Optimal Look-back Horizon for Time Series Forecasting in Federated Learning

THE UNIVERSITY OF SYDNEY

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Dahao Tang[1], Nan Yang[13], Yanli Li[1], Zhiyu Zhu[2], Zhibo Jin[2], Dong Yuan[13]

[1]University of Sydney

[2]University of Technology Sydney

[3]Corresponding authors: n.yang@sydney.edu.au, dong.yuan@sydney.edu.au

AAAI-26 / IAAI-26 / EAAI-26

JANUARY 20–27, 2026 | SINGAPORE

Introduction

Selecting the optimal look-back horizon in Federated Learning (FL) is challenging due to non-IID data, where client-specific dynamics render fixed global horizons suboptimal. While recent work has explored horizon selection by preserving forecasting-relevant information in an intrinsic space, they fail to account for decentralized heterogeneity [1]. To address this, we propose a principled framework leveraging a Synthetic Data Generator (SDG) to capture temporal structures (AR, seasonality, trend) and client heterogeneity. By mapping data to an intrinsic space with well-defined geometric and statistical properties, we decompose the forecasting loss into Bayesian (irreducible) and approximation terms. We prove that a fundamental trade-off exists under the non-IID FL settings: the total loss is minimized at the smallest horizon where the irreducible loss saturates.

Key Contributions

We propose a principled theoretical framework for horizon selection in non-IID federated learning:

- **Constructive Intrinsic Space via SDG:** We introduce a Synthetic Data Generator (SDG) that explicitly models temporal dynamics (autoregression, seasonality, trend) and client heterogeneity. This allows us to construct a geometry-preserving intrinsic space and derive a computable Intrinsic Dimension $d_{I,k}(H)$ based on signal saturation.
- **Federated Loss Decomposition:** We establish a tight decomposition of predictive loss into irreducible (Bayesian) and approximation components, each analytically tied to the structural elements of time series data (e.g., AR memory, seasonality, trend) and the look-back horizon. Our analysis uncovers the fundamental bias–variance trade-off that governs forecasting performance in federated settings.
- **Provably Optimal Horizons (H_k^* and H_{server}^*):** We prove that the total loss is unimodal with respect to the horizon length and identify the smallest sufficient horizon as its global minimizer. This result provides the first rigorous criterion for horizon selection in time series forecasting and introduces a new design principle for model construction under sample-limited, heterogeneous environments.

Synthetic Data Generator (SDG)

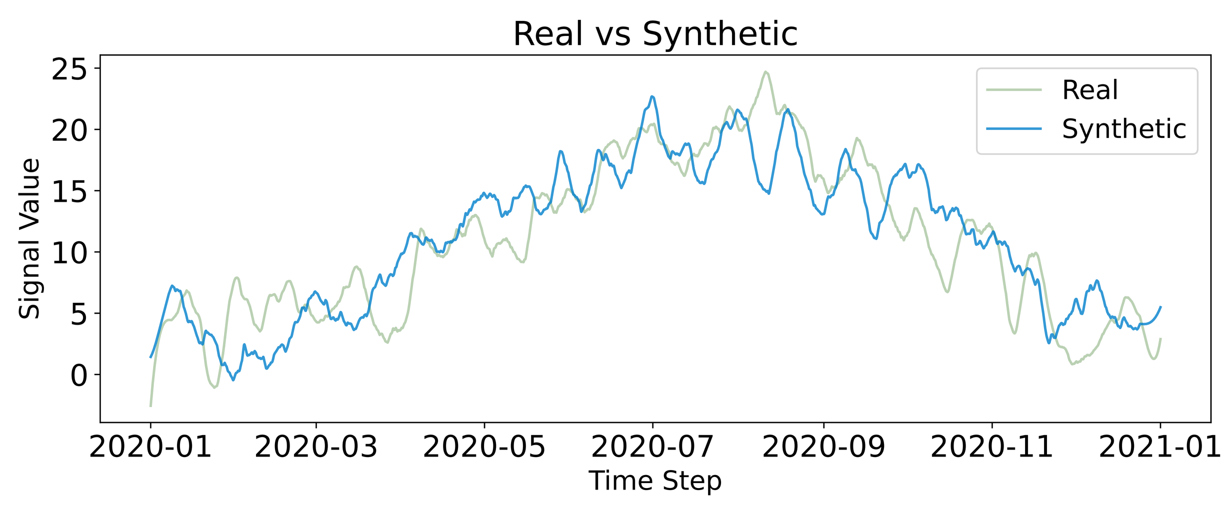

Figure: Comparison between real-world data and data generated by the SDG. The close alignment indicates that the SDG effectively captures the patterns present in real data.

