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
dahao.tang@sydney.edu.au, n.yang@sydney.edu.au, yanli.li@sydney.edu.au, zhiyu.zhu@student.uts.edu.au, zhibo.jin@student.uts.edu.au, dong.yang@sydney.edu.au

Abstract

Selecting an appropriate look-back horizon remains a fundamental challenge in time series forecasting (TSF), particularly in the federated learning scenarios where data is decentralized, heterogeneous, and often non-independent. While recent work has explored horizon selection by preserving forecasting-relevant information in an intrinsic space, these approaches are primarily restricted to centralized and independently distributed settings. This paper presents a principled framework for adaptive horizon selection in federated time series forecasting through an intrinsic space formulation. We introduce a synthetic data generator (SDG) that captures essential temporal structures in client data, including autoregressive dependencies, seasonality, and trend, while incorporating client-specific heterogeneity. Building on this model, we define a transformation that maps time series windows into an intrinsic representation space with well-defined geometric and statistical properties. We then derive a decomposition of the forecasting loss into a Bayesian term, which reflects irreducible uncertainty, and an approximation term, which accounts for finite-sample effects and limited model capacity. Our analysis shows that while increasing the look-back horizon improves the identifiability of deterministic patterns, it also increases approximation error due to higher model complexity and reduced sample efficiency. We prove that the total forecasting loss is minimized at the smallest horizon where the irreducible loss starts to saturate, while the approximation loss continues to rise. This work provides a rigorous theoretical foundation for adaptive horizon selection for time series forecasting in federated learning.

Introduction

Time series forecasting (TSF) underpins numerous high-impact domains, including finance (Zivot and Wang 2006), healthcare (Futoma, Hariharan, and Heller 2017), and energy systems (Kong et al. 2019), where accurate prediction of future values from historical trends is crucial for informed decision-making and operational efficiency. A central modeling choice in TSF is the selection of the look-back horizon, defined as the number of past time steps used as input. This choice significantly influences model complexity, predictive accuracy, and generalization performance (Lim et al. 2021).

Traditionally, the look-back horizon is treated as a tunable hyperparameter, often selected via cross-validation or heuristic search. Recent theoretical advances offer a more principled perspective. Shi et al. (Shi et al. 2024) propose a scaling law theory based on a theoretical framework that embeds time series into an intrinsic representation space, allowing the forecasting loss to be decomposed into two components: Bayesian error, capturing irreducible uncertainty from noise and limited information, and approximation error, reflecting the model's capacity to learn the true mapping. This decomposition enables analytical reasoning about the optimal look-back horizon as a function of dataset size, model complexity, and intrinsic dimensionality (Sharma and Kaplan 2020; Bahri et al. 2024). Empirical results support these insights, showing that the optimal horizon grows with data availability and varies by model type. For example, channel-dependent models like iTransformer (Liu et al. 2024) benefit from shorter horizons under limited data, while linear models such as NLinear (Zeng et al. 2023) maintain performance with longer horizons due to smoother feature decay and lower intrinsic complexity (Xu, Zeng, and Xu 2024; Toner and Darlow 2024).

However, this framework relies on strong assumptions, including centralized data, independent identically distribution (IID), and homogeneous model architectures, which are often violated in real-world federated learning scenarios. In such decentralized settings, data is distributed across clients with diverse distributions, sequence lengths, and domain characteristics (Kairouz et al. 2021). Applying a globally fixed horizon in this context may lead to mismatches between local dynamics and model inputs, degrading forecasting performance (Edwards et al. 2024). Moreover, real-world data frequently exhibits feature sparsity, variable noise levels, and heterogeneous scaling behaviors, challenging the smooth manifold assumptions in intrinsic dimension theory (Levi and Oz 2024; Zador 1982). These limitations highlight the need for adaptive horizon strategies that account for both data heterogeneity and localized model constraints. Integrating hybrid architectures or meta-learning mechanisms with principled theoretical foundations, such as those introduced by Shi et al., presents a promising direction for addressing these challenges in federated time series forecasting.

This paper addresses the challenge of selecting the opti-

*Corresponding authors.

mal look-back horizon for time series forecasting in federated learning environments characterized by non-IID client data. We develop a principled framework that leverages a structured Synthetic Data Generator (SDG) to model core temporal patterns (e.g., autoregressive dynamics, seasonality, and trends), while capturing client-specific heterogeneity. Using the SDG as a foundation, we construct a data-aware transformation that maps time series windows into an intrinsic representation space with well-defined geometric and statistical properties. This enables a rigorous loss decomposition into irreducible (Bayesian) and approximation components, each tied to the underlying generative structure. Crucially, the formulation reveals how the informativeness of historical context and thus the optimal look-back horizon varies across clients depending on their local dynamics and data regimes. Our analysis shows that the total forecasting loss is minimized at the smallest horizon where the Bayesian error saturates and the approximation error begins to dominate, yielding a theoretically grounded, client-adaptive criterion for horizon selection in federated forecasting settings. Our contributions include the following:

- We propose a novel intrinsic space formulation that transforms heterogeneous, non-IID multivariate time series into a compact and geometry-preserving representation. This space is rigorously characterized by bi-Lipschitz continuity, intrinsic dimensionality saturation, and inter-horizon compatibility, enabling consistent comparison and reasoning across clients and temporal contexts.
- 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.
- 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.

Related Work

Horizon Selection and Intrinsic Representation

A central yet understudied question in time series forecasting (TSF) is how much historical context, i.e., look-back horizon, is truly needed for accurate prediction (Kim et al. 2025). Traditional statistical models such as ARIMA select lag length using information criteria like AIC (Akaike 1974; Box et al. 2015), which implicitly perform horizon selection under strong linearity assumptions. While interpretable, these methods struggle to capture nonlinear or long-range dependencies. Modern deep learning approaches, including LSTMs (Hochreiter and Schmidhuber 1997), hybrid models like LSTNet (Lai et al. 2018), and attention-based architectures such as the Temporal Fusion Transformer (Lim et al. 2021) or Informer (Zhou et al. 2021), have greatly improved modeling capacity. However, they still treat the input horizon as a tunable hyperparameter, typically set through validation or heuristics, without theoretical grounding. This empirical approach can lead to overfitting, underfitting, or inefficient use of data, particularly in settings with limited or distributed samples (Woo et al. 2023; Koparanov, Georgiev, and Shterev 2020).

Recent theoretical work has begun to formalize the horizon selection problem. Notably, Shi et al. (2024) analyze how forecasting error scales with input length, dataset size, and model complexity, revealing a trade-off: longer horizons can improve identifiability of temporal structure but also increase approximation error due to model limitations and finite data (Shi et al. 2024). Their framework introduces the notion of an intrinsic representation space, where the forecasting loss decomposes into two parts: a Bayesian (irreducible) error reflecting inherent unpredictability, and an approximation error arising from statistical and model constraints. This idea builds on Takens' embedding theorem (Takens, Young, and Rand 2006), which implies that a system's future behavior can be reconstructed from a finite number of past observations, defining an intrinsic dimension sufficient for prediction. However, Shi's work assumes centralized and IID data, limiting its relevance to modern federated learning scenarios where data is distributed, non-IID, and client-specific. We extend this theory to federated, non-IID settings by introducing an intrinsic representation that captures essential temporal structure across clients. This enables a principled approach to selecting the optimal look-back horizon in decentralized forecasting.

Time Series Forecasting in Federated Learning

Federated learning (FL) enables decentralized training across clients without sharing raw data. The foundational FedAvg algorithm introduced by McMahan et al. (2017) laid the groundwork for collaborative model training in privacy-sensitive environments (McMahan et al. 2017). However, FL under non-IID data poses major challenges, including model divergence, degraded generalization, and client imbalance. To address data heterogeneity, methods like FedProx (Li et al. 2020) introduce regularization terms that stabilize optimization across diverse client distributions. In time series forecasting specifically, recent works apply FL to real-world sequential tasks, such as traffic and energy demand prediction, but focus primarily on model architecture and aggregation (Perifanis et al. 2023). These systems rarely examine how temporal structure varies across clients or how such variation affects forecasting horizons. While personalization and communication efficiency have been explored, no prior work provides a theoretical framework for look-back horizon selection in federated TSF. Our paper addresses this gap by analyzing horizon choice through the lens of synthetic modeling and intrinsic representation under client heterogeneity.

