Short-term Load Forecasting of Distribution Transformer Supply Zones Based on Federated Model-Agnostic Meta Learning

Changsen Feng, *Member, IEEE*, Liang Shao*, Student Member*, *IEEE*, Jiaying Wang, Youbing Zhang, *Member, IEEE,* Fushuan Wen, *Fellow, IEEE*

[1]***Abstract*—With the increasing data privacy concerns raised by not only organizations but also individuals in distribution systems, traditional centralized data-driven forecasting approaches for short-term load forecasting (STLF) in distribution transformer supply zones are confronted with the predicament of isolated data island. To this end, a federated model-agnostic meta learning (FMAML) based STLF method is proposed. On the basis of federated learning (FL), model agnostic meta learning (MAML) is employed to build high-quality personalized models for clients, thereby significantly enhancing the personalization and compatibility of the Federated Learning FL, while easing data privacy concerns leveraging the feature of FL. The stochastic controlled averaging (SCA) algorithm is integrated as the federated aggregation algorithm to mitigate the impacts of client-drift (CD) phenomenon that causes slow convergence and even divergence during the training process, especially when the data is highly heterogeneous. Finally, numerical results verify the high accuracy and strong robustness to data heterogeneity and packet dropout of the proposed method.**

***Index Terms* —Short-term load forecasting, distribution transformer supply zones, federated learning, model-agnostic meta learning.**

I. Introduction

Short-term load forecasting (STLF) at distribution system level provides directly support to various applications in power system operation [1], e.g., economic dispatch [2], sizing and locating of distributed generation resources (DER) [3], voltage regulation [4], and state estimation [5]. However, with increasing amount of heterogeneous data from a variety of DERs and end users, traditional centralized data-driven STLF methods of distribution transformer supply zones can hardly meet the requirement in terms of load forecasting accuracy, let alone the privacy concern raised by individuals and organizations.

The existing data-driven STLF methods can be broadly classified into two categories: 1) statistical analysis methods represented by the exponential smoothing forecasting model [6] and the autoregressive integrated moving average model [7]-[8]; and 2) machine learning-based methods with techniques such as support vector machine [9], random forest [10], extreme gradient boosting [11], and deep network [12]. The former has the advantages of being simple and easy to implement but bears the weakness in capturing the highly nonlinear impacts of some factors, such as economic, political, and meteorological elements. As for the latter, the machine learning-based methods exhibit excellent fitting capability especially in the nonlinear scenario. However, the drawback lies in the requisite fine-tuning of specific parameters. In recent years, the deep network based model as well as its advanced variants, such as long short-term memory (LSTM) [12], bi-directional long short-term memory (Bi-LSTM) [13], two-dimensional convolutional neural network (2D-CNN) [14], and hybrid model combining CNN and LSTM (CNN-LSTM) [15], are favored by researchers because of their excellent computational efficiency, strong fitting ability to nonlinearity, abnormal data resiliency, and powerful parallel computing capability. However, the deep network-based method typically necessitates a considerable amount of training data, and lack of data probably results in overfitting issues that could eventually compromise the forecasting accuracy [16]. Unfortunately, end users in the distribution transformer supply zones, especially those who have newly installed smart meters usually lack sufficient historical load data.

To this end, reference [17] uses the spatial information, e.g., the load profiles of neighboring households, to compensate insufficient temporal information, so as to mitigate the overfitting issue. Reference [18] aggregates historical residential load data from different electricity consumption scenarios to augment the volume and variety of the training dataset, thereby improving the generalization ability of the forecasting model. Both of the two methods implicitly assume that the historical load data can be uploaded to a central server for training a global model. This, however, will inevitably arouse data security and privacy concerns in today's social and legal environment especially after the "General Data Protection Regulation" promulgated by the European Union in 2018 imposes strict requirements on data security and privacy.

In the industry field, Google originated Federated Learning (FL) in 2016 to increase the volume and diversity of the training dataset while guaranteeing users' data security and privacy. Specifically, FL trains a global model with datasets isolated across multiple local servers avoiding explicit data exchange [19]. This is realized by sharing and aggregating parameters of local models trained by the clients (local servers) with aggregation algorithms such as federated averaging (Avg). Some works in the academic field introduced FL in

This work is supported by National Natural Science Foundation of China (No. 52107129 and U22B20116).

C. Feng and Y. Zhang are both with College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: fcs@zjut.edu.cn, youbingzhang@zjut.edu.cn)

L. Shao and F. Wen are with the School of Electrical Engineering, Zhejiang University, Hangzhou 310027, China, and F. Wen is also with the Hainan Institute, Zhejiang University, Sanya 572000, China (e-mail: shaol8849@gmail.com, fushuan.wen@gmail.com).

J. Wang is with State Grid Zhejiang Marketing Service Center, Hangzhou, 311100, China (e-mail: wangjiayingee@zju.edu.cn)

resolving some problems of the power system field, such as electricity consumption pattern identification [20], renewable energy power forecasting [21], voltage forecasting in distribution network [22], residential load forecasting [23], non-intrusive load monitoring [24].

Although the superiority of FL in resolving the dilemma of isolated data island, i.e., datasets are isolated across multiple local servers, has been demonstrated by the references mentioned above, the following two problems still exist in STLF of distribution transformer supply zones:

1) FL focuses on exploring the general features but ignores the unique features of individual clients. Specifically, the electricity consumption patterns of end users in different distribution transformer supply zones are highly likely to be heterogeneous, which makes the obtained global model in FL usually not fully suitable for individual clients [25].

2) The widely used aggregation algorithm in the above references, namely Avg, relies on the assumption that training data is identically and independently distributed, i.e., i.i.d. [26]. However, this assumption can hardly be satisfied by end user data from different distribution transformer supply zones, which could result in slow convergence and even non-convergence during the federated training process.

To address the first problem, reference [27]-[29] combined FL and load clustering techniques: apply the K-means to divide the user group into multiple clusters based on their historical consumption patterns, then apply FL for each cluster. However, the load clustering operations in [27]-[29] need users' data to be centrally aggregated for clustering which conflicts with the initiative of FL. Reference [30] proposes a personalized FL-based STLF approach (FLPer) that consists of two steps: clients in FL cooperatively train a global model in a distributed manner, and then each client fine-tunes the global model with their local data to personalize its own STLF model with a higher forecasting accuracy. However, the fine-tuning process is easy to generate overfitting personalized model, which is because the fine-tuning process is completely independent from the federated training process and the lack of necessary interactions between the two processes would inevitably bring an overfitting personalized model.

To this end, we introduced the core idea of MAML and proposed a FL framework with a pre-training mechanism, which is called federated model-agnostic meta-learning (FMAML). MAML is a personalized deep-learning technique that trains a pre-trained meta-model that can quickly adapt to different data distributions by considering the effects of fine-tuning operations in advance. In FMAML, the pre-trained meta-model mentioned in MAML will be generated in FL manner to protect the data privacy. Firstly, the STLF system enables multiple distribution transformer supply zones with accessible historical load data to collaboratively train a pre-trained global meta-model that involves the common load features of all distribution transformer supply zones and can be easily adapted to heterogeneous electricity consumption patterns (i.e., different peak load durations, various typical daily load curves and so on). Then all distribution transformer supply zones use their own historical load data to customize the pre-trained global meta-model into personalized STLF models with high forecasting accuracy by just a few rounds of gradient descent, thus avoiding the overfitting issue mentioned above and enhancing the forecasting accuracy.

Moreover, the FMAML framework is scenario-agnostic, and it can be easily applied in any other deep learning-based scenarios which need privacy protection, such as generating renewable energy forecasting models between operators [21], training non-intrusive load monitoring model utilizing residents' private data [24].

As for the second problem, the non-i.i.d. data from clients makes the Avg aggregation algorithm easily susceptible to Client-Drift (CD) phenomenon, which would cause slow convergence or even non-convergence [33]. To this end, we introduce the stochastic controlled averaging (SCA) algorithm [33] as the aggregation algorithm. Specifically, the SCA aggregation algorithm introduces two auxiliary variables respectively containing global gradient information and local gradient information to rectify the update direction in the federated training process. By continuously correcting the update direction with the SCA aggregation algorithm, the effects of CD phenomenon could be mitigated, leading to a global model that is much closer to the global optimum in comparison with the Avg aggregation algorithm.

TABLE I
COMPARISON OF THE CONSIDERED FACTORS IN FORECASTING METHODS IN THE EXISTING LITERATURE AND THIS PAPER

Refs.	shortfall of data	Privacy protects	Personalized clients' models	Mitigating CD phenomenon
[12], [13], [14], [15], [16]	×	×	×	×
[17], [18]	✓	×	×	×
[20], [21], [22]	✓	✓	×	×
[27]	✓	✓	✓	×
This paper	✓	✓	✓	✓

In this paper, a FMAML based STLF method is proposed to address the problems of isolated data island and data heterogeneity in distribution transformer supply zones. Specifically, it can utilize all users' local data to generate personalized models for each user without violating the data privacy. Comparisons among the key features considered in the existing literature against our proposed method are presented in Table TABLE I

. The contributions of this paper are threefold:

1) A FL based STLF framework is proposed to resolve the dilemma of isolated data island in distribution transformer supply zones while ensuring the data privacy.
2) A novel FMAML approach is proposed, with which the model fine-tuning process and the federated meta-training process mutually feedback, avoiding the overfitting problem. Compared to the existing methods, the proposed approach generates a meta model which could adapt quickly to specific features of clients' local data, thereby achieving a higher forecasting accuracy.
3) SCA aggregation algorithm is integrated into the proposed FMAML framework to effectively mitigate the CD phenomenon, greatly improving not only the forecasting accuracy but also the robustness against data heterogeneity in STLF.

The rest of the paper is organized as follows: Section II process the FMAML based STLF method for distribution transformer supply zones. Section III introduces the SCA algorithm and integrates it into the proposed STLF method.

Numerical case studies are conducted in Section IV, and the conclusions are drawn in Section V.

II. The Federated Model-Agnostic Meta Learning Method

The proposed FMAML method is illustrated in Fig. 1. Its framework comprises of a central server, several clients, i.e., the servers that are in charge of corresponding power supply zones. Each client collects data from the distribution transformer as well as the supply zone and trains a personalized model under the coordination of the central server. The entire process can be chronologically divided into five steps as in Fig. 1. The first four steps constitute the federated meta-training process, and the last step constitutes the federated meta-transfer process.

In the federated meta-training process, the central server and clients collaborate to train a global meta model based on an iterative process. It is worthwhile to emphasize that in this process, the data exchanged between the clients and the central server is indeed encrypted model parameters rather than the locally stored historical load data of clients, which indeed eases the data privacy concern. Thereafter, in the federated meta-transfer process, the central server will issue the global model with well-initialized parameters to the clients, and then each client leverages its own historical load data to further fine-tune the global model and ultimately obtains a personalized load forecasting model.

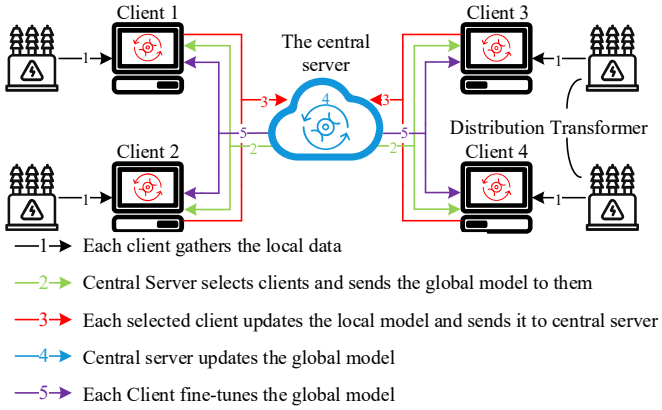

Fig. 1. The proposed FMAML method.

