

# Transfer Learning-based Model Training for Short-term Load Forecasting

Bozhen Jiang, Hongyuan Yang, Yidi Wang, Qin Wang, *Senior Member, IEEE*, and Hua Geng, *Fellow, IEEE*

**Abstract**—The smart grid infrastructure has recorded extensive real-time electricity consumption data, particularly at the levels of distribution transformers and below for short-term load forecasting (STLF). However, training individual short-term load forecasting model (SLFM) for each STLF scenario at these levels substantially increases the computational costs. To address this challenge, this paper proposes a transfer learning-based model training method for STLF. The proposed method is rooted in transfer learning principles and tailored to the unique characteristics of the aforementioned levels, incorporating several key steps. First, an approach for extracting key peak and valley points based on peak width and peak prominence is proposed for simplifying the evaluation of load sequence similarity. Subsequently, these key points are clustered using a density-based spatial clustering of applications with noise approach to ensure proper alignment along the time axis. Secondly, temporal and distribution similarity metrics are introduced to establish a performance guarantee for the transferred SLFM. Subsequently, a hierarchical clustering method groups load sequences, utilizing temporal similarity to quantify distances among sequences and distribution similarity to optimize cluster number selection. To minimize generalization error and further reduce computational costs, a modified bagging method is proposed and applied during the transferred SLFM fine-tuning. Empirical evidence from a study conducted in Guiyang, China demonstrates that the proposed method maintains the SLFM performance without degradation and significantly reduces computational costs by a minimum of 92.23% across multiple scenarios.

**Index Terms**—Smart grid, short-term load forecasting, artificial neural network, transfer learning, similarity, clustering, bagging.

Manuscript received: September 2, 2024; revised: January 27, 2025; accepted: July 23, 2025. Date of CrossCheck: July 23, 2025. Date of online publication: August 29, 2025.

This work was supported by the National Key Research and Development Program of China (No. 2024YFB4207200).

This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

B. Jiang and H. Geng (corresponding author) are with the Automation Department of Tsinghua University, Beijing, China, and B. Jiang is also with the Department of Electrical and Electronic Engineering of The Hong Kong Polytechnic University, Hong Kong, China (e-mail: bozhen.jiang@connect.polyu.hk; genghua@tsinghua.edu.cn).

H. Yang is with the Computer Science of Wesleyan University, Middletown, USA (e-mail: hyang01@wesleyan.edu).

Y. Wang is with the Department of Power Automation of China Electric Power Research Institute Co., Ltd., Beijing, China (e-mail: 1249042448@qq.com).

Q. Wang is with the Department of Electrical and Electronic Engineering of The Hong Kong Polytechnic University, Hong Kong, China (e-mail: qin-ee.wang@polyu.edu.hk).

DOI: 10.35833/MPCE.2024.000940

## I. INTRODUCTION

COUNTRIES worldwide are progressively deregulating their electricity markets, aiming for liberalization and increased competition. This shift has led to the rise of numerous load service entities (LSEs) operating in competitive environments [1]-[4]. In the day-ahead market, LSEs can establish agreements with power system operators or other market participants to offer load management and demand response services. These services rely on accurate forecasts of load demand and market conditions, ensuring grid stability while enhancing the integration of renewable energy sources [5]. Among LSEs, load aggregators play a specialized role by consolidating and managing loads from small-scale entities such as communities or localized regions. By providing flexible and predictable load services, these aggregators support grid reliability while engaging in market trading and bidding to secure economic benefits. Consequently, more accurate short-term load forecasting (STLF) technology is an effective way to improve the competitiveness of LSEs [6], [7].

Recent research has shown that machine learning-based short-term load forecasting models (SLFMs) outperform traditional statistical models (such as moving average and regression models) in day-ahead forecasting for distribution network and levels below, particularly in capturing multi-feature nonlinear relationships [8]. Especially, artificial neural network (ANN)-based SLFMs such as recurrent neural network and its variants [9], [10] demonstrate superior forecasting performance compared with fully connected (FC) [11] network and convolutional neural network (CNN), as they account for temporal coupling relationships within sequences [12]-[14]. Furthermore, [15] proposed a dynamic temporal dependency model (DTDM), a transformer-based architecture adapted from natural language processing (NLP) for STLF applications. Through attention heatmap visualizations and empirical results at both the national and distribution transformer levels, DTDM proves effective in identifying dynamic temporal patterns, leading to improved forecasting accuracy.

The application of ANNs for STLF at distribution transformer and levels below presents significant computational challenges due to the massive volume of data involved. Smart grid investments have significantly improved the flexibility, efficiency, and reliability of the power system through bidirectional power and information flows. A critical component of this infrastructure is the advanced metering system,

which provides LSEs with consumption data across multiple temporal resolutions (from minute-level to hour-level) and spatial scales (from transformer-level down to individual consumers) [16]. Reference [14] has demonstrated that load aggregation at the distribution transformer level substantially improves the accuracy of day-ahead STLF. This challenge is particularly evident in large urban areas like Guiyang, China (covering 8034 km<sup>2</sup> and comprising six urban districts), where approximately 380000 distribution transformers serve residential complexes. The resulting diversity in load patterns creates numerous forecasting scenarios, with each transformer representing a unique forecasting case. Developing and maintaining individual ANN-based forecasting models for each transformer will impose prohibitive computational costs on LSEs, especially considering the need for continuous model updates to maintain accuracy. Consequently, there is an urgent need to develop computationally efficient day-ahead STLF methods capable of handling the substantial computational requirements of ANN-based modeling at distribution transformer and levels below while maintaining forecasting accuracy.

The demand for STLF at distribution transformer and levels below is often concentrated within local areas (referring to many building complexes), leading to a clustering characteristic [8], [17]. This implies that numerous scenarios share great similarities in terms of climate and electricity consumption behavior due to geographical proximity. Firstly, the differences in factors such as temperature, humidity, and wind speed within a local area are not substantial. In practice, climate data from a single meteorological station are used to represent the local climate. Therefore, it is justifiable to reuse the results of feature selection in similar scenarios. For example, if the load series of locations  $\{A, B\}$  exhibit high similarities and feature selection has already been performed for location  $A$ , the results of that feature selection can be directly applied to location  $B$ . Secondly, historical origins and geographical location contribute to extensive communication and interaction among residents in local areas, resulting in similar lifestyle habits. Consequently, the load sequences exhibit obvious similarities, enabling the reuse of SLFMs. For instance, if the load series of locations  $\{A, B\}$  exhibit high similarities and an SLFM has already been trained for location  $A$ , the parameters of that model can be transferred to location  $B$ .

Transfer learning (TL) involves reusing model parameters acquired from one task to enhance performance in a related task, and has achieved considerable success in the field of NLP [18], [19]. The general process of TL in NLP involves pre-training a source domain model on an extensive corpus database, followed by fine-tuning the output layers of the target domain model on the corpus database specific to its downstream tasks. This approach significantly reduces computational costs. The effectiveness of TL can be attributed to several reasons. Firstly, the source and target tasks exhibit similarity, which implies that they may share parameters or prior distributions of the hyperparameters of the model [20],

[21]. Secondly, the shallow layers of the model generally learn features that are not task-specific, allowing the parameters learned by these layers to be applied to many similar tasks [22]. Thirdly, sufficient dataset from the target task is required to fine-tune the source domain model, enabling the model to perform better in the target task. Therefore, applying TL to STLF scenarios at distribution transformer and levels below holds great potential.

Reference [23] develops a useful TL framework for STLF that adapts a model pre-trained on multiple distribution nodes to new but similar scenarios. Specifically, outlier-insensitive clustering based temporal similarity evaluation method is employed to group similar distribution nodes into clusters, followed by training a base model between these clusters. The model is then transferred to other scenarios, with all layers fine-tuned during the transfer process. However, there are several improvements that need to be made when adapting TL to STLF scenarios. Firstly, the similarities among load sequences are not only based on temporal similarity but also on distribution similarity, which jointly determine the performance of the transferred model in similar STLF scenarios. Therefore, a comprehensive consideration of both aspects is necessary. Secondly, load sequences are typically long, and their similarity can be evaluated by identifying key points that carry clear physical meaning [24]. Meanwhile, since load sequences are discrete, computational distribution similarity using probability density functions is not feasible. Thus, efficient methods should be developed to reduce temporal similarity evaluation time and measure distribution similarity. Thirdly, fine-tuning all layers can be time-consuming for large amount of STLF scenarios, while fine-tuning only a subset of the output layers may lead to overfitting, thereby increasing the generalization error of the transferred model. Therefore, improvement is also required in the fine-tuning.