Preliminary

In this section, we defined the basic settings for time series forecasting in the federated learning scenario. More specifi-

cally, we propose a synthetic data generator (SDG) that well describes real-world non-IID data and implement a step-by-step transformation that converts the time series data described by the SDG into an intrinsic space that represents the information carried by a time series.

Time Series Forecasting in Federated Learning

We study S-step forecasting from a length-H look-back window in a federated setting with K clients. Client $k \in \{1, \dots, K\}$ holds a multivariate time series $\{x_t^{(k)}\}_{t=1}^{L_k}$ with F features, $x_t^{(k)} \in \mathbb{R}^F$. For time index $t \in \{H, \dots, L_k - S\}$, define the input window and S-step target block as:

$$X_{t,k}^{(H)} = \left[x_{t-H+1}^{(k)}, \dots, x_t^{(k)}\right] \in \mathbb{R}^{F \times H}, \tag{1}$$

$$Y_{t,k}^{(S)} = \left[x_{t+1}^{(k)}, \dots, x_{t+S}^{(k)}\right] \in \mathbb{R}^{F \times S}. \tag{2}$$

Training proceeds in rounds via standard FL aggregation (e.g., FedAvg). On client k, overlapping windows yield D_k training samples; due to overlap, the number of effectively independent samples scales as D_k/H.

Intrinsic Space Formulation

We adopt the concept of intrinsic space to represent the information carried by a time series. The intrinsic dimension d_I of the intrinsic space is defined as the minimum number of dimensions required to represent the time series without losing significant information. To gain a deeper understanding of the intrinsic space, we investigate the typical structure of non-IID time series data in the federated learning scenario and propose a synthetic data generator that is both theoretically and empirically proven to be sound in describing the structure of the focused non-IID time series data.

Synthetic Data Generator A Synthetic Data Generator (SDG) is a parametric model designed to simulate univariate time series data, which often exhibits structural patterns characterized by seasonality, temporal dependence (AR memory), and trend (Kim et al. 2025).

For a given client k, feature f, and time step t, the synthetic observation $\hat{x}_{f,t,k}$ is defined as:

$$\begin{aligned}\hat{x}_{f,t,k} &= \text{Seasonal}(A_{f,j,k}, T_{f,j,k}, \Theta_{f,j,k}) + \text{AR}_{p,k}(\phi_k) \\ &\quad + \text{Trend}(\beta_{f,k}) + \epsilon_{f,t,k} \\ &= \sum_{j=1}^{J} A_{f,j,k} \cdot \sin\left(\frac{2\pi t}{T_{f,j,k}} + \theta_{f,j,k}\right) \\ &\quad + \sum_{i=1}^{p} \phi_{k,i}\, x_{f,t-i,k} + \beta_{f,k}\, t + \epsilon_{f,t,k}.\end{aligned} \tag{3}$$

Here, seasonality is represented by a sum of sinusoids, parameterized by amplitude $A_{f,j,k}$, period $T_{f,j,k}$, and phase shift $\theta_{f,j,k}$. Temporal dependence is modeled via an autoregressive process $\text{AR}_{p,k}(\phi_k) = \sum_{i=1}^{p} \phi_{k,i}\, x_{f,t-i,k}$, where $\phi_{k,i}$ are the lag coefficients specific to client k. The trend is captured by a linear component $\text{Trend}(\beta_{f,k}) = \beta_{f,k}\, t$. The additive noise term is drawn from a Gaussian distribution: $\epsilon_{f,t,k} \sim \mathcal{N}(\mu_{f,k},\ \sigma_{f,k}^2)$.

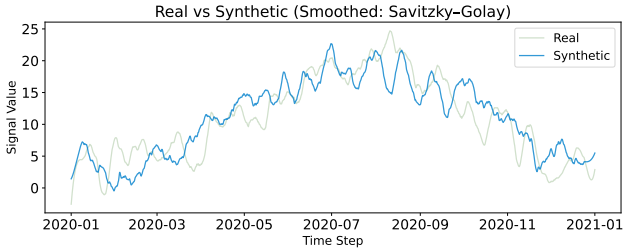

Figure 1: Comparison between real-world data and data generated by the SDG. The close alignment indicates that the SDG effectively captures the patterns present in real data.

We also provide empirical studies to demonstrate the validity of the SDG, as illustrated in Figure 1. Please refer to the Extended Version for more details.

Feature Skewness Formulation In the federated learning scenario, each client tends to observe a different distribution of the same features in time series data, simulating feature skew (Wu et al. 2024). We apply a customized skewness partitioning method to create feature heterogeneity.

To be more precise, we construct an affine transformation for each data point of the SDG: for client k, feature k:

$$x_{f,t,k} = \Lambda_{f,k} \tilde{x}_{f,t,k} + \delta_{f,k} \tag{4}$$

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 At a high level, 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, which explicitly models autoregressive dependencies, seasonal cycles, and linear trends, and serves as a unifying scaffold for both analytical reasoning and empirical evaluation across heterogeneous clients.

Our construction is supported by a set of structural assumptions. These include: (i) compactness of the intrinsic image to ensure bounded representation norms; (ii) bi-Lipschitz continuity to preserve distances and guarantee stable inverses; (iii) a horizon-indexed intrinsic dimension that increases monotonically and saturates once all relevant temporal structure is captured; (iv) compatibility of representations across horizons via stable linear projections; (v) approximate commutativity between truncation and projection, ensuring robustness under input length variation; and (vi) a power-law spectrum of the intrinsic covariance, which enables efficient dimensionality reduction. These assumptions reflect statistical regularities commonly observed in time series data and enable a clean separation between modeling complexity and representational geometry.

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; and (5) *Projection into intrinsic space* via principal components. Specifically, the intrinsic dimension for client k is approximated as:

$$d_{I,k}(H) \approx F \cdot (\min\{H, \ell_{\mathrm{AR},k}\} + g_k(H) + 1)\,. \tag{5}$$

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

$$\ell_{\mathrm{AR},k} = \left\lceil \frac{\ln(1/(1-\epsilon))}{-\ln \rho_k} \right\rceil,\ \epsilon \in (0,1) \tag{6}$$

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

$$g_k(H) = 2\sum_{j=1}^{J} w_{j,k} \cdot \min\left(1, \frac{H}{T^*_{j,k}}\right), \tag{7}$$

$$w_{j,k} = \frac{\sum_{f=1}^{F} A^2_{f,j,k}}{\sum_{f=1}^{F}\sum_{j=1}^{J} A^2_{f,j,k}}. \tag{8}$$

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. Please refer to the Extended Version for more details.

Loss Analysis

Before we proceed to analyze how the look-back horizon H affects forecasting performance in the federated setting, we explore decomposition of the prediction loss into two components: an intrinsic, irreducible term that captures the uncertainty of the data-generating process, and an approximation term that captures the difficulty of learning a mapping in an H-dimensional input space using a finite-capacity model trained on finite and heterogeneous data.

Overall Loss Analysis

Consider the forecasting task of predicting the next S values from the previous H observations. Under the intrinsic-space representation, this corresponds to learning a mapping $m : \mathcal{M}(H) \to \mathcal{M}(S)$. For a given client distribution and any measurable predictor m, the squared loss can be decomposed into

$$L(H, S; m) = L_{\mathrm{Bayes}}(H, S) + L_{\mathrm{approx}}(H, S; m) \tag{9}$$

where:

- **Bayesian loss** $L_{\mathrm{Bayes}}(H, S)$ is the irreducible error incurred even by an ideal predictor with full knowledge of the data distribution.
- **Approximation loss** $L_{\mathrm{approx}}(H, S; m)$ captures the additional error due to using a finite-capacity predictor m trained on limited local data.

The formal derivation of (9) and the precise definitions of the two terms are given in the Extended Version.

We now formalize this intuition by establishing a 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 1 (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

$$L(H, S; m) := \mathbb{E}_{k\sim\pi}\Big[\mathbb{E}\big[\, \| V_k - m(U_k) \|^2\big]\Big] \tag{10}$$

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:

$$L(H, S; m) = L_{\mathrm{Bayes}}(H, S) + L_{\mathrm{approx}}(H, S; m), \tag{11}$$

where the federated Bayesian loss is

$$L_{\mathrm{Bayes}}(H, S) := \mathbb{E}_{k\sim\pi}\Big[\mathbb{E}\big[\, \| V_k - m^*_k(U_k) \,|^2\big]\Big], \tag{12}$$

and the federated approximation loss is

$$L_{\mathrm{approx}}(H, S; m) := \mathbb{E}_{k\sim\pi}\Big[\mathbb{E}\big[\, \| m^*_k(U_k) - m(U_k) \|^2\big]\Big]. \tag{13}$$

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. Please refer to the Extended Version for the proof.

