxplots, and each dataset panel uses an independent logarithmic vertical scale. Client counts range from $N = 2$ on ETH to $N = 862$ on Traffic, with Electricity ($N = 321$), TW5K ($N = 200$), and SolarEnergy ($N = 137$) in between, so the ETH panel collapses to two points per method while the larger panels resolve a continuous distribution.

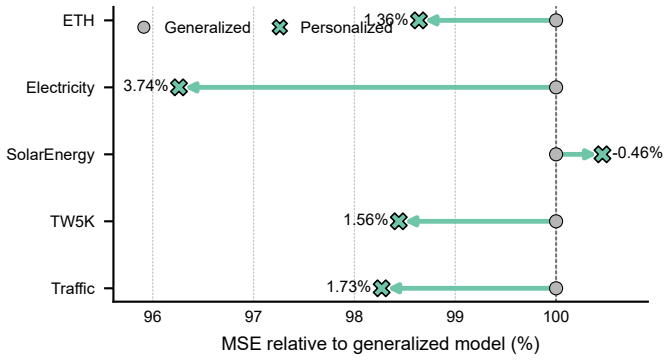


Figure E4: Change in MSE after personalization.

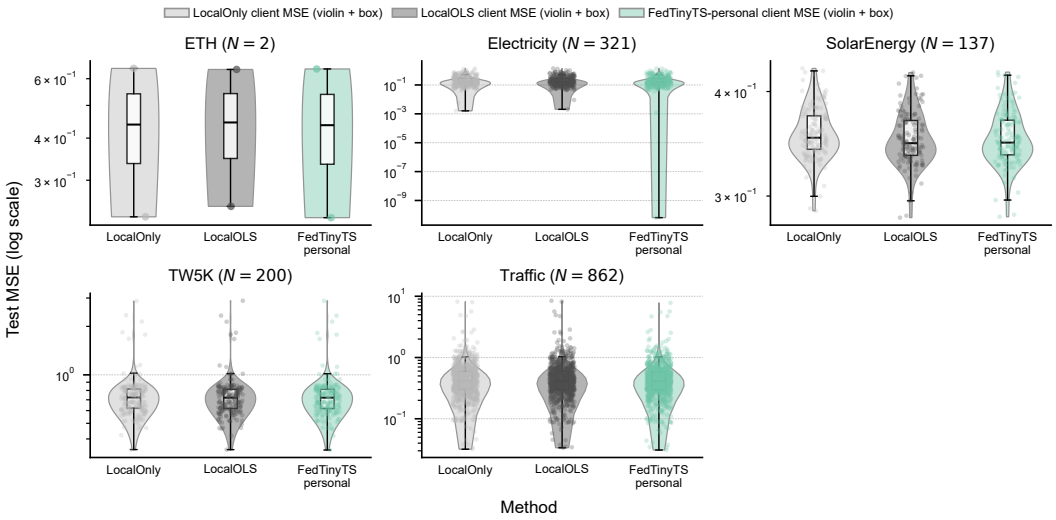


Figure E5: Client-level test-MSE distributions.

The distributions describe variation in absolute client losses, not paired changes from the global to the personalized model. Their overlap alone cannot establish which individual clients benefit from personalization.

Figure E6 reports the unweighted diagnostics $\varepsilon_{\mathcal{T}}^{(0)}$ and $\varepsilon_{\mathcal{H}}^{(0)}$ for every client on a logarithmic scale. These diagnostics precede robust reweighting and therefore differ from the terminal weighted errors in Table C2. The median reverse-Hankel residual exceeds the median Toeplitz residual on all five evaluated benchmarks. Their ratios are $3.48\times$ on ETH, $1.98\times$ on Electricity, $2.30\times$ on SolarEnergy, $5.46\times$ on TW5K, and $2.19\times$ on Traffic. TW5K has the smallest absolute $\varepsilon_{\mathcal{T}}^{(0)}$ median of the five datasets, while Traffic and ETH have the largest, so the ranking of Toeplitz approximation difficulty does not track client count. ETH has only two clients, so each box collapses to its two values.

Figure E7 shows the unweighted standardized training moments, their structured reconstructions, and the residuals for SolarEnergy client 0. The top row compares the input Gram moment $\mathcal{S}_{xx}$ with its Toeplitz reconstruction, while the bottom row compares $\mathcal{S}_{xy}$ with its reverse-Hankel reconstruction. Both comparisons show the lag patterns retained by the structured representation. The residual panels use their own color scale because their magnitudes are much smaller, making their structure visible. The reconstructions preserve the dominant lag patterns, while the nonzero residuals confirm that the compression is approximate rather than lossless. Nonzero residuals identify position-specific relations that the pooled lag representation does not retain. This single-client illustration does not establish the distribution of reconstruction errors across clients, which is reported separately in Figure E6.