To address the aforementioned challenges, this paper presents a TL-based model training method for reducing the computational cost of SLFM while maintaining the accurate performance. Based on TL principles and characteristics of STLF, the proposed method is designed to include five points. Firstly, it utilizes a key point identification method and Euclidean distance to measure temporal similarity. Meanwhile, it employs kernel density estimation (KDE) and the Kullback-Leibler (KL) divergence to evaluate distribution similarity. Then, the proposed method applies a hierarchical clustering that utilizes temporal similarity to measure load sequence distance and distribution similarity to determine the optimal cluster number. Additionally, it selects features that are common across multiple STLF scenarios as inputs for all scenarios within the same cluster. Finally, the proposed method incorporates a fine-tuning process using a bagging method, which significantly reduces model computational costs and improves generalization performance.

The remainder of this paper is organized as follows. Section II investigates the transfer conditions when applying TL to STLF and further analyzes generalization error to improve

fine-tuned SLFM performance. Section III introduces the proposed method. Section IV presents and discusses the experimental results to verify the effectiveness of the proposed method. Finally, conclusions are drawn in Section V.

## II. TRANSFER CONDITION WHEN APPLYING TL TO STLF AND GENERALIZATION ERROR ANALYSIS

### A. SLFM

The ANN-based SLFM generally uses the mean squared error (MSE) as the loss function, and the optimization problem can be drawn as:

$$\min \frac{1}{num} \sum_{i=1}^{num} (Load_{pre,i} - Load_{actual,i})^2 \quad (1)$$

$$Load_{pre,i} = FC(F(x_i)) \quad (2)$$

where  $num$  is sampling number of history loads;  $Load_{pre,i}$  is the predicted load at the  $i^{\text{th}}$  step;  $Load_{actual,i}$  is the actual load at the  $i^{\text{th}}$  step;  $x_i$  is the input at the  $i^{\text{th}}$  step;  $F(\cdot)$  denotes an ANN such as FC network, CNN [25], [26], long short-term memory (LSTM) neural network [27], [28], bi-directional LSTM (Bi-LSTM) neural network [29] - [31], and DTDM [32]; and  $FC(\cdot)$  denotes an FC layer.

### B. Load Analysis at Distribution Transformer and Levels Below

The demand for STLF at distribution transformer and levels below is primarily concentrated within local areas, where numerous load consumers are in close proximity to each other and exhibit apparent similarities in their load behaviors, as shown in Fig. 1. Figure 1 presents the load curves of distribution transformers with the same rated capacity in Guiyang, China, where each curve represents a distribution transformer. Clearly, certain curves exhibit similar trends, with peaks and valleys coinciding at corresponding time points. This can be attributed to two main reasons. Firstly, the geographical proximity plays a significant role. In local areas, the differences in factors such as temperature, humidity, and wind speed are not substantial. Moreover, the extensive communication and interaction among residents in these areas result in similar lifestyle habits. Secondly, in practical applications, climate data from a single meteorological station are commonly utilized to represent the local climate. As a result, various local STLF scenarios tend to share the same climate factors. Thus, it becomes justifiable to reuse feature selection results and SLFMs at distribution transformer and levels below.

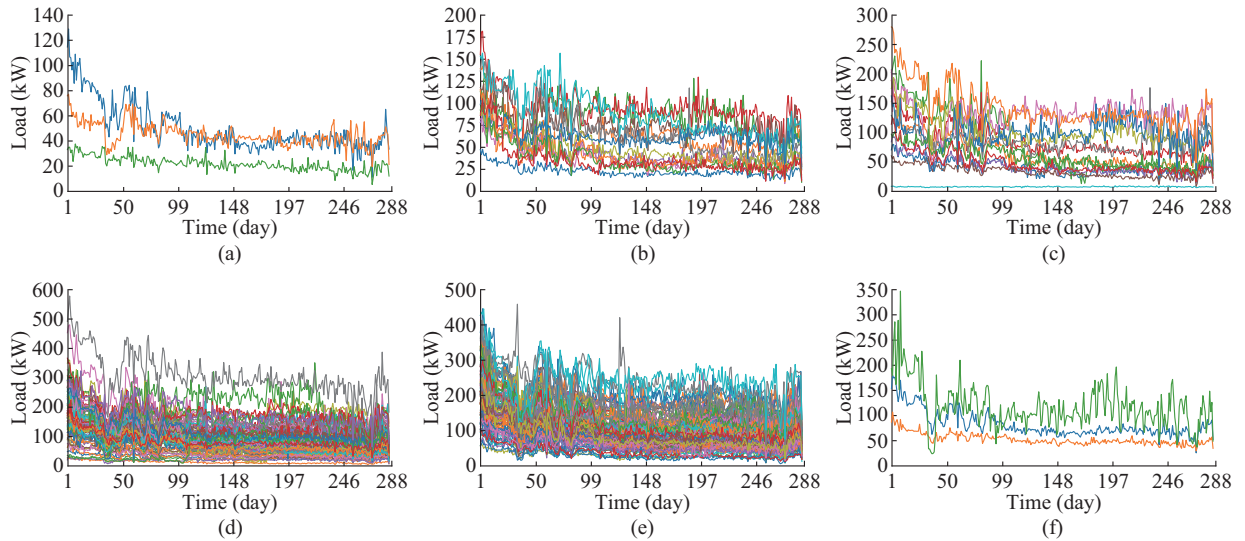


Fig. 1. Load curves of distribution transformers with same rated capacity in Guiyang, China. (a) Rated capacity of 160 kVA. (b) Rated capacity of 200 kVA. (c) Rated capacity of 250 kVA. (d) Rated capacity of 315 kVA. (e) Rated capacity of 400 kVA. (f) Rated capacity of 500 kVA.

### C. Transfer Condition Analysis

High-level layers, including the output layer and certain hidden layers, are not appropriate for transfer as they are more closely tied to specific tasks [22]. For the purpose of analysis, we make the assumption that only the output layer will be fine-tuned.

An SLFM is trained in a reference scenario through the back-propagation algorithm, and its loss function can be written as  $J_{ref}(\theta) = (\theta X' - Y_{ref})^T (\theta X' - Y_{ref}) / 2$ , where  $\theta$  is the vector of output-layer parameters;  $Y_{ref}$  is the load sequence in the reference scenario; and  $X'$  is the  $F(X)$  in (2),  $X = [x_1, x_2, \dots, x_s]$ , and  $s$  is the length of  $Y_{ref}$ . Given that these lay-

ers process same input features, the loss function resulting from direct transfer of the last layer to a similar scenario is  $J_{tra}(\theta) = (\theta X' - Y_{tra})^T (\theta X' - Y_{tra}) / 2$ , which can be drawn as:

$$J_{tra}(\theta) = J_{ref}(\theta) + \frac{1}{2} (Y_{ref} - Y_{tra})^T (2\theta X' - Y_{ref} - Y_{tra}) \quad (3)$$

where  $Y_{tra}$  is the load sequence in a similar transferred scenario. Furthermore, by manipulating and simplifying (3), we can obtain:

$$J_{tra}(\theta) - J_{ref}(\theta) = (Y_{ref} - Y_{tra})^T \theta X' + \frac{1}{2} (Y_{tra}^T Y_{tra} - Y_{ref}^T Y_{ref}) \quad (4)$$

By taking modulus values on both sides of (4) and according to Schwarz inequality, we can obtain:

$$|J_{tra}(\boldsymbol{\theta}) - J_{ref}(\boldsymbol{\theta})| \leq |\boldsymbol{\theta} \mathbf{X}'| |\mathbf{Y}_{ref} - \mathbf{Y}_{tra}| + \frac{1}{2} (|\mathbf{Y}_{tra}|^2 - |\mathbf{Y}_{ref}|^2) \quad (5)$$

It can be observed that the transferred objective function have an upper bound, which consists of two parts:  $|\mathbf{Y}_{ref} - \mathbf{Y}_{tra}|$  and  $|\mathbf{Y}_{tra}|^2 - |\mathbf{Y}_{ref}|^2$ .