Server–client interpretation The global model m is hosted on the central server and evaluated on clients sampled according to $k \sim \pi$. The term $L_{\mathrm{Bayes}}(H, S)$ captures the irreducible uncertainty within each client's local data-generating process, averaged over clients, while $L_{\mathrm{approx}}(H, S; m)$ measures the discrepancy between the global server model and the collection of Bayes-optimal per-client predictors $\{m^*_k\}_{k=1}^K$. Although both components are defined via client-side distributions, the total loss $L(H, S; m)$ represents the expected prediction error of the server's global model.

In the remainder of this section, we investigate these two components respectively.

Bayesian (Irreducible) Loss

We first characterize the irreducible component of predictive loss for each client using the structure of the SDG.

Theorem 2 (Client-wise Bayesian Loss)**.** *According to the SDG model in Equation* (3) *for client k: each feature is generated as an additive sum of (i) an autoregressive component, (ii) a seasonal component, (iii) a linear trend, and (iv)*

an innovation noise term, with the innovations independent across time and independent of the deterministic seasonal and trend components. Then the client-wise Bayesian loss admits the exact decomposition:

$$L^{(k)}_{\mathrm{Bayes}}(H,S) = L^{(k)}_{\mathrm{AR}}(S) + L^{(k)}_{\mathrm{seas}}(H) + L^{(k)}_{\mathrm{trend}}(H) \quad (14)$$

where each term is the contribution of the corresponding SDG component to the conditional mean-squared error under a horizon-H Bayesian predictor:

$$L^{(k)}_{\mathrm{AR}}(S) := \mathbb{E}\big[\| Y^{(S)}_{\mathrm{AR},k} - \mathbb{E}[Y^{(S)}_{\mathrm{AR},k} \mid X^{(H)}_k] \|_2^2\big], \quad (15)$$

$$L^{(k)}_{\mathrm{seas}}(H) := \mathbb{E}\big[\| Y^{(S)}_{\mathrm{seas},k} - \mathbb{E}[Y^{(S)}_{\mathrm{seas},k} \mid X^{(H)}_k] \|_2^2\big], \quad (16)$$

$$L^{(k)}_{\mathrm{trend}}(H) := \mathbb{E}\big[\| Y^{(S)}_{\mathrm{trend},k} - \mathbb{E}[Y^{(S)}_{\mathrm{trend},k} \mid X^{(H)}_k] \|_2^2\big]. \quad (17)$$

Here $X^{(H)}_k$ and $Y^{(S)}_k$ denote the input window and S-step forecast block for client k, and $Y^{(S)}_{\mathrm{AR},k}$, $Y^{(S)}_{\mathrm{seas},k}$, $Y^{(S)}_{\mathrm{trend},k}$ are the corresponding SDG components of the future block $Y^{(S)}_k$. Please refer to the Extended Version for the component-wise characterization and bounds.

Remark 1. *For each client k, the Bayesian loss $L^{(k)}_{\mathrm{Bayes}}(H,S)$ decreases with the look-back horizon H as longer histories improve identifiability of seasonal structure and (where present) trend components. The loss increases in the forecast horizon S, reflecting the accumulation of autoregressive innovations. Once the dominant seasonal cycles and the client's effective AR memory are covered, further increasing H yields only negligible improvement: the Bayesian loss has reached its horizon-dependent saturation level.*

The irreducible uncertainty perceived by the server is the weighted combination of these client-level Bayesian losses.

Lemma 1 (Server-level Bayesian Loss Aggregation)**.** *Let $\pi = (\pi_1, \ldots, \pi_K)$ be any probability distribution over the K clients. The population-level Bayesian loss is*

$$L^{(\mathrm{server})}_{\mathrm{Bayes}}(H,S) = \sum_{k=1}^{K} \pi_k \, L^{(k)}_{\mathrm{Bayes}}(H,S), \quad (18)$$

a quantity determined by the client data-generating processes (U_k, V_k) and independent of any global predictor. It aggregates the client-wise irreducible components (autoregressive variation, seasonal residuals, and optional trend terms), each of which exhibits a distinct dependence on the horizon H according to the client's temporal dynamics.

Approximation Loss

We now analyze the approximation loss in the federated setting, where a global model m is trained on a central server using client-local updates. This loss arises from the discrepancy between the global model and the Bayes-optimal predictor on each client.

Theorem 3 (Client-wise Approximation Loss)**.** *For client k, let $m^*_k(X)$ be the Bayesian predictor and m be any learned predictor. The approximation loss at horizon (H,S) is*

$$L^{(k)}_{\mathrm{approx}}(H,S;m) := \mathbb{E}\left[\, \| m(X) - m^*_k(X) \|_2^2 \,\right]. \quad (19)$$

Assume the Bayesian predictor m^*_k is twice differentiable on the intrinsic representation space with bounded curvature, and that m is a piecewise-affine model defined on the intrinsic manifold of dimension $d_{I,k}(H)$. Let D_k denote the number of training windows on client k.

Then the approximation loss admits the intrinsic-dimension–dependent bound

$$L^{(k)}_{\mathrm{approx}}(H,S;m) \lesssim \Big(K_2^2\, d_{I,k}(H)^2\Big)^{\frac{d_{I,k}(H)}{4+d_{I,k}(H)}} + \Big(\frac{d_{I,k}(H)\, H}{D_k}\Big)^{\frac{4}{4+d_{I,k}(H)}} \quad (20)$$

where K_2 is a curvature constant depending only on m^*_k. The first term reflects the geometric complexity of the intrinsic manifold, and the second term quantifies finite-sample limitations due to the effective sample size D_k/H. Full technical derivation is provided in the Extended Version.

The client-wise approximation losses aggregate to form the global loss on the server:

Lemma 2 (Server-level Approximation Loss Aggregation)**.** *Let $\pi = (\pi_1, \ldots, \pi_K)$ be the client-sampling distribution used by the server, with $\pi_k \geq 0$ and $\sum_k \pi_k = 1$. The global approximation loss under the server-side predictor m is the weighted aggregation of the client-wise approximation losses:*

$$L^{(\mathrm{server})}_{\mathrm{approx}}(H,S;m) = \sum_{k=1}^{K} \pi_k \, L^{(k)}_{\mathrm{approx}}(H,S;m). \quad (21)$$

Remark 2. *Because the intrinsic dimension $d_{I,k}(H)$ typically increases with the look-back horizon H, and the number of effectively independent samples scales as D_k/H, both the curvature-driven bias term and the finite-sample variance term in $L^{(k)}_{\mathrm{approx}}(H,S;m)$ grow with H. Consequently, each client exhibits a horizon beyond which approximation error begins to dominate. The server-level approximation loss inherits this behavior via the mixture weights π, reflecting how rising intrinsic complexity and diminishing effective sample size jointly amplify the approximation error.*

In this section, we decompose the forecasting error into its fundamental components and characterize how each behaves as a function of the look-back horizon H and forecasting span S. We begin by expressing the population prediction error as the sum of (i) the Bayes loss, which reflects irreducible uncertainty in the SDG, and (ii) the approximation loss, which arises from learning a nonlinear predictor from finite data. This yields the client-wise decomposition with the global loss obtained by aggregation across clients.

The combined loss, therefore, exhibits a fundamental tradeoff: the Bayesian loss decreases and eventually plateaus, while the approximation loss increases with H. Their interaction induces a unimodal structure in the total loss $L^{(k)}(H,S)$ and yields a client-specific optimal look-back horizon that balances signal coverage with statistical efficiency. This analysis forms the theoretical basis for the optimal horizon selection framework developed in the next section.

Optimal Horizon H

This section formalizes the choice of the look-back horizon H as an explicit optimization over the total loss

$$L(H, S; m) = L_{\text{Bayes}}(H, S) + L_{\text{approx}}(H, S; m) \quad (22)$$

where L_{Bayes} and L_{approx} are defined through the intrinsic-space formulation and federated loss decomposition in the previous sections. Since the trained model implicitly depends on the horizon, we slightly abuse notation and write $L(H, S)$ and $L_{\text{approx}}(H, S)$ for $L(H, S; m)$ and $L_{\text{approx}}(H, S; m)$, suppressing the dependence on m to avoid clutter.

