# System Design and Convergence Analysis for Decentralized Federated Fine-Tuning on LEO Satellite Networks

Zhigang Yan*, Guangxu Zhu†, Yao Tang‡, Haoyuan Pan§, Nikolaos Pappas¶, Tse-Tin Chan*
* Dept. of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong, China
† Shenzhen Research Institute of Big Data, Shenzhen, China
‡ Dept. of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
§ College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
¶ Dept. of Computer and Information Science, Linköping University, Linköping, Sweden
E-mails: yanz@eduhk.hk, gxzhu@sribd.cn, ytang@ie.cuhk.edu.hk,
hypan@szu.edu.cn, nikolaos.pappas@liu.se, tsetinchan@eduhk.hk

***Abstract*—With low-Earth orbit (LEO) satellites emerging as edge nodes for Federated Learning (FL), their high deployment cost and limited onboard resources make efficient model training crucial. This paper addresses the challenge of global weather forecasting using a network of LEO satellites, where individual satellites have limited regional coverage and insufficient access to global data. To overcome this, we propose a Decentralized FL (DFL) framework that incorporates both observed and self-generated data for training. Each satellite locally fine-tunes its weather forecasting model using its own observations and synthetic data generated from its model, and then collaboratively updates model parameters across the network using DFL to achieve global knowledge sharing without transmitting raw data. Furthermore, we provide a theoretical convergence bound for the proposed framework, and simulation results demonstrate that it achieves accurate forecasting while ensuring robust convergence.**

***Index Terms*—Decentralized federated learning, Low Earth orbit satellites, satellite communication, fine-tuning.**

## I. Introduction

With the rapid growth of Artificial Intelligence (AI), traditional centralized model training on cloud servers faces increasing challenges related to communication overhead and data privacy [1]. Edge intelligence has emerged as a promising paradigm, enabling local data processing and training near the data source, thereby improving efficiency and security [2]. Federated Learning (FL) further enhances this approach by allowing model training through the exchange of parameters, rather than raw data [3]. However, centralized FL (CFL) depends on a central server for aggregation, which limits scalability in highly dynamic environments such as low-Earth orbit (LEO) satellite networks. In contrast, decentralized FL (DFL) enables direct model aggregation among edge nodes without a central coordinator, offering greater robustness and scalability [4]. Recent advances in LEO satellite technology, such as inter-satellite links (ISLs) [5] and onboard AI capabilities [6], allow satellites to act as intelligent edge nodes, supporting distributed learning and efficient global model updates. Consequently, DFL over LEO satellite networks has become a promising architecture for enabling large-scale, privacy-preserving, and communication-efficient AI in space-based systems [7].

Early studies on FL in LEO satellite networks mainly explored CFL frameworks [8]–[10], with [8] first proposing the use of LEO satellites as FL edge nodes. Given the limited efficiency of ground-to-satellite links (GSL) and the increasing stability of intra-plane ISLs, later works shifted toward aggregation schemes that reduce dependence on GSL by leveraging inter-satellite communication [9], [10]. Inter-plane ISLs are typically categorized as connecting either crossing orbital planes (COP) or neighboring orbital planes (NOP), with the latter offering greater stability due to smoother relative motion [11]. Consequently, routing strategies that combine intra- and inter-plane ISLs have been proposed [12], [13], enabling fully ISL-based model aggregation, particularly in constellations without COP ISLs, such as the Walker Star. Furthermore, several DFL frameworks have been introduced to enhance efficiency in LEO networks [14]–[16], focusing on techniques such as satellite number control and aggregation sequencing to reduce latency [14], [15], as well as optimized power and bandwidth allocation to minimize energy consumption [16].

While prior studies have advanced DFL in LEO satellite networks, two major practical challenges remain. *a) Impracticality of training from scratch*. In real-world scenarios, LEO satellites are expected to perform data collection, training, and inference simultaneously. Starting model training solely from local data is unrealistic, so fine-tuning pre-trained models for specific downstream tasks is more practical. *b) Unrealistic labeling assumptions*. Most existing works assume fully labeled datasets and focus on supervised learning. In contrast, LEO satellites typically collect unlabeled data and lack mechanisms for reliable labeling, making them better suited for self-supervised or unsupervised learning paradigms.

To address these issues, this paper introduces a DFL framework for fine-tuning a pre-trained global weather forecasting model across a LEO satellite network, which is a representative scenario of self-supervised learning. Each satellite fine-tunes its model using locally observed data, while data from unobserved regions are generated using its current model to enhance training completeness. The combination of real and generated data supports local learning, and model parameters are collaboratively aggregated through DFL, enabling global knowledge sharing without the need for a central coordinator. Additionally, a theoretical convergence bound is derived to

This work was supported in part by the National Natural Science Foundation of China under Grants 62371302, 62371313, 62522118, in part by Guangdong Young Talent Research Project under Grant 2023TQ07A708, in part by Shenzhen Science and Technology Program under Grant JCYJ20250604181549064, in part by the Excellence Center at Linköping - Lund in Information Technology (ELLIIT), and in part by the Research Matching Grant Scheme from the Research Grants Council of Hong Kong. (*Corresponding author: Tse-Tin Chan.*)

assess the performance of the proposed framework.

## II. Proposed DFL Framework

In this section, we first introduce the DFL framework for fine-tuning a global weather forecasting model over LEO networks. Then, the system models of the entire fine-tuning process on DFL are presented.

### A. Pre-trained Weather Forecasting Models

Recent advances in deep learning have led to the development of pre-trained weather forecasting models that capture the spatiotemporal dynamics of the Earth system using historical meteorological data. Notable examples include FourCastNet [17] and GraphCast [18], which are built on large-scale neural network architectures and trained in a self-supervised autoregressive manner to predict future atmospheric states from past observations.