$|\mathbf{Y}_{ref} - \mathbf{Y}_{tra}|$  is used to calculate the Euclidean distance between the two load sequences, which can be understood as the temporal similarity.  $|\mathbf{Y}_{tra}|^2 - |\mathbf{Y}_{ref}|^2$  is used to calculate the distribution difference between the two load sequences. It means that only if the distribution difference between the two load sequences is consistent, can the modulus of the two load sequences be equal. For example, for  $\mathbf{Y}_{tra} = [0.5, 0.1, 0.3]$  and  $\mathbf{Y}_{ref} = [0.3, 0.5, 0.1]$ , although the element positions are different, the modulus of the two load sequences is equal. Therefore, it is necessary to take both temporal similarity and distribution similarity into consideration.

There are three technical challenges that need to be addressed. Firstly, due to the subjective nature of human behavior, load sequences are often characterized by fluctuations, with numerous peaks and valleys. Some of these fluctuations hold significant physical meanings, while others are influenced by random factors. Therefore, there is a need for peak and valley point identification, especially those that correspond to significant physiological load behaviors. Secondly, since load is typically recorded at fixed time intervals, load sequences are discrete and the direct acquisition of their probability density functions is not feasible. Therefore, it becomes necessary to estimate the probability density function in order to calculate the distribution similarity. Thirdly, (3) indicates that the coefficient is determined by  $\boldsymbol{\theta} \mathbf{X}'$  and the constant 1/2. However, in practice,  $\boldsymbol{\theta}$  is subject to fine-tuning. Consequently, it is essential to integrate both temporal similarity and distribution similarity to comprehensively evaluate sequence similarity in a more reasonable manner.

#### D. Generalization Error Analysis

The purpose of the generalization error is to analyze the performance of the trained model when faced with previously unseen data [33]. The expected generalization error  $E(f; D)$  can be decomposed into the sum of variance *Var*, bias *Bias*, and noise  $\epsilon$  [33], as shown in (6).

$$E(f; D) = \mathbb{E}_D [(f(\mathbf{x}; D) - \hat{f}(\mathbf{x}))^2] + (\hat{f}(\mathbf{x}) - y)^2 + \mathbb{E}_D [(y_D - y)^2] = \text{Var}(\mathbf{x}) + \text{Bias}^2(\mathbf{x}) + \epsilon^2 \quad (6)$$

where  $\mathbf{x}$  is the input vector;  $y_D$  is the label corresponding to  $\mathbf{x}$  in the training dataset  $D$ ;  $y$  is the true label of  $\mathbf{x}$ ;  $f(\mathbf{x}; D)$  is the predicted output of model  $f$  learned on training set  $D$ ;  $\mathbb{E}_D[\cdot]$  denotes taking expectation over the training dataset  $D$ ;  $\hat{f}(\cdot) = E(f; D)$  denotes the expected prediction of the model trained on the training set  $D$ ;  $\text{Var}(\mathbf{x}) = \mathbb{E}_D [(f(\mathbf{x}; D) - \hat{f}(\mathbf{x}))^2]$ ;  $\text{Bias}^2(\mathbf{x}) = (\hat{f}(\mathbf{x}) - y)^2$ ; and  $\epsilon^2 = \mathbb{E}_D [(y_D - y)^2]$ .

During fine-tuning the output layer of the transferred SLFM, the bias tends to decrease while the variance may increase. This phenomenon can be attributed to two reasons. Firstly, despite the similarity between the target and source load sequences, the target load sequence still possesses unique characteristics. The frozen shallow networks are trained solely in the source STLF scenario, and the acquired

knowledge may not be entirely suitable for the target STLF scenario. Secondly, fine-tuning only specific layers instead of the entire forecasting model can potentially lead to overfitting. Considering these factors, further improvements are required for fine-tuning output layer to reduce variance.

### III. PROPOSED METHOD

#### A. Temporal Similarity Measurement

The purpose of clustering is to divide the samples in a dataset into several (usually disjoint) subsets, each subset called a ‘‘cluster’’. In the context of SLFM transfer, clustering serves two objectives. One is to cluster peak and valley points from the time axis to align key peak and valley points. Another is to cluster a sequence of key peak and valley points from the scenario axis to discover similar power consumption patterns.

##### 1) Extraction of Peak and Valley Points

Reducing the dimension by preserving peak and valley points in the time series is a promising method [24], [34]. However, in cases where the load sequences present numerous minor fluctuations, a large number of peak and valley points will be identified. Some of these small fluctuations in the peak and valley points could be attributed to random factors impacting the load behavior, rather than indicating shared living habits. To address this, it is beneficial to establish appropriate criteria for the extraction of peak and valley points. Specifically, peak width and peak prominence can be utilized to limit the selection of these points. Peak width serves as a metric to measure the width of a peak point, while the height of the peak point above the baseline is referred to as peak prominence. Peak prominence indicates the prominence of a peak point relative to other peak points. Generally, half of the peak prominence is chosen as the peak width [35].

##### 2) Identification of Key Peak and Valley Points Based on Density-based Spatial Clustering of Applications with Noise (DBSCAN)

Despite small fluctuation points are filtered out using peak width and peak prominence criteria, a considerable amount of noise points that do not align with the other data points remain. DBSCAN, a density clustering method, offers an effective solution by considering the continuity between samples, enabling the identification of clusters with arbitrary morphological distributions [36]. The calculation process of DBSCAN exhibits low time complexity and requires only a small number of parameters. Additionally, DBSCAN demonstrates robustness against noise points. Therefore, DBSCAN is well-suited for the task of identifying peak and valley points, aligning them along the time axis. However, as shown in Fig. 1, the horizontal axis represents the time while the vertical axis represents the load. This leads to an inconsistency in units between the two axes, particularly since the distribution of data points on the time axis is typically much denser than that on the load axis. This inconsistency may cause DBSCAN to cluster data along the time axis. Therefore, we artificially multiply the time axis by a coefficient of 200, which helps prevent clustering along the time axis and ensures proper alignment. In other words, DB-

SCAN will only perform along the vertical direction of the time axis.

### 3) Hierarchical Clustering

Obtaining key points aims to discover similar power consumption patterns, which can then be used for SLFM transfer. Hierarchical clustering is a clustering method that divides data sets at different levels to form a tree-shaped clustering structure [37]. Since the process of aggregating local load sequences is from bottom to top, a bottom-up hierarchical clustering is employed. Hierarchical clustering requires setting a certain “distance” for measuring the “distance” of the set, as well as determining the appropriate cluster number. The commonly used “distances” include minimum distance, maximum distance, and average distance [38]. This study uses the average distance  $d_{avg}$  among key points, as shown in (7).

$$d_{avg}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{z \in C_j} d_E(x, z) \quad (7)$$

where  $C_i$  and  $C_j$  are the different clusters; and  $d_E(x, z)$  is the Euclidean distance between  $x$  and  $z$ .

### B. Distribution Similarity Measurement

As mentioned earlier, hierarchical clustering necessitates the selection of an appropriate cluster number. Distribution similarity measurement approach is proposed to assist in determining the appropriate cluster number  $p$ . By identifying the appropriate cluster number, the temporal and distribution similarities within the same cluster can be maximized, while those across different clusters can be maximized. However, it is important to note that achieving optimal values for both temporal and distribution similarities simultaneously is not always possible, as it involves striking a tradeoff between the two factors.

#### 1) KDE

Since the probability density distribution of discrete load sequences is unknown, KDE is employed to estimate it, which is expressed as:

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \quad (8)$$

where  $x_i$  is the sample point;  $n$  is the number of sample points;  $x$  is the point to be estimated;  $f_h(x)$  is the estimated probability density;  $K(\cdot)$  is the kernel function; and  $h$  is the bandwidth.

KDE is a non-parametric estimation method that does not need any prior knowledge and can approximate density functions of arbitrary shapes, given sufficient samples [39].

#### 2) KL Divergence

KL divergence  $D_{KL}(P||Q)$  describes the similarity between two probability distributions  $P$  and  $Q$  [40], as shown in (9).

$$D_{KL}(P||Q) = \sum_i P(x_i) \ln \frac{P(x_i)}{Q(x_i)} \quad (9)$$

where  $P(x_i)$  and  $Q(x_i)$  are the probabilities of  $x_i$  occurrence under the two probability distributions  $P$  and  $Q$ , respectively. The smaller the KL divergence of the two probability distributions, the more similar they are. The KDE is used to estimate the probability distribution of any two load sequences, and then the KL divergence is used to evaluate their proba-

bility distribution similarity.