A. Federated Meta-Training Process

We consider N distribution transformer supply zones and each corresponds to a client in FL. They are denoted as $\boldsymbol{C} = [\boldsymbol{C}_0, \ldots, \boldsymbol{C}_i, \ldots, \boldsymbol{C}_{N-1}]$ and indexed by i. The historical load datasets for these clients, namely distribution transformer supply zones, are denoted as $[\boldsymbol{D}_0, \ldots, \boldsymbol{D}_i, \ldots, \boldsymbol{D}_{N-1}]$.

To obtain a global model, the optimization objective of a traditional FL, as formulated in (1), is to minimize the mean value of all clients' losses, i.e., $F'(\boldsymbol{\omega}^{\text{Global}})$,

$$\min F'(\boldsymbol{\omega}^{\text{Global}}) = \frac{1}{N}\sum_{i=1}^{N} L_i(\boldsymbol{\omega}^{\text{Global}}, \boldsymbol{D}_i) \quad (1)$$

where $L_i(\cdot)$ represents the loss function of client $\boldsymbol{C}_i$ and $\boldsymbol{\omega}^{\text{Global}}$ represents the global model parameter.

The global model trained with the optimization objective of minimizing (1) is merely capable of capturing the general features of all clients but ignores the unique features of individual clients. This, despite the prominent generalization ability, could result in forecasting accuracy losses. Different from traditional FL that merely focuses on the performance of the global model, this paper is dedicated to realizing a meta learning model with well-trained parameters that can quickly adapt to different clients' specific data with different data distributions. To this end, the optimization objective is modified to incorporate the effects of fine-tuning the local models beforehand. With the FMAML method [32], fine-tuning the local models may involve multiple rounds, but usually a single round would be sufficient to accomplish the personalization. The parameters of a personalized model after a single round fine-tuning can be described as:

$$\boldsymbol{\omega}_i^{\text{Per}} = \boldsymbol{\omega}^{\text{Global}} - \alpha^{\text{Per}} \nabla L_i\left(\boldsymbol{\omega}^{\text{Global}}, \boldsymbol{D}_i\right) \quad (2)$$

where $\boldsymbol{\omega}_i^{\text{Per}}$ represents the final personalized model for client $\boldsymbol{C}_i$; α^{Per} is the learning rate of model fine-tuning; ∇L_i is the gradient of loss function L_i. The optimization objective (1) is replaced by (3) incorporating $\boldsymbol{\omega}_i^{\text{Per}}$ from (2).

$$\min F(\boldsymbol{\omega}^{\text{Global}}) = \frac{1}{N}\sum_{i=1}^{N} L_i(\boldsymbol{\omega}_i^{\text{Per}}, \boldsymbol{D}_i) \quad (3)$$

The federated meta-training process iterates between the central server and the clients to minimize the optimization objective (3) and terminates until the preset maximum number of global model update iterations (K) is reached. In the k-th iteration of the outer layer, the central server and the clients will implement the following three steps:

Step1: The central server randomly selects some of the clients, collected by set S_k, and sends the global model parameters $\boldsymbol{\omega}_k^{\text{Global}}$ to each client in S_k, namely $\boldsymbol{C}_i \in S_k$.

Step2: Each client in S_k receives the parameters from the central server and initializes the parameters of its local model as in (4),

$$\boldsymbol{\omega}_{i,k,0} = \boldsymbol{\omega}_k^{\text{Global}} \quad (4)$$

where $\boldsymbol{\omega}_{i,k,0}$ represents the initial parameters of the local model for the client $\boldsymbol{C}_i$.

Hereafter, each client in S_k starts to update its local model with its own data. The optimization objective of the client $\boldsymbol{C}_i$ in the t-th iteration is defined as in (5) and the gradient can be calculated as in (6),

$$\min F_{i,k,t}\left(\boldsymbol{\omega}_{i,k,t}\right) = L_i(\boldsymbol{\omega}_{i,k,t} - \alpha^{\text{Per}} \nabla L_i(\boldsymbol{\omega}_{i,k,t}, \boldsymbol{D}_i), \boldsymbol{D}_i) \quad (5)$$

$$\nabla F_{i,k,t}\left(\boldsymbol{\omega}_{i,k,t}\right) = \left(\boldsymbol{I} - \alpha^{\text{Per}} \nabla^2 L_i\left(\boldsymbol{\omega}_{i,k,t}, \boldsymbol{D}_i\right)\right) \times \mu_{i,k,t} \quad (6)$$

where $\mu_{i,k,t} = \nabla L_i(\boldsymbol{\omega}_{i,k,t} - \alpha^{\text{Per}} \nabla L_i(\boldsymbol{\omega}_{i,k,t}, \boldsymbol{D}_i), \boldsymbol{D}_i)$ represents an auxiliary variable; $\boldsymbol{I}$ represents an identity matrix; $F_{i,k,t}$ is the objective function of the client $\boldsymbol{C}_i$ in the t-th iteration; $\nabla F_{i,k,t}$ is the gradient of $F_{i,k,t}$; and $\nabla^2 L_i(\cdot)$ is the Hessian matrix of L_i.

The parameters of the local model of the client $\boldsymbol{C}_i$ can be updated as:

$$\boldsymbol{\omega}_{i,k,t+1} = \boldsymbol{\omega}_{i,k,t} - \beta \nabla F_{i,k,t}\left(\boldsymbol{\omega}_{i,k,t}, \boldsymbol{D}_i\right) \quad (7)$$

where β represents the learning rate of local training.

The local model is iteratively updated with (6)-(7) as in the inner layer shown in Fig. 2 until the preset maximum number of iterations is reached. After the local model updating iteration terminates, the local model parameters $\boldsymbol{\omega}_{i,k}$ for client i in k-th global model updating iteration can be obtained.

Step3: Each client uploads its $\boldsymbol{\omega}_{i,k}$ to the central server, with which the central server uses an aggregation algorithm to

obtain the new parameters $\boldsymbol{\omega}_{k+1}^{\text{Global}}$ for the next iteration $k+1$. The aggregation algorithms will be detailed in the next section.

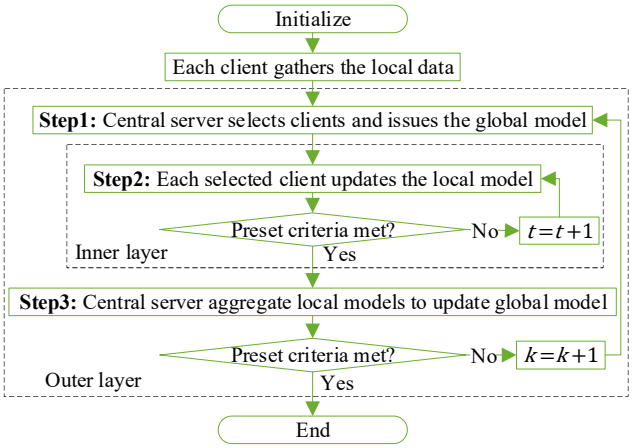

Fig. 2. Flowchart of the federated meta-training process.

Generally, the federated meta-training process is a two-layer iterative process as illustrated in Fig. 2. In the inner layer, i.e., Step2, clients initialize their local model with the parameters from the global model, and then iteratively update their local models with their own data to minimize their individual loss functions. In the outer layer, the central server aggregates the local models trained in the inner layer to update the global model and then issue the latest parameters of the global model to the clients.

In the iterative process, it is worthwhile to mention that the heavy computational burden of calculating the Hessian matrix $\nabla^2 L_i(\boldsymbol{\omega}_{i,k,t}, \boldsymbol{D}_i)$ and the vector $\mu_{i,k,t}$ in (6) necessitates a computationally friendly approximation manner. For any function $f(\boldsymbol{x})$, the product of its Hessian matrix $\nabla^2 f(\boldsymbol{x})$ and a vector $\boldsymbol{v}$ can be approximated as in (8)

$$\nabla^2 f(\boldsymbol{x})\boldsymbol{v} \approx \frac{\nabla f(\boldsymbol{x}+\delta\boldsymbol{v}) - \nabla f(\boldsymbol{x}-\delta\boldsymbol{v})}{2\delta} \tag{8}$$

with an error bounded by $\rho\delta\|\boldsymbol{v}\|^2$, where $\delta > 0$ indicates the approximation accuracy and ρ represents the Lipschitz continuous constant, i.e., $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \le \rho\|\boldsymbol{x}-\boldsymbol{y}\|$.

On this basis, equation (6) can be simplified as (9),

$$\nabla F_{i,k,t}\left(\boldsymbol{\omega}_{i,k,t}\right) = \mu_{i,k,t} - \alpha^{\text{Per}}\boldsymbol{d}_{i,k,t} \tag{9}$$

where $\boldsymbol{d}_{i,k,t}$ approximates $\nabla^2 L_i\left(\boldsymbol{\omega}_{i,k,t}, \boldsymbol{D}_i\right) \times \mu_{i,k,t}$ and can be calculated as in (10).

$$\boldsymbol{d}_{i,k,t} = \frac{\left(\nabla L_i\left(\boldsymbol{\omega}_{i,k,t}+\delta\mu_{i,k,t}, \boldsymbol{D}_i\right) - \nabla L_i\left(\boldsymbol{\omega}_{i,k,t}-\delta\mu_{i,k,t}, \boldsymbol{D}_i\right)\right)}{2\delta} \tag{10}$$

In this way, the computation cost of FMAML can decrease from $\mathcal{O}(d^2)$ to $\mathcal{O}(d)$ [34], where d indicates the dimension of deep learning model.

B. Federated Meta-Transfer Process

The purpose of the federated meta-transfer process is to fine-tune the global model considering the unique features of the heterogenous data from individual clients, so as to build personalized models. The federated meta-transfer process is as follows:

1) The central server issues the global model parameters $\boldsymbol{\omega}_K^{\text{Global}}$ to each client. $\boldsymbol{\omega}_K^{\text{Global}}$ represent the parameters of a well-trained global as K indicated the last global model update iteration.
2) Each client uses its local data to personalize the received global model with one fine-tuning round as:

$$\boldsymbol{\omega}_i^{\text{Per}} = \boldsymbol{\omega}_K^{\text{Global}} - \alpha^{\text{Per}}\nabla L_i\left(\boldsymbol{\omega}_K^{\text{Global}}, \boldsymbol{D}_i\right) \tag{11}$$

III. Stochastic Control Averaging Algorithm Integrated Federated Meta-Training Process

In this section, a widely used aggregation algorithm, Avg, together with the CD phenomenon it encounters will be first discussed. On this basis, we propose the SCA algorithm focusing on resolving the challenge of the CD phenomenon.

A. Client-Drift Phenomenon

The Avg aggregation algorithm can be straightforwardly expressed as (12),

$$\boldsymbol{\omega}_{k+1}^{\text{Global}} = \sum_{i\in S_k} \frac{p_{i,k}}{P_k}\boldsymbol{\omega}_{i,k} \tag{12}$$

where $p_{i,k}$ represents the length of training data of client i in k-th iteration and $P_k = \sum_{i\in S_k} p_{i,k}$. Due to the simplicity and low communication cost, the Avg aggregation algorithm has been applied in many fields [20]-[24]. As an example, shown in Fig. 3, two sets of concentric circles represent the contours of the loss function of two clients. In iteration, the two client models (the blue and red balls) will respectively move towards the two client optimums and the global model (the green ball) will be generated by aggregating the two client models.

However, because of the data heterogeneity of two clients, the global model will favor one of the clients and stray from the global optimum (the black triangle which minimizes the sum of the squares of the loss in two clients) which will cause the oscillations during iterations, reduce the convergence of global models and ultimately discourage client from participating in training the federated model.

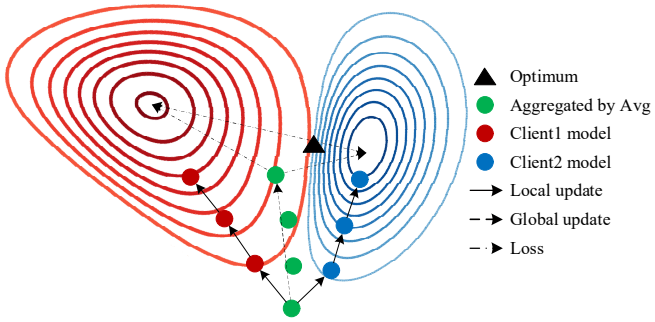

Fig. 3. Schematic graph of client-drift phenomenon

B. SCA Aggregation Algorithm

From the above example, it can be concluded that non-i.i.d data from individual clients is the cause of the CD phenomenon. To this end, the SCA aggregation algorithm is applied instead, and it introduces control variables to the central server side to correct the updating direction of the local models, which is similar to adding inertia in gradient descent optimization algorithm.