An SDG is a parametric model designed to simulate univariate time series data characterized by seasonality, temporal dependence (AR memory), and trend. For a given client k, feature f, and time step t, the observation $\hat{x}_{f,t,k}$ is:

$$\begin{aligned}\hat{x}_{f,t,k} &= \mathrm{Seasonal}(A_{f,j,k}, T_{f,j,k}, \Theta_{f,j,k}) + \mathrm{AR}_{p,k}(\phi_k) \\ &\quad + \mathrm{Trend}(\beta_{f,k}) + \epsilon_{f,t,k} \\ &= \sum_{j=1}^{J} A_{f,j,k} \cdot \sin\left(\frac{2\pi t}{T_{f,j,k}} + \theta_{f,j,k}\right) \\ &\quad + \sum_{i=1}^{p} \phi_{k,i}\, x_{f,t-i,k} + \beta_{f,k}\, t + \epsilon_{f,t,k}.\end{aligned} \tag{1}$$

Feature Skewness Formulation

In federated learning, each client observes a different distribution of the same features (feature skew). We model this heterogeneity as:

$$x_{f,t,k} = \Lambda_{f,k}\tilde{x}_{f,t,k} + \delta_{f,k} \tag{2}$$

where Λ_{fk} is the linear scale, which controls how the variance of the feature f, σ_f^2, changes for client k; δ_{fk} is the mean shift, which changes the mean of the feature f, μ_f, for the client k. Note that, though the univariate SDG is able to describe each feature, each client is allowed to observe a subset of all the features.

Intrinsic Space Construction

We construct a geometry-aware representation space that captures the essential temporal structure of non-IID time series. The transformation pipeline involves: (1) Client-wise normalization; (2) Window flattening; (3) Global covariance estimation; (4) Intrinsic dimension estimation; and (5) Projection. The intrinsic dimension for client k is approximated as:

$$d_{I,k}(H) \approx F \cdot \left(\min\{H, \ell_{\mathrm{AR},k}\} + g_k(H) + 1\right). \tag{3}$$

Here, $\ell_{\mathrm{AR},k}$ denotes the effective AR memory:

$$\ell_{\mathrm{AR},k} = \left\lceil \frac{\ln(1/(1-\epsilon))}{-\ln \rho_k} \right\rceil, \quad \epsilon \in (0,1) \tag{4}$$

where $\rho_k \in (0,1)$ is the spectral radius of the AR companion matrix. $g_k(H)$ reflects the resolved seasonal complexity:

$$g_k(H) = 2\sum_{j=1}^{J} w_{j,k} \cdot \min\left(1, \frac{H}{T_{j,k}^*}\right), \tag{5}$$

$$w_{j,k} = \frac{\sum_{f=1}^{F} A_{f,j,k}^2}{\sum_{f=1}^{F}\sum_{j=1}^{J} A_{f,j,k}^2}. \tag{6}$$

This formulation yields a compact and information-preserving representation that enables a precise loss decomposition and supports optimal horizon analysis under federated, non-IID settings.

Loss Decomposition

We now formalize this precise decomposition of the prediction loss in the federated setting, showing how the Bayesian and approximation components arise directly from the client-specific data-generating distributions and the server-side evaluation protocol.

Theorem (Federated Loss Decomposition)

For each client $k \in \{1,\ldots,K\}$, let (U_k, V_k) denote its data-generating pair, where U_k takes values in a measurable input space $\mathcal{M}(H)$ and V_k in an output space $\mathcal{M}(S)$, both embedded in a real Hilbert space $(\mathcal{H}, \|\cdot\|)$ with the associated Borel σ-algebras.
Let $m_k^*(u) := \mathbb{E}[\, V_k \mid U_k = u\,]$ be the client-specific Bayesian predictor, defined P_{U_k}–almost everywhere. For any measurable, square-integrable predictor $m : \mathcal{M}(H) \to \mathcal{M}(S)$, the server's global predictive loss is

$$L(H,S;m) := \mathbb{E}_{k\sim\pi}\Big[\mathbb{E}\big[\, \| V_k - m(U_k) \|^2\big]\Big] \tag{7}$$

where $\pi = (\pi_1,\ldots,\pi_K)$ is any distribution over clients and the inner expectation is over (U_k, V_k) under client k's distribution. Then the loss decomposes as:

$$L(H,S;m) = L_{\mathrm{Bayes}}(H,S) + L_{\mathrm{approx}}(H,S;m), \tag{8}$$

where the federated Bayesian loss is

$$L_{\mathrm{Bayes}}(H,S) := \mathbb{E}_{k\sim\pi}\Big[\mathbb{E}\big[\, \| V_k - m_k^*(U_k) \|^2\big]\Big], \tag{9}$$

and the federated approximation loss is

$$L_{\mathrm{approx}}(H,S;m) := \mathbb{E}_{k\sim\pi}\Big[\mathbb{E}\big[\, \| m_k^*(U_k) - m(U_k) \|^2\big]\Big]. \tag{10}$$

In particular, the total loss separates into the expected irreducible (client-wise Bayes) component and the expected approximation error of the global predictor relative to each client's Bayes-optimal rule.