Formally, let $\mathbf{S}_{t_1-t_0:t_1}$ denote a sequence of past $t_0$ weather states observed over the interval $[t_1 - t_0, t_1]$, where each $\mathbf{s}_t \in \mathbb{R}^d$ represents a multi-dimensional atmospheric variable vector (e.g., temperature, pressure, and wind). The goal is to predict future states $\mathbf{S}_{t_1+t_2}$ for some lead time $t_2 > 0$. An autoregressive forecasting model $f_\theta(\cdot)$ parameterized by $\boldsymbol{\theta}$ learns the prediction

$$\hat{\mathbf{S}}_{t_1+t_2} = f_\theta(\mathbf{S}_{t_1-t_0:t_1}; \boldsymbol{\theta}), \tag{1}$$

where $\hat{\mathbf{S}}_{t_1+t_2}$ denotes the forecast results of $\mathbf{S}_{t_1+t_2}$.

These models are typically trained on large-scale reanalysis datasets (e.g., ERA5) using a sequence-to-sequence learning framework, where the objective is to minimize the discrepancy between predicted and actual future states across multiple forecast horizons. The training input consists of global atmospheric states sampled at regular intervals, including physical variables such as temperature, pressure, wind velocity, and humidity across different altitudes and locations. These inputs form multi-channel spatiotemporal fields representing the entire Earth, typically derived from reanalysis datasets that assimilate observations from various sources. Let $\mathcal{S}_{\text{pr}} = \{\mathbf{S}_1, \cdots, \mathbf{S}_{T_p}\}$ denote the dataset used for pre-training. The loss function is

$$L(\boldsymbol{\theta}, \mathcal{S}_{\text{pr}}) = \frac{1}{T_p} \sum_{t_1=1}^{T_p} \left\| \hat{\mathbf{S}}_{t_1} - \mathbf{S}_{t_1} \right\|_2^2, \tag{2}$$

where $T_p$ is the total training iterations for pre-training. This formulation does not require additional labels, as the future states within the dataset serve as supervisory labels. This self-supervised learning approach is suitable for LEO satellites, as their observed data often lacks labels.

### B. DFL for Fine-tuning Weather Forecasting Model on LEO Satellite Network

We consider the Walker Star constellation, which offers remarkable advantages in stable ISL communication and global coverage, as illustrated in Fig. 1. It has been widely applied in various communication and LEO systems, including OneWeb. We assume the Walker Star constellation, which consists of $M$ orbits, with each orbit containing $N$ satellites.

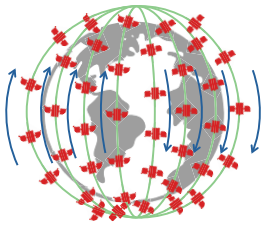

Fig. 1. Diagram of a Walker Star constellation.

In DFL over the LEO satellite network, the goal of model training is to seek the optimal global model $\boldsymbol{\theta}^*$ by all participating satellites collaboratively, where $\boldsymbol{\theta}^*$ satisfies that

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathcal{S}), \tag{3}$$

where $L(\cdot)$ is the global loss function of DFL, $\boldsymbol{\theta}$ and $\mathcal{S}$ are the global model parameter and the sum of datasets on all satellites, respectively. Different from traditional machine learning models, which train iteratively in a centralized manner to minimize their loss function, the global loss function of DFL can only be minimized in a distributed manner, as all local data cannot be shared. Thus, (3) is rewritten as

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} L_{ij}(\boldsymbol{\theta}, \mathcal{S}_{ij}), \tag{4}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{11}, \cdots, \boldsymbol{\theta}_{MN}\}$, $L_{ij}(\cdot)$ is the local loss function of the $i$-th satellite in the $j$-th orbit, $\boldsymbol{\theta}_{ij}$ and $\mathcal{S}_{ij}$ are the local parameter and dataset, respectively. We assume that $\mathcal{S}_{ij}$ for all $i, j$ have $K$ samples, which are $\mathbf{S}_{ij}^{(k)}$, $k = 1, \cdots, K$, so we have $\mathcal{S}_{ij} = \{\mathbf{S}_{ij}^{(1)}, \cdots, \mathbf{S}_{ij}^{(K)}\}$ and $\mathcal{S} = \mathcal{S}_{11} \cup \cdots \cup \mathcal{S}_{MN}$. Thus, $L_{ij}(\cdot)$ is given by

$$L_{ij}(\boldsymbol{\theta}_{ij}, \mathcal{S}_{ij}) = \frac{1}{K} \sum_{k=1}^{K} \left\| \hat{\mathbf{S}}_{ij}^{(k)} - \mathbf{S}_{ij}^{(k)} \right\|_2^2. \tag{5}$$

One of the most common optimization algorithms in FL is FedAvg [19], which includes local training and aggregation in all iterations from $t$=1 to $T$ to update the global model parameters until the updating process converges. In DFL for fine-tuning weather forecasting models on the LEO satellites, five steps are implemented in each iteration as follows:

**Step 1: Get training data:** The pre-trained model is trained on the data of the global observation, but a single LEO satellite cannot provide such comprehensive data coverage due to the limited swath width and revisit frequency. Thus, each sample of the local dataset used for fine-tuning on LEO satellites is a combination of its observation data and generated data for the missing parts outside its observational coverage. It means that $\mathbf{S}_{ij}^{(k)}$=[$\mathbf{S}_{ij,k}^{\text{ob}}, \mathbf{S}_{ij,k}^{\text{ge}}$], where $\mathbf{S}_{ij,k}^{\text{ob}}$ and $\mathbf{S}_{ij,k}^{\text{ge}}$ are the observed and generated part of sample $\mathbf{S}_{ij}^{(k)}$ respectively. Therefore, different from the pre-training process, $\mathbf{S}_{ij}^{(k)}$ does not represent the ground truth at any moment, as it includes $\mathbf{S}_{ij,k}^{\text{ge}}$.