We work client-wise (suppressing the client index when clear) and treat $H \in \mathbb{N}$, using forward differences

$$\Delta f(H) := f(H{+}1) - f(H) \quad (23)$$

to study how each loss component changes when one additional time step is added to the look-back window. Intuitively, the Bayesian loss should decrease as more history becomes available, while the approximation loss should increase due to higher intrinsic dimensionality and lower effective sample size. The remainder of this section quantifies this trade-off and characterizes the minimizer H^*.

Bayesian–Approximation Loss Dynamics

We study how the total client-wise prediction loss

$$L^{(k)}(H) = L^{(k)}_{\text{Bayes}}(H) + L^{(k)}_{\text{approx}}(H; m) \quad (24)$$

varies with the look-back horizon H. We use the discrete forward difference $\Delta f(H) := f(H{+}1) - f(H)$ to analyze whether adding one additional step of history decreases or increases the loss.

Bayesian Loss Behavior Under the SDG generative model (Eq. (3)), the irreducible loss decomposes into AR, seasonal, and (optionally zero) trend components. As H increases, $\Delta L^{(k)}_{\text{Bayes}}(H) \le 0$, and $\Delta L^{(k)}_{\text{Bayes}}(H) \to 0$.

Proof sketch. If $H_2 > H_1$, then $\sigma(X_{1:H_1}) \subseteq \sigma(X_{1:H_2})$, and thus $\text{Var}(Y \mid X_{1:H_2}) \le \text{Var}(Y \mid X_{1:H_1})$ almost surely; taking expectations yields $L^{(k)}_{\text{Bayes}}(H_2) \le L^{(k)}_{\text{Bayes}}(H_1)$. Under the SDG, the future depends on finite AR memory $\ell_{\text{AR},k}$, finite seasonal periods $T_{f,j,k}$, and a linear trend; therefore, the conditional expectation becomes invariant once $H \ge H^*_k$, implying $\Delta L^{(k)}_{\text{Bayes}}(H) \to 0$. The monotonic decrease, therefore, follows from the conditional-variance identity, and the eventual plateau follows from the finite dependency structure of the SDG. Thus, the Bayesian loss enters a plateau where additional history yields negligible improvement.

Approximation Loss Behavior The learned model must approximate the intrinsic mapping from past to future. From the dimension-dependent bound (Theorem 3),

$$L^{(k)}_{\text{approx}}(H; m) \lesssim \left(K_2^2\, d_{I,k}(H)^2\right)^{\frac{d_{I,k}(H)}{4+d_{I,k}(H)}} + \left(\frac{d_{I,k}(H)\, H}{D_k}\right)^{\frac{4}{4+d_{I,k}(H)}}, \quad (25)$$

where $d_{I,k}(H)$ denotes the intrinsic dimension of the input window and D_k is the number of overlapping training windows on client k.

Proof sketch. Both terms in (25) worsen as H grows. First, $d_{I,k}(H)$ is nondecreasing by construction, so the model must approximate a higher-dimensional function class; the dimension-dependent term $\left(K_2^2 d_{I,k}(H)^2\right)^{d_{I,k}(H)/(4+d_{I,k}(H))}$ therefore increases with H. Second, due to window overlap, the number of effectively independent samples scales as D_k/H; hence, the statistical term $\left(d_{I,k}(H)H/D_k\right)^{4/(4+d_{I,k}(H))}$ also increases with H since the numerator grows and the effective sample size shrinks. Thus, for sufficiently large H, $\Delta L^{(k)}_{\text{approx}}(H; m) > 0$, showing that the approximation cost eventually worsens once the horizon is long enough.

Hence, the approximation loss exhibits the opposite trend of the Bayesian loss: increasing window size ultimately leads to higher approximation error.

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.

Formaly, for any tolerance $\delta > 0$, define the smallest sufficient horizon as

$$H^*_k(\delta) := \min\{H : |\Delta L^{(k)}_{\text{Bayes}}(H)| \le \delta\}, \quad (26)$$

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.

Theorem 4 (Unimodality and Optimal Horizon)**.** *If for a given $\delta > 0$ the Bayesian loss satisfies $\Delta L^{(k)}_{\text{Bayes}}(H) \le -\delta$ for all $H < H^*_k(\delta)$, and the approximation loss satisfies $\Delta L^{(k)}_{\text{approx}}(H; m) \ge \delta$ 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

$$\Delta L_{\text{Bayes}}(H, S) < 0 \quad (H < H_0), \quad (27)$$

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

$$\Delta L_{\text{Bayes}}(H, S) \ge -\delta \quad (H \ge H_0). \quad (28)$$

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

$$\Delta L_{\text{approx}}(H, S) \ge \eta \quad (H \ge H_0). \quad (29)$$

Fix any $\delta \in (0, \eta)$ and define $H^*(\delta)$ as the smallest $H \geq H_0$ with $\Delta L_{\text{Bayes}}(H, S) \geq -\delta$. Then for $H < H^*(\delta)$, we have $\Delta L_{\text{Bayes}}(H, S) < -\delta$ and $\Delta L_{\text{approx}}(H, S) \geq 0$, so $\Delta L(H, S) = \Delta L_{\text{Bayes}}(H, S) + \Delta L_{\text{approx}}(H, S) < -\delta < 0$, and $L(H, S)$ is strictly decreasing. For $H \geq H^*(\delta)$, we have $\Delta L_{\text{Bayes}}(H, S) \geq -\delta$ and $\Delta L_{\text{approx}}(H, S) \geq \eta$, hence $\Delta L(H, S) \geq -\delta + \eta > 0$, 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.

□

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.

Seasonal Coverage and Horizon Selection The tolerance δ can be linked to an interpretable signal structure via seasonal coverage. Let $A_k^2 = \sum_{f,j} A_{f,j,k}^2$ denote the total seasonal energy of client k, and define the τ-coverage horizon $T_k^{(\tau)}$ as the smallest H for which the unresolved seasonal energy beyond H obeys:

$$\sum_{f,j:T_{f,j,k}>H} A_{f,j,k}^2 \;\leq\; (1-\tau)\, A_k^2, \tag{30}$$

assuming the residual seasonal loss satisfies $L_{\text{seas}}^{(k)}(H) \leq A_k^2 r(H, T)$ with $r(\cdot, T)$ decreasing in H.

Corollary 1 (Coverage–Tolerance Mapping)**.** *If a coverage level τ is chosen so that $(1-\tau)A_k^2 \leq \delta$, then every $H \geq T_k^{(\tau)}$ satisfies $|\Delta L_{\text{Bayes}}^{(k)}(H)| \leq \delta$. Thus, the optimal horizon is given by*

$$H_k^*(\delta) = \max\{\ell_{\text{AR},k},\, T_k^{(\tau)}\}. \tag{31}$$

This provides a direct way to set the horizon using interpretable signal parameters: choose a desired seasonal coverage τ, infer the corresponding δ, and compute $H_k^*(\delta)$.

Federated Horizon Aggregation In federated learning, the server must choose a single global horizon H_{server} despite heterogeneous client optima $\{H_k^*(\delta)\}$. Because extreme clients (e.g., very large horizons) can substantially reduce effective sample sizes for all participants, robustness is crucial. Let $w_k \propto n_k$ be data-proportional weights normalized so that $\sum_k w_k = 1$.

Robust Federated Horizon The global horizon can be defined via the weighted trimmed mean:

$$H_{\text{server}}^* = \text{TrimMean}_\alpha\left(\{H_k^*(\delta)\}_{k=1}^K; \{w_k\}_{k=1}^K\right) \tag{32}$$

which discards an α-fraction of the smallest and largest client-specific horizons (by weight) and averages the remainder.

This estimator is equivalent to minimizing a convex Huber-type aggregation objective and yields a horizon that balances most clients while avoiding inflation by a small number of extreme ones.

Discussion and Conclusion

Limitation and Discussion

This work proposes a principled framework for federated time-series forecasting under non-IID conditions, grounded in a structured synthetic data generator (SDG) and an intrinsic space formulation. The framework enables a precise decomposition of forecasting error and leads to a provably optimal look-back horizon. To make the analysis tractable and the theoretical guarantees possible, several assumptions are made that define the scope of applicability.