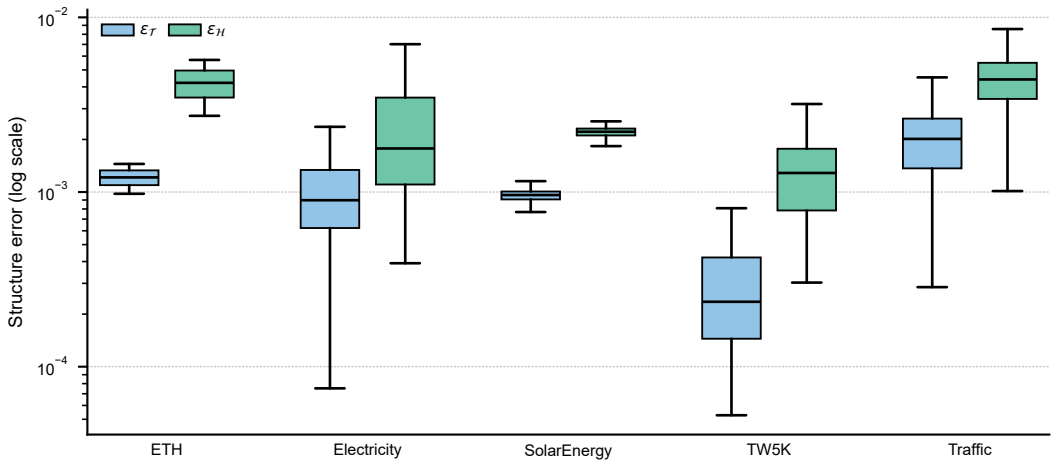


Figure E6: Per-client unweighted structural-error distributions.

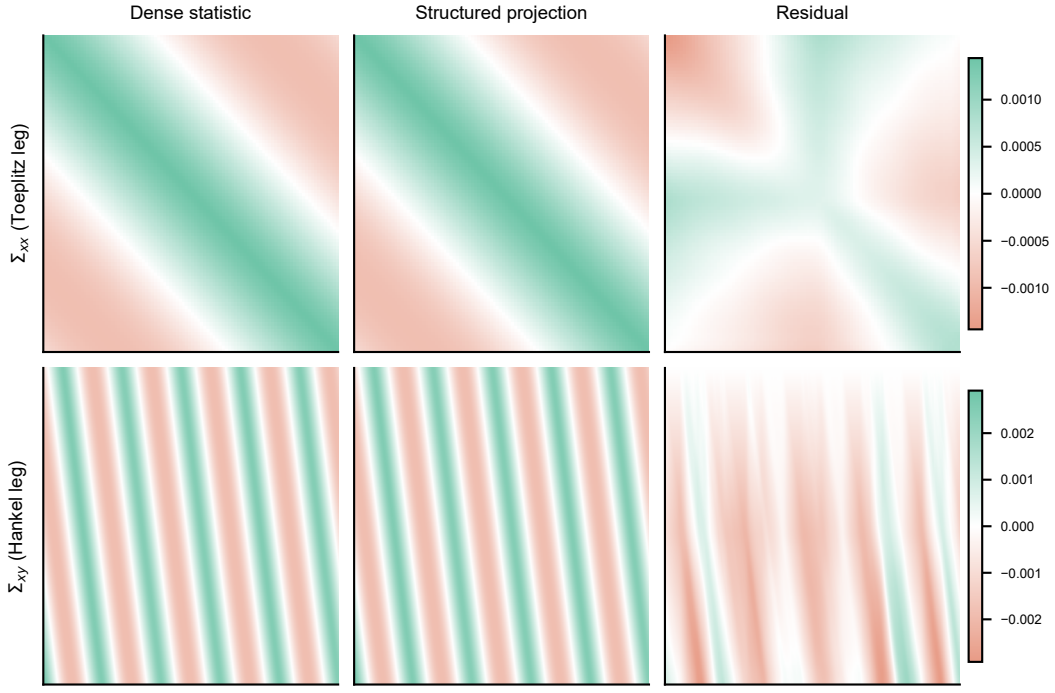


Figure E7: Unweighted moments and structured reconstructions.

E.2 Full Per-Horizon Results

Table E1: Payload-dimension complexity and message-payload ratios to FedTinyTS, aggregated over the full grid and at full participation. q_s is QATFL's quantized upload; $\bar{u}$ and $\bar{\iota}$ are FedLUAR's mean omitted parameters and index overhead.

Method	Up/client	Down/client	Clients	Rounds	Total	Ratio (all H)	Ratio $M{=}N$ $H{=}720$
FedAvg	$P+1$	P	M	R	$RM(2P+1)$	756.7×	8,549×
QATFL	q_s	P	M	R	$RM(q_s+P)$	437.5×	4,942×
FedLUAR	$P-\bar{u}+1$	$P+\bar{\iota}$	M	R	$RM(2P-\bar{u}+\bar{\iota}+1)$	477.1×	4,319×
DLSA	L^2+LH+1	P	N	1	$N(L^2+LH+P+1)$	85.53×	90.69×
FedRidge	L^2+LH	P	N	1	$N(L^2+LH+P)$	85.52×	90.68×
Ours	$L+H+2$	$L+H$	N	1	$N(2L+2H+2)$	1.00×	1.00×

Table E2: Per-horizon message payload.