**Step 2: Local training:** Each satellite calculates the stochastic gradient of $L_{ij}(\cdot)$, then updates its local model through $\tau$ rounds of Stochastic Gradient Descent (SGD) in parallel. For the $i$-th satellite on the $j$-th orbit, in $t$-th iteration, its local model parameter is updated by

$$\boldsymbol{\theta}_{ij,t+1} = \boldsymbol{\theta}_{ij,t} - \eta \mathbf{g}_{ij,t}, \tag{6}$$

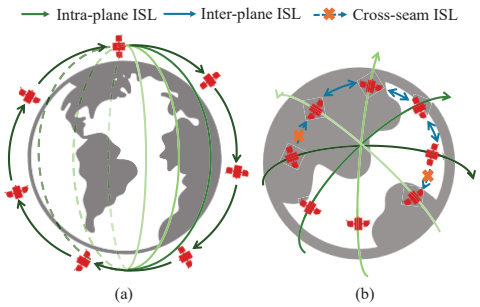


Fig. 2. Diagram of intra-orbit and inter-orbit aggregation.

where $\eta$ is the learning rate and $\mathbf{g}_{ij,t}$ is the accumulated stochastic gradient of $L_{ij}(\cdot)$ of $\tau$ round updates from $\boldsymbol{\theta}_{ij,t}$ to $\boldsymbol{\theta}_{ij,t+1}$. Therefore, $\mathbf{g}_{ij,t} = \sum_{k=0}^{\tau-1} \nabla L_{ij}(\boldsymbol{\theta}_{ij,t+\frac{k}{\tau}}, \xi_{ij,t+\frac{k}{\tau}})$, where $\xi_{ij,t+\frac{k}{\tau}}$ is sampled from $\mathcal{S}_{ij}$ and $\nabla L_{ij}(\cdot, \xi_{ij,t+\frac{k}{\tau}})$ is the stochastic gradient of $L_{ij}(\cdot, \mathcal{S}_{ij})$.

**Step 3: Intra-orbit aggregation:** After all satellites finish their local training steps, because of the stable intra-plane ISLs, all local parameters of satellites in the same orbit are aggregated first. For simplicity, we consider the simple average for aggregation. However, the proposed framework could be extended to the weighted average aggregation.

We consider the Ring-AllReduce algorithm for completing Intra-orbit aggregation among satellites, a common scheme for aggregating parameters in a ring topology [9]. The goal of Intra-orbit aggregation is to obtain the averaged model parameters of all satellites in this orbit, e.g., in the orbit $j$,

$$\boldsymbol{\theta}_{j,t} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\theta}_{ij,t}, \tag{7}$$

To achieve this goal, the Ring-AllReduce algorithm requires all satellites to split their local parameters into $N$ chunks, and each satellite sends and receives chunks from its neighbors, summing partial results along the ring. After $N-1$ steps, each satellite holds the sum of one unique chunk. Then, satellites share their chunk of the summed result with neighbors. Likewise, after $N-1$ steps, each satellite has a complete aggregate parameter. Thus, $\boldsymbol{\theta}_{ij,t}$ for all $i$ are aggregated among all satellites in the same orbit with $2N-2$ rounds of communication, and a diagram of this step is given by Fig. 2(a).

**Step 4: Inter-orbit aggregation:** This step aims to average the aggregated parameters of all orbits, which means

$$\boldsymbol{\theta}_{t} = \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{\theta}_{j,t}, \tag{8}$$

However, the inter-plane ISLs between the first and last orbit are cross-seam ISLs with significant Doppler shift, since the movable direction of satellites in these two orbits is opposite [7]. Consequently, satellites in different orbits cannot be connected in a ring topology. Therefore, we consider a sequential aggregation scheme for this step. Specifically, one satellite in the first orbit transmits $\boldsymbol{\theta}_{1,t}$ to one satellite in the second orbit, until all $\boldsymbol{\theta}_{j,t}$ are transmitted to one satellite in the $M$-th orbit and aggregated. Then this aggregated $\boldsymbol{\theta}_t$ is sent to all orbits by the reverse path. The entire process of Step 4 requires $2M-2$ rounds of communication, and a diagram of this process is shown in Fig. 2(b).

**Step 5: Intra-orbit synchronization:** Upon completing Step 4, aggregated parameters $\boldsymbol{\theta}_{j,t}$ for all $j$ have been combined across all orbits to form $\boldsymbol{\theta}_t$, and each orbit has one satellite that has obtained $\boldsymbol{\theta}_t$. Then, these satellites synchronize $\boldsymbol{\theta}_t$ to the other satellites in the same orbit via sequential transmission, involving $N-1$ rounds of communication, and this process is performed in parallel in each orbit. Finally, after Steps 3 to 5, all satellites obtain the aggregated parameter $\boldsymbol{\theta}_t$, which is calculated by

$$\boldsymbol{\theta}_{t} = \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} \boldsymbol{\theta}_{ij,t}. \tag{9}$$

Moreover, the global loss function is defined as

$$L(\boldsymbol{\theta}_t, \mathcal{S}) \triangleq \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} L_{ij}(\boldsymbol{\theta}_t, \mathcal{S}_{ij}), \tag{10}$$

where $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_M$ and $\mathcal{S}_j = \mathcal{S}_{1j} \cup \cdots \cup \mathcal{S}_{Nj}$.

