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

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

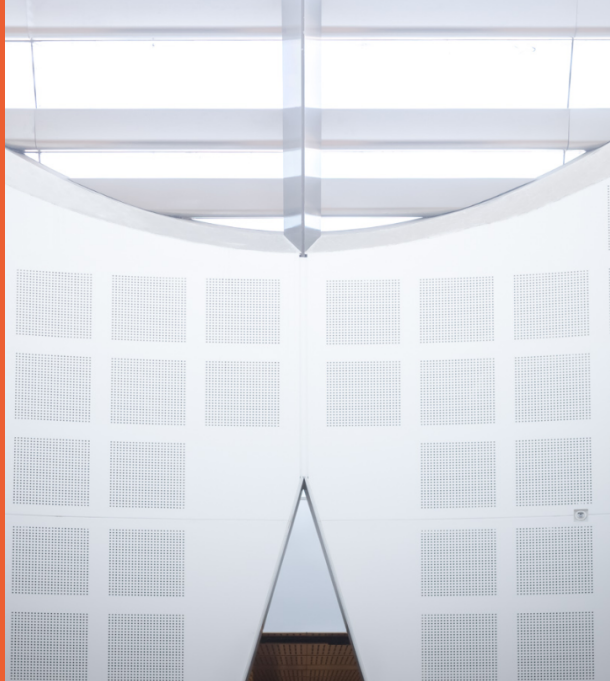
[1]University of Sydney

[2]University of Technology Sydney

THE UNIVERSITY OF SYDNEY

AAAI-26 / IAAI-26 / EAAI-26

JANUARY 20-27, 2026 | SINGAPORE



Introduction

Selecting an appropriate look-back horizon remains a fundamental challenge in time series forecasting (TSF), particularly in the federated learning (FL) scenarios where data is decentralized, heterogeneous, and often non-independent.

Problem Gap: Horizon Selection in Federated TSF

1. Horizon Selection in Time Series Forecasting (TSF)

- Traditional methods (e.g., ARIMA, LSTMs, Transformers) treat the look-back horizon as a tunable hyperparameter set via heuristics or validation, often leading to overfitting or inefficient data use [1], [2], [3], [4], [5], [6], [7].

2. Intrinsic Representation

- Shi et al (2024) introduced the idea of an intrinsic space framework for dynamic horizon selection, but assumes centralized, IID data [8]. The framework is lack of step-by-step transformation from time series into the intrinsic space.

3. Federated Learning (FL) for TSF

- Prior FL research focuses on aggregation algorithms (FedAvg, FedProx) or model architectures but does not address horizon selection, especially under the non-IID client dynamic settings [9], [10], [11].

Research Goal

Find the optimal look-back horizon for time series forecasting in the non-IID data setting under the federated learning scenario.

Solution Pipeline & 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 an FL-based decomposition of predictive loss into Bayesian and Approximation components, each analytically tied to the structural elements of time series data and the look-back horizon. Our analysis shows 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 client-wise total loss is (conditionally) unimodal with respect to the horizon length and identify the smallest sufficient horizon as its global minimiser. This result provides the first rigorous criterion for horizon selection in time series forecasting under sample-limited, heterogeneous environments.

SDG: Modeling Temporal Structures

We model the observation $\hat{x}_{f,t,k}$ for feature *f*, time *t*, and client *k* as:

$$\hat{x}_{f,t,k} = \underbrace{\sum_{j=1}^{J} A_{f,j,k} \cdot \sin\left(\frac{2\pi t}{T_{f,j,k}} + \theta_{f,j,k}\right)}_{\text{Seasonality}} + \underbrace{\sum_{i=1}^{p} \phi_{k,i}\, x_{f,t-i,k}}_{\text{AR Memory}} + \underbrace{\beta_{f,k}\, t}_{\text{Trend}} + \underbrace{\epsilon_{f,t,k}}_{\text{Noise}} \qquad (1)$$

- **Seasonality:** Defined by amplitude $A_{f,j,k}$, period $T_{f,j,k}$, and phase θ.
- **AR Memory:** Autoregressive process with client-specific lag coefficients $\phi_{k,i}$.
- **Trend:** Linear component slope $\beta_{f,k}$.
- **Noise:** Gaussian innovation $\epsilon \sim \mathcal{N}(\mu_{f,k}, \sigma^2_{f,k})$.