The SDG models additive components, e.g., trend, autoregressive memory, and seasonality, with Gaussian innovations. While this structure captures core dynamics observed in real-world data, it does not account for regime switches, nonlinear seasonal patterns, or cross-feature interactions beyond what is implicitly represented through PCA. The analysis assumes local stationarity and a stable autoregressive structure, which may be challenged in long-memory or near-unit-root settings. Additionally, estimating global covariance in a federated context requires secure or privacy-aware aggregation, and our sample efficiency analysis treats overlapping windows as approximately independent, which may overstate the effective sample size under certain data regimes.

These assumptions are common in theoretical work and are intentionally chosen to isolate the role of horizon length and data heterogeneity. Importantly, they enable the first provable characterization of optimal look-back windows in federated forecasting, providing a foundation for future extensions that relax these constraints.

Conclusion

This paper introduces a principled framework for horizon selection in federated time series forecasting under non-IID conditions, grounded in a synthetic data generator (SDG) that models key temporal structures, trend, autoregressive memory, and seasonality, along with client-specific heterogeneity. By embedding time-series windows into a geometry-preserving intrinsic space, we enable a precise decomposition of forecasting loss into irreducible Bayesian error and model-dependent approximation error, each tied to the underlying statistical structure and data distribution. Our analysis reveals a fundamental trade-off: while the Bayesian loss decreases with horizon length as more temporal structure becomes identifiable, the approximation loss increases due to growing intrinsic dimension and reduced sample efficiency. This yields a provable result that the total loss is minimized at the smallest sufficient horizon H^*, where additional history no longer improves identifiability but exacerbates overfitting. Furthermore, we propose a robust aggregation strategy to identify a global horizon across clients. Together, these contributions establish the first theoretically grounded criterion for adaptive horizon selection in federated settings, offering practical guidance for model design, deployment, and benchmarking in decentralized, heterogeneous environments.

Acknowledgments

This work is partly supported by the Australian Research Council Linkage Project (Grant No. LP220200893).

References

Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723.

Bahri, Y.; Dyer, E.; Kaplan, J.; Lee, J.; and Sharma, U. 2024. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27): e2311878121.

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

Edwards, T. D. P.; Alvey, J.; Alsing, J.; Nguyen, N. H.; and Wandelt, B. D. 2024. Scaling-laws for Large Time-Series Models. *arXiv preprint arXiv:2405.13867*.

Futoma, J.; Hariharan, S.; and Heller, K. 2017. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In *International conference on machine learning*, 1174–1182. PMLR.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210.

Kim, J.; Kim, H.; Kim, H.; Lee, D.; and Yoon, S. 2025. A Comprehensive Survey of Deep Learning for Time Series Forecasting: Architectural Diversity and Open Challenges. *Artificial Intelligence Review*, 58(7).

Kong, W.; Dong, Z. Y.; Jia, Y.; Hill, D. J.; Xu, Y.; and Zhang, Y. 2019. Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid*, 10(1): 841–851.

Koparanov, K. A.; Georgiev, K. K.; and Shterev, V. A. 2020. 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)*, 61–64. Sofia, Bulgaria: IEEE.

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 95–104. New York, NY, USA: ACM.

Levi, N.; and Oz, Y. 2024. The Underlying Scaling Laws and Universal Statistical Structure of Complex Datasets. *arXiv preprint arXiv:2403.09756*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Lim, B.; Arık, S. Ö.; Loeff, N.; and Pfister, T. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4): 1748–1764.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers are Effective for Time Series Forecasting. *arXiv preprint arXiv:2402.08372*.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Perifanis, V.; Pavlidis, N.; Koutsiamanis, R.-A.; and Efraimidis, P. S. 2023. Federated Learning for 5G Base Station Traffic Forecasting. *Computer Networks*, 235(109950).

Sharma, U.; and Kaplan, J. 2020. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*.

Shi, J.; Ma, Q.; Ma, H.; and Li, L. 2024. Scaling Law for Time Series Forecasting. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 83314–83344. Curran Associates, Inc.

Takens, F.; Young, L.-S.; and Rand, D. 2006. Detecting Strange Attractors in Turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, 366–381. Berlin, Heidelberg: Springer Berlin Heidelberg.

Toner, W.; and Darlow, L. 2024. An Analysis of Linear Time Series Forecasting Models. *arXiv preprint arXiv:2403.14587*.

Woo, G.; Liu, C.; Sahoo, D.; Kumar, A.; and Hoi, S. 2023. Learning deep time-index models for time series forecasting. In *International Conference on Machine Learning*, 37217–37237. PMLR.

Wu, C.; Wang, H.; Zhang, X.; Fang, Z.; and Bu, J. 2024. Spatio-Temporal Heterogeneous Federated Learning for Time Series Classification with Multi-View Orthogonal Training. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, 2613–2622. New York, NY, USA: ACM.

Xu, Z.; Zeng, A.; and Xu, Q. 2024. FiTS: Modeling Time Series with 10K Parameters. *arXiv preprint arXiv:2307.03756*.

Zador, P. L. 1982. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2): 139–149.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 11121–11128.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.

Zivot, E.; and Wang, J. 2006. *Modeling Financial Time Series with S-Plus*. New York, NY: Springer, 2nd edition.

Appendix

Appendix A: SDG Experimental Validation

We validate the SDG by decomposing a real series into interpretable components (linear trend, sinusoidal seasonality, and AR(p) residuals), then re-synthesizing a signal with the fitted pieces and comparing it back to the original across marginal, temporal, and spectral views.

Dataset The temperature time series used in this study was extracted from the Max Planck Institute for Biogeochemistry's weather portal (https://www.bgc-jena.mpg.de/wetter/), which provides 10-minute resolution meteorological data recorded in Jena, Germany, throughout the year 2020. The full dataset includes 21 weather indicators (e.g., air temperature, humidity, wind speed), but only the air temperature series was used in this experiment. A temperature time series at 10-minute resolution ($N = 52{,}696$) is parsed as y_t, $t = 1, \ldots, N$.

Hyperparameters Trend is estimated via OLS ($y_t \approx \beta t + c$); dominant periods are selected from the periodogram (daily peak $T_1 = 144$ plus an ultra–low-frequency component tied to record length); detrended residuals are fit with AR models, with order chosen by AIC/BIC (adopt $p = 30$) and Gaussian innovation scale σ taken from the residual standard deviation.

Metrics (1) Mean/variance gaps to check marginal alignment; (2) ACF L^2 gap up to 30 lags for short-range dependence; (3) normalized PSD L^2 gap for spectral structure; (4) two-sample KS statistic for distributional shape; (5) window-level (50-step) real-vs-synthetic discrimination accuracy of a random-forest classifier to probe downstream indistinguishability.

Result The SDG reproduces the original series closely: $\Delta\mu \approx 3.6 \times 10^{-3}$, ACF $L^2 \approx 3.7 \times 10^{-6}$, normalized PSD $L^2 \approx 8.2 \times 10^{-3}$, KS $= 0.042$ (large-N significance), and a RF accuracy of 0.892 on 50-step windows, indicating strong agreement in dependence/spectral structure with minor residual cues useful for further tuning.

Appendix B: Intrinsic Space Construction

In this section, we provide a step-by-step transformation to convert a non-IID time series data of length L, which can be well described by the SDG, into a vector z in an intrinsic space $M(L)$.

Let $O(H) \subset \mathbb{R}^{FH}$ be the space of normalized windows where H is the window length for one sample of a time series data. We construct $\Phi_H : O(H) \to M(H)$ satisfying the following assumptions:

Assumption 1 (Compact image with uniform radius control)**.** *For each horizon $H \in \mathbb{N}$ there exists a finite radius R_H such that*

$$\sup_{x \in O(H)} \|\Phi_H(x)\|_2 \;\le\; R_H, \tag{33}$$

so that $M(H) := \Phi_H(O(H)) \subseteq \overline{B}_{R_H}(0) \subset \mathbb{R}^{d_I(H)}$. $\overline{B}_{R_H}(0)$ is the closed Euclidean ball of the radius R_H:

$$\overline{B}_{R_H}(0) := \left\{ y \in \mathbb{R}^{d_I(H)} : \|y\|_2 \le R_H \right\} \tag{34}$$

Remark 3. *This keeps all intrinsic representations within a controlled region. It ensures basic quantities (norms, averages, covariances) are well-behaved and comparable across horizons.*

Assumption 2 (Bi-Lipschitz embedding with stable inverse)**.** *For each horizon H there exist constants $\alpha_H > 0$ and $\beta_H < \infty$ such that for all $x, y \in O(H)$*

$$\alpha_H \|x - y\|_2 \;\le\; \|\Phi_H(x) - \Phi_H(y)\|_2 \;\le\; \beta_H \|x - y\|_2. \tag{35}$$

Consequently, Φ_H is injective and continuous, and its inverse on $M(H)$ is Lipschitz with constant at most $1/\alpha_H$. Define the distortion $\kappa_H := \beta_H / \alpha_H$.