These five steps iteratively repeat until convergence to optimize $\boldsymbol{\theta}_t$, which minimizes the global loss function $L(\boldsymbol{\theta}_t, \mathcal{S})$. The whole process of DFL for fine-tuning weather forecasting models on the LEO satellite network is shown in Fig. 3.

## III. Convergence Analysis

In this section, we analyze the convergence behavior of the proposed DFL-based fine-tuning framework for global weather forecasting over LEO satellite networks. While existing studies have extensively examined FL convergence, they mainly consider training from scratch with perfectly labeled data. In contrast, our framework involves a self-supervised fine-tuning task where the labels are synthetically generated due to the limited coverage of each satellite. Therefore, it is essential to extend traditional DFL convergence analysis to account for imperfect labels and the distributional discrepancy between synthetic fine-tuning data and pre-trained datasets, providing deeper insight into the framework's convergence.

### A. Assumptions and Preliminaries

To complete the analysis of the convergence of DFL for the proposed self-supervised fine-tuning task, we make the following assumptions about the local and global loss functions and their gradients. They are adopted in the previous works to simplify the analysis, such as [4], [9], [19]–[22], and are commonly used in stochastic optimization. In the following assumptions, $L(\boldsymbol{\theta})$ and $L_{ij}(\boldsymbol{\theta}_{ij})$ are short for $L(\boldsymbol{\theta}, \mathcal{S})$ and $L_{ij}(\boldsymbol{\theta}_{ij}, \mathcal{S}_{ij})$. $\nabla L_{ij}$ and $\nabla^2 L_{ij}$ are the gradient and Hessian matrix of $L_{ij}(\boldsymbol{\theta}_{ij})$. In addition, $\mathbb{E}[\cdot]$ is short for $\mathbb{E}_{\xi_{ij,t} \sim \mathcal{S}_{ij}}[\cdot]$.

*Assumption 1:* For all local loss functions $L_{ij}(\boldsymbol{\theta}_{ij})$, we assume that they are

- ***m*-nonconvex [21], [22]**: All the eigenvalues of $\nabla^2 L_{ij}$ are larger than $-m$, where $m$ is a constant and $m > 0$.

*Remark 1: If we replace the local loss function with a specific regularized version, it can be transformed into an $m$-strongly convex function.*

*Assumption 2:* For all local stochastic gradients $\mathbf{g}_{ij}$, we assume that they are

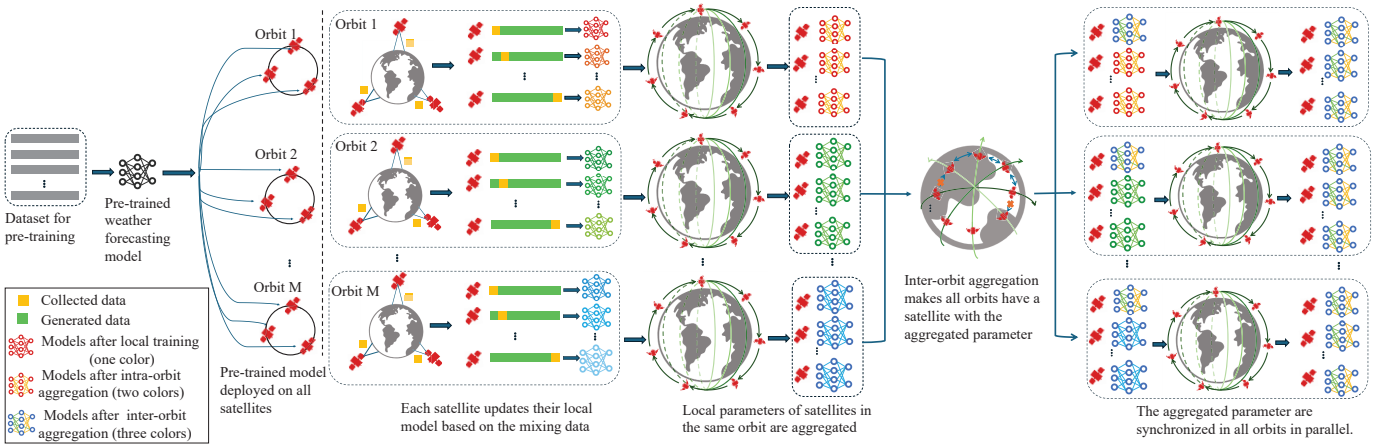


Fig. 3. Illustration of the proposed DFL framework for fine-tuning a weather forecasting model in a LEO satellite network.

- **$\rho$-Lipschitz [4], [9]**: For any pair of $\boldsymbol{\theta}_{ij,t_1}$ and $\boldsymbol{\theta}_{ij,t_2}$, they have $\|\mathbf{g}_{ij,t_1} - \mathbf{g}_{ij,t_2}\|_2 \leq \rho \|\boldsymbol{\theta}_{ij,t_1} - \boldsymbol{\theta}_{ij,t_2}\|_2$.
- ***G*-bounded [19], [21]**: $\mathbb{E}[\|\mathbf{g}_{ij}\|_2^2] \leq G, \quad \forall i, j,$

where $\rho$ and $G$ are positive constants.

*Remark 2: From Assumptions 1, 2, and triangle inequality, it is easy to check that the global loss $L(\boldsymbol{\theta})$ and its stochastic gradient are also $m$–nonconvex, $\rho$-Lipschitz, and $G$-bounded.*

*Assumption 3:* For relationships between $\mathbf{g}_{ij}$ and $\nabla L_{ij}$

- **No bias of $\nabla L_{ij}$**: $\mathbb{E}[\mathbf{g}_{ij}] = \nabla L_{ij}, \quad \forall i, j.$
- **$\sigma$-bounded variance [21]**: $\mathbb{E}[\|\mathbf{g}_{ij} - \nabla L_{ij}\|_2^2] \leq \sigma, \quad \forall i, j,$

where $\sigma$ is a positive constant.