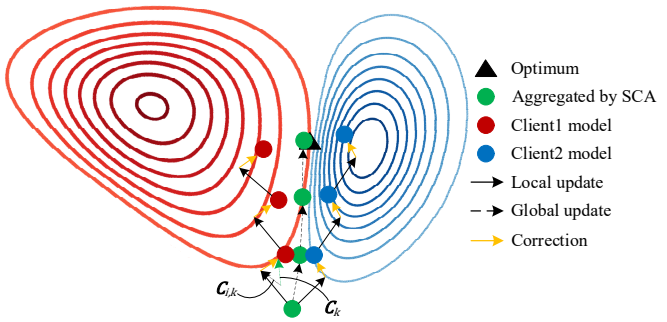

Fig. 4. Schematic graph of SCA aggregation algorithm

As shown in Fig. 4, we introduced two auxiliary variables: 1) $\boldsymbol{c}_{i,k}$, represents the gradient directions in the local updates in the (k-1)th iteration which can be approximately calculated as $\boldsymbol{c}_{i,k-1} - \boldsymbol{c}_{k-1} + \frac{1}{T\beta}(\boldsymbol{\omega}_{k-1}^{Global} - \boldsymbol{\omega}_{i,k-1})$; 2) $\boldsymbol{c}_k$, represent the updated direction in the (k-1)th iteration of the global update, which can be calculated by $\frac{1}{N}\sum_{i=0}^{N-1}\boldsymbol{c}_{i,k-1}$.

Based on the above definitions, $\boldsymbol{c}_k - \boldsymbol{c}_{i,k}$ can be interpreted as the discrepancy in the updated direction of the client $\boldsymbol{C}_i$ from the global optimal direction in the (k-1)th iteration of the global update. SCA aggregation algorithm integrates this discrepancy, i.e., $\boldsymbol{c}_k - \boldsymbol{c}_{i,k}$, to the local model update iteration in the k-th iteration of the global update (as denoted by the yellow arrow in Fig. 4), making the aggregated model (the green ball) closer to the optimum (the black triangle).

At the beginning of the k-th iteration of the outer layer, the central server issues the global model parameter $\boldsymbol{\omega}_k^{\text{Global}}$ together with the auxiliary $\boldsymbol{c}_k$ to each client $\boldsymbol{C}_i \in S_k$. Once receiving these parameters, client $\boldsymbol{C}_i \in S_k$ initializes its local model with $\boldsymbol{\omega}_k^{\text{Global}}$ as in (4).

In t-th iteration of the inner layer iteration, different from (7), $\boldsymbol{\omega}_{i,k,t}$ is now updated with (13),

$$\boldsymbol{\omega}_{i,k,t+1} = \boldsymbol{\omega}_{i,k,t} - \beta(\nabla F_{i,k,t}(\boldsymbol{\omega}_{i,k,t}, \boldsymbol{D}_i) + \boldsymbol{c}_k - \boldsymbol{c}_{i,k}) \quad (13)$$

where $\boldsymbol{c}_k - \boldsymbol{c}_{i,k}$ constitutes the correction term.

The inner layer iteration will terminate until the preset maximum number of iterations is reached. Then, the local control variable $\boldsymbol{c}_{i,k+1}$ can be calculated via (14),

$$\boldsymbol{c}_{i,k+1} = \boldsymbol{c}_{i,k} - \boldsymbol{c}_k + \frac{1}{T\beta}(\boldsymbol{\omega}_k^{\text{Global}} - \boldsymbol{\omega}_{i,k}) \quad (14)$$

where T the preset maximum number of local model update iterations.

Thereafter, the clients calculate the variation of local control variables in two consecutive iterations, i.e., $\Delta\boldsymbol{c}_{i,k}$, and the variation between the local and global model parameters, i.e., $\Delta\boldsymbol{\omega}_{i,k}$, respectively as in (15) and (16),

$$\Delta\boldsymbol{c}_{i,k} = \boldsymbol{c}_{i,k+1} - \boldsymbol{c}_{i,k} \quad (15)$$

$$\Delta\boldsymbol{\omega}_{i,k} = \boldsymbol{\omega}_{i,k} - \boldsymbol{\omega}_k^{\text{Global}} \quad (16)$$

$\Delta\boldsymbol{c}_{i,k,t}$ and $\Delta\boldsymbol{\omega}_{i,k,t}$ are then uploaded to the central server, with which the central server updates the global control variables and the global parameters via (17) and (18),

$$\boldsymbol{c}_{k+1} = \boldsymbol{c}_k + \frac{1}{N}\sum_{i\in S_k}\Delta\boldsymbol{c}_{i,k} \quad (17)$$

$$\boldsymbol{\omega}_{k+1}^{\text{Global}} = \boldsymbol{\omega}_k^{\text{Global}} + \frac{\gamma}{|S_k|}\Delta\boldsymbol{\omega}_{i,k} \quad (18)$$

where γ is the global learning rate and $|S_k|$ is the cardinality of the set S_k.

The whole process of FMAML-SCA algorithm is summarized in TABLE II. Note that a convergence analysis of FL-AVG is given in the appendix section. The case of FMAML-SCA is conceptually similar but necessarily more complex. We refer the interested reader to [33].

Table II
THE PROCESS OF FMAML-SCA

Algorithm: FMAML-SCA

Initialize: $N, K, T, \gamma, \beta, \delta, \alpha^{\text{Per}}, \boldsymbol{\omega}_0^{\text{Global}}, \boldsymbol{c}_0$,
$\boldsymbol{C} = [\boldsymbol{C}_0, \dots, \boldsymbol{C}_i, \dots, \boldsymbol{C}_{N-1}]$,
$\boldsymbol{D} = [\boldsymbol{D}_0, \dots, \boldsymbol{D}_i, \dots, \boldsymbol{D}_{N-1}]$,
$\boldsymbol{c} = [\boldsymbol{c}_{0,0}, \dots, \boldsymbol{c}_{i,0}, \dots, \boldsymbol{c}_{N-1,0}]$;

for $k = 0$ to $K-1$
  The **central server** randomly selects clients as S_k and broadcasts $\boldsymbol{\omega}_k^{\text{Global}}$ and $\boldsymbol{c}_k$ to all selected clients.
  for client $\boldsymbol{C}_i \in S_k$
    Initializes the local model as $\boldsymbol{\omega}_{i,k,0} = \boldsymbol{\omega}_k^{\text{Global}}$.
    for $t = 0{:}T-1$
      Computes the approximation of hessian-vector product via (8) and computes the local model gradient via (9).
      Updates the local model as in (13).
    end
    Updates local control variables as in (14).
    Computes $\Delta\boldsymbol{c}_{i,k}$ and $\Delta\boldsymbol{\omega}_{i,k}$ via (15) and (16).
    Exchanges $\Delta\boldsymbol{c}_{i,k}$ and $\Delta\boldsymbol{\omega}_{i,k}$ with the **central server**.
  end
  Updates the control variables as in (17).
  Updates the global meta model as in (18).
end
The **central server** broadcasts $\boldsymbol{\omega}_k^{\text{Global}}$ to all clients.
for $i = 0$ to $N-1$
  Client $\boldsymbol{C}_i$ updates the personalized model as in (11).
Output: personalized model for all clients.
$\boldsymbol{\omega}_i^{\text{Per}} = [\boldsymbol{\omega}_0^{\text{Per}}, \boldsymbol{\omega}_1^{\text{Per}}, \dots, \boldsymbol{\omega}_{N-1}^{\text{Per}}]$

IV. CASE STUDIES

The adopted dataset consists of historical load data from ten different distribution transformer supply zones located in different cities of China. The data has strong heterogeneity. The historical load data has a time resolution of 15 minutes. 75% of the data forms the training dataset while the remaining 25% forms the test dataset. To achieve the higher STLF accuracy, the weather data is introduced as the concomitant variable for STLF. For the sake of weather data accessibility, this paper uses a numerical weather prediction-free method (i.e., using the measured weather data instead of numerical weather prediction data) [35] and the measured weather data is from National Climatic Data Center (NCDC) [36]. Besides, to mitigate the effects of temporal resolution mismatch between load data and measured weather data, an approximate alternative method [37] is used in this paper. Two metrics, Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), are used to evaluate the performance of the proposed method as well as those compared methods. Specifically, RMSE quantifies the absolute error while MAPE quantifies the relative error in the form of percentages. Two metrics are expressed as

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{y}_m - y_m)^2} \quad (19)$$

$$\text{MAPE} = \frac{1}{M}\sum_{m=1}^{M}\left|\frac{\hat{y}_m - y_m}{y_m}\right| \quad (20)$$

where y_m represents the forecast value and $\hat{y}_m$ represents the actual value. M is the length of forecasting values. Intuitively, smaller RMSE and MAPE values indicate a better performance of the load forecasting method.

All the programs are implemented with Python 3.7 and PyTorch 1.13, and cases are executed on a PC equipped with Intel Core i9-12900K CPU, NVIDIA GeForce RTX 3090 GPU, and 128GB RAM.

A. Comparisons of Different Training Modes

To verify the effectiveness of the proposed method in terms of the forecasting accuracy and generalization ability, it is compared with three different deep learning models, including ANN, LSTM and CNN-LSTM, with different training modes, including centralized training mode, fully local training mode, FL-Avg training mode, and FMAML-SCA training mode as detailed in the following:

Centralized training mode (Cen): A central server utilizes the historical load data of all clients to centrally train a global forecasting model for all clients.

Fully local training mode (Loc): Each client only uses its own load data to train the forecasting model.

FL-Avg training mode (FL-Avg): A global forecasting model is trained with the FL framework using the Avg aggregation algorithm.

FMAML-SCA training mode (FMAML-SCA): Personalized models for individual clients are trained with the proposed FMAML method and the SCA aggregation algorithm.

The parameter settings are shown in TABLE III. The results are summarized in TABLE IV and the forecasted load curves of six clients with the four training modes are shown in Fig. 5.

TABLE III
PARAMETER SETTINGS OF DIFFERENT TRAINING MODE

Mode	α	β	γ	δ	T	K
Cen	-	0.001	-	-	-	400
Loc	-	0.001	-	-	400	-
FL-Avg	-	0.001	-	-	4	100
FMAML-SCA	0.01	0.001	1	1e-6	4	100

The Cen mode exhibits the best performance in most cases, as it can directly obtain global information from all distribution transformer supply zones to train the load forecasting model. Not surprisingly, the Loc training mode exhibits the worst performance. It is reasonable because limited training data would result in overfitting and compromises the forecasting accuracy. Due to the lack of considerations for personalization and the impact of CD phenomenon, the forecasting accuracy of the FL-Avg training mode shows a large gap to that of centralized training mode and the FMAML-SCA training mode. Furthermore, the historical load data at the distribution transformer supply zones contain strong noises, which also magnify the impact of CD phenomenon and further deteriorates the forecasting accuracy of the FL-Avg training mode. The FMAML-SCA training mode performs as well as the centralized training mode, demonstrating its effectiveness in terms of forecasting accuracy while being capable of ensuring data privacy.