Capturing Feature Heterogeneity

Real-world FL data exhibits feature skew. We model this via affine transformations:

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

Explanation of Terms:

- $\Lambda_{f,k}$ (Linear Scale): Controls the variance σ_f^2 for feature *f* on client *k*.
- $\delta_{f,k}$ (Mean Shift): Adjusts the mean μ_f for feature *f* on client *k*.

Significance: This allows the framework to explicitly account for non-IID distributions where clients observe different scales of the same underlying features.

Intrinsic Space Transformation

We construct a geometry-aware representation space that captures the essential temporal structure of non-IID time series through a transformation grounded in the SDG. The transformation pipeline proceeds in five steps:

1. Client-wise normalization to remove affine feature skew and align marginal distributions
2. Window flattening to convert each normalized time-series segment into a fixed-length vector
3. Global covariance estimation and eigendecomposition to identify dominant axes of variation
4. Intrinsic dimension estimation based on the SDG and empirical spectrum
5. Projection into intrinsic space via principal components

Intrinsic Dimension Estimation

We map windows to an intrinsic space. The dimension $d_{l,k}(H)$ approximates the effective degrees of freedom:

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

Components:

- F: Number of features.
- $\ell_{\mathrm{AR},k}$: Effective AR memory length (how far back history matters).
- $g_k(H)$: Resolved seasonal complexity.
- $+1$: Accounts for the linear trend component.

Detailed Intrinsic Components

1. Effective AR Memory ($\ell_{\mathrm{AR},k}$):

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

- ρ_k: Spectral radius of the AR companion matrix (stability metric).
- Represents the time steps needed for impulse response to decay.

2. Seasonal Complexity ($g_k(H)$):

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

- Measures how many full seasonal cycles fit within horizon H.
- Saturates when H exceeds the period $T^*_{j,k}$.

Theorem 1: Federated Loss Decomposition

The total prediction loss for a predictor m decomposes into two distinct sources of error for both the client and server side:

$$L(H, S; m) = \underbrace{L_{\text{Bayes}}(H, S)}_{\text{Irreducible}} + \underbrace{L_{\text{approx}}(H, S; m)}_{\text{Reducible}} \tag{6}$$

- **Bayesian Loss (L_{Bayes}):** The error of an ideal predictor with full knowledge of the data distribution. It reflects inherent uncertainty (noise).
- **Approximation Loss (L_{approx}):** The excess error due to using a finite-capacity model trained on finite samples, relative to the Bayes-optimal predictor.

Theorem 2: Client-wise Bayesian Loss

For client k, the irreducible loss is the sum of component-wise errors:

$$L_{\text{Bayes}}^{(k)}(H, S) = L_{\text{AR}}^{(k)}(S) + L_{\text{seas}}^{(k)}(H) + L_{\text{trend}}^{(k)}(H) \tag{7}$$

Behaviour with respect to Horizon (H):

- **Decreases:** As H increases, we resolve more seasonal phases and trend direction.
- **Saturates:** Once H covers the AR memory and seasonal periods, adding more history provides **zero** additional information gain.
- **Limit:** $\Delta L_{\text{Bayes}}^{(k)}(H) \to 0$ for large H.

Theorem 3: Approximation Loss Bound

The approximation error is bounded by curvature (bias) and variance terms:

$$L_{\text{approx}}^{(k)}(H, S; m) \lesssim \underbrace{\left(K_2^2\, d_{l,k}(H)^2\right)^{\frac{d_{l,k}(H)}{4+d_{l,k}(H)}}}_{\text{Curvature / Bias Term}} + \underbrace{\left(\frac{d_{l,k}(H)\, H}{D_k}\right)^{\frac{4}{4+d_{l,k}(H)}}}_{\text{Variance / Finite Sample Term}} \tag{8}$$

Why does this increase with H?

1. **Intrinsic Dimension ($d_{l,k}$):** Grows with H, making the function harder to learn (Curse of Dimensionality).
2. **Effective Samples (D_k/H):** As window length H grows, the number of independent samples in a fixed dataset decreases, increasing variance.

The Fundamental Trade-off

The total loss $L^{(k)}(H)$ is (conditionally) minimised at the smallest sufficient horizon H^*.