Remark 4. *This preserves distances up to a fixed factor when moving between window space and intrinsic space. It lets us transfer geometric and statistical statements between the two spaces with controlled distortion.*

Assumption 3 (Intrinsic dimension: regularity, monotonicity, and saturation)**.** *For each horizon $H \in \mathbb{N}$, the image $M(H) := \Phi_H(O(H))$ is assumed to be a connected, continuously differentiable (C^1) submanifold of $\mathbb{R}^{d_I(H)}$ with intrinsic dimension equal to the embedding dimension, i.e., $\dim M(H) = d_I(H)$. Moreover, the intrinsic dimension is non-decreasing with horizon, satisfying $d_I(H+1) \ge d_I(H)$ for all H. Finally, there exists a finite* saturation horizon H_{id} *such that for all $H \ge H_{\mathrm{id}}$, the intrinsic dimension stabilizes: $d_I(H) = d_I(H_{\mathrm{id}})$.*

Remark 5. *This assumption encodes a structural view of how complexity in the representation evolves with increasing horizon: initially, as H grows, the map Φ_H incorporates more information and thus the intrinsic dimensionality of its image increases. Once a sufficient horizon is reached, where it captures all essential long-term structure, the complexity no longer grows. The saturation point H_{id} marks the minimal horizon beyond which the representation is informationally complete.*

Assumption 4 (Inter-horizon compatibility via a stable linear map)**.** *For any $H_1 \le H_2$, there exists a linear map $P[H_2, H_1] : \mathbb{R}^{d_I(H_2)} \to \mathbb{R}^{d_I(H_1)}$ that preserves distances on $M(H_2)$ up to a uniform distortion factor. Specifically, there exists a constant $C_{\mathrm{iso}} \ge 1$, independent of H_1 and H_2, such that for all $z_1, z_2 \in M(H_2)$,*

$$\begin{aligned} C_{\mathrm{iso}}^{-1} \|z_1 - z_2\|_2 \;&\le\; \|P[H_2, H_1](z_1) - P[H_2, H_1](z_2)\|_2 \\ &\le\; C_{\mathrm{iso}} \|z_1 - z_2\|_2. \end{aligned} \tag{36}$$

Remark 6. *This gives a simple, stable way to relate representations at different horizons. It supports comparing statistics and constructions across H using a single linear operator.*

Assumption 5 (Truncation behaves like projection up to a small error)**.** *Let $t_p[H_2, H_1] : O(H_2) \to O(H_1)$ denote the truncation map that discards the last $H_2 - H_1$ time steps. Assume that the representation map $\Phi_{H_2} : O(H_2) \to M(H_2)$*

is a bijection onto its image. Then, define the induced intrinsic truncation map:

$$\mathcal{T}_{H_2 \to H_1} := \Phi_{H_1} \circ t_p[H_2, H_1] \circ \Phi_{H_2}^{-1} : M(H_2) \to \mathbb{R}^{d_I(H_1)} \tag{37}$$

There exist constants $c_{\text{err}} > 0$ and $\gamma \in [1, 2]$ such that for all $z \in M(H_2)$,

$$\|\mathcal{T}_{H_2 \to H_1}(z) - P[H_2, H_1]\, z\|_2 \;\le\; c_{\text{err}}\, H_1^{-\gamma} \tag{38}$$

where $P[H_2, H_1] : \mathbb{R}^{d_I(H_2)} \to \mathbb{R}^{d_I(H_1)}$ is the stable linear projection map from Assumption 4.

Remark 7. *This assumption states that truncating the input sequence and then applying the lower-horizon representation map is approximately equivalent to linearly projecting the higher-horizon representation. The approximation error decays with horizon length, reflecting the idea that the truncation becomes less destructive as H_1 grows.*

Assumption 6 (Power-law decay of intrinsic covariance eigenvalues)**.** *Let Σ_H be the intrinsic covariance on $M(H)$ with eigenvalues $\lambda_1(H) \ge \lambda_2(H) \ge \cdots$. There exist constants $\alpha_Z \in (1, 2]$ and $C_Z > 0$ such that for all $i \ge 1$ and all H*

$$\lambda_i(H) \;\le\; C_Z\, i^{-\alpha_Z}. \tag{39}$$

Equivalently, the tail energy beyond the top d components decays on the order of $d^{\,1-\alpha_Z}$, uniformly in H.

Remark 8. *This says most variance sits in a few leading directions, making low-dimensional summaries effective. It supports later bounds that depend on how quickly energy concentrates in the intrinsic representation.*

Now we propose the step-by-step transformation converting a non-IID time series data described by the SDG into an intrinsic space:

1. **Client-wise Normalization:** To reduce the impact of feature skew caused by the affine skew, we apply the client-wise local normalization, allowing data from different clients to be aligned into a common representation space. For each client k and feature f, compute:

$$\mu_{f,k} = \frac{1}{H} \sum_{i=1}^{H} x_{f,i,k}, \tag{40}$$

$$\sigma_{f,k} = \sqrt{\frac{1}{H} \sum_{i=1}^{H} (x_{f,i,k} - \mu_{f,k})^2} \tag{41}$$

and normalize:

$$\hat{x}_{f,i,k} = \frac{x_{f,i,k} - \mu_{f,k}}{\sigma_{f,k}} \quad \text{for all } f, i \tag{42}$$

2. **Flatten Windows into Vectors:** To analyze the global structure, we flatten the windows into vectors, which is suitable for further PCA analysis. Each normalized window $\hat{X}_{f,t,k}^{(H)} \in \mathbb{R}^{F \cdot H}$ is flattened (e.g., row-wise) into:

$$\hat{x}_{t,k}^{\text{flat}} = \text{vec}(\hat{X}_{t,k}) \in \mathbb{R}^{F \cdot H} \tag{43}$$

3. **Global Covariance and Eigendecomposition:** To identify principal directions of variation in the normalized time series data, we compute the empirical covariance matrix of the flattened windows. Let $X \in \mathbb{R}^{N \times (F \cdot H)}$ be the matrix whose rows are the centered, flattened vectors $\hat{x}_{t,k}^{\text{flat}}$. We compute:

$$\Sigma = \frac{1}{N} X^\top X \tag{44}$$

and perform eigendecomposition:

$$\Sigma = U \Lambda U^\top \tag{45}$$

The matrix U contains the principal directions of variation. When the data lies near a low-dimensional linear (or locally linear) submanifold, this eigenspace approximates the intrinsic structure. The eigenvalues in Λ reflect the energy captured by each direction.

4. **Estimate Intrinsic Dimension $d_{I,k}(H)$:** The intrinsic dimension $d_I(H)$ reflects the number of effective degrees of freedom in a time series window of size H for F features. Since the SDG is composed of a small number of structured, decoupled components (including AR memory, seasonality, and trend), the observed data must lie near a low-dimensional manifold whose dimension is approximately the sum of degrees of freedom required to encode each component. Hence, we estimate the intrinsic dimension $d_I(H)$ based on the modeled components of the SDG:

AR memory. Let $\rho_k \in (0, 1)$ be the spectral radius of the AR companion matrix. Define the effective memory length

$$\ell_{\text{AR},k} = \tau_{e,k} = \left\lceil \frac{\ln(1/(1-\epsilon))}{-\ln \rho_k} \right\rceil,\ \epsilon \in (0, 1) \tag{46}$$

The e-folding time for the impulse response to decay to a fraction ϵ. (Empirically, the ACF $1/e$ crossing is a similar estimate.)