		FedAvg		QATFL		FedLUAR		DLSA		FedRidge		**Ours**	
Dataset	H	Round	Total	Round	Total	Round	Total	Round	Total	Round	Total	Round	Total
ETH	96	0.1421	14.2097	0.0822	8.2153	0.0878	8.7796	0.2117	0.2117	0.2117	0.2117	**0.0029449**	**0.0029449**
	192	0.2842	28.4187	0.1643	16.4299	0.1751	17.5120	0.3530	0.3530	0.3530	0.3530	**0.0044098**	**0.0044098**
	336	0.4973	49.7322	0.2875	28.7518	0.3138	31.3760	0.5651	0.5651	0.5651	0.5651	**0.0066071**	**0.0066071**
	720	1.0657	106.5681	0.6161	61.6100	0.7036	70.3646	1.1305	1.1305	1.1305	1.1305	**0.0124664**	**0.0124664**
Elec	96	2.2736	227.3560	1.3145	131.4453	1.4270	142.7001	33.9742	33.9742	33.9730	33.9730	**0.4727**	**0.4727**
	192	4.5470	454.6997	2.6288	262.8784	2.8019	280.1923	56.6621	56.6621	56.6609	56.6609	**0.7078**	**0.7078**
	336	7.9572	795.7153	4.6003	460.0281	4.9292	492.9247	90.6939	90.6939	90.6927	90.6927	**1.0604**	**1.0604**
	720	17.0509	1705.0903	9.8576	985.7605	10.6182	1061.8192	181.4454	181.4454	181.4442	181.4442	**2.0009**	**2.0009**
Solar	96	0.9236	92.3634	0.5340	53.3997	0.5707	57.0674	14.4999	14.4999	14.4994	14.4994	**0.2017**	**0.2017**
	192	1.8472	184.7218	1.0679	106.7944	1.1684	116.8432	24.1829	24.1829	24.1824	24.1824	**0.3021**	**0.3021**
	336	3.2326	323.2594	1.8689	186.8864	1.9867	198.6677	38.7074	38.7074	38.7068	38.7068	**0.4526**	**0.4526**
	720	6.9269	692.6929	4.0047	400.4652	4.3136	431.3641	77.4393	77.4393	77.4388	77.4388	**0.8540**	**0.8540**
TW5K	96	1.4210	142.0975	0.8215	82.1533	0.8942	89.4195	21.1678	21.1678	21.1670	21.1670	**0.2945**	**0.2945**
	192	2.8419	284.1873	1.6430	164.2990	1.7651	176.5118	35.3035	35.3035	35.3027	35.3027	**0.4410**	**0.4410**
	336	4.9732	497.3221	2.8752	287.5175	3.2756	327.5604	56.5071	56.5071	56.5063	56.5063	**0.6607**	**0.6607**
	720	10.6568	1065.6815	6.1610	616.1003	6.8451	684.5110	113.0501	113.0501	113.0493	113.0493	**1.2466**	**1.2466**
Traffic	96	6.1102	611.0191	3.5326	353.2593	3.9148	391.4850	91.2330	91.2330	91.2297	91.2297	**1.2693**	**1.2693**
	192	12.2201	1222.0055	7.0649	706.4857	7.7296	772.9630	152.1581	152.1581	152.1548	152.1548	**1.9006**	**1.9006**
	336	21.3848	2138.4850	12.3633	1236.3255	13.1426	1314.2633	243.5456	243.5456	243.5424	243.5424	**2.8476**	**2.8476**
	720	45.8243	4582.4303	26.4923	2649.2313	29.2844	2928.4378	487.2458	487.2458	487.2426	487.2426	**5.3730**	**5.3730**
#1st		0		0		0		0		0		40	
Avg Rank		5.00		3.00		4.00		4.40		3.60		1.00	

Tables E2, E3, E4, E5, and E6 report the complete communication, runtime, global-loss, and personalized-loss results. In every table below, bold and underline mark the best and second-best values per row, respectively. Comparisons use means rounded to the displayed precision. Displayed ties share a mark and receive their average occupied rank. Each displayed mean is its unrounded value rounded once, and the percentages quoted in the main text are computed from the unrounded means rather than from the displayed ones, whereas the absolute differences quoted there are formed from the displayed values so that they can be checked against the tables directly. The #1st row counts per-row wins, while Avg Rank averages ranks across rows. Tables E3 and E4 measure each strategy's native end-to-end simulator path, including orchestration and one held-out evaluation per round. These are not isolated solver-kernel benchmarks. Round time is the across-round mean, while total time covers the complete experiment. Both report mean ± standard deviation (SD) over three archived seeded executions, including the analytic methods, because wall-clock time varies across runs. Tables E5 and E6 report mean ± standard deviation for gradient and iterative methods, taking for each seed the best test loss reached over its 100 archived rounds rather than the value at the last round. For analytic and one-shot methods, they report the three-seed mean without a repeat standard deviation. Across the 200 analytic dataset-horizon-loss cells in these two tables, the seeded repeats agree to within floating-point tolerance in all but one, the largest spread being 2.5×10^{-7} among the 120 global cells and 4.8×10^{-6} among the 80 personalized ones. The exception is the personalized solve on Electricity at $H{=}96$, whose three repeats span 1.6×10^{-2} in MSE and 1.5×10^{-3} in MAE, and that cell carries its repeat standard deviation in Table E6. The pooled quantities entering that

solve agree across repeats to a relative spread of 2.6×10^{-7} and the global loss on the same cell to 2.3×10^{-8}, so the personalized spread reflects the conditioning of that client-level system rather than a stochastic component. The omitted deviations therefore indicate a deterministic solve on a fixed split rather than an absence of variability in the complete execution pipeline, whose wall-clock time does vary.

Table E1 separates message dimensions from the number of clients and rounds contributing to the total. Here $P = H(L+1)$ counts the implemented model parameters, M is the iterative cohort size, and R is the number of rounds. Dense analytic methods avoid repeated exchanges but retain the L^2+LH upload, whereas our profiles reduce the complete per-client exchange to $2L+2H+2$ values. The two ratio columns distinguish aggregation over the evaluated grid from the full-participation comparison at $H = 720$; they should not be read as ratios for an individual dataset-horizon row. These dimensions explain the communication trend, while the declared-size accounting in Section D determines the reported MB.