Moreover, in our proposed scenario, each satellite trains on a dataset where part of the data are true observations from that satellite, and the rest are generated to fill gaps in the global view. Each local dataset represents the globe rather than the region directly observed by the satellite. Thus, the data is closer to being independent and identically distributed (i.i.d.) with lower heterogeneity in the proposed DFL framework. In previous works, the non-i.i.d. level of DFL is evaluated by some metrics, such as a) the upper bound of the gradient divergence (e.g., $L_2$ norm of the gap between the local and global gradient) [21], b) the gap between the global optimum and weights sum of local optimums [19], and c) the upper bound of the ratio of the $L_2$ norm of local gradients sum to the $L_2$ norm of global gradient [20]. Thus, in the proposed DFL framework, we have the following lemma.

*Lemma 1: The datasets among all satellites are closer to i.i.d., which means that*

$$\mathbb{E}\left[\left\|\nabla L_{ij} - \frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}\nabla L_{ij}\right\|_2^2\right] \approx 0, \ \forall i, j. \tag{11}$$

Some convergence bounds of FL have been derived based on the above assumptions. [1] Then, substitute (11) into these bounds, they can be summarized in the following form.

*Lemma 2: Convergence bound of DFL in i.i.d. case*

$$\mathbb{E}[L(\boldsymbol{\theta}_T, \mathcal{S}) - L(\boldsymbol{\theta}^*, \mathcal{S})] \leq \mathcal{O}\left(\frac{L(\boldsymbol{\theta}_0, \mathcal{S}) - L(\boldsymbol{\theta}^*, \mathcal{S})}{T}\right). \tag{12}$$

[1]Not all bounds are derived under all assumptions. For example, Theorem 2 in [20] uses only Assumption 2. [19] derives Theorem 1 using Assumptions 2 and 3, while Theorem 1 in [21] is proved under Assumptions 1 to 3.

where $T$ is the total iteration of the training process of DFL.

*Proof: By substituting the i.i.d. settings of these works into their derived bounds, we complete the proof.* □

Note that the distance between the initial parameter and the optimal parameter affects the performance of model training by the DFL framework. For tasks trained from scratch, the initial parameters are usually random, and it is difficult to obtain the global optimal value. Therefore, it is hard to optimize the model performance by controlling this distance. However, different from these previous works, model fine-tuning starts from the pre-trained model in the proposed DFL framework, so the distance reflects the performance of applying the pre-trained model that has not been fine-tuned on the new dataset. It denotes the generalization capacity of the pre-trained model on the dataset for fine-tuning.

### B. Performance Analysis

Although discussions of derived convergence bounds have been presented in previous works, these bounds cannot be used directly to analyze the performance of the proposed DFL framework. This is because we need to explore the impact of generalization errors in the pre-trained model on the convergence. To complete the performance analysis, we analyze how the generalization capacity of the pre-trained model on the fine-tuning dataset $\mathcal{S}_{\text{ft}}$ affect the distance between the initial parameter $\boldsymbol{\theta}_0^{\text{ft}}$ and the optimal parameter $\boldsymbol{\theta}_{\text{ft}}^*$ at first. If the pre-trained model is trained for $T_p$ rounds on the pre-trained dataset $\mathcal{S}_{\text{pr}}$, the upper bound of $\mathbb{E}[L(\boldsymbol{\theta}_0^{\text{ft}}, \mathcal{S}_{\text{ft}}) - L(\boldsymbol{\theta}_{\text{ft}}^*, \mathcal{S}_{\text{ft}})]$ is

$$\begin{aligned}&\mathbb{E}[L(\boldsymbol{\theta}_0^{\text{ft}}, \mathcal{S}_{\text{ft}}) - L(\boldsymbol{\theta}_{\text{ft}}^*, \mathcal{S}_{\text{ft}})] = \mathbb{E}[L(\boldsymbol{\theta}_{T_p}^{\text{pr}}, \mathcal{S}_{\text{ft}}) - L(\boldsymbol{\theta}_{\text{ft}}^*, \mathcal{S}_{\text{ft}})] \\ &\leq \mathbb{E}[|L(\boldsymbol{\theta}_{T_p}^{\text{pr}}, \mathcal{S}_{\text{ft}}) - L(\boldsymbol{\theta}_{T_p}^{\text{pr}}, \mathcal{S}_{\text{pr}})|] + L(\boldsymbol{\theta}_{T_p}^{\text{pr}}, \mathcal{S}_{\text{pr}}) - L(\boldsymbol{\theta}_{\text{ft}}^*, \mathcal{S}_{\text{ft}}),\end{aligned} \tag{13}$$

where $\boldsymbol{\theta}_{T_p}^{\text{pr}}$ is the trained parameter of the pre-trained model after $T_p$ iterations pre-training on dataset $\mathcal{S}_{\text{pr}}$. Since $L(\boldsymbol{\theta}_{T_p}^{\text{pr}}, \mathcal{S}_{\text{pr}})$ is fixed after pre-trained model training, and $L(\boldsymbol{\theta}_{\text{ft}}^*, \mathcal{S}_{\text{ft}})$ depends only on the expression of $L(\cdot)$ and the samples in $\mathcal{S}_{\text{ft}}$, the value of $L(\boldsymbol{\theta}_{T_p}^{\text{pr}}, \mathcal{S}_{\text{pr}}) - L(\boldsymbol{\theta}_{\text{ft}}^*, \mathcal{S}_{\text{ft}})$ is a constant. Thus, the upper bound in (13) is determined by the first term, which is the gap between the performance of the pre-trained model on the pre-training dataset $\mathcal{S}_{\text{pr}}$ and fine-tuning dataset $\mathcal{S}_{\text{ft}}$. To further analyze this performance gap, the upper bound of this term is given by the following lemma.