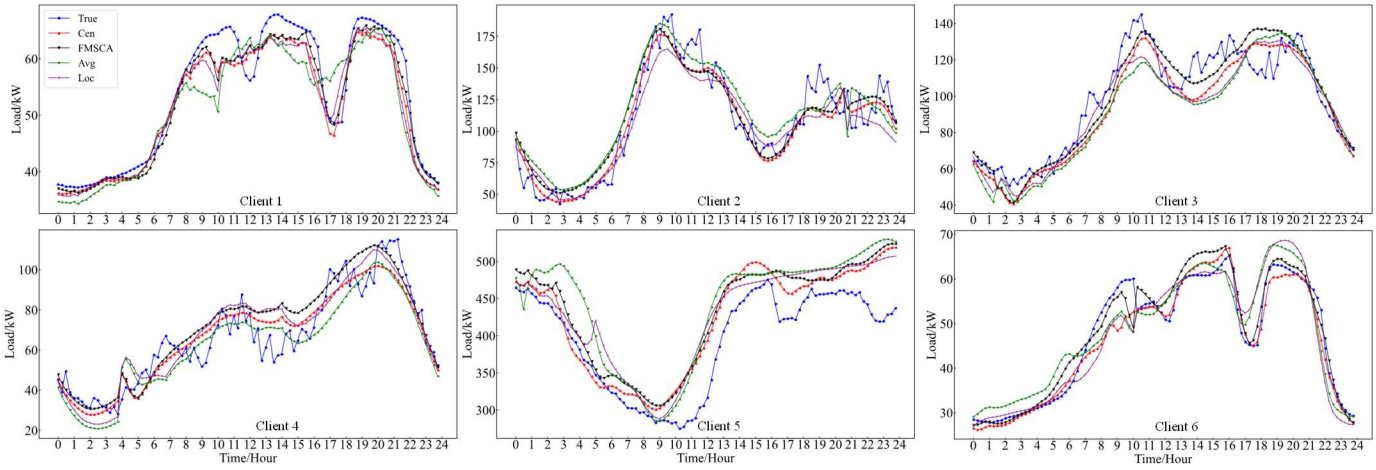

Fig. 5. Comparison of different training modes

TABLE IV
COMPARISON OF DIFFERENT TRAINING MODES

Deep Learning Model	Training Mode	ANN				LSTM				CNN-LSTM			
		Cen	Loc	FL-Avg	FMAML-SCA	Cen	Loc	FL-Avg	FMAML-SCA	Cen	Loc	FL-Avg	FMAML-SCA
1	RMSE/kW	66.76	71.32	41.15	45.08	34.99	70.44	44.33	37.12	39.44	82.61	38.15	34.94
	MAPE/%	20.25	30.76	15.06	23.20	10.86	32.29	20.54	12.00	12.28	35.39	16.91	10.84
2	RMSE/kW	38.77	38.14	51.38	39.26	13.18	43.08	37.43	15.20	34.13	18.44	43.43	27.97
	MAPE/%	19.33	13.98	37.44	17.54	10.73	17.76	19.36	15.42	16.77	10.62	22.06	14.16
3	RMSE/kW	19.33	32.17	11.65	11.32	11.48	21.30	15.20	12.53	8.70	16.90	28.77	19.42
	MAPE/%	8.65	18.51	13.26	3.13	5.65	13.92	9.46	8.09	3.97	17.09	20.55	12.51
4	RMSE/kW	66.49	105.74	72.12	42.82	29.71	97.46	65.22	37.08	39.47	148.98	81.84	31.67
	MAPE/%	47.39	64.63	49.12	35.42	23.20	51.94	40.84	33.27	30.02	74.74	46.21	32.84
5	RMSE/kW	54.26	52.66	53.66	46.22	75.43	63.27	60.04	51.87	42.25	65.40	77.33	44.66
	MAPE/%	19.58	26.88	13.63	17.22	30.77	30.53	29.28	16.09	15.20	27.15	34.80	12.74
6	RMSE/kW	5.08	21.65	45.34	23.05	19.53	36.36	25.60	12.83	11.61	39.06	15.15	10.47
	MAPE/%	6.41	18.61	31.67	15.71	29.54	24.22	19.08	15.59	11.98	28.16	11.30	14.19
7	RMSE/kW	38.66	68.96	71.94	17.38	9.51	50.96	34.49	33.57	6.61	42.26	28.42	27.37

This article has been accepted for publication in IEEE Transactions on Power Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2024.3393017

	MAPE/%	10.56	32.29	30.95	6.00	4.33	24.35	23.65	20.56	5.85	28.63	16.85	11.13
8	RMSE/kW	9.32	39.77	28.45	27.26	3.69	37.07	23.59	32.53	30.25	33.05	27.84	26.70
	MAPE/%	6.05	19.22	17.09	10.97	2.53	30.53	12.30	12.70	17.69	27.17	10.92	10.66
9	RMSE/kW	56.49	107.71	86.15	46.19	61.20	80.55	67.77	52.26	62.95	81.97	56.06	53.17
	MAPE/%	37.25	67.26	58.75	37.46	36.14	66.61	47.97	40.04	45.40	62.33	50.99	36.57
10	RMSE/kW	18.55	54.38	25.59	32.12	46.49	41.39	28.22	39.17	20.62	39.04	30.15	23.22
	MAPE/%	5.82	42.34	19.29	18.04	15.97	45.09	24.93	9.07	6.54	34.79	17.02	15.54
Average	RMSE/kW	37.38	59.25	48.75	33.06	30.52	54.19	40.19	32.41	29.60	56.77	42.89	29.96
	MAPE/%	18.13	33.45	28.63	18.47	16.97	33.72	25.74	18.28	16.57	34.61	24.66	17.11

B. Comparisons of Aggregation Algorithms.

TABLE V
SETTINGS OF DIFFERENT FEDERATED LEARNING METHODS

Method	α	β	γ	δ	T	K
FL-Avg	-	0.001	-	-	4	100
FL-SCA	-	0.001	1	-	4	100
FLPer-Avg	-	0.001	-	-	4	100
FLPer-SCA	-	0.001	1	-	4	100
FMAML-Avg	0.01	0.001	-	1e-6	4	100
FMAML-SCA	0.01	0.001	1	1e-6	4	100

To demonstrate the superiority of the proposed FMAML-SCA method over other existing FL-based methods, we compare it with five typical methods, including FL+Avg/SCA, FLPer+Avg/SCA, and FMAML+Avg. The parameter settings of those methods are shown in TABLE V and the results of six clients are shown in TABLE VI.

As shown in TABLE VI, for FMAML-SCA and the five compared methods, i.e., FL-Avg, FL-SCA, FLPer-Avg, FLPer-SCA, and FMAML-Avg, the SCA aggregation algorithm performs much better than the Avg aggregation algorithm with the same settings to T and K. This shows the higher optimality of the model trained with SCA aggregation algorithm compared with the model trained with the Avg aggregation algorithm within the same number of iterations and, as well, demonstrates that SCA aggregation algorithm has a better convergence ability than the Avg aggregation algorithm.

In addition, either in the group of FL-Avg, FLPer-Avg, and FMAML-Avg or in the group of FL-SCA, FLPer-SCA, and FMAML-SCA, the proposed FMAML method shows its ability of considerably enhancing the forecasting performance with all three studied deep network models. The forecasting performance with FMAML method is significantly better than that of FL since the latter misses the personalized fine-tuning process for individual clients. Furthermore, regardless of combining with Avg or SCA as the aggregation algorithm and adopting any of the three deep learning models, the performance of FLPer method is not obviously better than that of FL, although FLPer enables a personalized fine-tuning process. This is because the personalized fine-tuning process is completely disconnected with the federated training process, which eventually lowers its forecasting accuracy, and, in some cases, even to a degree worse than the FL. In sum, the proposed FMAML framework outperforms due to integration of the novel personalization technique, MAML, and the SCA aggregation algorithm.

TABLE VI
COMPARISON OF DIFFERENT FEDERATED LEARNING METHODS (MAPE/%)

Model	Mode	1	2	3	4	5	6	7	8	9	10	Average
ANN	FL-Avg	15.06	37.44	13.26	49.12	13.63	31.67	30.95	17.09	58.75	19.29	28.63
	FL-SCA	15.55	23.90	2.66	38.17	23.79	27.30	9.95	28.66	43.80	16.62	23.04
	FPer-Avg	24.27	37.63	19.78	38.30	26.87	20.77	10.08	26.51	69.08	20.42	29.37
	FPer-SCA	13.78	28.07	5.33	39.29	40.75	23.37	8.81	21.08	55.12	17.68	25.33
	FMAML-Avg	16.98	31.91	5.01	33.73	26.60	17.76	7.79	12.70	47.87	16.86	21.72
	FMAML-SCA	23.20	17.54	3.13	35.42	17.22	15.71	6.00	10.97	37.46	18.04	18.47
LSTM	FL-Avg	20.54	19.36	9.46	40.84	29.28	19.08	23.65	12.30	47.97	24.93	25.74
	FL-SCA	22.14	23.73	15.81	44.56	15.79	13.91	19.13	19.94	48.26	18.50	24.18
	FPer-Avg	27.72	14.16	9.19	60.13	19.18	11.36	10.20	22.23	56.98	11.67	24.28
	FPer-SCA	15.01	27.73	17.16	29.85	16.39	17.73	18.94	19.42	43.24	10.19	21.57
	FMAML-Avg	10.44	24.75	14.13	30.78	18.41	23.70	17.63	17.42	58.85	15.98	23.21
	FMAML-SCA	12.00	15.42	8.09	33.27	16.09	15.59	20.56	12.70	40.04	9.07	18.28
CNN-LSTN	FL-Avg	38.15	22.06	20.55	46.21	34.80	11.30	16.85	10.92	50.99	17.02	24.66
	FL-SCA	16.95	21.20	23.37	29.19	18.73	19.87	23.00	25.98	38.46	20.05	23.68
	FPer-Avg	11.66	18.42	26.56	39.32	23.92	18.41	25.18	22.49	63.27	19.71	26.89
	FPer-SCA	23.42	15.08	16.22	29.78	14.86	15.43	16.64	17.73	38.62	10.63	19.84
	FMAML-Avg	10.52	18.34	8.38	25.89	15.65	15.06	19.86	18.99	58.35	17.02	20.81
	FMAML-SCA	10.84	14.16	12.51	30.84	12.74	14.19	11.13	12.66	36.57	15.54	17.11

C. Robustness to Heterogeneous Data

The convergence of various FL based methods is easily affected by data heterogeneity. To validate the robustness of the proposed FMAML method against data heterogeneity, it is further compared with FL and FLPer methods under different degrees of data heterogeneity in the training dataset.

First, the heterogeneity degrees of the datasets from different distribution transformer supply zones need to be quantitatively analyzed. The maximum mean discrepancy (MMD) is taken as the metric to evaluate the heterogeneity degree. MMD is a widely used loss function in the field of transfer learning, which can quantify the distance between two different data distributions in a regenerated Hilbert space. A larger value of MMD indicates stronger heterogeneity between

the two data distributions, while zero MMD means the two sets of data are identically distributed. MMD can be expressed as in (21),

$$\mathrm{MMD}(\boldsymbol{X},\boldsymbol{Y}) = \left\| \frac{1}{A^2}\sum_{a=1}^{A}\sum_{b=1}^{A} k(\boldsymbol{x}_a,\boldsymbol{x}_b) - \frac{1}{2AB}\sum_{a=1}^{A}\sum_{b=1}^{B} k(\boldsymbol{x}_a,\boldsymbol{y}_b) + \frac{1}{B^2}\sum_{a=1}^{B}\sum_{b=1}^{B} k(\boldsymbol{y}_a,\boldsymbol{y}_b) \right\| \quad (21)$$

where A and B are respectively the numbers of rows of matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$; $k(\boldsymbol{x}_a,\boldsymbol{x}_b) = e^{\frac{-\|x_a - x_b\|^2}{\sigma}}$ is a Gaussian kernel function with σ being a hyperparameter.

Without loss of generality, we set $\sigma = [0.25, 0.5, 1, 2, 4]$, and calculate all the output values of the Gaussian kernel function corresponding to each of these values in σ and take the mean value as the ultimate MMD. The MMD for the training datasets from 10 distribution transformer supply zones are depicted in a heat map as Fig. 6. The diagonal elements are always zeros representing the distances from oneself.