- **Small H:** High Bayesian Loss (Underfitting the dynamics).
- **Large H:** High Approximation Loss (Overfitting / Insufficient samples).

Smallest Sufficient Horizon (H_k^*)

Defined as the smallest H where Bayesian loss improvements saturate within a tolerance δ:

$$H_k^*(\delta) := \min\{H : |\Delta L_{\text{Bayes}}^{(k)}(H)| \leq \delta\} \tag{9}$$

Theorem 4 (Margin Conditions). If there exists $\delta > 0$ such that

$$\Delta L_{\text{Bayes}}^{(k)}(H) \leq -\delta, \ \forall H < H_k^*(\delta), \qquad \Delta L_{\text{approx}}^{(k)}(H) \geq \delta, \ \forall H \geq H_k^*(\delta), \tag{10}$$

then the total loss $L^{(k)}(H)$ is **unimodal** and is **minimized at** $H_k^*(\delta)$ (up to integer ties).

Practical Selection: Seasonal Coverage

We can link the tolerance δ to interpretable seasonal coverage:

$$H_k^*(\delta) = \max\{\ell_{\mathrm{AR},k},\ T_k^{(\tau)}\} \tag{11}$$

Interpretation:

- The optimal horizon is simply the maximum of the **effective AR memory** and the **seasonal period** required to capture $(1-\tau)$ of the signal energy.
- This provides a theoretically grounded, calculable target for each client.

Global Horizon

In Federated Learning, we need a single global horizon H_{server} despite heterogeneous local optima. We use a weighted trimmed mean:

$$H^*_{\text{server}} = \text{TrimMean}_\alpha \left(\{H^*_k(\delta)\}_{k=1}^K ; \{w_k\}_{k=1}^K \right) \tag{12}$$

- **Weights w_k:** Proportional to client data size (n_k).
- **Trimmed Mean:** Discards the top/bottom α-fraction of extreme horizons.
- **Benefit:** Robustness. Prevents a single client with an extremely long seasonality (requiring huge H) from degrading the sample efficiency for all other clients.

Limitations & Future Work

Limitations:

- Assumes a structured SDG, which may not capture complex nonlinear or regime-switching real-world dynamics.
- Validated on limited real-world settings, so the generality of the optimal horizon theory across diverse federated datasets remains untested.

Future work:

- Extend the SDG and theory to handle nonlinear, non-stationary, or multivariate temporal behaviors beyond the current additive model.
- Validate the framework across broader and more challenging federated forecasting benchmarks, including irregular sampling and strong distribution drift.

Conclusion

This work establishes the first theoretically grounded criterion for adaptive horizon selection, offering practical guidance for model design and benchmarking in decentralized, heterogeneous environments.

- **Novel Framework:** This work introduces a principled approach for selecting input horizons in non-IID federated time series using a Synthetic Data Generator (SDG).
- **Federated Loss Decomposition:** We show that the decomposition of the forecasting loss into Bayesian error (decreases with horizon) and approximation error (increases with horizon) works under the FL settings.
- **Optimal Horizon:** We prove that the client-wise total loss is (conditionally) minimised at the smallest sufficient horizon H^*, balancing structure identification against the curse of dimensionality. We also propose a robust aggregation mechanism to identify a single, effective horizon across heterogeneous clients.

Thank You

Questions?

References I

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974. DOI: 10.1109/TAC.1974.1100705.

[2] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th. Hoboken: Wiley, 2015.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[4] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International journal of forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[5] H. Zhou et al., "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 11 106–11 115.

[6] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi, "Learning deep time-index models for time series forecasting," in *International Conference on Machine Learning*, PMLR, 2023, pp. 37 217–37 237.

[7] K. A. Koparanov, K. K. Georgiev, and V. A. Shterev, "Lookback period, epochs and hidden states effect on time series prediction using a lstm based neural network," in *2020 28th National Conference with International Participation (TELECOM)*, Sofia, Bulgaria: IEEE, 2020, pp. 61–64. DOI: 10.1109/TELECOM50385.2020.9299551.

[8] 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.

[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication□efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

References II

[10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[11] V. Perifanis, N. Pavlidis, R.-A. Koutsiamanis, and P. S. Efraimidis, "Federated learning for 5g base station traffic forecasting," *Computer Networks*, vol. 235, no. 109950, 2023. DOI: 10.1016/j.comnet.2023.109950.