Table E2 sums the required uplink and downlink fields over participating clients and reports both one-round and full-experiment message payload in MB. Non-federated references are omitted because they communicate nothing. FedTinyTS logs the smallest payload in every one of the 20 dataset-horizon rows, in both the Round and Total columns. The Round-column ratio against FedAvg ranges from approximately $4.58\times$ to $8.55\times$ across Electricity, SolarEnergy, TW5K, and Traffic. On ETH, the only two-client dataset, it ranges from $48.3\times$ to $85.5\times$. The Total-column ratios are approximately 100 times the corresponding Round-column ratios because the iterative baselines repeat communication over $R = 100$ rounds. For Traffic at $H = 720$, the displayed values give $45.8263/5.3730 \approx 8.53$ per round and $4582.6271/5.3730 \approx 853$ over the experiment. Total divided by Round is therefore one for FedTinyTS, DLSA, and FedRidge, and approximately 100 for the iterative baselines.

Across all 20 dataset-horizon rows, FedTinyTS records the smallest one-round and total logged payload. The advantage over iterative methods increases in the total columns because their per-round payload repeats over 100 rounds.

Table E3 reports end-to-end simulator time per round, including held-out test-window evaluation, on the five federated benchmarks. FedTinyTS never wins a per-round comparison and is frequently the slowest method in the row, for example on Traffic at $H{=}720$ where its $140.5\,\mathrm{s}$ round exceeds every other method, including the iterative baselines FedAvg ($37.3\,\mathrm{s}$) and QATFL ($53.1\,\mathrm{s}$). The one-shot round includes robust moment construction, server-side personalized solves, and evaluation, as disclosed in Section D. The simulator computes personalization during its sole aggregation, not over repeated training rounds.

Figure E8 compares total experiment time at $H{=}720$ against client count. The point near 200 clients is TW5K, whose clients each hold five channels rather than one. Channels contribute rows to a shared fit, not separate forecasters. Because dataset lengths and channel counts also vary, this cross-dataset plot does not isolate scaling with client count.

FedTinyTS records no per-round runtime win and has an average rank of 7.85. This result separates communication volume from computation time because a smaller one-shot payload does not imply the shortest simulated round.

Table E4 reports complete experiment time and includes one held-out test-window evaluation per round. Total time reverses the per-round picture because FedTinyTS executes one round while the iterative baselines repeat theirs $R = 100$ times. We calculate speedup as the baseline's displayed mean total time divided by the corresponding mean for FedTinyTS. Across the 20 rows, this gives approximately $10.5\times$-$39.3\times$ relative to FedAvg. For Traffic at $H = 720$, the mean times are $149.283\,\mathrm{s}$ for FedTinyTS, $3741.450\,\mathrm{s}$ for FedAvg, and $3988.280\,\mathrm{s}$ for LocalOnly. The corresponding ratios are $25.1\times$ and $26.7\times$, despite FedTinyTS being slower per round on that row. These ratios compare mean simulator times. They do not measure network acceleration or average per-seed speedups. FedTinyTS remains slower in total time than LocalOLS, DLSA, and FedRidge on every row.

Complete experiment time reverses the per-round comparison because the analytic methods execute once. DLSA and FedRidge have the two lowest average ranks, while FedTinyTS ranks ahead of every 100-round method.

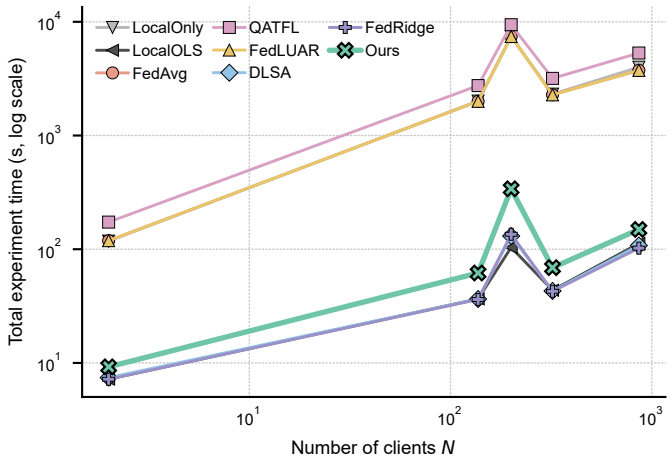


Figure E8: Total experiment time at $H{=}720$ versus client count on logarithmic axes.

Table E3: Per-horizon simulator round time.