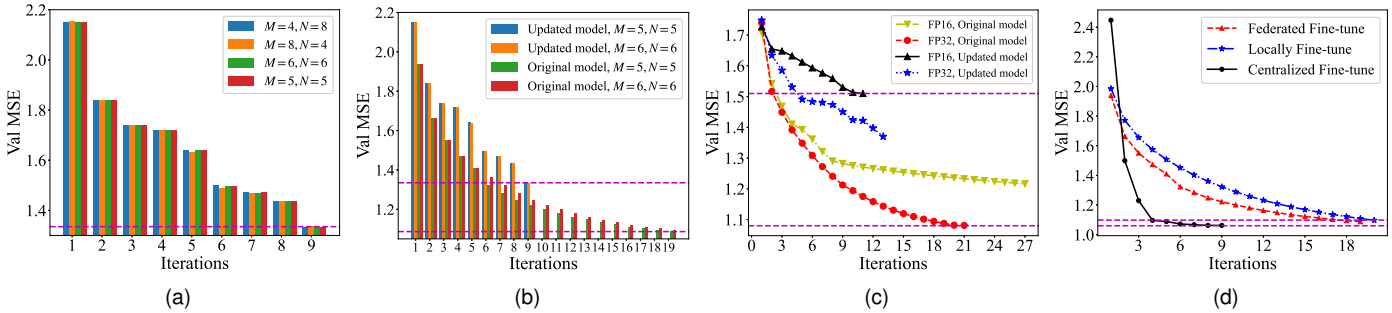


Fig. 4. Convergence behavior of model fine-tuning under the proposed DFL framework with different settings. (a) Impact of the number of satellites. (b) Impact of different local data generation schemes for missing regions. (c) Impact of model quantization schemes. (d) Comparison with different benchmarks.

*Lemma 3: This upper bound can be expressed by the sum of the expected generalization errors of the model trained on $\mathcal{S}_{\rm pr}$ and $\mathcal{S}_{\rm ft}$. If we assume that $\mathcal{S}_{\rm pr}$ and $\mathcal{S}_{\rm ft}$ are sampled from the population $\mathcal{S}_0$ and $\eta = \frac{c}{1+t}$, $c > 0$, we have*

$$\mathbb{E}[|L(\boldsymbol{\theta}^{\rm pr}_{T_p}, \mathcal{S}_{\rm ft}) - L(\boldsymbol{\theta}^{\rm pr}_{T_p}, \mathcal{S}_{\rm pr})|] \leq \mathcal{O}\Big(\frac{1}{n_{\rm ft}} + \frac{T_p^{\alpha}}{n_{\rm pr}}\Big), \quad (14)$$

where $n_{\rm pr}$ and $n_{\rm ft}$ are the number of samples in $\mathcal{S}_{\rm pr}$ and $\mathcal{S}_{\rm ft}$, and $\alpha = \frac{\rho c}{\rho c+1}$.

*Proof: See Appendix A.* □

Notice that, for the loss function that satisfies Assumptions 1 to 3, (12) reveals that increasing the number of training iterations improves the model's performance on the training dataset. On the contrary, (14) indicates that a too large $T_p$ results in a more severe generalization error on other datasets, as it may overfit the training dataset. Then, by substituting (13) and (14) into (12), the convergence bound of the proposed DFL framework is obtained as follows:

*Theorem 1: Convergence bound of the proposed DFL is*

$$\mathbb{E}[L(\boldsymbol{\theta}^{\rm ft}_T, \mathcal{S}_{\rm ft}) - L(\boldsymbol{\theta}^{*}_{\rm ft}, \mathcal{S}_{\rm ft})] \leq \mathcal{O}\Big(\Big(\frac{1}{n_{\rm ft}} + \frac{T_p^{\alpha}}{n_{\rm pr}} + \frac{\Delta L}{T_p} + C\Big) \cdot \frac{1}{T}\Big), \quad (15)$$

where $\boldsymbol{\theta}^{\rm ft}_T$ is the trained parameter after $T$ iterations of fine-tuning on $\mathcal{S}_{\rm ft}$, $\Delta L = L(\boldsymbol{\theta}^{\rm pr}_0, \mathcal{S}_{\rm pr}) - L(\boldsymbol{\theta}^{*}_{\rm pr}, \mathcal{S}_{\rm pr})$ and $C = L(\boldsymbol{\theta}^{*}_{\rm pr}, \mathcal{S}_{\rm pr}) - L(\boldsymbol{\theta}^{*}_{\rm ft}, \mathcal{S}_{\rm ft})$. $\boldsymbol{\theta}^{*}_{\rm ft}$ is the optimal parameter of $L(\cdot)$ on $\mathcal{S}_{\rm ft}$, so $C$ is a non-negative constant.

*Proof: See Appendix B.* □

Theorem 1 shows that the performance of the pre-trained model only affects the convergence rate by determining the value of the coefficient of $T^{-1}$. From this coefficient, we find that, although increasing $T_p$ improves the pre-trained model on $\mathcal{S}_{\rm pr}$, it does not necessarily mean a better pre-trained model for our model fine-tuning task. This is because a too large $T_p$ may over-fit the $\mathcal{S}_{\rm pr}$, which makes it not suitable for $\mathcal{S}_{\rm ft}$. This conflict appears as a trade-off in $T_p$ between the second and third items of this coefficient. However, whatever the performance of the pre-trained model on $\mathcal{S}_{\rm pr}$ is, the process of model fine-tuning on proposed DFL framework can converge.