Seasonality (amplitude-weighted). Each sinusoid contributes two degrees of freedom once *one full cycle* is observed. Let

$$g_k(H) = 2 \sum_{j=1}^{J} w_{j,k}\, \min\Big(1, \frac{H}{T_{j,k}^*}\Big) \tag{47}$$

$$w_{j,k} = \frac{\sum_{f=1}^{F} A_{f,j,k}^2}{\sum_{f=1}^{F} \sum_{j=1}^{J} A_{f,j,k}^2} \tag{48}$$

where $T_{j,k}^*$ is the (feature-aggregated) period for component j on client k.

Trend and total intrinsic dimension. Add $+1$ for linear trend. The resulting intrinsic dimension is

$$d_{I,k}(H) \approx F \cdot \Big(\min\{H, \ell_{\text{AR},k}\} + g_k(H) + 1 \Big), \tag{49}$$

which is smooth, monotone in H, and saturates once $H \ge \ell_{\text{AR},k}$ and $H \ge \max\{T_{j,k}^* : w_{j,k} \text{ significant}\}$.

The intrinsic dimension $d_I(H)$ can also be estimated empirically to verify that the theoretical $d_I(H)$ agrees with the actual spectrum and provides a quick plug-in value:

$$\frac{\sum_{i=1}^{d_I} \lambda_i}{\sum_{j=1}^{F \cdot H} \lambda_j} \geq \eta \quad \text{(e.g., } \eta = 0.99\text{)} \tag{50}$$

This step provides a principled way to compute d_I from the generative process structure, ensuring that the low-dimensional representation retains core dynamics.

5. **Step 5: Project to Intrinsic Space** Let $U_{d_I} \in \mathbb{R}^{F \cdot H \times d_I}$ be the matrix of top eigenvectors. We project the flattened time series data into an intrinsic space:

$$\begin{aligned} z_{t,k} &= \Phi_H(x_{t,k}) \\ &= U_{d_I}^\top \cdot \hat{x}_{t,k}^{\text{flat}} \in \mathbb{R}^{d_I} \\ &= U_{d_I}^\top \cdot \text{vec}\left(\frac{X_{t,k} - \mu_k}{\sigma_k}\right) \end{aligned} \tag{51}$$

Now we have shown the detailed steps of transforming typical real-world non-IID time series data described by the SDG into an intrinsic space. In words, the intrinsic space formulation offers a unified geometric and statistical foundation for representing time series data in federated settings.

Appendix C: Loss Analysis

In this section, we include supplementary materials for the Section Loss Analysis, starting with the formal definition of the Bayesian loss and the Approximation loss.

Definition 1 (Bayesian (irreducible) loss)**.** *Fix a horizon pair $(H, S) \in \mathbb{N}^2$. Let (U, V) be square-integrable random elements with $U \in \mathcal{M}(H)$, $V \in \mathcal{M}(S)$, defined on a common probability space and induced by the data-generating process (e.g., the SDG).*

Since $V \in L^2$ and U is measurable, the conditional expectation $\mathbb{E}[V \mid U]$ exists in L^2 and admits a measurable version defined almost everywhere. We denote any such version by the Bayes predictor*:*

$$m^*(u) := \mathbb{E}[\, V \mid U = u \,], \quad \textit{a.e. } u. \tag{52}$$

The Bayesian (irreducible) loss *at (H, S) is*

$$L_{\text{Bayes}}(H, S) := \mathbb{E}\left[\, \| \, V - m^*(U) \, \|_2^2 \,\right]. \tag{53}$$

Equivalently, $L_{\text{Bayes}}(H, S)$ is the minimum achievable risk under squared loss over all measurable predictors $m : \mathcal{M}(H) \to \mathcal{M}(S)$:

$$L_{\text{Bayes}}(H, S) = \inf_m \mathbb{E}\left[\, \|V - m(U)\|_2^2 \,\right]. \tag{54}$$

For client k, define $L_{\text{Bayes}}^{(k)}(H, S)$ analogously with (U, V) replaced by (U_k, V_k).

Definition 2 (Approximation (excess) loss)**.** *Let $m : \mathcal{M}(H) \to \mathcal{M}(S)$ be any square-integrable predictor. The* approximation loss *of m at (H, S) is the expected squared distance to the Bayes predictor:*

$$L_{\text{approx}}(H, S; m) := \mathbb{E}\left[\, \| \, m(U) - m^*(U) \, \|_2^2 \,\right] \tag{55}$$

where $m^*(U) := \mathbb{E}[V \mid U]$ is the Bayes predictor from Definition 1. This quantity measures the excess risk of m over the Bayes-optimal predictor:

$$\mathbb{E}\big[\|V - m(U)\|_2^2\big] = L_{\text{Bayes}}(H, S) + L_{\text{approx}}(H, S; m) \tag{56}$$

For client k, the client-specific approximation loss is

$$L_{\text{approx}}^{(k)}(H, S; m) := \mathbb{E}\big[\, \| \, m(U_k) - m_k^*(U_k) \, \|_2^2 \,\big] \tag{57}$$

where $m_k^*(U_k) := \mathbb{E}[V_k \mid U_k]$ is the client-wise Bayes predictor.

The following proof will show that the total loss can be decomposed into the Bayesian loss and the approximation loss.

Proof of Total Loss Decomposition. Let X be the model input and Y be the optimal output. By the nature of time series data, with $Y \in L^2$, and let $m : \mathbb{R}^d \to \mathbb{R}^p$ be any (possibly learned) predictor such that $m(X) \in L^2$. Define the Bayes predictor as $m^*(X) := \mathbb{E}[Y \mid X]$, which is the *orthogonal projection* of Y onto the closed subspace of L^2 consisting of functions of X. In Hilbert space terms, this means $Y = m^*(X) + R$, where the residual $R := Y - m^*(X)$ satisfies:

$$\mathbb{E}[\langle R,\ f(X)\rangle] = 0 \quad \text{for all } f(X) \in L^2. \tag{58}$$

In particular, for any predictor $m(X)$, the difference $m^*(X) - m(X)$ is itself a function of X, so we have:

$$\mathbb{E}\left[\langle Y - m^*(X),\ m^*(X) - m(X)\rangle\right] = 0. \tag{59}$$

Now, using the identity:

$$\|Y - m(X)\|^2 = \|Y - m^*(X) + m^*(X) - m(X)\|^2. \tag{60}$$

We expand the squared norm:

$$\begin{aligned} \|Y - m(X)\|^2 = \|Y - m^*(X)\|^2 + \|m^*(X) - m(X)\|^2 \\ + 2\langle Y - m^*(X),\ m^*(X) - m(X)\rangle. \end{aligned} \tag{61}$$

Taking expectations on both sides and applying the orthogonality condition above, we obtain:

$$\begin{aligned} &\mathbb{E}[\|Y - m(X)\|^2] \\ &\quad = \mathbb{E}[\|Y - m^*(X)\|^2] + \mathbb{E}[\|m^*(X) - m(X)\|^2] \\ &\qquad + 2\,\mathbb{E}\left[\langle Y - m^*(X),\ m^*(X) - m(X)\rangle\right] \\ &\quad = \mathbb{E}[\|Y - m^*(X)\|^2] + \mathbb{E}[\|m^*(X) - m(X)\|^2]. \end{aligned} \tag{62}$$

Thus, the total predictive loss decomposes into:

$$L = L_{\text{Bayes}} + L_{\text{approx}}, \tag{63}$$

where

$$L_{\text{Bayes}} := \mathbb{E}[\|Y - m^*(X)\|^2], \tag{64}$$

$$L_{\text{approx}} := \mathbb{E}[\|m^*(X) - m(X)\|^2]. \tag{65}$$

□

The following proof will show the representation of the Bayesian loss under the setting of the SDG and the intrinsic space.

Proof of Bayesian Loss Decomposition (Theorem 2). Let $L^{(k)}_{\text{Bayes}}(H, S)$ denote the irreducible forecasting loss for client k when predicting S future steps from a look-back window of length H. Under the SDG model (Equation 3), the time series consists of three additive and statistically independent components:

1. An autoregressive (AR) component with order p and client-specific coefficients $\phi_{k,i}$,
2. A sum of deterministic sinusoidal (seasonal) signals,
3. A deterministic linear trend (possibly optional).