Smallest Sufficient Horizon

Now we define a key concept, the smallest sufficient horizon, which serves as the optimal look-back horizon that minimizes the forecasting loss. Formally, for any tolerance $\delta > 0$, define the smallest sufficient horizon as

$$H_k^*(\delta) := \min\{H : |\Delta L_{\mathrm{Bayes}}^{(k)}(H)| \le \delta\}, \tag{11}$$

at which the Bayesian loss has effectively saturated: further historical context improves the irreducible loss by at most δ. Together, these monotonicity properties imply a unimodal structure for the total loss.

Unimodality and Optimal Horizon

If for a given $\delta > 0$ the Bayesian loss satisfies $\Delta L_{\mathrm{Bayes}}^{(k)}(H) \le -\delta$ for all $H < H_k^*(\delta)$, and the approximation loss satisfies $\Delta L_{\mathrm{approx}}^{(k)}(H;m) \ge \delta$ for all $H \ge H_k^*(\delta)$, then the combined loss obeys that $L^{(k)}(H)$ decreases on $[1, H_k^*(\delta)]$, and $L^{(k)}(H)$ increases on $[H_k^*(\delta), \infty)$. Consequently, $H_k^*(\delta) \in \arg\min_{H\in\mathbb{N}} L^{(k)}(H)$ with uniqueness up to integer ties.

Proof.

From the Bayesian loss analysis, increasing H reduces seasonal/phase ambiguity and uncovers AR structure, but only up to a finite coverage horizon. Hence, there exists H_0 such that

$$\Delta L_{\mathrm{Bayes}}(H,S) < 0 \quad (H < H_0), \tag{12}$$

while for any $\delta > 0$ we can choose H_0 large enough so that

$$\Delta L_{\mathrm{Bayes}}(H,S) \ge -\delta \quad (H \ge H_0). \tag{13}$$

For the approximation term, the curvature–variance bound on the intrinsic manifold shows that the error grows with both the intrinsic dimension $d_I(H)$ and the factor H/D coming from the effective sample size per window ($\propto D/(HN)$). Since $d_I(H)$ is non-decreasing and eventually saturated, while H/D grows linearly, there exists $\eta > 0$, independent of H, such that

$$\Delta L_{\mathrm{approx}}(H,S) \ge \eta \quad (H \ge H_0). \tag{14}$$

Fix any $\delta \in (0,\eta)$ and define $H^*(\delta)$ as the smallest $H \ge H_0$ with $\Delta L_{\mathrm{Bayes}}(H,S) \ge -\delta$. Then for $H < H^*(\delta)$, we have $\Delta L_{\mathrm{Bayes}}(H,S) < -\delta$ and $\Delta L_{\mathrm{approx}}(H,S) \ge 0$, so $\Delta L(H,S) = \Delta L_{\mathrm{Bayes}}(H,S) + \Delta L_{\mathrm{approx}}(H,S) < -\delta < 0$, and $L(H,S)$ is strictly decreasing. For $H \ge H^*(\delta)$, we have $\Delta L_{\mathrm{Bayes}}(H,S) \ge -\delta$ and $\Delta L_{\mathrm{approx}}(H,S) \ge \eta$, hence $\Delta L(H,S) \ge -\delta + \eta > 0$, so $L(H,S)$ is strictly increasing.
Thus $L(H,S)$ decreases up to $H^*(\delta)$ and increases thereafter, so it is unimodal in H and attains its unique minimum at $H^*(\delta)$ (up to trivial ties), as claimed. $\square$

Hence, before $H_k^*(\delta)$, the reduction in irreducible error outweighs the increase in approximation error; afterwards, the opposite holds. The total loss thus has a single optimal basin, and the smallest sufficient horizon attains the minimum.

Limitation and Discussion

This work introduces a principled framework for federated time-series forecasting under non-IID data, built on a structured synthetic data generator (SDG) and an intrinsic representation space. The formulation enables a clean decomposition of forecasting error and yields a provably optimal look-back horizon.

The SDG captures key components, but assumes Gaussian innovations, local stationarity, and stable AR structure, limiting its ability to represent regime shifts, nonlinear patterns, or strong feature interactions. Estimating global covariance also requires privacy-aware aggregation, and treating overlapping windows as independent may overstate sample size.

References

[1] J. Shi, Q. Ma, H. Ma, and L. Li, "Scaling law for time series forecasting," in *Advances in Neural Information Processing Systems*, A. Globerson et al., Eds., vol. 37, Curran Associates, Inc., 2024, pp. 83 314–83 344. DOI: 10.52202/079017-2650. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/97c2f0fac182353062d304d0322ae285-Paper-Conference.pdf.