## IV. Simulation Results

In this section, we verify the feasibility of the proposed DFL framework for LEO satellite networks in a weather prediction task using a fine-tuned model. To investigate the learning performance of the proposed DFL framework on the weather prediction model fine-tuning task, we consider fine-tuning the pre-trained FourCastNet model and observe the convergence behavior of the fine-tuning process and the prediction capacity after completion. Its pre-trained weights are provided by High-Flyer AI, which were trained on the ERA5 data from Jan. 1979 to Dec. 2022. Thus, in our simulation, we investigate the convergence behavior in terms of the validation loss over the entire iteration of the model fine-tuning process on the proposed DFL framework with various settings. The convergence results are shown in Fig. 4.

*1) Impact of number of satellites:* Fig. 4(a) shows the convergence behavior with different orbit numbers and satellite numbers on each orbit. It is observed that, whatever the value of $M$ and $N$ is, after the fine-tuning process converges, their convergence results are almost the same (error is about $10^{-3}$). This result verifies the theoretical convergence analyses in Theorem 1. This is because the training samples on all satellites are almost i.i.d., as they all include the entire global weather information, which is obtained through their own observations and locally generated. In this part, this local data generation task is completed by the fine-tuning model itself.

*2) Impact of local generation models:* Except for using the fine-tuned model to generate missing data, using a standalone model for this generation task is another option. Thus, we compare these two generation modes in Fig. 4(b), where the standalone model is the original pre-trained FourCastNet without fine-tuning. From Fig. 4(b), we observe that using a standalone model to generate the missing data out of their observation scopes achieves a better performance than using the fine-tuned model. This suggests that the missing data generated by the fine-tuned model may cause bias accumulation during the fine-tuning process. Specifically, the predictions of the fine-tuned model may exhibit systematic biases, as it is still in the process of learning. Those biases get reinforced during training by a self-training feedback loop. By contrast, a frozen standalone model trained on full true global data provides the generated missing data with more physical consistency.

*3) Impact of model quantization:* Considering limitations of the computation capacity of the onboard payloads on LEO satellites and requirements of digital communication, it is common to quantize local models during fine-tuning. Thus, we compare the convergence behavior between the local model with/without quantization in Fig. 4(c), which demonstrates that the fine-tuning process can still converge when the models are quantized from FP32 to FP16. The performance gap between the FP32 and FP16 models is due to quantization

error, and quantizing models also slows down the convergence. However, fine-tuning FP16 models with the standalone model for generation also outperforms FP32 models with the fine-tuned model. It means that the impact of bias accumulation is more significant than quantization errors on convergence.

*4) Comparison among different fine-tuning schemes:* In Fig. 4(d), we compare the convergence behavior with two benchmarks, **fine-tuning model locally on satellite** and **centralized fine-tuning on the ground**. These two benchmarks denote fine-tuning the model without any communications and sending all data to the ground and using it to fine-tune the model, respectively. The gap between centralized fine-tuning and local fine-tuning is caused by errors in data generation, and the better performance of centralized fine-tuning proves the necessity of aggregation.

## V. Conclusion

This work has presented a DFL framework tailored for weather forecasting using LEO satellite networks. By leveraging both observational and self-generated data, the proposed approach enables each satellite to refine its local model while contributing to global information sharing without compromising data privacy or communication efficiency. Theoretical analysis provides convergence guarantees. Simulation results demonstrate that the framework achieves accurate forecasting.

## Appendix A
## Proof of Lemma 3

$$\begin{aligned}
&\mathbb{E}[|L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{ft}})-L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{pr}})|]\\
&=\mathbb{E}\big[\big|\big(L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{ft}})-L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_0)\big)-\big(L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{pr}})-L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_0)\big)\big|\big]\\
&\overset{(a)}{\leq}\mathbb{E}\big[\big|L(\boldsymbol{\theta}_0^{\mathrm{ft}},\mathcal{S}_{\mathrm{ft}})-L(\boldsymbol{\theta}_0^{\mathrm{ft}},\mathcal{S}_0)\big|\big]+\mathbb{E}\big[\big|L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{pr}})-L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_0)\big|\big]\\
&\overset{(b)}{\leq}\mathcal{O}\Big(\frac{1}{n_{\mathrm{ft}}}+\frac{T_p^{\alpha}}{n_{\mathrm{pr}}}\Big),
\end{aligned}\tag{16}$$

Step $(a)$ and $(b)$ are based on triangle inequality and Theorem 4.1 in [23], respectively.