Dataset	H	LocalOnly	LocalOLS	FedAvg	QATFL	FedLUAR	DLSA	FedRidge	**Ours**
ETH	96	0.693 ± 0.003	0.357 ± 0.029	0.690 ± 0.007	1.085 ± 0.002	0.692 ± 0.005	0.344 ± 0.003	**0.341 ± 0.011**	0.807 ± 0.007
	192	0.793 ± 0.004	**0.415 ± 0.022**	0.787 ± 0.002	1.260 ± 0.001	0.788 ± 0.001	0.420 ± 0.010	0.424 ± 0.032	1.124 ± 0.003
	336	0.881 ± 0.002	0.521 ± 0.031	0.883 ± 0.002	1.431 ± 0.001	0.878 ± 0.001	**0.512 ± 0.004**	0.522 ± 0.024	1.557 ± 0.005
	720	1.113 ± 0.003	**0.785 ± 0.007**	1.124 ± 0.001	1.667 ± 0.004	1.124 ± 0.002	0.809 ± 0.006	0.790 ± 0.005	2.787 ± 0.033
Elec	96	14.752 ± 0.148	22.424 ± 7.763	13.435 ± 0.033	21.643 ± 0.047	**13.391 ± 0.028**	17.281 ± 0.161	23.436 ± 8.910	26.346 ± 0.026
	192	15.431 ± 0.054	26.470 ± 10.458	13.959 ± 0.004	22.385 ± 0.022	**13.950 ± 0.012**	19.305 ± 0.033	26.454 ± 10.545	30.344 ± 0.025
	336	23.218 ± 0.110	38.832 ± 16.190	**21.804 ± 0.017**	30.665 ± 0.071	21.841 ± 0.049	27.586 ± 0.075	38.839 ± 16.379	46.760 ± 0.027
	720	23.050 ± 0.050	36.383 ± 0.035	22.750 ± 0.016	31.744 ± 0.012	**22.704 ± 0.011**	36.585 ± 0.091	36.271 ± 0.089	61.614 ± 0.120
Solar	96	11.666 ± 0.033	14.176 ± 0.991	**11.068 ± 0.015**	17.796 ± 0.007	11.081 ± 0.018	14.160 ± 0.776	14.421 ± 1.326	22.613 ± 0.021
	192	12.380 ± 0.017	17.816 ± 3.878	**11.721 ± 0.008**	18.446 ± 0.010	11.738 ± 0.015	17.557 ± 3.111	17.837 ± 3.697	27.016 ± 1.189
	336	19.279 ± 0.018	28.626 ± 8.585	**18.670 ± 0.024**	25.802 ± 0.051	18.679 ± 0.015	27.655 ± 6.935	28.557 ± 8.463	42.778 ± 2.252
	720	20.022 ± 0.010	29.649 ± 0.018	19.917 ± 0.015	27.408 ± 0.017	**19.915 ± 0.029**	29.833 ± 0.039	29.590 ± 0.028	54.863 ± 0.110
TW5K	96	34.153 ± 0.024	41.578 ± 0.174	33.713 ± 0.018	51.213 ± 0.109	**33.661 ± 0.048**	41.857 ± 0.131	41.677 ± 0.296	96.032 ± 0.043
	192	36.626 ± 0.094	49.591 ± 0.065	**36.009 ± 0.036**	53.678 ± 0.018	36.040 ± 0.033	50.322 ± 0.118	50.060 ± 0.370	123.026 ± 0.102
	336	70.138 ± 0.136	84.998 ± 0.566	**69.622 ± 0.042**	89.028 ± 0.093	69.748 ± 0.056	85.851 ± 0.651	85.252 ± 0.130	210.348 ± 0.401
	720	74.310 ± 0.072	128.465 ± 10.444	74.264 ± 0.028	94.223 ± 0.035	**74.142 ± 0.116**	122.793 ± 1.311	124.697 ± 5.159	316.583 ± 2.666
Traffic	96	28.585 ± 0.167	37.026 ± 7.771	**23.463 ± 0.037**	38.136 ± 0.014	23.553 ± 0.056	31.909 ± 0.038	38.194 ± 9.212	45.799 ± 0.037
	192	29.514 ± 0.051	44.841 ± 14.901	**24.284 ± 0.038**	38.992 ± 0.047	24.300 ± 0.011	34.880 ± 0.106	45.670 ± 15.806	51.522 ± 0.162
	336	42.772 ± 0.014	81.448 ± 47.148	**37.329 ± 0.125**	52.971 ± 0.018	37.360 ± 0.097	48.726 ± 0.017	48.213 ± 0.059	77.695 ± 0.130
	720	39.783 ± 0.122	103.770 ± 61.486	37.336 ± 0.016	53.071 ± 0.035	**37.328 ± 0.038**	101.340 ± 57.422	95.704 ± 50.404	140.503 ± 60.454
#1st		0	2	9	0	7	1	1	0
Avg Rank		3.45	5.00	2.10	5.85	2.25	4.45	5.05	7.85

Table E4: Per-horizon simulator experiment time.