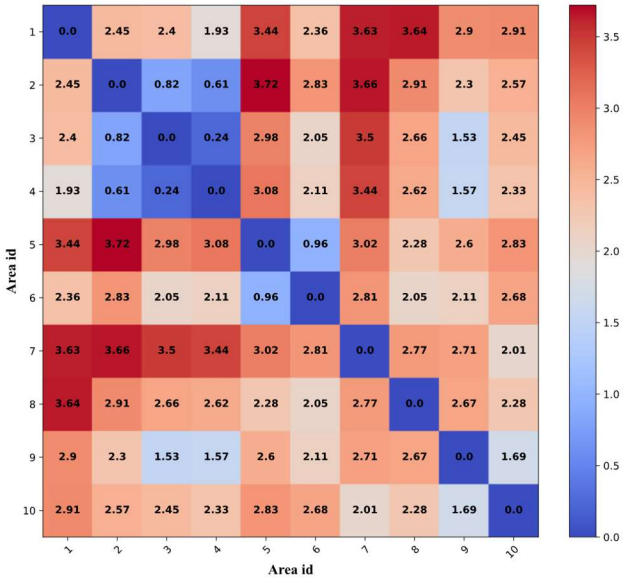

Fig. 6. MMD heatmap of distribution transformer supply zones

Four typical pairs are chosen from Fig. 6. They are shown below:

Pair 1: Datasets of area 3 and area 4 (MMD = 0.24)
Pair 2: Datasets of area 3 and area 9 (MMD = 1.53)
Pair 3: Datasets of area 6 and area 9 (MMD = 2.11)
Pair 4: Datasets of area 1 and area 5 (MMD = 3.44)

We take the CNN-LSTM as the deep learning model and set MAPE $< 50\%$ and MAPE $< 30\%$ as the stopping criteria. The numbers of the global model update iterations taken the six methods to meet the stopping criteria with different learning methods are compared in TABLE VII and Fig. 7-8.

TABLE VII
SENSITIVITY ANALYSIS OF HETEROGENEITY

Method	Pair	MAPE <50%	MAPE <30%
FL-Avg [20]-[23]	Pair 1	11 (1.0×)	32 (1.0×)
	Pair 2	70 (6.4×)	218 (6.8×)
	Pair 3	91 (8.3×)	500+ (15.6+×)
	Pair 4	173 (15.7×)	500+ (15.6+×)
FL-SCA	Pair 1	10 (1.0×)	19 (1.0×)
	Pair 2	41 (4.1×)	63 (3.3×)
	Pair 3	48 (4.8×)	120 (6.3×)
	Pair 4	83 (8.3×)	364 (19.2×)
FLPer-Avg [24]	Pair 1	10 (1.0×)	22 (1.0×)
	Pair 2	58 (5.8×)	174 (7.9×)
	Pair 3	87 (8.7×)	500+ (22.7+×)
	Pair 4	167 (16.7×)	500+ (22.7+×)
FLPer-SCA	Pair 1	9 (1.0×)	18 (1.0×)
	Pair 2	35 (3.9×)	76 (4.2×)
	Pair 3	52 (5.8×)	118 (6.6×)
	Pair 4	139 (15.4×)	383 (22.3×)
FMAML-Avg	Pair 1	10 (1.0×)	29 (1.0×)
	Pair 2	49 (4.9×)	162 (5.6×)
	Pair 3	75 (7.5×)	193 (6.7×)
	Pair 4	162 (16.2×)	500+ (17.2+×)
FMAML-SCA	Pair 1	9 (1.0×)	18 (1.0×)
	Pair 2	25 (2.8×)	57 (3.2×)
	Pair 3	30 (3.3×)	103 (5.7×)
	Pair 4	69 (7.6×)	263 (14.6×)

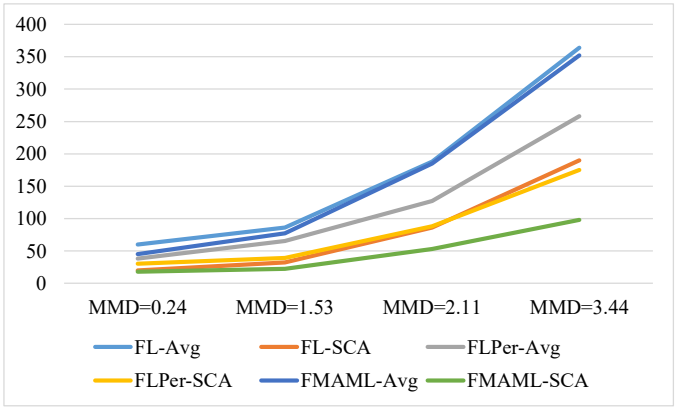

Fig. 7. Sensitivity analysis of heterogeneity (MAPE<50%)

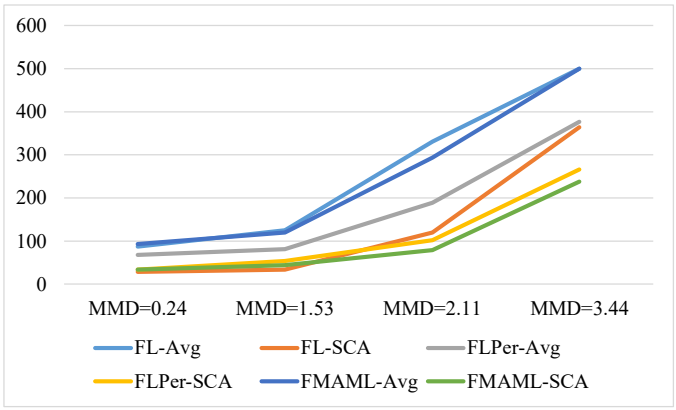

Fig. 8. Sensitivity analysis of heterogeneity (MAPE<30%)

It can be seen from TABLE VII and Fig. 7-8, the proposed FMAML-SCA method is evidently superior to the other compared methods regardless of the degree of data heterogeneity. As the degree of data heterogeneity increases, the number of iterations taken by six methods to satisfy the stopping criteria increases as well. However, the proposed FMAML-SCA method performs the slowest growing with the least number of iterations, followed by FL-SCA and FLPer-SCA. This indicates that the SCA aggregation algorithm is significantly more resistant to heterogeneous data than the Avg aggregation algorithm. Moreover, the FLPer framework performs even worse than the FL framework in some cases, while the FMAML framework considerably outperforms it, which further verifies the excellent personalization capabilities of MAML. In summary, the proposed FMAML-SCA based distributed STLF method for distribution transformer supply zones presents stronger robustness against data heterogeneity than the methods in the existing works [20]-[24].

D. Robustness to Packet Dropout

In this section, the packet dropout is taken as an example to study the robustness of the proposed training frameworks when encountering anomalous communication. The packet dropout can be modeled as a stochastic process considering its unpredictability. Gilbert-Elliott model is used to model the packet dropout event.

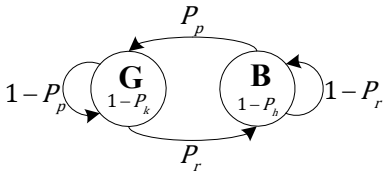

Fig. 9. The Gilbert-Elliott model

As is shown in Fig. 9, there are two states in Gilbert-Elliott model, i.e., the good (G) and the bad (B). P_r and P_p respectively denote the probabilities of transferring form B to G and from G to B. $(1-P_k)$ and $(1-P_h)$ are respectively the probabilities of packet dropout staying state G and B. The Gilbert-Elliot model is essentially a Markov chain that consists of states B and G. The state transition matrix P can be expressed as:

$$P = \begin{Bmatrix} 1-P_p & P_p \\ P_r & 1-P_r \end{Bmatrix} \tag{22}$$

When $P_p \in (0,1)$ and $P_r \in (0,1)$, the Markov chain can reach the stationary state. Let P_G and P_B respectively denote the probabilities of G and B in the stationary state. According to $SP = S$ where $S = (P_G, P_B)$ and $P_G + P_B = 1$, the P_G and P_B can be respectively calculated as:

$$P_G = \frac{P_r}{P_p+P_r} \text{ and } P_B = \frac{P_p}{P_p+P_r} \tag{23}$$

Thereby, the probability of packet dropout in the stationary state P_E can be calculated as (24) [38].

$$P_E = P_G(1-P_k) + P_B(1-P_h) \tag{24}$$

We use P_p=0.00253 and P_r=0.25 for all cases [38]. As shown in TABLE VIII, we set P_h=0.5 and P_k can be calculated to obtain the desired P_E.

TABLE VIII
PARAMETER SETTINGS FOR DIFFERENT PACKET DROPOUT RATE

Parameters	Case 0	Case 1	Case 2	Case 3	Case 4
P_k	1	0.995	0.955	0.904	0.80
P_h	0.5	0.5	0.5	0.5	0.5
P_E	0%	1%	5%	10%	20%

TABLE IX
ROBUSTNESS ANALYSIS OF PACKET DROPOUT

Method	Case	MAPE<50%	MAPE<30%
FL-Avg [20]-[23]	Case 0	51 (1.0×)	81 (1.0×)
	Case 1	60 (1.2×)	87 (1.1×)
	Case 2	86 (1.7×)	125 (1.5×)
	Case 3	188 (3.7×)	331 (4.1×)
	Case 4	364 (7.1×)	500+ (6.2+×)
FL-SCA	Case 0	19 (1.0×)	25 (1.0×)
	Case 1	20 (1.0×)	28 (1.1×)
	Case 2	32 (1.7×)	34 (1.4×)
	Case 3	86 (4.5×)	120 (4.8×)
	Case 4	190 (10.0×)	364 (14.6×)
FLPer-Avg [24]	Case 0	32 (1.0×)	46 (1.0×)
	Case 1	38 (1.2×)	68 (1.5×)
	Case 2	65 (2.0×)	81 (1.8×)
	Case 3	127 (4.0×)	189 (4.1×)
	Case 4	258 (8.1×)	377 (8.2×)
FLPer-SCA	Case 0	24 (1.0×)	35 (1.0×)
	Case 1	30 (1.3×)	34 (1.0×)
	Case 2	39 (1.6×)	54 (1.5×)
	Case 3	88 (3.7×)	102 (2.9×)
	Case 4	175 (7.3×)	266 (7.6×)
FMAML-Avg	Case 0	42 (1.0×)	75 (1.0×)
	Case 1	45 (1.1×)	93 (1.2×)
	Case 2	77 (1.8×)	120 (1.6×)
	Case 3	185 (4.4×)	293 (3.9×)
	Case 4	352 (8.4×)	500+ (6.7+×)
FMAML-SCA	Case 0	16 (1.0×)	33 (1.0×)
	Case 1	18 (1.1×)	34 (1.0×)
	Case 2	22 (1.4×)	44 (1.5×)
	Case 3	53 (3.3×)	79 (2.4×)
	Case 4	98 (6.1×)	238 (7.2×)

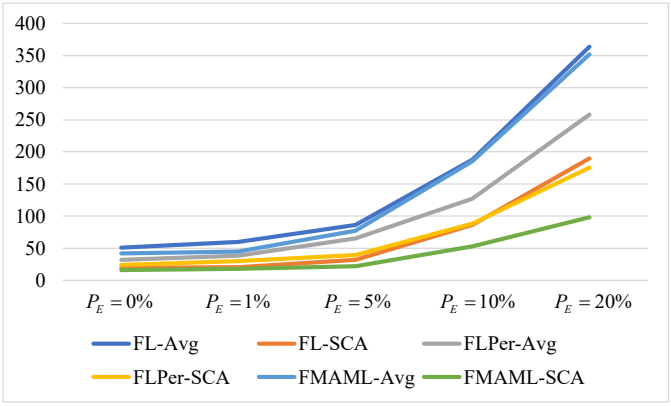

Fig. 10. Sensitivity analysis of packet dropout (MAPE<50%)

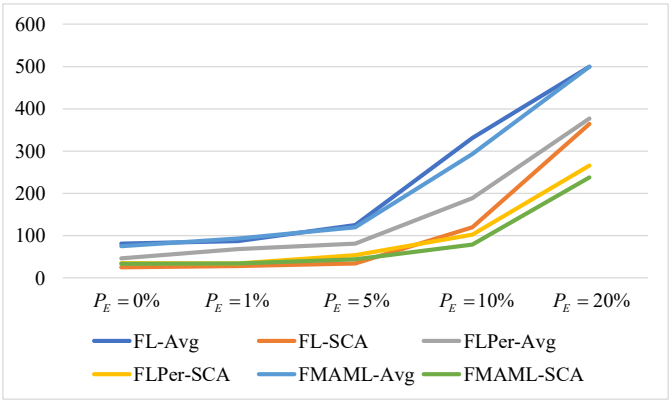

Fig. 11. Sensitivity analysis of packet dropout (MAPE<30%)

It is assumed that packet dropout rates of the central server and the clients are P_E. Specifically, if the central server fails to receive information from client $\boldsymbol{C}_i$, it will utilize the information of $\boldsymbol{C}_i$ from the previous iteration to substitute the missing information and continue with the interaction process. The CNN-LSTM is taken as the deep learning model and the stopping criteria is the same as in Section V.C. The number of iterations taken by the six methods to reach the stopping criteria are shown in TABLE IX and Fig. 10-11.