### 3) Distribution Similarity Indicator

The indicators for evaluating the quality of clustering include sum of squared errors (SSE) and silhouette coefficients. However, as demonstrated in Section II-C, the upper bounds of transfer performance are constrained by the distribution similarity, which is overlooked by the aforementioned two indicators. Thus, a distribution similarity indicator  $S_{dis}$  is proposed in this paper to evaluate the quality of clustering, as shown in (10).

$$S_{dis} = \frac{D_{KL,same}}{D_{KL,all} - D_{KL,same}} \quad (10)$$

where  $D_{KL,same} = \sum_i \sum_{x_p, x_q \in C_i} D_{KL}(P_{i,p}(x_p)||Q_{i,q}(x_q))$  is the sum of KL divergence in the same cluster,  $x_p$  and  $x_q$  are the load sequences, and  $P_{i,p}(\cdot)$  and  $Q_{i,q}(\cdot)$  are the probabilities of the  $p^{\text{th}}$  and  $q^{\text{th}}$  sequences in the  $i^{\text{th}}$  cluster occurrence under the two probability distributions  $P$  and  $Q$ , respectively; and  $D_{KL,all} = \sum_i \sum_j \sum_{x_p \in C_i, x_q \in C_j} D_{KL}(P_{i,p}(x_p)||Q_{j,q}(x_q))$  is the sum of KL divergence among all clusters. Equation (10) is strictly decreasing, meaning that a smaller value indicates a better clustering effect.

The elbow method is commonly used for determining the best cluster number [23], [41]. In the elbow method, different cluster numbers are considered, and the SSE is recorded and plotted on a curve. The cluster number corresponding to the point where the rate of decrease in SSE is the highest is considered the most reasonable cluster number. This is because prior to this cluster number, the SSE significantly changes with the cluster number, indicating the presence of points that are still far from the cluster centers and suggesting the need for additional clustering. Beyond this cluster number, further increasing the cluster number does not significantly improve the SSE. Therefore, the elbow point in the curve is considered the most appropriate cluster number. Similarly, in this study, we determine the appropriate cluster number based on significant changes in  $S_{dis}$  with varying cluster numbers.

### C. Bagging Method

Bagging method is an ensemble learning method commonly employed to reduce the aforementioned variance within noisy datasets [42]. In this method, multiple training subsets are created by randomly sampling the original dataset with replacement. These subsets are then used to train multiple base learners, and their predictions are combined, typically through voting or averaging, to yield the final ensemble prediction.

An average of  $l$  random variables  $\{Z_1, Z_2, \dots, Z_l\}$ , each with variance  $\sigma^2$ , has variance  $\sigma^2/l$ . If the variables are simply (identically distributed, but not necessarily independent) with positive pairwise correlation  $\rho$ , the variance of the average is [33]:

$$\text{Var}\left(\frac{1}{l} \sum_{i=1}^l Z_i\right) = \rho\sigma^2 + \frac{1-\rho}{l}\sigma^2 \quad (11)$$

For  $m$  well-trained SLFMs, the errors between their load

forecasts and actual loads can be denoted as  $\{Z_1, Z_2, \dots, Z_m\}$ . Two primary conclusions can be drawn from (11). Firstly, as the number of SLFMs increases, the variance of the average SLFM error shows a decline, which will reach a minimum of  $\rho\sigma^2$ . Therefore, the bagging method can reduce generalization error and improve SLFM performance. Secondly, if the correlation between SLFMs is excessively high, the variance of the average SLFM error increases. This suggests that it may not be necessary to individually fine-tune each SLFM.

The clustering process inherently generates numerous similar scenarios. Leveraging this characteristic, the bagging method employed in this study can be simplified. Specifically, since the load sequences have undergone preprocessing including temporal and distribution similarity clustering, these steps can effectively be viewed as generating multiple distinct training subsets. Furthermore, multiple reference transformers within the same cluster are selected for training, serving as multiple base learners. Finally, these base learners are integrated to produce the final load forecast, with their weights optimized using least squares method.

#### D. SLFM Transfer Stage

After obtaining the clustering results, the SLFM transfer process is implemented through three key steps: determining feature subsets, selecting reference transformers, and identifying SLFM components requiring fine-tuning. Since different transformers may have distinct feature subsets after feature selection, the common feature subset for each cluster is derived by taking the intersection of all feature subsets from transformers within the same cluster, as analyzed in Section II. For each cluster, several reference transformers are randomly selected and trained using the Adam optimization algorithm to obtain corresponding reference SLFMs [43]. The outputs of these reference SLFMs are then aggregated through a bagging method, where the averaging weights are optimized using the least squares method.

The overall process of the proposed method, illustrated in Fig. 2, consists of two main stages: load sequence clustering and SLFM transfer. In the load sequence clustering stage, key points are first extracted from load sequences sharing the same rated capacity, followed by hierarchical clustering to assign cluster labels. In the SLFM transfer stage, several load sequences from each cluster are randomly selected to train reference SLFMs using Adam optimization algorithm. These trained models are then aggregated to construct a transferred SLFM for the remaining load sequences in the same cluster, using the bagging method to fine-tune the output layer.

### IV. CASE STUDY

#### A. Data Set Description

Experimental study is conducted on a private data set to verify the performance of the proposed method. The private data set, which is provided by Electric Power Science Research Institute of Guizhou Power Grid Co., Ltd. and National Meteorological Information Center of China, consists of two parts: the load data and features.

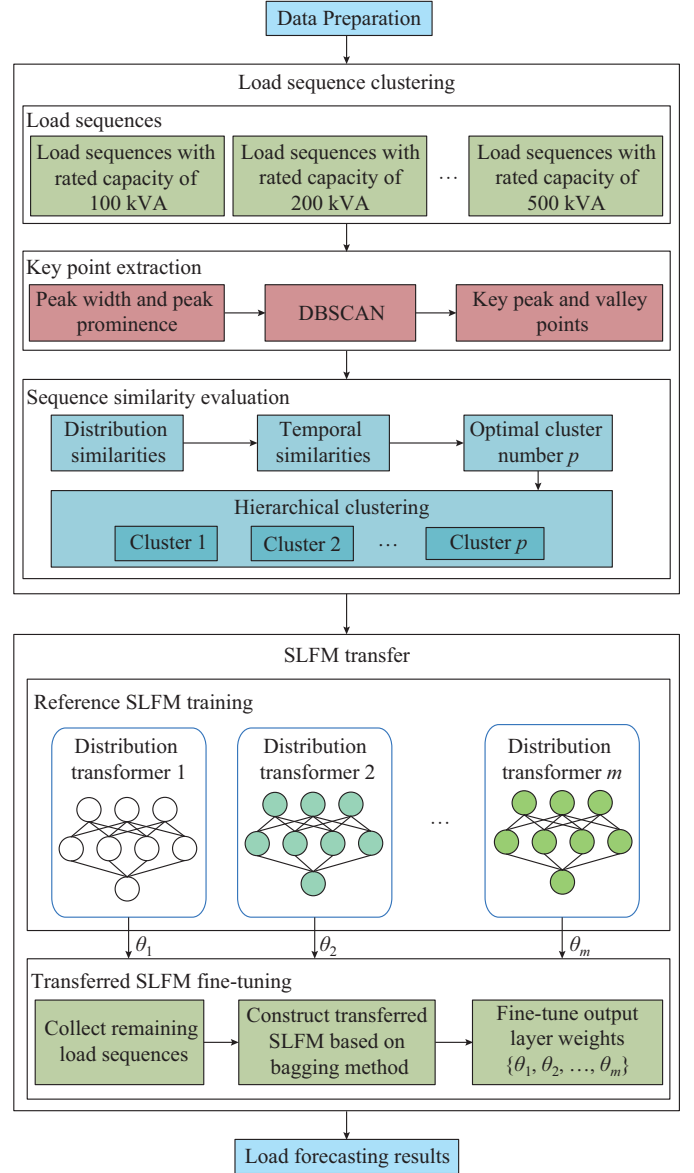


Fig. 2. Overall process of proposed method.

As of November 2021, there are about 380000 distribution transformers in Guiyang, China. Each distribution transformer is equipped with remote terminal units (RTUs) that continuously monitor real-time data such as load, voltage, and current. These RTUs periodically transmit the collected data to the data center for storage at intervals of 5 min. The load data comprises 5-min monitoring data from 220 distribution transformers in 6 districts of Guiyang. These distribution transformers exclusively supply power to residential users, with each transformer covering multiple buildings within its power supply area. The candidate features include the daily climate data (including temperature, rainfall, air pressure, wind speed, humidity, sunshine hours, etc.) and the time stamps (including month and day) from January 9, 2021 to November 21, 2021. Due to the private data set limitation, the experiment is conducted on a daily basis, which means the load is daily-averaged.