Dataset	H	LocalOnly	LocalOLS	FedAvg	QATFL	FedLUAR	DLSA	FedRidge	**Ours**
ETH	96	75.763 ± 0.407	**6.613 ± 0.233**	75.187 ± 0.618	114.733 ± 0.296	75.457 ± 0.462	6.647 ± 0.283	6.637 ± 0.039	7.140 ± 0.290
	192	85.753 ± 0.415	**6.573 ± 0.045**	85.160 ± 0.243	132.250 ± 0.185	85.180 ± 0.241	6.727 ± 0.071	6.807 ± 0.119	7.590 ± 0.110
	336	94.623 ± 0.247	**6.723 ± 0.205**	94.617 ± 0.304	149.513 ± 0.305	94.190 ± 0.078	6.760 ± 0.120	6.950 ± 0.205	7.890 ± 0.042
	720	117.813 ± 0.399	7.203 ± 0.200	118.900 ± 0.079	173.190 ± 0.260	118.897 ± 0.224	7.427 ± 0.250	**7.187 ± 0.172**	9.240 ± 0.106
Elec	96	1484.707 ± 14.761	32.423 ± 7.875	1350.217 ± 3.251	2170.997 ± 4.743	1345.720 ± 2.835	**23.427 ± 0.230**	29.657 ± 8.997	36.363 ± 0.167
	192	1552.650 ± 5.401	36.727 ± 10.284	1402.487 ± 0.474	2245.203 ± 2.178	1401.710 ± 1.147	**25.500 ± 0.094**	32.663 ± 10.668	40.613 ± 0.054
	336	2331.227 ± 11.059	49.240 ± 16.399	2187.153 ± 1.841	3073.210 ± 7.097	2190.870 ± 4.802	**33.863 ± 0.217**	45.100 ± 16.469	57.013 ± 0.170
	720	2312.450 ± 4.985	43.557 ± 0.013	2281.980 ± 1.593	3181.333 ± 1.254	2277.283 ± 0.996	42.963 ± 0.198	**42.663 ± 0.083**	68.833 ± 0.181
Solar	96	1174.320 ± 3.398	21.980 ± 1.089	1113.210 ± 1.382	1785.993 ± 0.741	1114.587 ± 1.825	**20.367 ± 0.893**	20.637 ± 1.353	30.497 ± 0.056
	192	1245.810 ± 1.621	25.863 ± 3.944	1178.653 ± 0.913	1851.100 ± 0.889	1180.240 ± 1.569	**23.797 ± 3.171**	24.093 ± 3.823	35.017 ± 1.177
	336	1935.900 ± 1.694	36.567 ± 8.700	1873.653 ± 2.377	2586.793 ± 5.085	1874.447 ± 1.461	**33.937 ± 7.059**	34.923 ± 8.660	50.863 ± 2.282
	720	2009.233 ± 0.947	36.553 ± 0.172	1998.537 ± 1.549	2747.667 ± 1.586	1998.327 ± 2.937	36.467 ± 0.116	**36.160 ± 0.093**	61.780 ± 0.037
TW5K	96	3423.753 ± 2.316	50.647 ± 0.306	3378.123 ± 1.703	5128.233 ± 10.857	3373.050 ± 4.754	48.340 ± 0.196	**48.180 ± 0.358**	105.057 ± 0.128
	192	3671.393 ± 9.483	58.777 ± 0.151	3607.990 ± 3.578	5374.890 ± 1.891	3611.027 ± 3.279	57.047 ± 0.037	**56.777 ± 0.565**	132.230 ± 0.116
	336	7022.703 ± 13.719	94.400 ± 0.577	6969.940 ± 4.411	8910.197 ± 9.325	6982.067 ± 5.526	93.993 ± 1.332	**92.153 ± 0.262**	220.067 ± 0.363
	720	7439.463 ± 7.241	139.353 ± 9.977	7434.623 ± 2.750	9430.420 ± 3.525	7422.407 ± 11.671	**130.567 ± 1.390**	135.367 ± 5.106	339.823 ± 15.144
Traffic	96	2872.673 ± 16.752	53.903 ± 8.062	2353.587 ± 3.747	3821.370 ± 1.796	2362.587 ± 5.553	**38.160 ± 0.131**	44.500 ± 9.313	62.307 ± 0.176
	192	2965.600 ± 5.164	61.947 ± 15.050	2435.753 ± 3.815	3906.713 ± 4.623	2437.417 ± 1.100	**41.517 ± 0.655**	51.997 ± 15.948	68.473 ± 0.439
	336	4291.610 ± 1.426	98.733 ± 46.985	3740.347 ± 12.619	5304.483 ± 1.724	3743.360 ± 9.688	55.017 ± 0.074	**54.830 ± 0.482**	95.150 ± 0.766
	720	3988.280 ± 12.239	112.293 ± 61.523	3741.450 ± 1.530	5315.027 ± 3.354	3740.607 ± 3.767	107.810 ± 57.509	**102.197 ± 50.540**	149.283 ± 60.526
#1st		0	3	0	0	0	9	8	0
Avg Rank		6.90	2.70	5.50	8.00	5.60	1.65	1.70	3.95

Table E5 reports per-horizon generalization MSE and MAE on all five benchmarks, where lower values are better. The largest single-row MSE gap in the table is on ETH at $H{=}720$, where QATFL's 0.5345 is roughly 20% higher than FedTinyTS's best-in-row 0.4458; QATFL's MAE on the same row is 0.5541 against FedTinyTS's 0.4710, its worst relative showing on either metric. FedAvg's

Table E5: Per-horizon generalization MSE and MAE.