From TABLE IX and Fig. 10-11, the proposed FMAML-SCA method is obviously superior to the other compared methods regardless of the packet dropout rate. As the packet dropout rate increases, the needed iterations for the six methods to reach the stopping criteria also rises. However, the proposed FMAML-SCA method is least affected by the increasing packet dropout rate. In sum, the proposed FMAML-SCA based distributed STLF method for distribution transformer supply zones presents stronger robustness against

packet dropout than the methods in the existing works [20]-[24].

V. Conclusion

For distribution transformer supply zones, a FMAML based STLF approach is proposed to improve the forecasting accuracy while ensuring the data privacy. Specifically, MAML is combined with FL to build high-quality personalized models for the clients and the SCA aggregation algorithm is applied to mitigate the impact of CD phenomenon. Numerical results show the high forecasting accuracy of the proposed method and demonstrate its stronger robustness to data heterogeneity and packet dropout than the existing methods. In the future work, we intend to improve the proposed FMAML framework and provide tailor-made forecasting models with different structures for clients with the focus of further enhancing the adaptability and accuracy.

Appendix

A. Assumptions

Without loss of generality, several assumptions are made to simplify the follow-up analysis.

(Assumption 1) Bounded gradient dissimilarity:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2 \le G^2 + B^2\left\|\nabla F(\boldsymbol{\omega}_k^{\text{Global}})\right\|^2 \quad \text{(A1)}$$

where G and B are constants and $G \ge 0$, $B \ge 1$.

If F_i is l-smooth:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla F_i\left(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i\right)\right\|^2 \le G^2 + 2lB^2\left(F\left(\boldsymbol{\omega}_k^{\text{Global}}\right) - F(\boldsymbol{\omega}^*)\right) \quad \text{(A2)}$$

(Assumption 2) Bounded Hessian dissimilarity:

$$\left\|\nabla^2 F_i(\boldsymbol{\omega}_k^{\text{Global}}) - \nabla^2 F(\boldsymbol{\omega}_k^{\text{Global}})\right\| \le \delta \quad \text{(A3)}$$

where δ is a constant.

(Assumption 3) F_i is μ-convex: for any i, $\boldsymbol{x}$, $\boldsymbol{y}$, there exists a constant $\mu \ge 0$ let

$$F_i(\boldsymbol{x}) \ge F_i(\boldsymbol{y}) + (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\nabla F_i(\boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \quad \text{(A4)}$$

(Assumption 4) $g_i(x) = \nabla F_i(\boldsymbol{x};\boldsymbol{D}_i)$ is the unbiased stochastic gradient of F_i with bounded variance σ:

$$\mathbb{E}[\|g_i(\boldsymbol{x}) - \nabla F_i(\boldsymbol{x})\|^2] \le \sigma^2 \quad \text{(A5)}$$

(Assumption 5) F_i is l-smooth: for any i, $\boldsymbol{x}$, $\boldsymbol{y}$

$$F_i(\boldsymbol{x}) \le F_i(\boldsymbol{y}) + (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\nabla F_i(\boldsymbol{y}) + \frac{l}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \quad \text{(A6)}$$

$$\|\nabla F_i(\boldsymbol{x}) - \nabla F_i(\boldsymbol{y})\| \le l\|\boldsymbol{x}-\boldsymbol{y}\| \quad \text{(A7)}$$

(Assumption 6) For any μ-convex and l-smooth function F_i, it yields: for any i, $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$

$$(\boldsymbol{z}-\boldsymbol{y})^{\mathrm{T}}\nabla F_i(\boldsymbol{x}) \ge F_i(\boldsymbol{z}) - F_i(\boldsymbol{y}) + \frac{\mu}{4}\|\boldsymbol{y}-\boldsymbol{z}\|^2 - l\|\boldsymbol{z}-\boldsymbol{x}\|^2 \quad \text{(A8)}$$

B. Lemmas

Next, several lemmas will be given or proved to support the convergence analysis.

(Lemma 1 [33]**)** For τ random variables with conditional mean $\xi_i = \mathbb{E}[\Xi_i|\Xi_{i-1},\dots\Xi_1]$ and bounded variances $\sigma^2 \ge \mathbb{E}[\|\Xi_i - \xi_i\|^2]$, we have:

$$\mathbb{E}[\|\textstyle\sum_{i=1}^{\tau}\Xi_i\|^2] \le \|\sum_{i=1}^{\tau}\xi_i\|^2 + \tau^2\sigma^2 \quad \text{(A9)}$$

$$\mathbb{E}[\|\textstyle\sum_{i=1}^{\tau}\Xi_i\|^2] \le 2\|\sum_{i=1}^{\tau}\xi_i\|^2 + 2\tau\sigma^2 \quad \text{(A10)}$$

(Lemma 2) For any $8(1+B^2)lT\beta\gamma \le 1$, we have:

$$\mathbb{E}\left\|\boldsymbol{\omega}_k^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 \le \left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 + \frac{1}{T|S|}\eta^2\sigma^2 + \left(1-\frac{|S|}{N}\right)\frac{4\eta^2}{S}G^2 - \eta\left(\mathbb{E}\left[F\left(\boldsymbol{\omega}_k^{\text{Global}}\right)\right] - F(\boldsymbol{\omega}^*)\right) + 3l\eta\varepsilon_k \quad \text{(A11)}$$

where $\eta = T\beta\gamma$, $\varepsilon_k = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t} - \boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]$, and $\boldsymbol{\omega}^*$ is the optimal solution.

Proof of **Lemma 2:**

According to (7)(16)(18) in the sections II and III, we have:

$$\Delta\boldsymbol{\omega}_k = -\frac{\eta}{T|S|}\sum_{i\in S}\sum_{t=1}^{T} g_i(\boldsymbol{\omega}_{i,k,t-1}) \quad \text{(A12)}$$

$$\mathbb{E}[\Delta\boldsymbol{\omega}_k] = -\frac{\eta}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\nabla F(\boldsymbol{\omega}_{i,k,t-1},\boldsymbol{D}_i) \quad \text{(A13)}$$

Thus, we have (A14).

Then, applying the **Lemma 1** to (A14), it yields as in (A15).

$$\mathbb{E}\left[\left\|\boldsymbol{\omega}_k^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2\right] = \mathbb{E}\left[\left\|\boldsymbol{\omega}_k^{\text{Global}} + \Delta\boldsymbol{\omega}_k - \boldsymbol{\omega}^*\right\|^2\right] = \left\|\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 + \chi_1 + \eta^2\mathbb{E}\left\|\frac{1}{T|S|}\sum_{i\in S}\sum_{t=1}^{T} g_i(\boldsymbol{\omega}_{i,k-1,t})\right\|^2 \quad \text{(A14)}$$

where $\chi_1 = -\frac{2\eta}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\langle\nabla F_i\left(\boldsymbol{\omega}_{i,k-1,t},D_i\right),\left(\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right)\right\rangle$.

$$\mathbb{E}\left[\left\|\boldsymbol{\omega}_k^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2\right] = \mathbb{E}\left[\left\|\boldsymbol{\omega}_k^{\text{Global}} + \Delta\boldsymbol{\omega}_k - \boldsymbol{\omega}^*\right\|^2\right] = \left\|\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 + \chi_1 + \eta^2\mathbb{E}\left\|\frac{1}{T|S|}\sum_{i\in S}\sum_{t=1}^{T} g_i\left(\boldsymbol{\omega}_{i,k-1,t}\right)\right\|^2 \le \left\|\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 + \chi_1 + \chi_2 + \frac{\eta^2\sigma^2}{T|S|} \quad \text{(A15)}$$

where $\chi_2 = \eta^2\mathbb{E}\left\|\frac{1}{T|S|}\sum_{i\in S}\sum_{t=1}^{T}\nabla F_i(\boldsymbol{\omega}_{i,k-1,t},\boldsymbol{D}_i)\right\|^2$.

For χ_1, we use the (**Assumption 6**) and it yields:

$$\begin{aligned}\chi_1 &= -\frac{2\eta}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\langle\nabla F_i\left(\boldsymbol{\omega}_{i,k-1,t},D_i\right),\left(\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right)\right\rangle \\ &\le \frac{2\eta}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} F_i(\boldsymbol{\omega}^*,\boldsymbol{D}_i) - F_i\left(\boldsymbol{\omega}_{k-1}^{\text{Global}},\boldsymbol{D}_i\right) + l\left\|\boldsymbol{\omega}_{i,k-1,t} - \boldsymbol{\omega}_{k-1}^{\text{Global}}\right\|^2 - \frac{\mu}{4}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 \\ &= -2\eta\left(F(\boldsymbol{\omega}^*) - F(\boldsymbol{\omega}_{k-1}^{\text{Global}}) + \frac{\mu}{4}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2\right) + 2l\eta\varepsilon_{k-1}\end{aligned} \quad \text{(A16)}$$

For χ_2, we have:

$$\begin{aligned}\chi_2 &= \eta^2\mathbb{E}\left\|\frac{1}{T|S|}\sum_{i\in S}\sum_{t=1}^{T}\nabla F_i(\boldsymbol{\omega}_{i,k-1,t},\boldsymbol{D}_i)\right\|^2 \\ &\le \eta^2\mathbb{E}\left\|\frac{1}{T|S|}\sum_{i\in S}\sum_{t=1}^{T}\nabla F_i\left(\boldsymbol{\omega}_{i,k-1,t},\boldsymbol{D}_i\right) - \nabla F_i\left(\boldsymbol{\omega}_{k-1}^{\text{Global}},\boldsymbol{D}_i\right) + \nabla F_i\left(\boldsymbol{\omega}_{k-1}^{\text{Global}},\boldsymbol{D}_i\right)\right\|^2 \\ &\le \eta^2\mathbb{E}\left\|\frac{1}{T|S|}\sum_{i\in S}\sum_{t=1}^{T}\nabla F_i\left(\boldsymbol{\omega}_{i,k-1,t},\boldsymbol{D}_i\right) - \nabla F_i\left(\boldsymbol{\omega}_{k-1}^{\text{Global}},\boldsymbol{D}_i\right)\right\|^2 + 2\eta^2\mathbb{E}\left\|\frac{1}{|S|}\sum_{i\in S}\nabla F_i\left(\boldsymbol{\omega}_{k-1}^{\text{Global}},\boldsymbol{D}_i\right)\right\|^2 \\ &\le \frac{2\eta^2}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left\|\nabla F_i\left(\boldsymbol{\omega}_{i,k-1,t},\boldsymbol{D}_i\right) - \nabla F_i\left(\boldsymbol{\omega}_{k-1}^{\text{Global}},\boldsymbol{D}_i\right)\right\|^2 + 2\eta^2\mathbb{E}\left\|\frac{1}{|S|}\sum_{i\in S}\nabla F_i\left(\boldsymbol{\omega}_{k-1}^{\text{Global}},\boldsymbol{D}_i\right) - \nabla F\left(\boldsymbol{\omega}_{k-1}^{\text{Global}}\right) + \nabla F\left(\boldsymbol{\omega}_{k-1}^{\text{Global}}\right)\right\|^2 \\ &\le \frac{2\eta^2 l^2}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left\|\boldsymbol{\omega}_{i,k-1,t} - \boldsymbol{\omega}_{k-1}^{\text{Global}}\right\|^2 + 2\eta^2\left\|\nabla F\left(\boldsymbol{\omega}_{k-1}^{\text{Global}}\right)\right\|^2 + 4\left(1-\frac{|S|}{N}\right)\eta^2\frac{1}{N|S|}\sum_{i=1}^{N}\left\|\nabla F_i(\boldsymbol{\omega}_{k-1}^{\text{Global}},\boldsymbol{D}_i)\right\|^2 \\ &\le 2\eta^2 l^2\varepsilon_{k-1} + 8\eta^2 l\left(B^2+1\right)\left(F\left(\boldsymbol{\omega}_{k-1}^{\text{Global}}\right) - F(\boldsymbol{\omega}^*)\right) + \frac{4\eta^2}{|S|}\left(1-\frac{|S|}{N}\right)G^2\end{aligned} \quad \text{(A17)}$$

where we use **Lemma 1** in the second and fourth inequality, **Assumption 1** in the fifth inequality.