The data are normalized using  $z$ -score normalization and

divided into two subsets: a training set and a testing set. The training set spans from January 9, 2021 to November 1, 2021. We apply 3-fold cross-validation to systematically evaluate and select the optimal model parameters in the training set. Specifically, the training set is further divided into three parts. In each fold, two parts are used to train the SLFMs, while the remaining part is used to evaluate their performance. Finally, the average error across the folds is calculated to select the optimal hyperparameters. The holistic feature selection method [44] is employed to select the optimal feature subset, which can consider the correlation between candidate features and load, the redundancy among candidate features, and the interaction among candidate features. Three measurements are used to evaluate the performance of SLFMs: root MSE (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

### B. Capacity Classification

The distribution transformers in the Guiyang dataset have rated capacities of 160 kVA, 200 kVA, 250 kVA, 315 kVA, 400 kVA, and 500 kVA. The load curves are classified by capacity and visualized in Fig. 1. Figure 3, extracted from Fig. 1(d), illustrates similar load sequences observed in each distribution transformer with a rated capacity of 315 kVA. For example, the load sequences extracted from day 30 to day 40 are characterized by a decline followed by an ascent, as shown in Fig. 3(a). Additionally, similarities in peak or valley characteristics can be observed around days 81, 238, and 271. These findings suggest that load behavior in distribution transformers exhibits similarities, potentially influenced by factors such as cultural exchange and geographical proximity, as discussed in Section II.

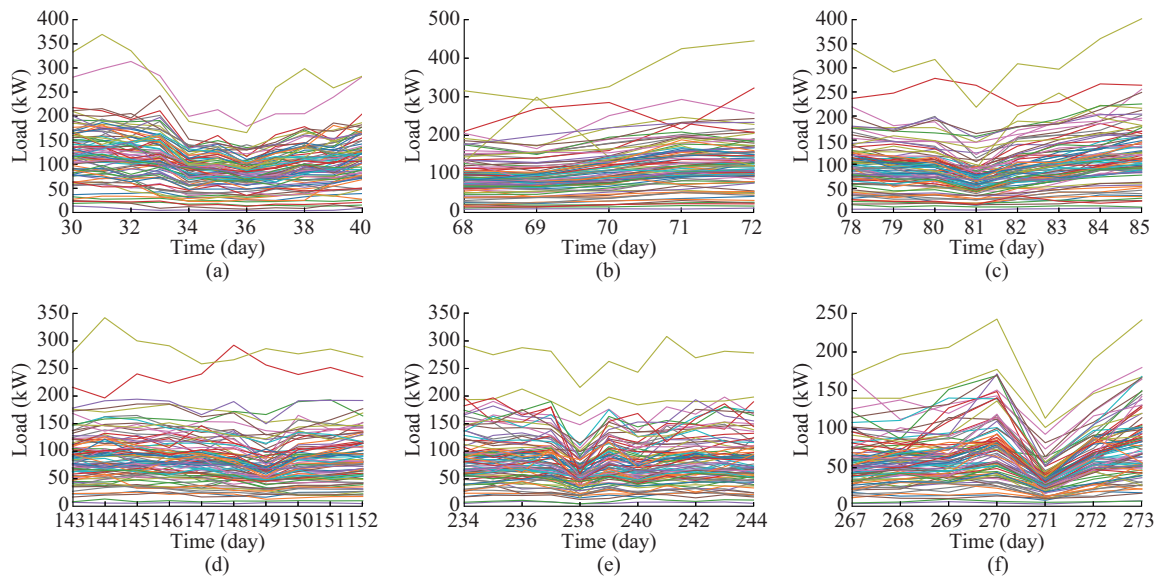


Fig. 3. Load sequences of distribution transformers with rated capacity of 315 kVA. (a) Load sequences around day 35. (b) Load sequences around day 70. (c) Load sequences around day 81. (d) Load sequences around day 148. (e) Load sequences around day 238. (f) Load sequences around day 271.

### C. Key Peak and Valley Point Extraction and Visualization

Take distribution transformers with rated capacity of 315 kVA as an example. Figures 4 and 5 show the extraction process of peak and valley points, respectively. The scatter plots of peak and valley points directly extracted are shown in Figs. 4(a) and 5(a), respectively. The scatter plots of key peak and valley points extracted based on peak width and peak prominence are shown in Figs. 4(b) and 5(b), respectively. The scatter plots of key peak and valley points extracted based on peak width, peak prominence, and DB-SCAN are shown in Figs. 4(c) and 5(c), respectively. The key peak and valley points are drawn back to the original load curve, as shown in Figs. 4(d) and 5(d), respectively. It can be observed that, firstly, the proposed method effectively aligns the key peak and valley points with the timeline, indicating its ability to uncover similar electricity consumption patterns among residents. Secondly, while there are variations in the values of key peak and valley points, there is a notable similarity in the timing of their appearance. This suggests a broader consistency in load behavior across diverse

regions.

### D. Hierarchical Clustering Process

For bandwidth selection in KDE, Silverman's rule is employed [45], given by:

$$h = \sigma \sqrt{\frac{4}{3n}} \quad (12)$$

where  $\sigma$  is the sample standard deviation; and  $n$  is the sample size. Since the data are standardized using  $z$ -score normalization ( $\sigma=1$ ), the resulting bandwidth  $h \approx 0.1$  for the training dataset. The distribution similarity  $S_{dis}$  will be calculated and collected according to (10) by traversing cluster number. Furthermore,  $S_{dis}$  will be plotted for selecting an appropriate cluster number according to elbow point. Figure 6 illustrates the hierarchical clustering process. The characteristic elbow-shaped curve of  $S_{dis}$  suggests an optimal cluster number of 9 for distribution transformers with a rated capacity of 315 kVA. Figure 7 shows load sequences consisting only of key peak and valley points.

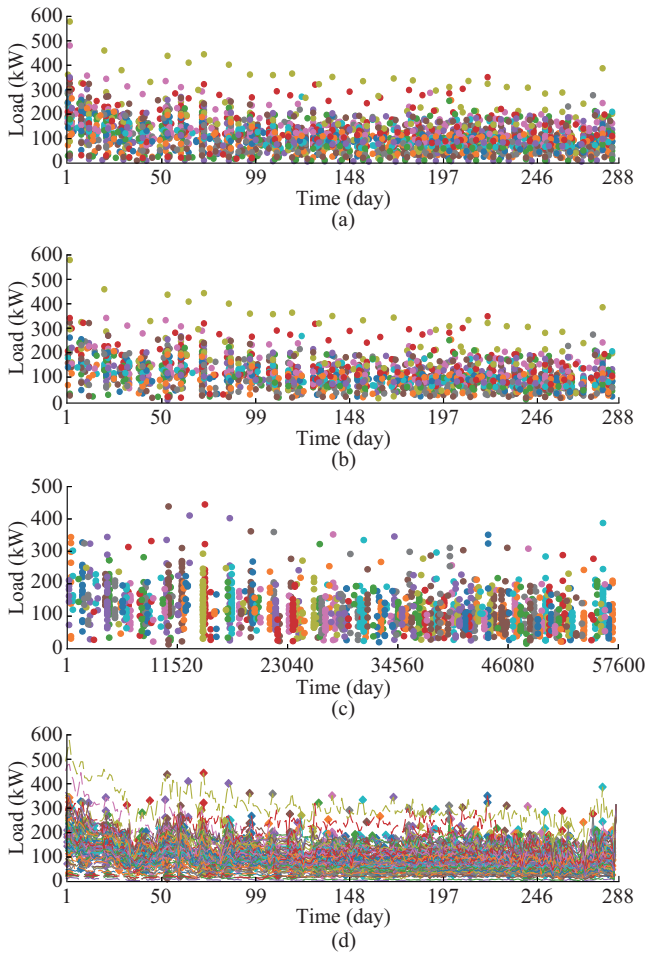


Fig. 4. Extraction process of peak points from load curve of distribution transformers with rated capacity of 315 kVA. (a) Peak points. (b) Peak points considering peak width and peak prominence. (c) Peak points considering peak width, peak prominence, and DBSCAN. (d) Peak points in original load curve.