Because these components are independent, their contribution to the total irreducible error (i.e., Bayes loss) is additive:

$$L^{(k)}_{\text{Bayes}}(H, S) = L^{(k)}_{\text{AR,Bayes}}(S) + L^{(k)}_{\text{seas,Bayes}}(H) + L^{(k)}_{\text{trend}}. \tag{66}$$

We analyze each term individually:

(1) Autoregressive (AR) Component. Let $\Phi_k(z) = 1 - \sum_{i=1}^{p} \phi_{k,i} z^i$ denote the AR characteristic polynomial, and let $\psi_{k,s}$ be the impulse response of the equivalent MA(∞) process resulting from AR inversion:

$$\psi_{k,0} = 1, \quad \psi_{k,s} = \sum_{i=1}^{p} \phi_{k,i} \psi_{k,s-i} \quad \text{for } s > 0. \tag{67}$$

Once the AR coefficients $\phi_{k,i}$ are identified (i.e., $H \geq p$), the residual uncertainty in future values arises purely from S future innovations, each filtered through the MA coefficients. Thus, the AR contribution to the Bayes loss is:

$$L^{(k)}_{\text{AR,Bayes}}(S) = \sum_{f=1}^{F} \sigma^2_{f,k} \sum_{s=0}^{S-1} \psi^2_{k,s}, \tag{68}$$

where $\sigma^2_{f,k}$ is the innovation variance for feature f on client k.

This term grows with S, but remains independent of H once the AR parameters are recovered. Using geometric decay of the impulse response ($\psi_{k,s} \lesssim \rho_k^s$), we obtain the bound:

$$L^{(k)}_{\text{AR,Bayes}}(S) \leq \sum_{f=1}^{F} \sigma^2_{f,k} \cdot \frac{1 - \rho_k^{2S}}{1 - \rho_k^2}, \tag{69}$$

where ρ_k is the spectral radius of the AR companion matrix.

(2) Seasonal Component. Each seasonal term is modeled as a sinusoid:

$$A_{f,j,k} \cdot \sin\left(\frac{2\pi t}{T_{f,j,k}} + \theta_{f,j,k}\right). \tag{70}$$

With infinite context, these deterministic components can be predicted perfectly. However, with a finite horizon H, there is residual uncertainty due to phase and frequency aliasing (i.e., spectral leakage). The residual energy for each sinusoid decays as the window covers more of its period:

$$L^{(k)}_{\text{seas,Bayes}}(H) \leq \sum_{f=1}^{F} \sum_{j=1}^{J} A^2_{f,j,k} \cdot r(H, T_{f,j,k}), \tag{71}$$

where the decay function satisfies $r(H, T) \leq c \cdot \min(1, (T/H)^\gamma)$ for some $\gamma \in [1, 2]$, depending on the estimator's spectral resolution. As H increases, this term decreases monotonically and saturates when $H \geq T_{f,j,k}$ for all dominant components.

(3) Trend Component. The linear trend $\beta_{f,k} t$ is fully deterministic and perfectly predictable when modeled. Thus, if the trend is included explicitly in the predictor class, it contributes no irreducible loss:

$$L^{(k)}_{\text{trend}} = 0. \tag{72}$$

Otherwise, if left unmodeled, it contributes a fixed additive term proportional to the variance of the time index t over the prediction window.

Conclusion. Summing the three components yields the upper bound:

$$\begin{aligned} L^{(k)}_{\text{Bayes}}(H, S) \leq & \sum_{f=1}^{F} \sigma^2_{f,k} \cdot \frac{1 - \rho_k^{2S}}{1 - \rho_k^2} \\ & + \sum_{f=1}^{F} \sum_{j=1}^{J} A^2_{f,j,k} \cdot r(H, T_{f,j,k}) \\ & + L^{(k)}_{\text{trend}}. \end{aligned} \tag{73}$$

This confirms the decomposition in Theorem 2, where each term reflects a distinct structural contribution from the SDG.

□

The following proof demonstrates the decomposition of the approximation loss into the curvature and variance components in the federated scenario.

Proof of Approximation Loss Decomposition and Bound. Fix a client k. Recall that the approximation loss is

$$L^{(k)}_{\text{approx}}(H, S; m) := \mathbb{E}\big[\| m_k^\star(U_k) - m(U_k)\|_2^2\big] \tag{74}$$

where $m_k^\star := \mathbb{E}[V_k \mid U_k]$ is the client-wise Bayes predictor. Let $\mathcal{H}$ be the hypothesis class in which the learned model m lies.

1. Decomposition via the best in-class predictor. Define the in-class oracle

$$\tilde{m}_k \in \arg\min_{h \in \mathcal{H}} \mathbb{E}\big[\| m_k^\star(U_k) - h(U_k)\|_2^2\big]. \tag{75}$$

This is the (unique up to null sets) metric projection of the Bayes rule onto the closed convex set $\{h(U_k) : h \in \mathcal{H}\} \subset L^2$. By the Hilbert-space projection theorem, the residual $m_k^\star(U_k) - \tilde{m}_k(U_k)$ is orthogonal to the error $\tilde{m}_k(U_k) - m(U_k)$. Thus,

$$\mathbb{E}[\langle\, m_k^\star(U_k) - \tilde{m}_k(U_k),\ \tilde{m}_k(U_k) - m(U_k)\rangle] = 0, \quad (76)$$

without requiring any unbiasedness assumptions. Expanding the square and using this orthogonality yields the exact decomposition

$$\begin{aligned} L_{\text{approx}}^{(k)}(H, S; m) = &\underbrace{\mathbb{E}\big[\| m_k^\star(U_k) - \tilde{m}_k(U_k)\|_2^2\big]}_{\text{approximation (curvature) error}} \\ &+ \underbrace{\mathbb{E}\big[\| \tilde{m}_k(U_k) - m(U_k)\|_2^2\big]}_{\text{estimation (variance) error}}. \end{aligned} \quad (77)$$

2. Bounding the curvature term. Assume the Bayes rule is twice differentiable on the intrinsic manifold $\mathcal{M}(H)$, with uniformly bounded Hessian:

$$\|\nabla^2 m_k^\star(u)\|_{\text{op}} \le K_{2,k}, \quad \forall\, u \in \mathcal{M}(H). \quad (78)$$

Let $K_2 := \max_k K_{2,k}$. Partition the intrinsic space into N_k cells of diameter $r_k \asymp N_k^{-1/d_{I,k}(H)}$. On each cell, the second-order Taylor remainder gives

$$\|m_k^\star(u) - m_k^\star(u_0)\| \lesssim K_2\, r_k^2, \quad (79)$$

hence the best piecewise-constant or piecewise-affine approximation over cell diameter r_k incurs squared error $\lesssim K_2^2 r_k^4$. Integrating over the manifold yields the curvature bound

$$L_{\text{curv}}^{(k)}(H) \le C_{\text{curv}}\, K_2^2\, d_{I,k}(H)^2\, N_k^{-4/d_{I,k}(H)}, \quad (80)$$

for some constant $C_{\text{curv}} > 0$.

3. Bounding the sampling variance term. Client k has D_k samples. Because overlapping windows reduce the number of statistically independent windows by a factor of H, the effective sample size is $\asymp D_k/H$. If the N_k regions receive (on average) $\bar{n}_k \approx D_k/(HN_k)$ samples, standard nonparametric concentration arguments give

$$L_{\text{var}}^{(k)}(H, S) \le C_{\text{sample}}\, \frac{d_{I,k}(H)\, H\, N_k}{D_k}, \quad (81)$$

for a constant $C_{\text{sample}} > 0$. (This term depends on S only through the output dimension.)

4. Optimizing over the resolution N_k. The total approximation loss satisfies

$$\begin{aligned} L_{\text{approx}}^{(k)}(H, S; m) \lesssim\ & K_2^2 d_{I,k}(H)^2 N_k^{-4/d_{I,k}(H)} \\ &+ \frac{d_{I,k}(H)\, H\, N_k}{D_k}. \end{aligned} \quad (82)$$

Balancing the two terms in N_k yields

$$\begin{aligned} L_{\text{approx}}^{(k)}(H, S; m) \lesssim\ & \Big(K_2^2 d_{I,k}(H)^2\Big)^{\frac{d_{I,k}(H)}{4+d_{I,k}(H)}} \\ &+ \Big(\frac{d_{I,k}(H)\, H}{D_k}\Big)^{\frac{4}{4+d_{I,k}(H)}}. \end{aligned} \quad (83)$$

This gives the stated scaling of the approximation loss in terms of the intrinsic dimension $d_{I,k}(H)$, the horizon H, and the available client sample size D_k. □