## Appendix B
## Proof of Theorem 1

Based on (12)-(14), we have

$$\begin{aligned}
&\mathbb{E}[L(\boldsymbol{\theta}_T^{\mathrm{ft}},\mathcal{S}_{\mathrm{ft}})-L(\boldsymbol{\theta}_{\mathrm{ft}}^{*},\mathcal{S}_{\mathrm{ft}})]\leq\mathcal{O}\Big(\frac{L(\boldsymbol{\theta}_0^{\mathrm{ft}},\mathcal{S}_{\mathrm{ft}})-L(\boldsymbol{\theta}_{\mathrm{ft}}^{*},\mathcal{S}_{\mathrm{ft}})}{T}\Big)\\
&\overset{(a)}{\leq}\mathcal{O}\Big(\mathbb{E}[|L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{ft}})-L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{pr}})|]\\
&\qquad\qquad+L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{pr}})-L(\boldsymbol{\theta}_{\mathrm{ft}}^{*},\mathcal{S}_{\mathrm{ft}})\Big)\cdot\frac{1}{T}\\
&\overset{(b)}{\leq}\mathcal{O}\Big(\frac{1}{n_{\mathrm{ft}}}+\frac{T_p^{\alpha}}{n_{\mathrm{pr}}}+L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{pr}})-L(\boldsymbol{\theta}_{\mathrm{ft}}^{*},\mathcal{S}_{\mathrm{ft}})\Big)\cdot\frac{1}{T}\\
&=\mathcal{O}\Big(\frac{1}{n_{\mathrm{ft}}}+\frac{T_p^{\alpha}}{n_{\mathrm{pr}}}+L(\boldsymbol{\theta}_{T_p}^{\mathrm{pr}},\mathcal{S}_{\mathrm{pr}})-L(\boldsymbol{\theta}_{\mathrm{pr}}^{*},\mathcal{S}_{\mathrm{pr}})\\
&\qquad\qquad+\underbrace{L(\boldsymbol{\theta}_{\mathrm{pr}}^{*},\mathcal{S}_{\mathrm{pr}})-L(\boldsymbol{\theta}_{\mathrm{ft}}^{*},\mathcal{S}_{\mathrm{ft}})}_{C}\Big)\cdot\frac{1}{T}\\
&\overset{(c)}{\leq}\mathcal{O}\Big(\Big(\frac{1}{n_{\mathrm{ft}}}+\frac{T_p^{\alpha}}{n_{\mathrm{pr}}}+\frac{\Delta L}{T_p}+C\Big)\cdot\frac{1}{T}\Big),
\end{aligned}\tag{17}$$

where steps $(a)$, $(b)$, and $(c)$ are based on (13), (14) and (12), respectively. Thus, we complete the proof.

## References


[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, Apr. 2017, pp. 1273–1282.

[2] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[3] G. Zhu *et al.*, "Pushing AI to wireless network edge: An overview on integrated sensing, communication, and computation towards 6G," *Sci. China Inf. Sci.*, vol. 66, no. 3, Feb. 2023, Art. no. 130301.

[4] Z. Yan and D. Li, "Performance analysis for resource constrained decentralized federated learning over wireless networks," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 4084–4100, Jul. 2024.

[5] M. Toyoshima, "Recent trends in space laser communications for small satellites and constellations," *J. Lightw. Technol.*, vol. 39, no. 3, pp. 693–699, Feb. 2021.

[6] Y. Li *et al.*, "LEO satellite constellation for global-scale remote sensing with on-orbit cloud AI computing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9369–9381, Sep. 2023.

[7] C. Wu, Y. Zhu, and F. Wang, "DSFL: Decentralized satellite federated learning for energy-aware LEO constellation computing," in *Proc. IEEE Int. Conf. Satell. Comput. (Satellite)*, Nov. 2022, pp. 25–30.

[8] N. Razmi, B. Matthiesen, A. Dekorsy, and P. Popovski, "Ground-assisted federated learning in LEO satellite constellations," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 717–721, Apr. 2022.

[9] Y. Shi *et al.*, "Satellite federated edge learning: Architecture design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 15 212–15 229, Oct. 2024.

[10] N. Razmi, B. Matthiesen, A. Dekorsy, and P. Popovski, "On-board federated learning for satellite clusters with inter-satellite links," *IEEE Trans. Commun.*, vol. 72, no. 6, pp. 3408–3424, Jun. 2024.

[11] A. U. Chaudhry and H. Yanikomeroglu, "Laser intersatellite links in a starlink constellation: A classification and analysis," *IEEE Veh. Technol. Mag.*, vol. 16, no. 2, pp. 48–56, Jun. 2021.

[12] R. Wang, M. A. Kishk, and M.-S. Alouini, "Ultra reliable low latency routing in LEO satellite constellations: A stochastic geometry approach," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 5, pp. 1231–1245, May 2024.

[13] Q. Chen *et al.*, "Shortest path in LEO satellite constellation networks: An explicit analytic approach," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 5, pp. 1175–1187, May 2024.

[14] Z. Yan and D. Li, "Convergence time optimization for decentralized federated learning with LEO satellites via number control," *IEEE Trans. Veh. Technol.*, vol. 73, no. 3, pp. 4517–4522, Mar. 2024.

[15] P. Qin *et al.*, "Decentralized federated learning in LEO satellite-based IoE communications: Latency optimization under reliability constraints," *IEEE Internet Things J.*, vol. 13, no. 1, pp. 24–38, Jan. 2026.

[16] F. Zhou, Z. Wang, Y. Shi, and Y. Zhou, "Decentralized satellite federated learning via intra- and inter-orbit communications," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2024, pp. 786–791.

[17] T. Kurth *et al.*, "FourCastNet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators," in *Proc. Platform Adv. Sci. Comput. Conf.*, Jun. 2023, pp. 1–11.

[18] R. Lam *et al.*, "Learning skillful medium-range global weather forecasting," *Science*, vol. 382, no. 6677, pp. 1416–1421, Nov. 2023.

[19] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2020, pp. 1–26.

[20] S. Wan *et al.*, "Convergence analysis and system design for federated learning over wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3622–3639, Dec. 2021.

[21] Z. Yan *et al.*, "Adaptive decentralized federated learning in energy and latency constrained wireless networks," *IEEE Trans. Mobile Comput.*, vol. 25, no. 2, pp. 2908–2926, Feb. 2026.

[22] Z. Yang *et al.*, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2021.

[23] B. Le Bars *et al.*, "Improved stability and generalization guarantees of the decentralized SGD algorithm," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2024, pp. 26 215–26 240.