There is a notable similarity in load fluctuations across different distribution transformers. This finding supports our earlier analysis in Section II, suggesting that geographical proximity may lead to synchronized energy consumption patterns due to shared lifestyle behaviors among local populations.

### E. Performance Comparison

The performance comparison of different SLFMs based on the Guiyang dataset is shown in Table I. The structures of different SLFMs in Guiyang dataset is shown in Table II, where the symbol -r means using all hidden states in the LSTM neural network. Furthermore, the linear regression (LR) model [46] and the support vector regression (SVR) model with different kernels including linear function (LF), polynomial function (PF), and radial basis function (RBF) are also taken into comparison. Additionally, we employ 3-fold cross-validation for optimal parameter selection. Meanwhile, 5 distribution transformers are randomly selected as reference transformers for SLFM transfer. Meanwhile, the proposed method without clustering (WC) is also directly tested for comparison. The results reveal several key findings. Firstly, the DTDMM maintains accurate performance compared with other SLFMs.

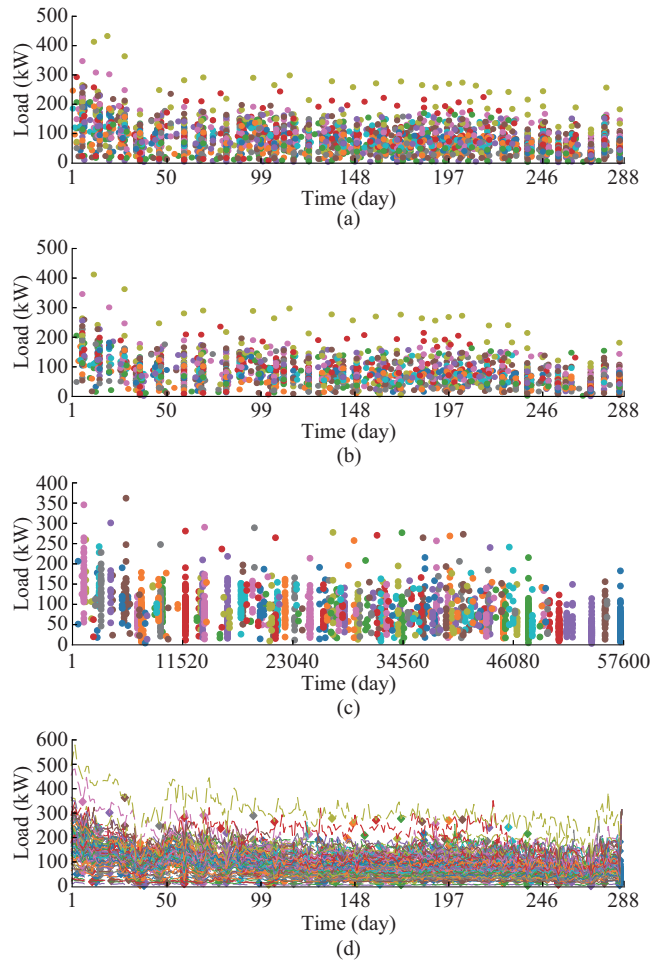


Fig. 5. Extraction process of valley points from load curve of distribution transformers with rated capacity of 315 kVA. (a) Valley points. (b) Valley points considering peak width and peak prominence. (c) Valley points considering peak width, peak prominence, and DBSCAN. (d) Valley points in original load curve.

Secondly, compared with direct training, the proposed method enhances the training efficiency of the DTDMM model, reducing training time costs by at least 92.23%. Thirdly, the forecasting performance of the model transfer after clustering is superior to that WC. These findings confirm that temporal similarity and distribution similarity can provide valuable prior information, ensuring better performance in the subsequent model transfer.

### F. Discussion on Method Promotion

The rapid expansion of renewable energy power plants has introduced significant challenges in managing multiple generation units and vast amounts of historical data. Each renewable energy unit (such as wind and photovoltaic) generates significant operational data, increasing the demand for forecasting of power generation [47]. While current forecasting technologies, particularly ANN-based methods, demonstrate strong predictive performance, their application to scenarios involving numerous generation units results in prohibitively high computational costs. Renewable energy stations are typically situated in regions where environmental factors such as wind speed, temperature, and solar radiation exhibit short-term consistency.

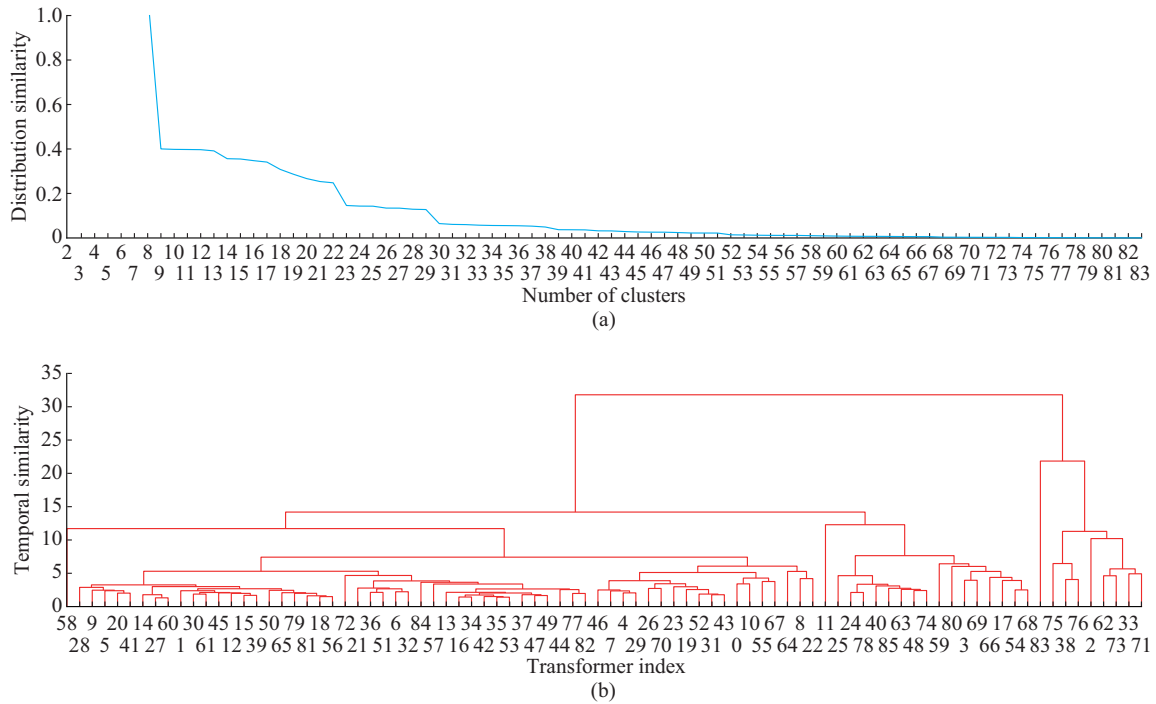


Fig. 6. Hierarchical clustering process of key points of distribution transformers with rated capacity of 315 kVA. (a) Curve of distribution similarity versus number of clusters. (b) Hierarchical clustering process.

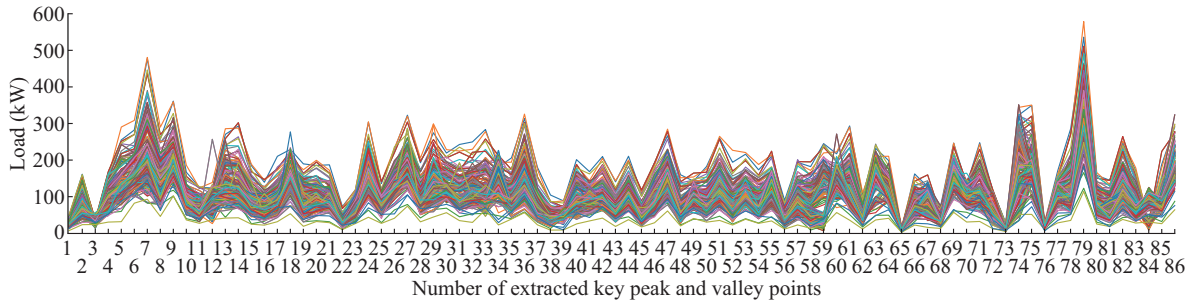


Fig. 7. Load sequences consisting only of peak and valley points of distribution transformers with rated capacity of 315 kVA.