Combining (A16) and (A17), we have:

$$\mathbb{E}\left[\left\|\boldsymbol{\omega}_k^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2\right] \le (1-\tfrac{\mu\eta}{2})\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2-(2\eta-8\eta^2 l(B^2+1))(F(\boldsymbol{\omega}_{k-1}^{\text{Global}})-F(\boldsymbol{\omega}^*))+2(1+\eta l)\eta l\varepsilon_{k-1}+\frac{\eta^2\sigma^2}{T|S|}+\frac{4\eta^2}{|S|}(1-\frac{|S|}{N})G^2 \quad \text{(A18)}$$

(Lemma 3) For any $8(1+B^2)lT\beta\gamma \le 1$, $2 \le T \le 65$ we have:

$$3\eta l\varepsilon_k \le \frac{2\eta}{3}\left(\mathbb{E}\left[F(\boldsymbol{\omega}_{k-1}^{\text{Global}})\right]-F(\boldsymbol{\omega}^*)\right)+\frac{\eta^2\sigma^2}{2T\gamma^2}+18l\eta^3G^2 \quad \text{(A19)}$$

Proof of **Lemma 3:**

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t+1}-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]\\
&=\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t}-\beta g_i(\boldsymbol{\omega}_{i,k,t})-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]\\
&\le\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t}-\beta\nabla F_i(\boldsymbol{\omega}_{i,k,t},\boldsymbol{D}_i)-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]+\beta^2\sigma^2\\
&\le\left(1+\frac{1}{T-1}\right)\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t}-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]+(1+T-1)\beta^2\left\|\nabla F_i(\boldsymbol{\omega}_{i,k,t},D_i)\right\|^2+\beta^2\sigma^2\\
&=\left(1+\frac{1}{T-1}\right)\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t}-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]+\frac{\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_{i,k,t},D_i)-\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)+\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{\eta^2\sigma^2}{T^2\gamma^2}\\
&\le(1+\frac{1}{T-1})\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t}-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]+\frac{2\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_{i,k,t},D_i)-\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{2\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{\eta^2\sigma^2}{T^2\gamma^2}\\
&\le\left(1+\frac{1}{T-1}+\frac{2\eta^2l^2}{T\gamma}\right)\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t}-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]+\frac{2\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{\eta^2\sigma^2}{T^2\gamma^2}\\
&\le\left(1+\frac{2}{T-1}\right)\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t}-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]+\frac{2\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{\eta^2\sigma^2}{T^2\gamma^2}\\
&\le\left(1+\frac{2}{T-1}\right)\left(\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t-1}-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]+\frac{2\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{\eta^2\sigma^2}{T^2\gamma^2}\right)+\frac{2\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{\eta^2\sigma^2}{T^2\gamma^2}\\
&\le\sum_{t=1}^{T}\left(1+\frac{2}{T-1}\right)^t\left(\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,0}-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]+\frac{2\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{\eta^2\sigma^2}{T^2\gamma^2}\right)\\
&=\sum_{t=1}^{T}(1+\frac{2}{T-1})^t(\frac{2\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{\eta^2\sigma^2}{T^2\gamma^2})\\
&\le 3T(\frac{2\eta^2}{T\gamma}\left\|\nabla F_i(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{\eta^2\sigma^2}{T^2\gamma^2}) \quad \text{(A20)}
\end{aligned}$$

where we use **Lemma 1** in the first and third inequality, **Assumption 5** in the fourth inequality.

Thus, it yields:

$$\begin{aligned}
\varepsilon_k&=\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\boldsymbol{\omega}_{i,k,t}-\boldsymbol{\omega}_k^{\text{Global}}\right\|^2\right]\\
&=\frac{1}{N}\sum_{i=1}^{N}\frac{6\eta^2}{\gamma}\left\|\nabla F(\boldsymbol{\omega}_k^{\text{Global}},\boldsymbol{D}_i)\right\|^2+\frac{3\eta^2\sigma^2}{T\gamma^2}\\
&\le\frac{3\eta^2\sigma^2}{T}+6\eta^2G^2+12\eta^2B^2(F(\boldsymbol{\omega}_{k-1}^{\text{Global}})-F(\boldsymbol{\omega}^*)) \quad \text{(A21)}
\end{aligned}$$

(Lemma 4 [39]**)** When $\mu>0$, for any $c\ge 0$, $0\le\eta\le\frac{1}{\mu}$ and non-negative sequence f_k, we have:

$$\frac{1}{W_K}\sum_{k=1}^{K+1}\left(\frac{w_k}{\eta}(1-\mu\eta)f_{k-1}-\frac{w_k}{\eta}f_k+c\eta w_k\right) \le 3\mu d_0\exp(-\mu\eta K)+c\eta \quad \text{(A22)}$$

where, $w_k=(1-\mu\eta)^{1-k}$ is the weight parameter and $W_K=\sum_{k=1}^{K+1}w_k$.

(Lemma 5 [40]**)** For any $c_1\ge 0$, $c_2\ge 0$, $\eta_{\max}\ge 0$ and non-negative sequence f_k, we have:

$$\frac{1}{K+1}\sum_{k=1}^{K+1}\left(\frac{f_{k-1}}{\eta}-\frac{f_k}{\eta}+c_1\eta+c_2\eta^2\right) \le\frac{d_0}{\eta_{\max}(K+1)}+2\sqrt{\frac{c_1d_0}{K+1}}+2\left(\frac{d_0}{K+1}\right)^{\frac{2}{3}}c_2^{\frac{1}{3}} \quad \text{(A23)}$$

C. Convergence Proof of Avg

Adding the statements of **Lemmas 2** and **3**, we have:

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{\omega}_k^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2\le\left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\left\|\boldsymbol{\omega}_k^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2+\frac{1}{T|S|}\eta^2\sigma^2+\left(1-\frac{|S|}{N}\right)\frac{4\eta^2}{|S|}G^2-\eta\left(\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right]-F(\boldsymbol{\omega}^*)\right)+3l\eta\varepsilon_r\\
&=(1-\frac{\mu\eta}{2})\mathbb{E}\left\|\boldsymbol{\omega}_k^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2+\frac{1}{T|S|}\eta^2\sigma^2+(1-\frac{|S|}{N})\frac{4\eta^2}{|S|}G^2-\eta(\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right]-F(\boldsymbol{\omega}^*))+\frac{2\eta}{3}(\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right])-F(\boldsymbol{\omega}^*)+\frac{\eta^2\sigma^2}{2T\gamma^2}+18l\eta^3G^2\\
&=(1-\frac{\mu\eta}{2})\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2-\frac{\eta}{3}(\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right]-F(\boldsymbol{\omega}^*))+\frac{1}{T|S|}\eta^2\sigma^2+\frac{\eta^2\sigma^2}{2T\gamma^2}+(1-\frac{|S|}{N})\frac{4\eta^2}{|S|}G^2+18l\eta^3G^2\\
&=(1-\frac{\mu\eta}{2})\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2-\frac{\eta}{3}(\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right]-F(\boldsymbol{\omega}^*))+\eta^2(\frac{\sigma^2}{T|S|}+\frac{\sigma^2}{2T\gamma^2}+(1-\frac{|S|}{N})\frac{4G^2}{|S|}+18l\eta G^2) \quad \text{(A24)}
\end{aligned}$$

Moving $\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right]-F(\boldsymbol{\omega}^*)$ term, it yields:

$$\begin{aligned}
&\frac{\eta}{3}\left(\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right]-F(\boldsymbol{\omega}^*)\right)\\
&\le(1-\frac{\mu\eta}{2})\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2-\mathbb{E}\left\|\boldsymbol{\omega}_k^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2+\eta^2(\frac{\sigma^2}{T|S|}+\frac{\sigma^2}{2T\gamma^2}+(1-\frac{|S|}{N})\frac{4G^2}{|S|}+18l\eta G^2) \quad \text{(A25)}
\end{aligned}$$

Then multiplying $\frac{3}{\eta}$, we have:

$$\begin{aligned}
\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right]-F(\boldsymbol{\omega}^*)&\le\frac{3}{\eta}(1-\frac{\mu\eta}{2})\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2-\frac{3}{\eta}\mathbb{E}\left\|\boldsymbol{\omega}_k^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2+3\eta(\frac{\sigma^2}{T|S|}+\frac{\sigma^2}{2T\gamma^2}+(1-\frac{|S|}{N})\frac{4G^2}{|S|}+18l\eta G^2) \quad \text{(A26)}
\end{aligned}$$

According to the convexity of the F_i, the following discussions are made.

(1) If $\mu>0$, we can add the weight $w_k=(1-\frac{\mu\eta}{2})^{1-k}$ and directly use **Lemma 4** and (A26) yields:

$$\begin{aligned}
&\frac{1}{W_K}\sum_{k=1}^{K+1}w_k\left(\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right]-F(\boldsymbol{\omega}^*)\right)\\
&=\frac{1}{W_K}\sum_{k=1}^{K+1}w_k\left(\frac{3}{\eta}(1-\frac{\mu\eta}{2})\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\omega^*\right\|^2-\frac{3}{\eta}\mathbb{E}\left\|\boldsymbol{\omega}_k^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2+3\eta(\frac{\sigma^2}{T|S|}+\frac{\sigma^2}{2T\gamma^2}+(1-\frac{|S|}{N})\frac{4G^2}{|S|}+18l\eta G^2)\right)\\
&\le 3\mu\left\|\boldsymbol{\omega}_0^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2\exp(-\frac{K\mu\eta}{2})+(\frac{2\sigma^2}{T|S|}+\frac{\sigma^2}{T\gamma^2}+(1-\frac{|S|}{N})\frac{8G^2}{|S|})\eta+36lG^2\eta^2 \quad \text{(A27)}
\end{aligned}$$

(2) If $\mu=0$, we have:

$$\begin{aligned}
&\mathbb{E}\left[F(\boldsymbol{\omega}_k^{\text{Global}})\right]-F(\boldsymbol{\omega}^*)\\
&\le\frac{3}{\eta}\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2-\frac{3}{\eta}\mathbb{E}\left\|\boldsymbol{\omega}_k^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2+3\eta(\frac{\sigma^2}{T|S|}+\frac{\sigma^2}{2T\gamma^2}+(1-\frac{|S|}{N})\frac{4G^2}{|S|}+18l\eta G^2)
\end{aligned}$$

$$= \frac{3\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{\eta} - \frac{3\mathbb{E}\left\|\boldsymbol{\omega}_{k}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{\eta} + 3\left(\frac{\sigma^2}{T|S|} + \frac{\sigma^2}{2T\gamma^2} + \left(1 - \frac{|S|}{N}\right)\frac{4G^2}{|S|}\right)\eta + 54lG^2\eta^2 \quad \text{(A28)}$$