Dataset	H	FedAvg		QATFL		FedLUAR		DLSA		FedRidge		**Ours**	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETH	96	0.2963 ± 0.0000	**0.3538 ± 0.0000**	0.3014 ± 0.0004	0.3589 ± 0.0002	0.2970 ± 0.0004	0.3550 ± 0.0004	0.2957	0.3589	0.2951	0.3585	**0.2949**	0.3576
	192	0.3336 ± 0.0001	**0.3845 ± 0.0000**	0.3422 ± 0.0007	0.3927 ± 0.0002	0.3345 ± 0.0004	0.3856 ± 0.0002	0.3339	0.3914	0.3338	0.3915	**0.3321**	0.3876
	336	0.3646 ± 0.0002	**0.4073 ± 0.0000**	0.3804 ± 0.0003	0.4224 ± 0.0005	0.3658 ± 0.0006	0.4098 ± 0.0019	0.3663	0.4182	0.3666	0.4186	**0.3624**	0.4114
	720	0.4491 ± 0.0003	**0.4660 ± 0.0001**	0.5345 ± 0.0014	0.5541 ± 0.0035	0.4504 ± 0.0007	0.4675 ± 0.0002	0.4533	0.4801	0.4535	0.4803	**0.4458**	0.4710
Elec	96	0.1996 ± 0.0003	**0.2752 ± 0.0001**	0.2007 ± 0.0002	0.2777 ± 0.0001	0.2008 ± 0.0003	0.2781 ± 0.0019	0.1989	0.2760	**0.1969**	0.2790	0.1970	0.2766
	192	0.1991 ± 0.0001	**0.2785 ± 0.0001**	0.2006 ± 0.0002	0.2819 ± 0.0001	0.2003 ± 0.0011	0.2809 ± 0.0009	0.1977	0.2797	0.1968	0.2821	**0.1967**	0.2794
	336	0.2128 ± 0.0003	**0.2942 ± 0.0001**	0.2149 ± 0.0001	0.2983 ± 0.0003	0.2140 ± 0.0004	0.2963 ± 0.0011	0.2103	0.2953	0.2092	0.2975	**0.2091**	0.2943
	720	0.2464 ± 0.0003	**0.3246 ± 0.0002**	0.2496 ± 0.0002	0.3297 ± 0.0006	0.2477 ± 0.0009	0.3253 ± 0.0003	0.2433	0.3262	0.2426	0.3289	**0.2422**	0.3247
Solar	96	**0.2827 ± 0.0003**	**0.2941 ± 0.0012**	0.2845 ± 0.0004	0.3223 ± 0.0002	0.2868 ± 0.0014	0.2957 ± 0.0009	0.2851	0.3720	0.2847	0.3716	0.2844	0.3648
	192	**0.3122 ± 0.0002**	**0.3111 ± 0.0009**	0.3137 ± 0.0001	0.3413 ± 0.0005	0.3125 ± 0.0011	0.3134 ± 0.0008	0.3137	0.3917	0.3134	0.3914	0.3131	0.3886
	336	**0.3458 ± 0.0004**	**0.3269 ± 0.0006**	0.3479 ± 0.0004	0.3589 ± 0.0006	0.3477 ± 0.0004	0.3282 ± 0.0011	0.3477	0.4105	0.3473	0.4102	0.3473	0.4087
	720	**0.3508 ± 0.0003**	**0.3214 ± 0.0006**	0.3534 ± 0.0003	0.3565 ± 0.0006	0.3524 ± 0.0005	0.3237 ± 0.0006	0.3517	0.4087	0.3513	0.4085	0.3512	0.4073
TW5K	96	0.6191 ± 0.0009	**0.5460 ± 0.0008**	0.6233 ± 0.0012	0.5496 ± 0.0008	0.6261 ± 0.0022	0.5466 ± 0.0004	**0.6156**	0.5489	**0.6156**	0.5496	0.6169	0.5488
	192	0.6808 ± 0.0014	0.5839 ± 0.0011	0.6856 ± 0.0013	0.5876 ± 0.0010	0.6855 ± 0.0041	**0.5836 ± 0.0002**	**0.6771**	0.5870	**0.6771**	0.5875	0.6778	0.5873
	336	0.7196 ± 0.0015	0.6080 ± 0.0011	0.7264 ± 0.0014	0.6124 ± 0.0009	0.7227 ± 0.0018	**0.6073 ± 0.0005**	**0.7161**	0.6110	**0.7161**	0.6115	0.7166	0.6113
	720	0.7632 ± 0.0012	**0.6353 ± 0.0011**	0.7742 ± 0.0009	0.6417 ± 0.0010	0.7654 ± 0.0011	**0.6353 ± 0.0005**	**0.7603**	0.6376	**0.7603**	0.6380	0.7609	0.6382
Traffic	96	0.5325 ± 0.0005	**0.3597 ± 0.0001**	0.5330 ± 0.0004	0.3672 ± 0.0006	0.5289 ± 0.0007	0.3599 ± 0.0005	0.5242	0.3828	0.5238	0.3840	**0.5232**	0.3807
	192	0.4932 ± 0.0002	**0.3368 ± 0.0001**	0.4937 ± 0.0001	0.3435 ± 0.0005	0.4929 ± 0.0015	0.3373 ± 0.0005	0.4888	0.3583	**0.4880**	0.3600	**0.4880**	0.3564
	336	0.4992 ± 0.0003	**0.3389 ± 0.0001**	0.4999 ± 0.0002	0.3458 ± 0.0005	0.5000 ± 0.0008	0.3391 ± 0.0003	0.4950	0.3605	0.4943	0.3621	**0.4940**	0.3584
	720	0.5289 ± 0.0004	**0.3575 ± 0.0000**	0.5302 ± 0.0004	0.3654 ± 0.0004	0.5292 ± 0.0005	0.3583 ± 0.0004	0.5241	0.3804	0.5235	0.3814	**0.5223**	0.3776
#1st		22		0		3		4		6		11	
Avg Rank		2.16		5.14		3.50		3.76		3.84		2.60	

Table E6: Per-horizon personalized MSE and MAE.