TABLE I  
PERFORMANCE COMPARISON OF DIFFERENT SLFMS BASED ON GUIYANG DATASET

SLFM	1-day-ahead forecasting				3-day-ahead forecasting				7-day-ahead forecasting			
	RMSE (kW)	MAE (kW)	MAPE (%)	Training time (s)	RMSE (kW)	MAE (kW)	MAPE (%)	Training time (s)	RMSE (kW)	MAE (kW)	MAPE (%)	Training time (s)
LR [46]	25.61	19.44	33.50	<0.10								
LF + SVR	27.56	20.95	36.56	<0.10								
PF + SVR	29.93	24.39	42.45	<0.10								
RBF + SVR	31.35	25.14	42.38	<0.10								
FC network [11]	26.12	19.71	35.27	29.01								
CNN [25], [26]	20.79	16.88	26.54	31.15								
LSTM [27], [28]	21.11	16.27	27.87	35.18	21.28	16.61	28.01	35.36	21.19	15.84	30.10	35.81
Bi-LSTM [29], [30]	22.19	18.03	29.73	45.93	22.57	17.11	29.07	47.54	22.78	16.81	30.11	48.46
Bi-LSTM-r	22.78	18.21	28.73	48.11	22.61	16.61	28.46	47.69	19.96	14.57	27.74	49.57
DTDM [32]	19.73	15.48	25.39	26.54	19.77	15.72	25.71	26.17	18.97	14.99	27.11	27.93
Proposed method WC + DTDM	20.46	16.11	27.07	1.71	20.46	16.49	27.67	1.80	19.88	17.02	30.37	2.17
Proposed method + DTDM	18.09	14.13	23.97	1.71	18.73	15.14	25.01	1.80	18.42	14.87	25.57	2.17

TABLE II  
STRUCTURES OF DIFFERENT SLFMs IN GUIYANG DATASET

SLFM	Layer	Parameter	Activation
FC	FC	Hidden units: [64, 128, 256]	Sigmoid
	FC	Hidden units: [64, 128, 256]	Sigmoid
	FC	Hidden unit: 1	Linear
CNN	CNN	Kernel sizes: [8, 16, 32]; steps: [1, 2, 4]	ReLU
	CNN	Kernel sizes: [8, 16, 32]; steps: [1, 2, 4]	ReLU
	Flatten & FC	Hidden unit: 1	Linear
LSTM	LSTM	Hidden units: [16, 32, 64, 128]	Sigmoid
	Droupout	Rates: [0.1, 0.2]	
	FC	Hidden unit: 1	Linear
Bi-LSTM	LSTM	Hidden units: [16, 32, 64, 128]	Sigmoid
	Droupout	Rates: [0.1, 0.2]	
	FC	Hidden unit: 1	Linear
Bi-LSTM-r	LSTM	Hidden units: [16, 32, 64, 128]	Sigmoid
	Droupout	Rates: [0.1, 0.2]	
	FC & FC	Hidden unit: 1	Linear
DTDM	Multi-heads	Heads: [4, 8, 10]	Linear
	Attention	Hidden units: [16, 32, 64]	Linear
	Flatten & FC	Hidden unit: 1	Linear

As a result, power generation patterns in these areas tend to show strong similarities, as detailed in Section II-B. For example, in photovoltaic power forecasting, power generation output is highly dependent on solar irradiation intensity. Within a confined spatial range, variations in irradiation and cloud movement are gradual and continuous, leading to correlated generation behaviors among neighboring photovoltaic units. To capitalize on this spatial-temporal correlation, the proposed method clusters historical photovoltaic power generation curves, trains forecasting models for each cluster, and then applies bagging method and TL to other generation units within the same clusters. This method facilitates efficient photovoltaic power forecasting across multiple generation units, substantially reducing computational expenses while improving predictive accuracy.

## V. CONCLUSION

In this paper, a TL-based model training method for STLF is proposed for reducing the computational cost of training numerous SLTMs at the distribution transformer and levels below. The proposed method applies TL principles while incorporating the unique characteristics of load sequence at these levels. Specifically, key peak and valley points are extracted for simplifying the evaluation of load sequence similarity. Then, both temporal and distribution similarities are taken into consideration to ensure the forecasting accuracy of the transferred SLFMs. Furthermore, high-similarity load sequences are grouped via hierarchical clustering, and a modified bagging technique is employed for efficient fine-tuning. Evaluations on the Guiyang dataset across various SLFMs show that the proposed method maintains accurate forecasting performance while significantly decreasing computational cost.

## REFERENCES

- [1] J. C. D. Prado and W. Qiao, "A stochastic bilevel model for an electricity retailer in a liberalized distributed renewable energy market," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 4, pp. 2803-2812, Oct. 2020.
- [2] C. Qin, W. Liu, Y. Yan *et al.*, "Designing personalized incentive-based demand response services based on smart meter data and NSGA-III-DE algorithm," *Energy*, vol. 334, p. 137454, Oct. 2025.
- [3] J. Zhu, Y. Miao, H. Dong *et al.*, "Short-term residential load forecasting based on  $k$ -shape clustering and domain adversarial transfer network," *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 4, pp. 1239-1249, Jul. 2024.
- [4] P. Zeng, C. Sheng, and M. Jin, "A learning framework based on weighted knowledge transfer for holiday load forecasting," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 2, pp. 329-339, Mar. 2019.
- [5] X. Zhu, G. Ruan, H. Geng *et al.*, "Multi-objective sizing optimization method of microgrid considering cost and carbon emissions," *IEEE Transactions on Industry Applications*, vol. 60, no. 4, pp. 5565-5576, Jul. 2024.
- [6] J. P. Carvallo, P. H. Larsen, A. H. Sanstad *et al.*, "Long term load forecasting accuracy in electric utility integrated resource planning," *Energy Policy*, vol. 119, pp. 410-422, Aug. 2018.
- [7] B. Jiang, Y. Wang, Q. Wang *et al.*, "A novel interpretable short-term load forecasting method based on Kolmogorov-Arnold networks," *IEEE Transactions on Power Systems*, vol. 40, no. 1, pp. 1180-1183, Jan. 2025.
- [8] M. Grabner, Y. Wang, Q. Wen *et al.*, "A global modeling framework for load forecasting in distribution networks," *IEEE Transactions on Smart Grid*, vol. 14, no. 6, pp. 4927-4941, Nov. 2023.
- [9] C. Jia, H. He, J. Zhou *et al.*, "A novel deep reinforcement learning-based predictive energy management for fuel cell buses integrating speed and passenger prediction," *International Journal of Hydrogen Energy*, vol. 100, pp. 456-465, Jan. 2025.
- [10] C. Jia, H. He, J. Zhou *et al.*, "Learning-based model predictive energy management for fuel cell hybrid electric bus with health-aware control," *Applied Energy*, vol. 355, p. 122228, Feb. 2024.
- [11] S. Rai and M. De, "NARX: contribution-factor-based short-term multi-nodal load forecasting for smart grid," *International Transactions on Electrical Energy Systems*, vol. 31, no. 9, p. e12726, Sept. 2021.
- [12] B. Jiang, Q. Wang, S. Wu *et al.*, "Advancements and future directions in the application of machine learning to AC optimal power flow: a critical review," *Energies*, vol. 17, no. 6, p. 1381, Mar. 2024.
- [13] B. Jiang, C. Qin, and Q. Wang, "An unsupervised physics-informed