Thus, we can directly use **Lemma 4** and it yields:

$$\sum_{k=1}^{K+1}\left(\mathbb{E}\left[F\left(\boldsymbol{\omega}_k^{\text{Global}}\right)\right] - F(\boldsymbol{\omega}^*)\right) \leq \sum_{k=1}^{K+1}\left(\frac{3\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{\eta} - \frac{3\mathbb{E}\left\|\boldsymbol{\omega}_{k}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{\eta} + 3\left(\frac{\sigma^2}{T|S|} + \frac{\sigma^2}{2T\gamma^2} + \left(1 - \frac{|S|}{N}\right)\frac{4G^2}{|S|}\right)\eta + 54lG^2\eta^2\right) \quad \text{(A29)}$$

(3) If $\mu < 0$, we have:

$$\mathbb{E}\left[F\left(\boldsymbol{\omega}_k^{\text{Global}}\right)\right] - F(\boldsymbol{\omega}^*) \leq \frac{3}{\eta}\left(1 - \frac{\mu\eta}{2}\right)\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 - \frac{3}{\eta}\mathbb{E}\left\|\boldsymbol{\omega}_{k}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 + 3\eta\left(\frac{\sigma^2}{T|S|} + \frac{\sigma^2}{2T\gamma^2} + \left(1 - \frac{|S|}{N}\right)\frac{4G^2}{|S|} + 18l\eta G^2\right)$$
$$\leq \frac{3}{\eta}\left(1 - \frac{\mu\eta}{2}\right)\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 - \frac{3}{\eta}\left(1 - \frac{\mu\eta}{2}\right)\mathbb{E}\left\|\boldsymbol{\omega}_{k}^{\text{Global}} - \boldsymbol{\omega}^*\right\|^2 + 3\eta\left(\frac{\sigma^2}{T|S|} + \frac{\sigma^2}{2T\gamma^2} + \left(1 - \frac{|S|}{N}\right)\frac{4G^2}{|S|} + 18l\eta G^2\right)$$
$$= \frac{3\left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{\eta} - \frac{3\left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\left\|\boldsymbol{\omega}_{k}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{\eta} + 3\left(\frac{\sigma^2}{T|S|} + \frac{\sigma^2}{2T\gamma^2} + \left(1 - \frac{|S|}{N}\right)\frac{4G^2}{|S|}\right)\eta + 54lG^2\eta^2 \quad \text{(A30)}$$

Thus, we can also directly use **Lemma 5** and it yields:

$$\sum_{k=1}^{K+1}\left(\mathbb{E}\left[F\left(\boldsymbol{\omega}_k^{\text{Global}}\right)\right] - F(\boldsymbol{\omega}^*)\right) \leq \sum_{k=1}^{K+1}\left(\frac{3\left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\left\|\boldsymbol{\omega}_{k-1}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{\eta} - \frac{3\left(1-\frac{\mu\eta}{2}\right)\mathbb{E}\left\|\boldsymbol{\omega}_{k}^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{\eta} + 3\left(\frac{\sigma^2}{T|S|} + \frac{\sigma^2}{2T\gamma^2} + \left(1 - \frac{|S|}{N}\right)\frac{4G^2}{|S|}\right)\eta + 54lG^2\eta^2\right)$$
$$\leq \frac{3\left(1-\frac{\mu\eta}{2}\right)\left\|\boldsymbol{\omega}_0^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{\eta_{\max}(K+1)} + 6\sqrt{\frac{\left(1-\frac{\mu\eta}{2}\right)\left\|\boldsymbol{\omega}_0^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2\left(\frac{\sigma^2}{T|S|}+\frac{\sigma^2}{2T\gamma^2}+\left(1-\frac{|S|}{N}\right)\frac{4G^2}{|S|}\right)}{K+1}} + 2\left(\frac{3\left(1-\frac{\mu\eta}{2}\right)\left\|\boldsymbol{\omega}_0^{\text{Global}}-\boldsymbol{\omega}^*\right\|^2}{K+1}\right)^{\frac{2}{3}}(54lG^2)^{\frac{1}{3}} \quad \text{(A31)}$$

Here we have proved the convergence of the FL-AVG

REFERENCES

[1] X. Sun, P. B. Luh, K. W. Cheung, et al., "An efficient approach to short-term load forecasting at the distribution level." *IEEE Trans. Power Syst*., vol. 1, no. 4, pp. 2526-2537, Jul. 2016.

[2] Y. Sasaki, N. Yorino, Y. Zoka, et al., "Robust stochastic dynamic load dispatch against uncertainties." *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 5535-5542, Nov. 2018.

[3] Y. Jiang, C. Wan, J. Wang, et al., "Stochastic receding horizon control of active distribution networks with distributed renewables." *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1325-1341, Mar. 2019.

[4] S. X. Chen, Y. S. Foo. Eddy, H. B. Gooi, et al., "A centralized reactive power compensation system for lv distribution networks." *IEEE Trans. Power Syst.*, vol. 30, no. 1, pp. 274-284, Jan. 2015.

[5] C. Wan, J. Lin, and W. Guo, "Maximum uncertainty boundary of volatile distributed generation in active distribution network." *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 2930-2942, Jul. 2018.

[6] E. Ostertagová and O. Ostertag, "Forecasting using simple exponential smoothing method." *Acta Electrotechnica Et Informatica*, vol. 12, no. 3, Jan. 2012.

[7] Shyh-Jier Huang and Kuang-Rong Shih, "Short-term load forecasting via ARMA model identification including non-gaussian process considerations." *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 673-679, May 2003.

[8] C.-L. Hor, S. J. Watson, and S. Majithia, "Daily load forecasting and maximum demand estimation using ARIMA and GARCH," in *2006 International Conference on Probabilistic Methods Applied to Power Systems*, Jun. 2006, pp. 1–6.

[9] Q. Li, Q. Meng, J. Cai and H. Yoshino, et al., "Applying support vector machine to predict hourly cooling load in the building," *Appl. Energy*, vol. 86, no. 10, pp. 2249-2256, Oct. 2009.

[10] Y. Cheng, P. P. K. Chan and Z. Qiu, "Random forest based ensemble system for short term load forecasting," in *2012 International Conference on Machine Learning and Cybernetics*, Jul. 2012, vol. 1, pp. 52–56.

[11] A. M. Pirbazari, E. Sharma and A. Chakravorty, et al., "An ensemble approach for multi-step ahead energy forecasting of household communities," *IEEE Access*, vol. 9, pp. 36218-36240, Mar. 2021.

[12] W. Kong, Z. Y. Dong and Y. Jia, et al., "Short-Term residential load forecasting based on LSTM Recurrent Neural Network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841-851, Jan. 2019.

[13] Y. Guo, Y. Li, and X. Qiao, et al., "BiLSTM multitask learning-based combined load forecasting considering the loads coupling relationship for multienergy system," *IEEE Trans. Smart Grid*, vol. 13, no. 5, pp. 3481-3492, Sept. 2022.

[14] N. Singh, C. Vyjayanthi and C. Modi, "Multi-step short-term electric load forecasting using 2D convolutional neural networks," in *2020 IEEE-HYDCON*, Sep. 2020, pp. 1-5.

[15] L. Jiang, X. Wang and W. Li, et al., "Hybrid multitask multi-information fusion deep learning for household short-term load forecasting," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5362-5372, Nov. 2021.

[16] M. Ribeiro, K. Grolinger and H.F. ElYamany, et al., "Transfer learning with seasonal and trend adjustment for cross-building energy forecasting," *Energy and Buildings*, vol. 165, pp.352-363, Apr. 2018.

[17] H. Shi, M. Xu and R. Li, "Deep learning for household load forecasting-a novel pooling deep RNN," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5271-5280, Sept. 2018.

[18] B. Stephen, X. Tang and P. R. Harvey, et al., "Incorporating practice theory in sub-profile models for short term aggregated residential load forecasting," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1591-1598, Jul. 2017.

[19] B. McMahan, E. Moore and D. Ramage, et al., "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Apr. 2017, pp. 1273–1282.

[20] Y. Wang, M. Jia and N. Gao, et al., "Federated clustering for electricity consumption pattern extraction," *IEEE Trans. Smart Grid*, vol. 13, no. 3, pp. 2425-2439, May 2022.

[21] Y. Li, J. Li and Y. Wang, "Privacy-preserving spatiotemporal scenario generation of renewable energies: a federated deep generative learning spproach," *IEEE Trans. Ind. Inf.*, vol. 18, no. 4, pp. 2310-2320, Apr. 2022.

[22] J. F. Toubeau, F. Teng and T. Morstyn, et al., "Privacy-preserving probabilistic voltage forecasting in local energy communities," *IEEE Trans. Smart Grid*, vol. 14, no. 1, pp. 798–809, Jan. 2023.

[23] A. Taïk and S. Cherkaoui, "Electrical load forecasting using edge computing and federated learning," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, Jun. 2020, pp. 1–6.

[24] H. Wang, C. Si, and J. Zhao, "A federated learning framework for non-intrusive load monitoring," *arXiv preprint.* Available: 2104.01618, 2021.

[25] Q. Wu, K. He, and X. Chen, "Personalized federated learning for intelligent IoT applications: A cloud-edge based framework," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 35-44, Apr. 2020.

[26] T. Li, A. K. Sahu and A. Talwalkar, et al., "Federated learning: challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, May 2020.

[27] N. Gholizadeh and P. Musilek, "Federated learning with hyper parameter based clustering for electrical load forecasting," *Internet Things*, vol. 17, no. 1, pp. 1–10, 2022.

[28] Y. He, F. Luo, G. Ranzi, et al., "Short-term residential load forecasting based on federated learning and load clustering," in Proc. *IEEE Int. Conf. Smart Grid Commun.*, 2021, pp. 77–82.

[29] Y. L. Tun, K. Thar, C. M. Thwal, and C. S. Hong, "Federated learning based energy demand prediction with clustered aggregation," *in Proc. IEEE Int. Conf. Big Data Smart Comput.,* 2021, pp. 164–167.

[30] Y. Wang, N. Gao and G. Hug, "Personalized federated learning for individual consumer load forecasting," *CSEE Journal of Power and Energy Systems*, vol. 9, no. 1, pp. 326–330, Jan. 2023.

[31] Y. Jiang, J. Konečný and K. Rush, et al., "Improving federated learning personalization via model agnostic meta learning," *arXiv preprint.* Available: 1909.12488, 2019.

[32] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, Jul. 2017, pp. 1126–1135.

[33] S. P. Karimireddy, S. Kale and M. Mohri, et al., "Stochastic controlled averaging for federated learning," in *Proceedings of the 37th International Conference on Machine Learning,* Nov. 2020, pp. 5132–5143.

[34] A. Fallah, A. Mokhtari, and A. Ozdaglar, "On the convergence theory of gradient-based model-agnostic meta-learning algorithms," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics,* Jun. 2020, pp. 1082–1092.

[35] C. Brester, V. Kallio-Myers, A. V. Lindfors, et al, "Evaluating neural network models in site-specific solar PV forecasting using numerical weather prediction data and weather observations," *Renew. Energ*., vol. 207, pp. 266-274, 2023.

[36] National Centers for Environmental Information (NCEI). "Climate Data Online: Dataset Discovery." *Datasets | Climate Data Online (CDO) | National Climatic Data Center (NCDC)*, Available: https://www.ncdc.noaa.gov/cdo-web/datasets.

[37] Z. Si, M. Yang, Y. Yu, et al. "Photovoltaic power forecast based on satellite images considering effects of solar position," *Appl. Energy*, vol. 302, no. 15, pp. 117514, Apr. 2021.

[38] C. Feng, B. Liang, Z. Li, W. Liu and F. Wen, "Peer-to-peer energy trading under network constraints based on generalized fast dual ascent," *IEEE Trans. Smart Grid*, vol. 14, no. 2, pp. 1441-1453, March 2023.

[39] S. U. Stich, "Unified optimal analysis of the (stochastic) gradient method," arXiv preprint. Available: 1907.04232, 2019.

[40] A. Kulunchakov, J. Mairal, "Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise," *J. Mach. Learn. Res*., vol. 21, no. 1, pp. 6184-6235, Jan. 2020.