Dataset	H	LocalOnly		LocalOLS		**Ours**	
		MSE	MAE	MSE	MAE	MSE	MAE
ETH	96	0.2929 ± 0.0002	**0.3507 ± 0.0001**	0.2933	0.3561	**0.2912**	0.3529
	192	0.3304 ± 0.0002	**0.3820 ± 0.0001**	0.3311	0.3886	**0.3291**	0.3844
	336	0.3609 ± 0.0001	**0.4053 ± 0.0001**	0.3625	0.4154	**0.3586**	0.4087
	720	0.4387 ± 0.0013	**0.4613 ± 0.0002**	0.4451	0.4755	**0.4368**	0.4661
Elec	96	0.1869 ± 0.0003	**0.2639 ± 0.0001**	**0.1845**	0.2707	0.1997	0.2722
	192	0.1882 ± 0.0000	**0.2677 ± 0.0001**	0.1857	0.2741	**0.1852**	0.2728
	336	0.2016 ± 0.0003	**0.2827 ± 0.0001**	0.1987	0.2896	**0.1981**	0.2880
	720	0.2339 ± 0.0003	**0.3123 ± 0.0000**	0.2308	0.3206	**0.2302**	0.3186
Solar	96	0.2912 ± 0.0002	**0.3062 ± 0.0001**	0.2862	0.3743	**0.2858**	0.3643
	192	0.3199 ± 0.0001	**0.3228 ± 0.0001**	0.3147	0.3940	**0.3146**	0.3910
	336	0.3534 ± 0.0001	**0.3371 ± 0.0001**	**0.3484**	0.4127	0.3487	0.4107
	720	0.3577 ± 0.0002	**0.3306 ± 0.0001**	**0.3523**	0.4110	0.3527	0.4093
TW5K	96	0.6059 ± 0.0000	**0.5331 ± 0.0000**	0.6035	0.5400	**0.6031**	0.5381
	192	0.6702 ± 0.0005	**0.5736 ± 0.0000**	0.6674	0.5802	**0.6670**	0.5789
	336	0.7103 ± 0.0004	**0.5988 ± 0.0000**	0.7076	0.6050	**0.7071**	0.6038
	720	0.7548 ± 0.0001	**0.6271 ± 0.0001**	0.7521	0.6320	**0.7517**	0.6312
Traffic	96	0.5214 ± 0.0002	**0.3546 ± 0.0000**	0.5169	0.3896	**0.5132**	0.3857
	192	0.4864 ± 0.0001	**0.3339 ± 0.0000**	0.4821	0.3630	**0.4789**	0.3604
	336	0.4936 ± 0.0001	**0.3365 ± 0.0001**	0.4893	0.3647	**0.4858**	0.3623
	720	0.5239 ± 0.0002	**0.3552 ± 0.0001**	0.5203	0.3838	**0.5145**	0.3810
#1st		20		3		17	
Avg Rank		1.88		2.50		1.62	

MAE is the row-best or row-second entry in all 20 dataset-horizon rows, which is the per-row evidence underlying its lowest cumulative MAE rank in Figure E3.

FedTinyTS has lower mean MSE than best-round FedAvg in 16 of the 20 dataset-horizon rows. FedAvg leads MAE across most rows, so the forecasting comparison remains metric dependent.

Table E6 reports per-horizon personalized MSE and MAE for LocalOnly, LocalOLS, and FedTinyTS, where lower values are better. LocalOnly follows the iterative reporting convention, and LocalOLS and FedTinyTS follow the analytic reporting convention. FedTinyTS wins the row on MSE in 17 of the 20 dataset-horizon combinations; the largest of the three exceptions is Electricity at H=96, where LocalOLS's 0.1845 beats FedTinyTS's 0.1997, a case where the purely local closed-form solve outperforms the personalized global-plus-local blend on a dataset-horizon pair with comparatively little to gain from borrowing global structure. FedTinyTS records the lowest personalized MSE in 17 of the 20 dataset-horizon rows. LocalOnly records the lowest personalized MAE in all 20 rows, which exposes the same metric-specific limitation as the global comparison.

Table E7 pairs generalization and personalization losses for each client in a single diagnostic run, separate from the three-repeat results above. The upper panel computes reduction from dataset-average MSE; the lower panel averages individual percentage reductions, each relative to that client's generalization MSE. Clients are classified using an absolute MSE tolerance of 10^{-8}; none are un-

Table E7: Client-level personalization outcomes.

		Mean MSE		Reduction	Clients, n (%)	
Dataset	N	Generalization	Personalization	(%)	Improved	Worsened
ETH	2	0.35883	0.35394	1.36	2 (100.0)	0 (0.0)
Electricity	321	0.21125	0.20097	4.86	283 (88.2)	38 (11.8)
SolarEnergy	137	0.32397	0.32545	-0.46	58 (42.3)	79 (57.7)
TW5K	200	0.69302	0.68224	1.56	189 (94.5)	11 (5.5)
Traffic	862	0.50689	0.49810	1.73	714 (82.8)	148 (17.2)

	MSE reduction (%), mean $\pm$ SD across clients		
Dataset	All	Improved	Worsened
ETH	1.41 ± 0.14	1.41 ± 0.14	—
Electricity	3.45 ± 18.98	5.92 ± 9.57	-14.92 ± 44.99
SolarEnergy	-0.48 ± 1.03	0.45 ± 0.32	-1.17 ± 0.80
TW5K	1.75 ± 1.55	1.94 ± 1.21	-1.54 ± 2.75
Traffic	2.85 ± 8.47	4.72 ± 5.78	-6.17 ± 12.64

changed. Personalization benefits most clients on four benchmarks, whereas most SolarEnergy clients worsen. The dispersion among harmed Electricity clients also shows that an average gain does not guarantee a benefit for every client. These results describe client-level outcomes for one seed and do not measure variability across repeated runs.