- neural network method for AC power flow calculations,” *IEEE Transactions on Power Systems*, vol. 40, no. 5, pp. 4407-4410, Sept. 2025.
- [14] W. Kong, Z. Y. Dong, Y. Jia *et al.*, “Short-term residential load forecasting based on LSTM recurrent neural network,” *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841-851, Jan. 2019.
- [15] B. Jiang, H. Yang, Y. Wang *et al.*, “Dynamic temporal dependency model for multiple steps ahead short-term load forecasting of power system,” *IEEE Transactions on Industry Applications*, vol. 60, no. 4, pp. 5244-5254, Jul. 2024.
- [16] M. B. Mollah, J. Zhao, D. Niyato *et al.*, “Blockchain for future smart grid: a comprehensive survey,” *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 18-43, Jan. 2021.
- [17] D. Syed, H. Abu-Rub, A. Ghayeb *et al.*, “Deep learning-based short-term load forecasting approach in smart grid with clustering and consumption pattern recognition,” *IEEE Access*, vol. 9, pp. 54992-55008, Apr. 2021.
- [18] J. Devlin, M.-W. Chang, K. Lee *et al.*, “Bert: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA, Jun. 2019, pp. 4171-4186.
- [19] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: a pretrained language model for scientific text,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, Nov. 2019, pp. 3613-3618.
- [20] G. Pinto, Z. Wang, A. Roy *et al.*, “Transfer learning for smart buildings: a critical review of algorithms, applications, and future perspectives,” *Advances in Applied Energy*, vol. 5, p. 100084, Feb. 2022.
- [21] S. Jain, H. Salman, A. Khaddaj *et al.*, “A data-based perspective on transfer learning,” in *Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, Jun. 2023, pp. 3613-3622.
- [22] Z. Zhang, P. Zhao, P. Wang *et al.*, “Transfer learning featured short-term combining forecasting model for residential loads with small sample sets,” *IEEE Transactions on Industry Applications*, vol. 58, no. 4, pp. 4279-4288, Jul. 2022.
- [23] D. Syed, A. Zainab, S. S. Refaat *et al.*, “Inductive transfer and deep neural network learning-based cross-model method for short-term load forecasting in smart grids,” *IEEE Canadian Journal of Electrical and Computer Engineering*, vol. 46, no. 2, pp. 157-169, May 2023.
- [24] T. Fu, “A review on time series data mining,” *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164-181, Feb. 2011.
- [25] C. Li, G. Li, K. Wang *et al.*, “A multi-energy load forecasting method based on parallel architecture CNN-GRU and transfer learning for data deficient integrated energy systems,” *Energy*, vol. 259, p. 124967, Nov. 2022.
- [26] M. Imani, “Electrical load-temperature CNN for residential load forecasting,” *Energy*, vol. 227, p. 120480, Jul. 2021.
- [27] H. Dong, J. Zhu, S. Li *et al.*, “Probabilistic residential load forecasting with sequence-to-sequence adversarial domain adaptation networks,” *Journal of Modern Power Systems and Clean Energy*, vol. 12, no. 5, pp. 1559-1571, Sept. 2024.
- [28] Neeraj, J. Mathew, and R. K. Behera, “EMD-Att-LSTM: a data-driven strategy combined with deep learning for short-term load forecasting,” *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 5, pp. 1229-1240, Sept. 2022.
- [29] Y. Guo, Y. Li, X. Qiao *et al.*, “BiLSTM multitask learning-based combined load forecasting considering the loads coupling relationship for multienergy system,” *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 3481-3492, Sept. 2022.
- [30] X. Zhang, S. Kuenzel, N. Colombo *et al.*, “Hybrid short-term load forecasting method based on empirical wavelet transform and bidirectional long short-term memory neural networks,” *Journal of Modern Power Systems and Clean Energy*, vol. 10, no. 5, pp. 1216-1228, Sept. 2022.
- [31] D. Syed, H. Abu-Rub, A. Ghayeb *et al.*, “Household-level energy forecasting in smart buildings using a novel hybrid deep learning model,” *IEEE Access*, vol. 9, pp. 33498-33511, Feb. 2021.
- [32] B. Jiang, Y. Liu, H. Geng *et al.*, “A transformer based method with wide attention range for enhanced short-term load forecasting,” in *Proceedings of 2022 4th International Conference on Smart Power & Internet Energy Systems*, Beijing, China, Dec. 2022, pp. 1684-1690.
- [33] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin: Springer, 2009.
- [34] P. W. P. Man and M. H. Wong, “Efficient and robust feature extraction and pattern matching of time series by a lattice structure,” in *Proceedings of the Tenth International Conference on Information and Knowledge Management*, New York, USA, Nov. 2001, pp. 271-278.
- [35] J. J. Baeza-Baeza, M. J. Ruiz-Ángel, M. C. García-Álvarez-Coque *et al.*, “Half-width plots, a simple tool to predict peak shape, reveal column kinetics and characterise chromatographic columns in liquid chromatography: state of the art and new results,” *Journal of Chromatography A*, vol. 1314, pp. 142-153, Nov. 2013.
- [36] Y. Chen, L. Zhou, S. Pei *et al.*, “KNN-BLOCK DBSCAN: fast clustering for large-scale data,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 6, pp. 3939-3953, Jun. 2021.
- [37] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn *et al.*, “Hierarchical clustering: objective functions and algorithms,” *Journal of the ACM*, vol. 66, no. 4, pp. 1-42, Jun. 2019.
- [38] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: an overview,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86-97, Jan. 2012.
- [39] W. J. Lee, G. P. Mendis, M. J. Triebe *et al.*, “Monitoring of a machining process using kernel principal component analysis and kernel density estimation,” *Journal of Intelligent Manufacturing*, vol. 31, no. 5, pp. 1175-1189, Jun. 2020.
- [40] B. Zhou, X. Ma, Y. Luo *et al.*, “Wind power prediction based on LSTM networks and nonparametric kernel density estimation,” *IEEE Access*, vol. 7, pp. 165279-165292, Nov. 2019.
- [41] M. A. Syakur, B. K. Khotimah, E. S. Rochman *et al.*, “Integration K-means clustering method and elbow method for identification of the best customer profile cluster,” *IOP Conference Series: Materials Science and Engineering*, vol. 336, p. 012017, Apr. 2018.
- [42] B. Ghogh and M. Crowley. (2019, May). The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. [Online]. Available: <https://arxiv.org/abs/1905.12787>
- [43] D. P. Kingma and J. Ba. (2014, Oct.). Adam: a method for stochastic optimization. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [44] B. Jiang, Y. Liu, H. Geng *et al.*, “A holistic feature selection method for enhanced short-term load forecasting of power system,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1-11, Nov. 2022.
- [45] R. R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*. Pittsburgh: Academic Press, 2012.
- [46] S. Rai and M. De, “Load forecasting using two-level heterogeneous ensemble method for smart metered distribution system,” *Scientia Iranica*, vol. 32, no. 1, pp. 1-13, Jan. 2025.
- [47] H. Zhang, J. Yang, S. Fan *et al.*, “An ultra-short-term distributed photovoltaic power forecasting method based on GPT,” *IEEE Transactions on Sustainable Energy*, vol. 16, no. 4, pp. 2746-2754, Oct. 2025.

**Bozhen Jiang** received the M.S. degree in electronic and information engineering from Tsinghua University, Beijing, China, in 2023. He is currently working toward the Ph.D. degree with The Hong Kong Polytechnic University, Hong Kong, China. His research interests include load forecasting, big data, and machine learning application to smart grid.

**Hongyuan Yang** received the B.S. degree in computer science and theater from Wesleyan University, Middletown, USA, in 2023. He is currently working toward the M.S. degree with University of North Carolina at Charlotte, Charlotte, USA. His research interests include natural language processing, renewable energy source, and load forecasting.

**Yidi Wang** received the M.S. degree in power system and its automation from China Electric Power Research Institute (CEPRI), Beijing, China, in 2023. She is currently working at the Department of Power Automation, CEPRI. Her research interests include application of deep learning and reinforcement learning in power systems.

**Qin Wang** received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2006, the M.S. degree from South China University of Technology, Guangzhou, China, in 2009, and the Ph.D. degree in electrical engineering from Iowa State University, Ames, USA, in 2013. He is currently an Associate Professor with The Hong Kong Polytechnic University, Hong Kong, China. His previous industry experiences include positions with Electric Power Research Institute, Palo Alto, USA; National Renewable Energy Laboratory, Golden, USA; Midcontinent ISO, Carmel, USA; and ISO New England, Holyoke, USA. His research interests include power system reliability and online security analysis, grid integration of renewable energy resources, smart distribution system, transactive energy, ve-

hicle-to-everything, and electricity market.

**Hua Geng** received the B.S. degree in electrical engineering from Huazhong University of Science and Technology, Wuhan, China, in 2003, and the Ph.D. degree in control theory and application from Tsinghua University, Beijing, China, in 2008. From 2008 to 2010, he was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada. He joined Automation Department of

Tsinghua University in June 2010 and is currently a Full Professor. He is the Editor-in-Chief of IEEE Transactions on Sustainable Energy. He served as General Chair, Track Chairs, and Session Chairs of several IEEE conferences. He is an IEEE Fellow and an IET Fellow, Convener of the Modeling Working Group in IEC SC 8A. His current research interests include advanced control on power electronics and renewable energy conversion systems and artificial intelligence (AI) for energy systems.